Verified Intent Development

A Methodology for the Age of AI-Augmented Software Development

Oscar Valenzuela

First Edition - December 2025

Contents

1 Preface

1.1 Who This Book Is For

You're using AI to generate code—or your team is, or soon will be. You've experienced the productivity surge: features that took days now take hours. But you've also felt the nagging uncertainty: *How do I know this code is actually correct?*

This book is for software developers, engineering leaders, and teams who recognize that AI-generated code needs systematic verification, not just deployment. Whether you're a junior engineer learning to work with AI assistants, a senior engineer mentoring others, or a manager building team practices, you'll find practical guidance here.

1.2 What You'll Learn

Core Skills: - How to specify intent before generating code (preventing misalignment) - How to calibrate verification depth to actual risk (avoiding both negligence and waste) - How to maintain understanding of code you didn't write (preserving long-term maintainability) - How to track provenance systematically (knowing what needs extra scrutiny)

Practical Outcomes: - Catch bugs in verification instead of production - Ship AI-generated code with genuine confidence - Maintain velocity without accumulating technical debt - Build verification skills that scale with AI capabilities

1.3 Acknowledgment on AI Assistance

AI language models supported portions of the editorial review, fact validation, and automated build process for this edition. Every AI-assisted contribution went through human verification, alignment with VID principles, and provenance tracking before inclusion.

1.4 How to Use This Book

If you're an individual developer: Start with Chapters 1-3 to understand the paradigm shift, then dive into the five principles (Chapters 4-8) and core practices (Chapters 9-12). The junior engineer curriculum (Chapter 13) provides a structured 12-week learning path regardless of experience level.

If you're a team lead or manager: Read the one-page overview and VID for Engineering Managers guide first, then focus on Chapters 15-16 for team adoption strategies. Use the appendices (D-F) for risk rubrics, decision trees, and ready-made checklists.

If you're short on time: Start with the one-page overview, then read Chapter 4 (Intent Before Generation) and Appendix D (Risk Scoring). Apply those two practices immediately—you'll see results within a week.

1.5 What This Book Is Not

This isn't a prompt engineering guide, an AI tutorial, or a replacement for your current development methodology. VID complements Agile, Scrum, and other processes by addressing the specific gap they don't cover: systematic verification of AI-generated code.

The goal is judgment, not checklists. You'll learn principles that apply whether you're using today's AI tools or whatever emerges in five years.

1.6 Chapter 1: The Inversion

1.6.1 The Old World

For fifty years, software development methodologies have solved the same fundamental problem: **code is expensive to produce**.

Writing code requires understanding the problem, designing a solution, translating that design into syntax, debugging errors, and iterating until it works. This process is slow, cognitively demanding, and error-prone. Seasoned developers historically measure meaningful progress in small, high-quality increments rather than raw line counts.

Every methodology you've heard of optimizes around this constraint:

Waterfall tried to minimize wasted coding effort by doing extensive upfront planning. If writing code is expensive, don't write code until you're certain what to write.

Agile accepted that requirements change, so it optimized for shorter feedback loops. Write less code at a time, verify it works, adjust course. Don't invest too much before validating direction.

Scrum added coordination structures around Agile. If coding is the expensive part, make sure everyone is working on the right code and not duplicating effort.

Extreme Programming recognized that code quality problems compound, so it front-loaded quality practices. Pair programming, test-driven development, continuous integration — all designed to catch problems before code goes too far.

These methodologies disagree on many things, but they share one assumption: the primary constraint is producing code.

1.6.2 The Inversion

In 2023, that assumption began to collapse. AI coding assistants crossed a threshold from "occasionally useful autocomplete" to "genuine code generation." By 2025, the shift was undeniable.

Consider what changed:

- A developer can describe a function in plain English and receive working code in seconds
- Entire modules can be scaffolded from a description
- Tests can be generated automatically
- Documentation can be written by describing what to document
- Boilerplate that took hours now takes minutes

The raw act of producing code is no longer the bottleneck.

But something else happened that the optimists didn't predict: **the problems didn't go away, they moved**.

Code that generates in seconds still needs to: - Actually solve the intended problem (not just a similar one) - Handle edge cases correctly - Be secure against adversarial input - Perform acceptably under load - Integrate with existing systems without breaking them - Be maintainable by humans who didn't write it - Not infringe on someone else's intellectual property - Comply with applicable regulations

AI can write code fast. AI cannot *guarantee* any of these things.

The bottleneck inverted. We went from:

```
[Hard] Writing Code → [Easy] Verifying Code
```

To:

```
[Easy] Writing Code → [Hard] Verifying Code
```

Every methodology designed for the old world optimizes for the wrong constraint.

1.6.3 The Danger Zone

This inversion creates a dangerous period. Teams can generate code faster than ever, but their verification capacity hasn't increased. The result is predictable:

Technical debt accumulates faster. Code is generated without full understanding, creating maintenance burdens that compound. GitHub's 2023 developer experience study found that 92%

of enterprise developers already use AI coding tools, meaning high-velocity generation is now the norm for most teams.[1]

Security vulnerabilities multiply. AI generates code that works but doesn't anticipate adversarial conditions. Stanford's 2023 study found that developers using AI assistants were more likely to introduce security vulnerabilities, particularly in authentication and input validation code.[2] NYU researchers likewise observed that roughly 40% of Copilot's security-sensitive suggestions contained exploitable weaknesses.[3]

Quality becomes invisible. When code appears instantly, it *feels* free. Teams skip verification because the generation was so easy. GitClear's 2024 analysis of 153 million lines of code showed "churn" (code written and reverted) rising sharply in the AI era, signaling hidden quality problems that escape early review.[4]

Understanding erodes. Developers accept code they don't fully understand. When problems arise, they lack the context to debug effectively. Stack Overflow's 2024–2025 Developer Surveys report that only 43% of developers trust AI accuracy and 20% feel less confident in their own problem-solving after relying on AI tools, reflecting diminished understanding.[5]

This isn't an argument against AI-assisted development. The productivity gains are real and substantial. This is an argument for **methodology that fits the new reality**.

1.6.4 What VID Is (and Isn't)

Before explaining what VID offers, it's important to understand what it is—and what it is not.

VID is NOT:

A replacement for Agile, Scrum, or other methodologies — VID is complementary. You can practice VID within Agile sprints, Scrum ceremonies, or any development process. VID addresses a specific gap: how to handle AI-generated code safely.

An AI-first development framework — VID isn't about maximizing AI usage or treating AI as the primary developer. It's about verification regardless of code source. (That said, AI-generated code often needs more verification.)

A prompt engineering guide — VID doesn't teach you how to write better prompts or get more from AI. It teaches you what to do *after* AI generates code. Prompt engineering is orthogonal to VID.

A productivity framework — VID's goal isn't shipping faster; it's shipping *safely* while using AI. Sometimes VID makes you slower (verification takes time). That's acceptable—the alternative

[1] GitHub. "Survey reveals AI's impact on the developer experience." June 13, 2023. https://github.blog/news-insights/research/survey-reveals-ais-impact-on-the-developer-experience/

[2] Perry, N., Srivastava, M., Kumar, D., & Boneh, D. (2023). "Do Users Write More Insecure Code with AI Assistants?" *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, pp. 2785–2799. https://doi.org/10.1145/3576915.3623157

[3] Pearce, H., Ahmad, B., Tan, B., Dolan-Gavitt, B., & Karri, R. (2022). "Asleep at the Keyboard? Assessing the Security of GitHub Copilot's Code Contributions." *Proceedings of the 43rd IEEE Symposium on Security and Privacy (SP 2022)*, pp. 754–768. https://doi.org/10.1109/SP46214.2022.9833571

[4] GitClear. "Coding on Copilot: 2023 Data Suggests Downward Pressure on Code Quality." January 2024. https://www.gitclear.com/coding_on_copilot_data_shows_ais_downward_pressure_on_code_quality

[5] Stack Overflow. "2024 Developer Survey" and "2025 Developer Survey." https://survey.stackoverflow.co/2024/ and https://survey.stackoverflow.co/2025/

is shipping bugs.

Test-Driven Development (TDD) — While VID encourages test-first thinking, TDD focuses on using tests to drive design. VID focuses on verification across multiple dimensions: functional correctness, security, maintainability, and provenance—not just passing tests.

A development environment or tool — VID is methodology, not software. You don't install VID. You practice it. (Though tools can help—see Appendix F and the `/templates` directory for integration examples.)

VID IS:

A verification-first methodology for AI-augmented development — When code generation is cheap, verification becomes the core skill. VID systematizes verification.

A framework for calibrating trust — Not all code deserves equal scrutiny. VID provides principles for matching verification depth to risk level.

A mental model for the inverted bottleneck — VID helps developers think clearly about what changed, why it matters, and how to adapt.

Complementary to existing processes — VID fits into your current workflow. It answers the question "How should we handle AI-generated code?" that other methodologies don't address.

Applicable regardless of AI usage — Even if you write all code by hand, VID's principles (intent specification, graduated trust, understanding over acceptance) improve quality. But VID becomes essential when AI generates code.

The Core Distinction:

Most methodologies optimize for *code production*. VID optimizes for *code verification*. In the AI era, this is the critical shift.

1.6.5 What VID Offers

Verified Intent Development is designed for this inverted world. Its core insight is simple:

When generation is cheap and fast, verification becomes the valuable skill.

VID doesn't slow down generation. It builds systematic verification into the development process so that speed doesn't come at the cost of quality, security, or maintainability.

The developer's role transforms. Instead of spending the bulk of time writing code and a fraction reviewing, VID practitioners flip that ratio: specify intent deliberately, then invest most of the effort in verification. This isn't a step backward — it's a recognition of where human judgment adds value.

VID provides:

1. **A mental model** for thinking about AI-augmented development
2. **Principles** that guide decision-making in novel situations
3. **Practices** that build verification into daily work
4. **A vocabulary** for discussing AI-related development decisions
5. **A framework** for adapting as AI capabilities evolve

The rest of this book teaches all five.

1.7 Chapter 2: Why Existing Approaches Fall Short

Before presenting VID, we should understand why you can't just apply existing methodologies to AI-augmented development. The failures aren't superficial — they're structural.

1.7.1 Agile's Blind Spot

Agile's core insight is that requirements change, so we should plan in short iterations and adapt based on feedback. This remains valid. But Agile assumes the expensive operation is writing code, so its ceremonies optimize for coordination and prioritization of coding effort.

Standups ask: "What did you code yesterday? What will you code today?"

Sprint planning asks: "How much can we code this sprint?"

Retrospectives ask: "How can we code more effectively?"

Notice what's missing: systematic attention to verification.

Agile has testing, of course. But testing in Agile is typically focused on "does the feature work?" not "is this code trustworthy?" When code is written by humans who understand what they wrote, this is often sufficient. The developer's understanding provides implicit verification.

When code is generated by AI, that implicit verification disappears. The developer may not fully understand the generated code. "It passes tests" is not the same as "I understand why it's correct."

Agile provides no framework for deciding how much verification a piece of code needs, tracking where code came from (human versus AI), adjusting the development process based on code provenance, or building verification skills as a core competency. These gaps become critical when AI generates significant portions of your codebase.

1.7.2 The "Just Add AI" Fallacy

Many teams try to integrate AI by simply adding it to their existing workflow:

"We do Scrum, but now developers use Copilot."

This approach fails because it treats AI as a faster typewriter rather than a fundamental shift in how code comes into existence. The methodology remains optimized for the old constraint while the actual constraint has changed.

The symptoms of this failure are predictable and often alarming. Developers generate more code than ever before, yet defect rates increase rather than decrease. Sprint velocity appears higher on paper, but production incidents rise correspondingly. Technical debt accumulates faster than before because generated code is accepted without deep understanding. Code reviews become rubber stamps—there's simply too much code to review carefully when AI multiplies output. Perhaps most dangerously, no one knows which parts of the codebase were AI-generated, making it impossible to apply appropriate scrutiny where it's needed most.

1.7.3 The Vibe Coding Trap

On the opposite extreme, some developers abandon methodology entirely. "Vibe coding" — generating code through natural language prompts with minimal structure — produces impressive demos but dangerous production systems.

Vibe coding fails because:

No verification criteria. Without explicit criteria for correctness, there's no way to know if generated code is right. "It seems to work" is not verification.

No risk awareness. All code is treated equally, whether it's a utility function or an authentication handler.

No learning loop. Without systematic tracking, teams can't learn which patterns produce good outcomes.

No accountability. When everyone is prompting AI, no one is responsible for understanding the result.

Vibe coding works for prototypes and experiments. It's actively dangerous for production systems.

1.7.4 Emerging AIDD Frameworks

Several "AI-Driven Development" frameworks have emerged attempting to address these gaps. They typically share these characteristics:

- Replace sprints with shorter cycles ("bolts")
- Emphasize specification as input to AI
- Add AI agents for various development tasks
- Provide prompt templates and workflows

These frameworks improve on naive AI adoption, but most share a fundamental flaw: **they optimize for generation velocity, not verification confidence**.

Critical questions remain unanswered: How do you know AI-generated code is correct? When should humans verify versus trust automation? How do you build verification skills in developers? How do you adapt verification intensity to different risk levels? Most fundamentally, how do you maintain understanding as AI writes increasingly more of your code?

VID addresses these questions directly.

1.8 Chapter 3: The Core Insight

1.8.1 Verification as the Skill

VID is built on a single core insight:

In AI-augmented development, verification is the core professional skill.

This might seem obvious, but its implications are profound. Consider what it means for:

Education: We should teach developers to verify code, not just write it. Reading code critically becomes more important than writing code fluently.

Hiring: Verification ability matters more than generation speed. A developer who can reliably identify when AI-generated code is wrong is more valuable than one who can prompt faster.

Process: Development process should optimize for verification throughput, not generation throughput.

Tools: We need tools that help verify, not just tools that help generate.

Career development: Senior developers are distinguished by judgment about what to verify and how deeply, not by typing speed or language knowledge.

1.8.2 The Verification Spectrum

Not all verification is equal. VID recognizes a spectrum from shallow to deep:

Syntactic verification: Does it compile/parse? Does it pass the linter? This is fully automatable and should always happen.

Functional verification: Does it produce correct output for known inputs? Automated tests cover this, but test quality matters enormously.

Semantic verification: Does it do what was *intended*, not just what was *specified*? This requires human judgment to bridge the gap between specification and intent.

Robustness verification: Does it handle edge cases, malformed input, resource constraints, and concurrent access correctly? Requires thinking about what could go wrong.

Security verification: Does it resist adversarial input and protect sensitive data? Requires threat modeling and security expertise.

Maintainability verification: Can future developers understand and modify this code? Requires projecting into the future.

Provenance verification: Where did this code come from? Does it infringe on others' rights? Requires understanding AI training and licensing.

AI-generated code needs all of these. Vibe coding provides none. Traditional Agile provides the first two, partially. VID provides a framework for all of them, calibrated to risk.

1.8.3 The Understanding Requirement

There's a deeper issue than verification mechanics: **understanding**.

When a human writes code, they understand what it does (at least at the moment of writing). This understanding enables: - Effective debugging when problems arise - Confident modification when requirements change - Accurate estimation of impact when changes are proposed - Reliable integration with other code

When AI generates code, this understanding doesn't automatically transfer to the human. The developer might accept code they don't fully understand. This creates a time bomb:

- Today: Code works, developer is happy
- Next month: Bug appears, developer struggles to debug code they don't understand

- Next quarter: Modification needed, developer afraid to change code they don't understand
- Next year: Code is legacy that no one will touch

VID addresses this through a principle we'll explore in detail: **No code without understanding.** This doesn't mean understanding every character — it means understanding at a level appropriate to the code's importance.

1.9 Chapter 4: Principle One — Intent Before Generation

1.9.1 The Principle

Never generate code without first articulating what you intend to build and how you will verify correctness.

This principle seems simple but violates common practice. Most developers using AI start with a vague idea, prompt the AI, look at the result, and decide if it seems right. This is backwards.

1.9.2 Why Intent First?

Without explicit intent, you cannot verify. You can only react. You look at generated code and ask "does this seem okay?" rather than "does this match what I needed?"

The difference is profound:

Reactive verification: "This function seems to do something with dates. It looks reasonable. I'll accept it."

Intent-driven verification: "I needed a function that calculates business days between two dates, excluding weekends and holidays from a provided list. Let me check if this implementation handles holidays, excludes weekends, and correctly counts the days without double-counting boundaries."

Reactive verification catches obvious errors. Intent-driven verification catches subtle ones — the ones that matter in production.

1.9.3 Articulating Intent

Intent should include:

Functional intent: What should this code do? What are the inputs, outputs, and transformation?

Quality intent: What are the performance requirements? Reliability requirements? What trade-offs are acceptable?

Boundary intent: What inputs are valid? What should happen with invalid inputs? What are the edge cases?

Integration intent: How does this interact with existing code? What contracts must it honor?

You don't need a formal document for every function. But you should be able to articulate these before generating code. If you can't, you're not ready to generate.

1.9.4 The Verification Criteria

Intent leads to verification criteria. Before generating, ask:

"How will I verify this is correct?"

If you can't answer, you can't verify. And if you can't verify, you shouldn't generate.

Verification criteria might be: - Specific test cases with expected outputs - Properties that must hold (e.g., "output is always sorted") - Security requirements (e.g., "rejects SQL injection attempts") - Performance requirements (e.g., "responds in under 100ms for 1000 items")

Write these down, even informally. They guide your verification after generation.

1.9.5 Practice: The Intent Moment

Before every generation, pause for an "intent moment." Ask yourself:

1. What exactly should this code do?
2. What are the edge cases?
3. How will I verify it's correct?

This pause takes seconds. It saves hours of debugging wrong code.

1.10 Chapter 5: Principle Two — Graduated Trust

1.10.1 The Principle

The level of verification should match the level of risk. Not all code deserves equal scrutiny.

This principle provides efficiency. Verifying everything deeply is impractical. Verifying nothing is dangerous. VID graduates verification based on risk.

1.10.2 Understanding Risk

Risk in software has multiple dimensions:

Impact if wrong: What happens if this code has a bug? - A typo in a log message: minimal impact - A bug in payment processing: financial loss, customer trust damage - A security flaw in authentication: potential breach

Reversibility: How easily can we recover from problems? - A bug in a stateless API: fix and redeploy - Data corruption: may require complex recovery - Deleted customer data: potentially unrecoverable

Exposure: How widely is this code used? - Internal tool for one team: limited exposure - Core library used everywhere: bugs affect everything - Public API: bugs affect external parties

Regulatory sensitivity: What compliance requirements apply? - Internal tooling: minimal regulatory concern - Healthcare data: HIPAA requirements - Financial transactions: PCI-DSS requirements

1.10.3 The Trust Spectrum

VID defines a spectrum of trust levels:

High Trust (minimal verification): AI output accepted with automated checks only. Appropriate for low-risk, easily reversible code.

Moderate Trust (standard verification): AI output reviewed for functional correctness. Appropriate for typical production code.

Guarded Trust (thorough verification): AI output reviewed in depth for correctness, security, and maintainability. Appropriate for important code.

Minimal Trust (intensive verification): AI used only for suggestions. Human writes or extensively rewrites all code. Multiple reviewers. Appropriate for high-risk code.

1.10.4 Scoring Before Trusting

Before selecting a trust level, score the work using the VID risk formula (detailed rubric in Appendix D):

```
Risk Score = (Impact × 3) + (Reversibility × 2) + (Exposure × 2) + Compliance
```

Dimension	Scale	Max Contribution
Impact	1-5	15
Reversibility	1-5	10
Exposure	1-5	10
Compliance	0-10	10

Trust levels map to total scores: - **0-10:** High Trust - **11-20:** Moderate Trust - **21-30:** Guarded Trust - **31-47:** Minimal Trust

Escalate to the next level if Impact, Reversibility, or Exposure scores 4 or higher, OR if Compliance is 6+ (regulated domains). The quick calculation keeps risk discussions objective in the moment while Appendix D provides the deeper rubric.

1.10.5 Calibrating Trust

For any code generation, ask:

1. What's the worst case if this code is wrong?
2. How easily can we recover from problems?
3. How widely is this code exposed?
4. What compliance requirements apply?

The highest-risk answer determines your trust level.

Example calibrations:

Utility function for formatting dates: - Worst case: Dates display incorrectly - Recovery: Easy redeploy - Exposure: Internal dashboard - Compliance: None - **Calibration: High Trust**

Input validation for a public API: - Worst case: Security vulnerability - Recovery: Depends on exploitation - Exposure: Internet-facing - Compliance: Potential - **Calibration: Minimal Trust**

1.10.6 The Trap of Uniform Treatment

Teams often fall into treating all code the same:

Uniform high scrutiny: Every function gets intensive review. This is unsustainable. Reviews become cursory because there's too much to review carefully.

Uniform low scrutiny: Everything gets quick acceptance. Risk accumulates invisibly until it manifests as incidents.

VID's graduated approach focuses human attention where it matters. It's not about less verification overall — it's about appropriate verification everywhere.

1.11 Chapter 6: Principle Three — Understanding Over Acceptance

1.11.1 The Principle

Never accept code you don't understand at an appropriate level of depth.

This is perhaps VID's most challenging principle. It requires discipline when AI makes acceptance frictionless.

1.11.2 The Appropriate Level

"Understanding" doesn't mean comprehending every character. It means understanding at a level appropriate to:

Your role: A senior engineer reviewing should understand more deeply than a junior engineer using a well-tested utility.

The code's risk: High-risk code requires deeper understanding. Low-risk code can be understood more superficially.

The code's stability: Code that will be modified requires deeper understanding than code that will be called but not changed.

1.11.3 Levels of Understanding

Surface understanding: "This function takes X and returns Y. I can use it correctly."

Appropriate for: Low-risk library functions, well-tested utilities, code you call but won't modify.

Functional understanding: "I understand the algorithm and why it produces correct results for various inputs."

Appropriate for: Most production code, code you might need to debug or modify.

Deep understanding: "I understand the implementation details, edge cases, performance characteristics, and security implications."

Appropriate for: High-risk code, security-sensitive code, code you'll maintain long-term.

1.11.4 Building Understanding

When AI generates code, actively build understanding:

1. **Read the code.** Not skimming — actually reading. Trace the logic.
2. **Question the choices.** Why this approach? What are the alternatives? What are the trade-offs?
3. **Test your understanding.** Can you explain what this code does without looking at it? Can you predict what happens with edge cases?
4. **Identify gaps.** What parts don't you understand? Are those acceptable gaps or do they need resolution?

1.11.5 The Understanding Debt

Accepting code you don't understand creates "understanding debt" — code that works but no one understands. This debt compounds:

- Debugging becomes harder
- Modifications become scary
- The codebase becomes a minefield of mystery code

Unlike technical debt, understanding debt can't be measured by tools. It only reveals itself when someone needs to modify code and discovers they can't.

VID prevents understanding debt through discipline: if you don't understand it at an appropriate level, you're not done.

1.12 Chapter 7: Principle Four — Provenance Awareness

1.12.1 The Principle

Always know where code came from and what that origin implies.

Code has provenance — a history of where it came from. In traditional development, provenance is straightforward: a human wrote it. In AI-augmented development, provenance becomes complex and important.

1.12.2 Why Provenance Matters

Debugging: When code fails, knowing its origin helps debug. AI-generated code might fail in ways specific to how AI generates code.

Modification: Understanding provenance helps assess modification risk. Code that no one understood when it was generated is harder to modify safely.

Legal protection: AI-generated code might inadvertently derive from copyrighted sources. Understanding provenance helps assess and manage this risk.

Quality assessment: Different origins have different quality characteristics. Knowing the origin helps calibrate verification.

1.12.3 Provenance Categories

VID recognizes several provenance categories:

Human original: Code written entirely by a human developer. Lowest risk — the developer understood what they wrote.

AI generated, human verified: Code generated by AI and carefully reviewed by a human who understood it. Risk depends on verification depth.

AI generated, lightly reviewed: Code generated by AI with superficial review. Higher risk — the human might not have caught subtle issues.

AI generated, unreviewed: Code accepted without meaningful review. Highest risk.

Mixed provenance: Code that combines human-written and AI-generated portions. Risk depends on how well the integration was verified.

1.12.4 Tracking Provenance

Provenance should be trackable:

- Can you identify which code in your codebase was AI-generated?
- Can you identify who reviewed it and how deeply?
- Can you identify what prompt generated it?

This tracking doesn't need to be elaborate. At minimum: - Commit messages should indicate AI-assisted generation - Code review records should indicate review depth - Teams should maintain awareness of which areas are heavily AI-generated

1.12.5 Using Provenance

Provenance informs decisions:

Incident investigation: "This function is failing. It was AI-generated and received minimal review. The original developer didn't understand the edge cases."

Modification planning: "This module is mostly AI-generated with light review. We should budget extra time for understanding before modification."

Risk assessment: "This area of the codebase has extensive AI generation with minimal documentation. It's higher risk than areas with clear provenance."

1.13 Chapter 8: Principle Five — Continuous Calibration

1.13.1 The Principle

Regularly assess whether your verification practices match actual risk and adjust accordingly.

This principle prevents VID from becoming rigid. No methodology gets calibration right initially. Continuous calibration adapts to reality.

1.13.2 The Feedback Loop

VID requires feedback from outcomes:

Positive feedback: Code that passed verification performs well in production. This validates the verification approach.

Negative feedback: Code that passed verification fails in production. This indicates verification was insufficient.

Both are valuable. The goal is verification that catches problems without excessive overhead — this requires empirical calibration.

1.13.3 Calibration Questions

Regularly ask:

Are we catching problems? - How often does verification reject code? - What kinds of problems are caught? - If verification rarely rejects anything, it might be too lenient.

Are we missing problems? - What problems escape to production? - Were they the kind verification should catch? - If problems consistently escape, verification needs strengthening.

Is verification effort appropriate? - Are we spending appropriate time on high-risk code? - Are we spending excessive time on low-risk code? - Adjust risk calibration if effort is misallocated.

Are our risk assessments accurate? - When problems occur, were they in code we classified as high-risk? - If high-risk problems come from "low-risk" code, our classification is wrong.

1.13.4 Calibration Triggers

Recalibrate when:

- Production incidents occur
- New types of problems appear
- AI capabilities change (new models, new tools)
- Team composition changes
- Regulatory requirements change
- Project characteristics change

1.13.5 The Anti-Pattern: Frozen Methodology

Some teams adopt VID (or any methodology) and never adjust. They follow practices that made sense initially but no longer fit.

Warning signs: - "We've always done it this way" - Verification levels that haven't changed in months - No discussion of whether practices are working - No adaptation to new AI capabilities

VID is a living methodology. It should evolve with your team and the technology.

1.14 Chapter 9: The Intent Specification Practice

1.14.1 From Principle to Practice

Principle One (Intent Before Generation) requires practical implementation. The intent specification practice provides it.

1.14.2 The Practice

Before generating code:

1. **Articulate the functional requirement.** What should this code do? Be specific about inputs, outputs, and transformation.
2. **Identify boundaries.** What inputs are valid? What should happen at boundaries? What should happen with invalid inputs?
3. **State success criteria.** How will you know this code is correct? What tests or checks will verify it?
4. **Assess risk.** What's the impact if this is wrong? This determines verification depth.

1.14.3 Intent Specification Formats

Intent specification can range from informal to formal:

Mental intent (lowest formality): For trivial code, a clear thought is sufficient: "I need a function that capitalizes the first letter of each word."

Comment intent: Write a comment describing the intent before generating:

```
// Function: Capitalize first letter of each word
// Input: String of words separated by spaces
// Output: Same string with each word capitalized
// Edge cases: Empty string returns empty, handles multiple spaces
```

Test-first intent: Write the tests before generating the implementation:

```
def test_capitalizes_words():
    assert capitalize_words("hello world") == "Hello World"

def test_handles_empty():
    assert capitalize_words("") == ""

def test_handles_single_word():
    assert capitalize_words("hello") == "Hello"
```

Formal specification: For complex or high-risk code, use structured specification documents. Here's a template:

```
## Intent Specification: [Feature/Function Name]
```

```
**Author:** [Your name]
**Date:** [Date]
**Risk Level:** [High Trust / Moderate / Guarded / Minimal]
**Risk Score:** [0-47] (See Appendix D)

### Purpose
[1-2 sentences: What problem does this code solve?]

### Functional Requirements
1. [Specific requirement with measurable outcome]
2. [Specific requirement with measurable outcome]
3. [...]

### Input Specification
- **Valid inputs:** [Types, ranges, formats]
- **Invalid inputs:** [What should be rejected and how]
- **Edge cases:** [Boundary values, empty inputs, maximum sizes]

### Output Specification
- **Success case:** [What the code returns/produces on success]
- **Error cases:** [What happens for each error condition]
- **Side effects:** [Database changes, API calls, file writes, etc.]

### Non-Functional Requirements
- **Performance:** [Response time, throughput requirements]
- **Security:** [Authentication, authorization, data protection]
- **Compliance:** [GDPR, HIPAA, SOX, etc.]
- **Scalability:** [Expected load, growth projections]

### Success Criteria
**The code is correct if:**
1. [Testable criterion]
2. [Testable criterion]
3. [...]

### Verification Plan
**Based on risk level [X], I will:**
- [ ] [Verification step matching trust level]
- [ ] [Verification step matching trust level]
- [ ] [Peer review required? Yes/No]
- [ ] [Security review required? Yes/No]

### Dependencies & Integration
- **Depends on:** [Other systems, services, libraries]
- **Used by:** [Callers, consumers]
- **Breaking changes:** [Any backwards compatibility concerns]
```

```
### Assumptions & Constraints
- [Assumption 1]
- [Constraint 1]
```

Example - Formal Specification:

```
## Intent Specification: User Authentication Middleware

**Author:** Sarah Chen
**Date:** 2025-01-02
**Risk Level:** Guarded Trust
**Risk Score:** 26 (Impact: 5, Reversibility: 3, Exposure: 4, Compliance: 2)

### Purpose
Authenticate incoming HTTP requests using JWT tokens, attach user context to request object, ar

### Functional Requirements
1. Extract JWT token from Authorization header (Bearer scheme)
2. Validate token signature using RS256 algorithm
3. Verify token has not expired (exp claim)
4. Extract user ID from token payload (sub claim)
5. Attach validated user object to request.user
6. Return 401 Unauthorized for invalid/missing tokens
7. Return 403 Forbidden for expired tokens

### Input Specification
- **Valid inputs:**
  - Authorization: Bearer [valid-jwt-token]
  - Token format: Header.Payload.Signature (RS256)
  - Token payload must include: sub (user ID), exp (expiration), iat (issued at)
- **Invalid inputs:**
  - Missing Authorization header → 401
  - Malformed token → 401
  - Invalid signature → 401
  - Expired token → 403
  - Token from revoked user → 403
- **Edge cases:**
  - Token expires during request processing
  - Clock skew between auth server and app server (allow 60s tolerance)
  - Very long tokens (>8KB should be rejected)

### Output Specification
- **Success case:**
  - request.user = {id, email, roles, permissions}
  - next() called to continue request
- **Error cases:**
  - 401: {error: "invalid_token", message: "..."}
  - 403: {error: "token_expired", message: "...", expired_at: timestamp}
```

```
- **Side effects:**
  - Log authentication attempts (successful and failed)
  - Increment Prometheus counter for auth_attempts_total{status="success|failure"}

### Non-Functional Requirements
- **Performance:** <10ms p99 latency (token validation is synchronous)
- **Security:**
  - Must use constant-time comparison for signatures (prevent timing attacks)
  - Must validate ALL required claims before granting access
  - Must not log tokens or sensitive user data
  - Public key must be loaded from secure key vault, not hardcoded
- **Compliance:**
  - PCI DSS (for payment routes): Must log all access attempts
- **Scalability:**
  - Stateless (no session storage)
  - Must handle 10,000 req/sec per instance

### Success Criteria
**The code is correct if:**
1. Valid tokens grant access and attach correct user context
2. Invalid tokens are rejected with appropriate status codes
3. Expired tokens return 403, not 401
4. Token signature validation uses constant-time comparison
5. No tokens or sensitive data appear in logs
6. Performance meets <10ms p99 requirement under load
7. All error cases return helpful messages without leaking security details

### Verification Plan
**Based on risk level Guarded Trust, I will:**
- [ ] Line-by-line code review of authentication logic
- [ ] Test all success and error paths with unit tests
- [ ] Security review: Check for timing attacks, token leakage in logs
- [ ] Integration test with real JWT tokens (valid, expired, malformed)
- [ ] Load test to verify <10ms p99 under 10k req/sec
- [ ] Peer review by security-focused senior engineer
- [ ] Code review checklist from Appendix F

### Dependencies & Integration
- **Depends on:**
  - jsonwebtoken library (v9.0.0+)
  - Key vault service for public key retrieval
  - User service for checking revoked users
- **Used by:**
  - All protected API routes (/api/*)
  - Must be registered before route handlers
- **Breaking changes:**
  - Changing token format will break all mobile clients
```

```
  - Must version any changes to claims structure

### Assumptions & Constraints
- Tokens are signed by trusted auth service using RS256
- Public key rotates monthly but old keys remain valid for 7 days
- Maximum of 3 concurrent sessions per user
- Token lifetime is 1 hour (non-configurable)
```

1.14.4 Matching Formality to Risk

- Trivial code: Mental intent is sufficient
- Typical code: Comment intent or test-first intent
- Important code: Test-first intent with edge cases
- Critical code: Formal specification with comprehensive test cases

1.14.5 Anti-Pattern: Retroactive Intent

Writing intent after generation defeats the purpose. If you generate first, you'll rationalize that the output matches your intent rather than critically evaluating it.

The discipline is: intent first, always.

1.15 Chapter 10: The Verification Ritual Practice

1.15.1 The Practice

After every code generation, perform a verification ritual. The depth varies by risk, but the practice is consistent.

1.15.2 The High Trust Ritual

For low-risk code (risk score 0-10):

1. **Read the code.** Actually read it, don't skim.
2. **Verify against intent.** Does this do what you specified?
3. **Run automated checks.** Linting, type checking, tests.
4. **Ask: "What could go wrong?"** Spend 30 seconds considering failure modes.

This takes 5-10 minutes. It's the minimum for any code.

1.15.3 The Moderate Trust Ritual

For typical production code (risk score 11-20):

1. **Complete the High Trust ritual.**
2. **Trace the logic.** Walk through the code path mentally.
3. **Test edge cases.** Run specific tests for boundary conditions.
4. **Check integration.** How does this interact with existing code?

5. **Document understanding.** Write a brief note on what this code does.

This takes 15-30 minutes depending on code complexity.

1.15.4 The Guarded Trust Ritual

For important code (risk score 21-30):

1. **Complete the Moderate Trust ritual.**
2. **Adversarial thinking.** Try to break the code. What inputs could cause problems?
3. **Security review.** Are there injection risks? Access control issues? Data exposure?
4. **Performance consideration.** What happens at scale? Are there expensive operations?
5. **Maintainability review.** Can someone else understand this? Should it be refactored for clarity?

This takes 30-60 minutes.

1.15.5 The Minimal Trust Ritual

For high-risk code (risk score 31-47):

1. **Complete the Guarded Trust ritual.**
2. **Independent review.** Have another person review independently.
3. **Formal verification.** Apply relevant security tools, static analysis, etc.
4. **Documentation.** Document the code's behavior, limitations, and security considerations.
5. **Provenance documentation.** Record how this code was generated and verified.

This takes 1-3+ hours and involves multiple people.

1.15.6 Making It a Ritual

The value of a ritual is consistency. It shouldn't require decision-making about whether to do it. After every generation:

"What's the risk level? Apply that verification ritual."

Make it automatic.

1.16 Chapter 11: The Learning Loop Practice

1.16.1 The Practice

Track what happens to your verified code. Use outcomes to improve verification.

1.16.2 The Tracking

Record:

- **What was generated and when**
- **What risk level was assigned**
- **What verification was performed**

- **What happened in production** (if problems occurred)

This doesn't require elaborate tooling. A simple log or notes suffice.

1.16.3 The Review

Periodically (weekly or monthly) review:

- **Problems that escaped verification:** What problems made it to production? Should verification have caught them? Why didn't it?
- **Verification that seems excessive:** Are you spending significant time verifying code that never has problems? Can that verification be reduced?
- **Risk calibration accuracy:** Do your "high risk" and "low risk" assessments match reality?

1.16.4 The Adjustment

Based on review, adjust:

- **Verification depth:** Increase for areas with escaping problems, decrease for areas with no problems
- **Risk calibration:** Update risk assessment criteria based on actual outcomes
- **Practices:** Add or remove verification steps based on what's working

1.16.5 The Anti-Pattern: No Feedback

Teams that don't track outcomes can't calibrate. They might: - Over-verify some areas while under-verifying others - Miss entire categories of problems - Stick with practices that don't work

The learning loop prevents methodology ossification.

1.17 Chapter 12: The Provenance Hygiene Practice

1.17.1 The Practice

Maintain awareness and documentation of where code comes from.

1.17.2 The Scaling Challenge

A common objection to provenance tracking: "This sounds good in theory, but marking code as AI-generated in commits and documentation becomes chaotic once the codebase reaches a certain size."

This is a valid concern. Manual discipline doesn't scale. But provenance tracking doesn't have to be purely manual.

The key insight: **provenance tracking should evolve with your team and codebase size**.

1.17.3 Provenance Tracking: Maturity Levels

Different teams need different approaches:

1.17.3.1 Level 1: Manual Commit Messages

Good for: Teams < 5 people, codebases < 10,000 LOC

Simple commit message conventions:

```
feat: add OAuth login (AI-generated, verified)
fix: handle edge case in parser (human-written)
```

This works when: - Team discipline is high - Code review is thorough - The codebase is small enough to remember what's what

Limitations: Breaks down with team growth, easy to forget, hard to audit.

1.17.3.2 Level 2: Structured Metadata

Good for: Teams 5-20 people, codebases 10,000-100,000 LOC

When manual tracking becomes error-prone, add structure:

Git commit templates that prompt for provenance **Code comment conventions** that are greppable **Periodic manual audits** to verify metadata accuracy

Example structured approach:

```
// @provenance ai-generated
// @tool gpt-4
// @verification deep
// @reviewer alice@example.com
// Human intent: Implement token refresh with exponential backoff
// Known limitations: Doesn't handle network partitions > 30s
export async function refreshToken(token: string) {
  // ...
}
```

This works when: - You need programmatic queries ("show me all AI-generated code with light verification") - Multiple people touch the same code - Incident investigation requires provenance data

Limitations: Requires consistent adoption, metadata can drift from reality over time.

1.17.3.3 Level 3: Automated Tracking

Good for: Teams 20+ people, codebases 100,000+ LOC

When you need programmatic enforcement:

Git hooks that prompt for provenance before commit **Metadata files** maintained automatically by tooling **Continuous audit scripts** that run in CI **Provenance dashboards** for team visibility

This works when: - Manual processes have failed - Compliance or legal requires audit trails - The codebase is too large for manual tracking

Limitations: Requires tooling investment, can be seen as process overhead.

1.17.4 Principles for Scalable Provenance Tracking

Regardless of maturity level, effective provenance tracking follows these principles:

1.17.4.1 Make It Hard to Forget

Don't rely on memory. Use automation:

- **Git hooks** that ask "Was this AI-generated?" before allowing commits
- **PR templates** with provenance checklist sections
- **CI checks** that warn when provenance markers are missing
- **Editor snippets** that make adding provenance metadata trivial

The goal: forgetting to document provenance should feel like friction.

1.17.4.2 Make It Easy to Query

Structure metadata for programmatic access:

- **Use consistent, greppable markers** in code comments
- **Maintain a single source of truth** (e.g., `.provenance.yml` or similar)
- **Provide scripts** to answer questions like:
  - "Which files are AI-generated with minimal verification?"
  - "What percentage of our codebase is AI-generated?"
  - "Which modules have unknown provenance?"

If you can't query it, you can't audit it.

1.17.4.3 Make It Maintainable

Automation should help, not burden:

- **Automation updates metadata**, not humans editing files manually
- **Provenance should degrade gracefully** (missing metadata = "unknown", not broken builds)
- **Periodic audits should be scheduled**, not ad-hoc fire drills
- **Keep it simple** - complex systems that nobody understands won't be maintained

1.17.4.4 Make It Culturally Normal

The hardest part isn't technical:

- **Lead by example** - senior engineers document provenance visibly
- **Make it part of code review** - reviewers ask "What's the provenance?" if unclear
- **Share provenance in incident reviews** - "This bug was in AI-generated code that had light verification"
- **Don't use it punitively** - provenance is for learning and risk management, not blame

1.17.5 Commit Hygiene

When committing AI-generated or AI-assisted code:

Indicate AI involvement. Your commit message or description should indicate that AI assisted in generating this code.

Note the verification level. Briefly indicate how this code was verified.

Document significant AI interactions. For complex generations, keep the prompt/response history.

Example commit message:

```
feat: implement JWT token refresh logic

Added automatic token refresh with exponential backoff retry.

Provenance:
- AI-generated initial implementation
- Human verification: deep (tested edge cases, reviewed algorithm)
- Tool: GPT-4 via Copilot
- Files: src/auth/token-refresh.ts, src/auth/token-refresh.test.ts
```

1.17.6 Code Comments

For non-trivial AI-generated code:

Document the intent. What was this code meant to do?

Note limitations. What does the AI-generated version not handle that a more complete implementation might?

Mark uncertainty. If you're uncertain about parts of the code, note that for future maintainers.

Use greppable markers. Structured comments enable programmatic auditing:

```python
# @provenance ai-generated
# @verification standard
# Human intent: Parse ISO 8601 dates with timezone support
# Known limitations: Doesn't handle leap seconds
def parse_datetime(date_string: str) -> datetime:
    # ...
```

1.17.7 Team Awareness

Teams should have shared awareness:

Which areas are heavily AI-generated? This affects modification planning and risk assessment.

What AI tools were used? Different tools have different characteristics.

What verification standards were applied? This affects confidence in different areas of the codebase.

How to maintain awareness:

- Regular team syncs that review provenance metrics
- Documentation that highlights AI-heavy modules
- Onboarding materials that explain provenance conventions
- Dashboards or reports showing provenance distribution

1.17.8 Practical Implementation

The `templates/` directory in this repository contains ready-to-use tools:

Git templates: - `git-commit-template.txt` - Commit message template with provenance prompts - `git-hooks/prepare-commit-msg` - Hook that adds provenance section to commits - `git-hooks/post-commit` - Hook for automated metadata tracking

Code templates: - `provenance-markers.code-snippets` - VS Code snippets for quick provenance comments

Audit tools: - `scripts/audit-provenance.sh` - Generate provenance reports from your codebase - `scripts/check-verification.sh` - Find high-risk AI-generated code

PR templates: - `pr-template.md` - Pull request template with provenance checklist

See the templates directory for implementation details and usage instructions.

1.17.9 The Anti-Pattern: Invisible AI

When AI-generated code is indistinguishable from human-written code:

- Future maintainers don't know to be cautious
- Problems are harder to diagnose
- Risk assessment is wrong because provenance is unknown

Make provenance visible.

1.17.10 The Anti-Pattern: Overly Complex Metadata

The opposite extreme is also problematic:

Complex JSON schemas that require manual editing:

```
// DON'T DO THIS - too complex, won't be maintained
{
  "files": {
    "src/feature.ts": {
      "provenance": "ai-generated",
      "tool": "gpt-4",
      "model_version": "gpt-4-0613",
      "temperature": 0.7,
      "prompt_hash": "sha256:abc123...",
      "verification": {
        "level": "deep",
        "reviewer": "alice@example.com",
        "date": "2024-01-15T10:30:00Z",
        "tests_added": true,
        "edge_cases_verified": ["null", "empty", "overflow"]
      }
    }
  }
}
```

This fails because: - Too much overhead to maintain - Requires tools that may not exist - Breaks when forgotten or outdated - Creates false confidence ("we have metadata!") when it's actually stale

Instead: Keep it simple. A greppable comment is better than an unmaintained JSON file.

1.17.11 Starting Point

If you're introducing provenance tracking to an existing team:

Week 1: Add a commit message template with provenance prompt. Make it optional.

Week 2-4: Lead by example - senior engineers document provenance in their commits and code.

Month 2: Make provenance documentation part of code review checklist.

Month 3: Add simple audit script to understand current state.

Month 6: Evaluate whether you need automated enforcement (git hooks, CI checks).

Start small. Build habits. Automate when manual processes break down.

1.18 Chapter 13: The Junior Engineer's Path to Verification Mastery

1.18.1 The Paradox You Face

You're learning to code at a strange moment in history. AI can generate code faster than you can type it. This creates a paradox:

The shortcut that helps you today will cripple you tomorrow.

If you let AI generate code you don't understand, you'll ship features quickly. You'll look productive. But you won't be learning. And when things break—they always break—you won't be able to fix them. You'll become dependent on AI without developing the judgment to know when AI is wrong.

The engineers who will thrive learned verification *first*. They used AI as a learning accelerator, not a learning bypass.

This chapter provides a path to becoming that engineer.

1.18.2 The 12-Week Curriculum

This curriculum assumes you have basic programming knowledge. You can write simple programs. You understand variables, functions, loops, and conditionals. You may have used AI coding assistants casually.

The goal is to transform you from someone who *uses* AI to generate code into someone who can *verify* AI-generated code reliably—and in the process, become a much stronger engineer than you would have been otherwise.

1.19 Phase 1: Foundations (Weeks 1-3)

1.19.1 Week 1: Learning to Read Code Critically

The Goal: Develop the habit of reading code carefully before accepting it.

The Problem You're Solving: Most developers skim code. They look for obvious errors and move on. This works when you wrote the code yourself (you already understand it). It fails catastrophically with AI-generated code.

Daily Practice (30-45 minutes):

Take a small piece of AI-generated code (one function, 10-30 lines) and analyze it completely.

1. **Read it aloud.** Literally speak each line. This forces you to actually read rather than skim.
2. **Trace execution.** Pick a specific input. Walk through line by line. What value does each variable have at each step? Write it down.
3. **Identify assumptions.** What does this code assume about its inputs? What happens if those assumptions are violated?
4. **Find the edges.** What inputs would cause this code to behave unexpectedly? Empty inputs? Very large inputs? Negative numbers? Null values?
5. **Explain it.** Write a 2-3 sentence explanation of what this code does and how. If you can't explain it clearly, you don't understand it.

Exercise 1.1: The Explanation Challenge

Generate 5 simple functions using AI (string manipulation, basic math, list operations). For each: - Write your explanation *before* running any tests - Then test with various inputs - Compare actual behavior to your explanation - Note any surprises

Track: How often did the code do exactly what you expected? Where were you surprised?

Exercise 1.2: Bug Hunting

Ask AI to generate functions with intentional prompts that might produce subtle bugs: - "Write a function to calculate average" (what if the list is empty?) - "Write a function to find the second-largest number" (what if there are duplicates?) - "Write a function to check if a string is a palindrome" (what about case? spaces?)

Your job: Find the bugs before running the code. Then verify.

Exercise 1.3: The Trace Journal

Keep a journal this week. For every AI-generated function you use: - Write the input/output trace for at least 2 inputs - Note one edge case you identified - Record whether your analysis found any issues

End of Week Assessment:

You should be able to: - Read a 20-line function and explain what it does without running it - Identify at least 2 edge cases for any function - Predict output for specific inputs before testing

Common Mistakes This Week:

The most common mistake beginners make is skimming code instead of truly reading it. If you can't explain what each line does, you skimmed rather than read. Force yourself to slow down and actually process what the code says.

Another trap is substituting testing for tracing. Running the code and seeing it produce correct output isn't the same as understanding how it works. Trace execution manually first, predicting what will happen, then test to verify your understanding.

Finally, resist accepting "it works" as sufficient understanding. Code that passes tests isn't necessarily understood code. You need to know not just that it works, but why it works and when it might not work.

1.19.2 Week 2: Intent Specification Practice

The Goal: Build the habit of knowing what you want before you ask AI for it.

The Problem You're Solving: Most AI interactions start with vague prompts. "Write a function to handle dates." Handle how? What dates? What should happen with invalid dates? Vague prompts produce code that might not match your actual need—and you won't notice because you weren't clear about your need.

Daily Practice (30-45 minutes):

Before any code generation:

1. **Write the function signature first.** What's the name? What are the parameters and their types? What's the return type?
2. **Write example calls.** Give 3-5 concrete examples of how you'd call this function and what it should return.
3. **Write the edge cases.** What happens with empty input? Null? Invalid values? Extremely large values?
4. **Write the error cases.** When should this function fail? How should it fail?

Only after writing all this do you prompt the AI.

Exercise 2.1: The Spec-First Challenge

Choose 5 programming problems (sorting, searching, string manipulation, etc.). For each:

1. Write a complete specification (signature, examples, edges, errors) — spend at least 10 minutes
2. Then prompt AI to implement it
3. Verify the implementation against your specification
4. Note: Did the AI match your spec? Where did it differ?

Exercise 2.2: Spec Critique

Look at these vague intents. Rewrite each as a complete specification:

Vague: "Function to validate email" - What makes an email valid? - What should return type be? Boolean? Error message? - What about edge cases like very long emails, unicode characters,

multiple @ signs?

Vague: "Function to split a bill" - Split how? Evenly? By item? By percentage? - What about tax? Tip? - What if amounts don't divide evenly? - What's the input format? Output format?

Vague: "Function to check password strength" - What criteria? Length? Character types? Common passwords? - What's the output? Boolean? Score? Specific feedback? - What are the thresholds?

Write complete specs for each. Compare with a peer if possible.

Exercise 2.3: The Retroactive Spec Test

Find code you wrote (or AI generated) last week without a clear spec. Now write the spec that *should* have existed. Then check: does the code actually meet this spec? Often you'll find it doesn't—because you weren't clear about what you needed.

End of Week Assessment:

You should be able to: - Write a complete spec (signature, examples, edges, errors) in 10 minutes - Identify 3+ ambiguities in any vague requirement - Verify implementation against spec systematically

Common Mistakes This Week:

Vague specifications plague beginners. Writing "Handle errors appropriately" isn't a specification—it's wishful thinking. How should errors be handled? Which errors? What's appropriate? Force yourself to be concrete and specific.

Edge cases are easy to forget when you're focused on the happy path. Make it a habit: empty inputs, null values, very large numbers, very small numbers, duplicates, negative values—run through this mental checklist for every specification you write.

The strongest temptation is skipping straight to prompting the AI. You know what you want (sort of), so why not just ask? Because vague prompts produce vague code. Resist the urge. Specification first, generation second, always.

1.19.3 Week 3: Calibrating Trust

The Goal: Develop judgment about when to verify deeply vs. when to verify lightly.

The Problem You're Solving: You can't verify everything deeply—there isn't time. But you can't verify everything lightly—too many bugs will escape. You need judgment about what deserves attention.

Daily Practice (30-45 minutes):

For every piece of code you work with:

1. **Assess impact.** What's the worst case if this code is wrong? Categorize: trivial / annoying / problematic / serious / severe
2. **Assess reversibility.** If this breaks in production, how easily can we fix it? Categorize: instant / easy / difficult / very hard / impossible

3. **Assess exposure.** Who/what is affected? Just me / my team / internal users / external users / everyone

4. **Decide verification depth.** Based on assessments, choose: quick check / standard review / deep analysis / intensive verification

Exercise 3.1: Risk Calibration Scenarios

For each scenario below, assess impact (what's the worst that could happen?), reversibility (how easily can we fix it if it breaks?), and exposure (who's affected?). Based on your assessment, decide how much verification each deserves.

Quick Reference: Scenarios at a Glance

Before reading the detailed analysis, try scoring these yourself:

#	Scenario	Your Impact (1-5)	Your Reversibility (1-5)	Your Exposure (1-5)	Your Trust Level
1	Logging function for debug messages	?	?	?	?
2	Input validation for public signup form	?	?	?	?
3	Calculation for displaying shipping dates	?	?	?	?
4	Password hashing for authentication	?	?	?	?
5	Archive old records to cold storage	?	?	?	?
6	Generate unique IDs for database records	?	?	?	?

#	Scenario	Your Impact (1-5)	Your Reversibility (1-5)	Your Exposure (1-5)	Your Trust Level
7	Endpoint that deletes user accounts	?	?	?	?
8	Currency formatting for display	?	?	?	?
9	Rate limiting logic for API	?	?	?	?
10	Report generator for internal dashboards	?	?	?	?

Fill this out before reading further. Then compare your assessments with the detailed analysis below.

Detailed Scenario Analysis:

Consider a **logging function that formats debug messages**. If this breaks, what happens? Debug output looks weird. Can you fix it quickly? Absolutely—change the code and redeploy. Who's affected? Primarily developers. This is genuinely lower risk, though you should still verify the basics to avoid wasting developer time on broken logs.

Now contrast that with **input validation for a public signup form**. If this breaks? Attackers might register with malicious data, or legitimate users might be blocked. Can you fix it easily? Yes, but damage might already be done. Who's affected? All potential users, and possibly your production database. This demands careful verification—security implications, edge cases, error handling.

Consider **password hashing for user authentication**. If this is wrong, user accounts could be compromised. Reversibility is irrelevant—the damage is permanent the moment someone's password is exposed. Exposure is total: every user. This requires intensive verification. You need to understand exactly how the hashing works, verify it matches security best practices, check for timing attacks, ensure salts are unique.

Think about **an endpoint that deletes user accounts**. This is a destructive action. If it accidentally deletes the wrong accounts, that data is gone. If it lacks proper authorization checks, any user could delete any account. Impact: severe. Reversibility: difficult or impossible without backups. Exposure: potentially all users. This is high-risk code that demands intensive verification.

Contrast with **a function that formats currency for display**. Wrong output is visible but not destructive. Users might see "$1234" instead of "$1,234.00"—annoying but fixable immediately. Lower risk, lighter verification, though you should still check edge cases like very large numbers and different currencies.

After working through these scenarios yourself, compare your assessments with this guide:

Risk Calibration Summary Table

Scenario	Impact	Reversibility	Exposure	Risk Score	Trust Level	Verification Time
1. Logging function	1-2 (minor)	1 (instant fix)	2 (developers)	6-8	High Trust	5-10 min
2. Input validation (signup)	4 (security/UX)	3 (damage done)	4 (all users)	26	Guarded Trust	30-60 min
3. Shipping date display	2 (UX issue)	2 (quick fix)	3 (customers)	12	Moderate Trust	15-30 min
4. Password hashing	5 (critical)	5 (irreversible)	5 (all users)	35	Minimal Trust	1-3+ hours
5. Archive old records	4 (data loss)	4 (very hard)	3 (internal)	26	Guarded Trust	30-60 min
6. Generate unique IDs	4 (data integrity)	4 (migrations)	4 (all users)	28	Guarded Trust	30-60 min
7. Delete user accounts	5 (data loss)	5 (impossible)	5 (users)	35	Minimal Trust	1-3+ hours
8. Currency formatting	2 (display bug)	1 (instant)	3 (customers)	10	High Trust	5-10 min
9. Rate limiting (API)	4 (DoS/abuse)	3 (redeploy)	4 (all users)	26	Guarded Trust	30-60 min
10. Internal reports	2 (wrong data)	2 (regenerate)	2 (internal)	10	High Trust	5-10 min

Formula Reminder: Risk Score = (Impact × 3) + (Reversibility × 2) + (Exposure × 2) + Compliance

Key Patterns: - **Destructive operations** (#4, #5, #7): Always high risk—data loss is irreversible - **Security boundaries** (#2, #4, #9): Require careful verification—mistakes have cascading effects - **Display/formatting** (#1, #3, #8, #10): Generally lower risk—annoying but fixable - **Data integrity** (#6): Moderate-to-high risk—affects every future operation

Exercise 3.2: Your Risk History

Think about code you wrote or used recently. For each piece: - What was your implicit risk assessment? - How much verification did you actually do? - In retrospect, was that appropriate?

Many developers discover they over-verify trivial code (it's easy) and under-verify important code (it's harder and they're tired).

Exercise 3.3: The Calibration Journal

This week, for every verification decision: - Record what you verified and how deeply - Record your reasoning - At end of week, review: Do your decisions make sense? Any patterns of over/under verification?

End of Week Assessment:

You should be able to: - Assess risk across impact, reversibility, and exposure quickly - Match verification depth to risk level consistently - Explain why you chose a particular verification depth

Common Mistakes This Week:

New engineers often treat everything as high risk because they lack confidence. But if everything is critical, nothing is. You must differentiate. A logging function genuinely carries less risk than authentication logic, and treating them the same wastes time and attention.

Familiarity bias cuts the other way: code that feels familiar feels safe, even when it's not. Unfamiliar patterns feel risky even when they're low-stakes. Remember that risk is about actual impact—what happens if this breaks—not about your comfort level with the code.

The most insidious mistake is rationalizing. Calling something "low risk" because you don't want to do the verification work is self-deception. Be ruthlessly honest with yourself about why you're making calibration decisions.

1.20 Phase 2: Building Verification Skills (Weeks 4-7)

You've built the foundations: you can read code critically, specify intent clearly, and calibrate verification depth to risk. Now you'll develop the specific verification skills that let you evaluate code across multiple dimensions. Each week in this phase focuses on a different aspect of verification—functional correctness, semantic alignment, security, and maintainability. By the end, you'll be able to assess code comprehensively, not just check that it runs without errors.

1.20.1 Week 4: Functional Verification

The Goal: Learn to verify that code does what it's supposed to do.

Cross-reference: This week introduces the functional verification techniques detailed in Chapter 20 (input partitioning, boundary analysis, category testing). Reference those sections when you need more depth.

The Problem You're Solving: "It works" is not verification. Code that works for one input might fail for others. Functional verification systematically checks that code meets its specification across its intended domain.

Daily Practice (45-60 minutes):

For each function you work with:

1. **Identify categories of inputs.** What are the meaningfully different kinds of inputs? (e.g., for a search: found vs. not found, empty list, one item, many items)
2. **Select representatives.** Pick at least one input from each category.
3. **Determine expected output.** Before running, what should the output be?
4. **Test and compare.** Run the code, compare to expectations.
5. **Investigate surprises.** If output differs from expectation, is your expectation wrong, or is the code wrong?

Exercise 4.1: Category Analysis

For each function type, list input categories that should be tested:

1. A function that finds the maximum value in a list
   - List with one item
   - List with multiple items (max at start, middle, end)
   - List with all same values
   - List with negative numbers
   - Empty list
   - List with very large numbers
2. A function that validates a phone number
   - (You identify the categories)
3. A function that calculates days between two dates
   - (You identify the categories)
4. A function that merges two sorted lists
   - (You identify the categories)

Exercise 4.2: Test Case Design

Take a function you've recently generated with AI. Write test cases that cover: - Normal operation (typical inputs) - Boundary conditions (edges of valid input ranges) - Error conditions (invalid inputs) - Special cases (empty, null, very large)

Aim for 8-12 test cases for a single function. Run them. Did any fail?

Exercise 4.3: The Mutation Game

This exercise builds intuition for what tests actually verify.

1. Take a working, tested function
2. Introduce a small bug intentionally (change a < to <=, change a + to -, off-by-one error)

3. Run your tests
4. Did any test catch the bug?

If no test caught it, your tests have a gap. What test would have caught it?

This teaches you that tests only verify what they test. Passing tests doesn't mean no bugs—it means no bugs *that the tests check for*.

End of Week Assessment:

You should be able to: - Identify 6+ input categories for any function - Design test cases that cover categories systematically - Recognize when tests are insufficient (gaps in coverage)

Common Mistakes This Week:

Beginners naturally focus on the happy path—testing only normal operation with typical inputs. This feels productive because tests pass. But most bugs live off the happy path, in edge cases and error conditions you didn't test.

Some developers react by throwing random inputs at code, hoping to find problems. This isn't systematic verification—it's chaos. Random testing might stumble on bugs, but it won't give you confidence. Categorize your inputs thoughtfully instead.

The most dangerous mistake is trusting test passage as proof of correctness. Tests only verify what they check. Passing tests mean the code works for the scenarios you tested, nothing more. The bugs you didn't test for can still be there, waiting.

1.20.2 Week 5: Semantic Verification

The Goal: Learn to verify that code does what was *intended*, not just what was *specified*.

> **Cross-reference:** This week builds on the semantic verification techniques in Chapter 20 (intent alignment, implicit expectations, context checking). Use those techniques as you practice.

The Problem You're Solving: Specifications are imperfect. They describe what we think we want, but we often have implicit expectations we didn't specify. Semantic verification bridges the gap between specification and true intent.

Daily Practice (45-60 minutes):

For each piece of functionality:

1. **Ask "why?"** Why does this code exist? What problem does it solve? What would the user expect?
2. **Check alignment.** Does the code's behavior align with the underlying purpose, not just the written spec?
3. **Consider context.** How will this be used in practice? Does behavior make sense in that context?
4. **Identify implicit expectations.** What behaviors would users/callers assume even if not specified?

Exercise 5.1: Intent vs. Spec Analysis

For each scenario, identify the gap between spec and intent:

Scenario 1: Spec: "Function returns user's age in years" Code: Returns the difference between current year and birth year Intent gap: What about someone born in December asked in January? They'd be "10" when they're actually still 9.

Scenario 2: Spec: "Function splits a string into words" Code: Splits on space character Intent gap: (What about multiple spaces? Tabs? Newlines?)

Scenario 3: Spec: "Function calculates shipping cost based on weight" Code: Returns weight * rate Intent gap: (What implicit expectations might exist?)

Scenario 4: Spec: "Function checks if username is available" Code: Returns true if username not in database Intent gap: (What about case sensitivity? Reserved names? SQL injection?)

Exercise 5.2: The User Story Test

Take a function you're working with. Write a brief user story: "As a [user type], I want to [action] so that [benefit]"

Now review the code through the user's eyes: - Does the behavior make sense for this user? - Would the user be surprised by any behavior? - What would the user assume that might not be true?

Exercise 5.3: The Peer Explanation

Find a peer (or imagine one). Explain what a piece of code is *for*, not just what it *does*. As you explain, you'll often notice gaps:

"This function validates emails… well, actually it just checks for an @ symbol… but for our purposes that's probably… hmm, actually someone could put 'x@x' and it would pass… is that okay?"

The act of explaining intent reveals gaps between code and purpose.

End of Week Assessment:

You should be able to: - Articulate the purpose behind any code you work with - Identify at least 2 implicit expectations for any function - Recognize when code technically meets spec but violates intent

Common Mistakes This Week:

Engineers trained on specifications naturally stop when code matches the spec. "It does what the spec says"—task complete! But specs are imperfect human artifacts. If the spec is incomplete or wrong, compliance with it isn't success.

Context gets ignored when you verify code in isolation. You check whether a function works on its own, forgetting that it exists within a system. Behavior that seems fine for an isolated function might be completely wrong in the context where it's actually used.

Perhaps the subtlest mistake is assuming your implicit expectations are shared by others. You think "of course it should handle unicode," but the person who wrote the spec assumed ASCII. Your assumptions and theirs diverge invisibly until something breaks. Make implicit expectations explicit.

1.20.3 Week 6: Security Verification

The Goal: Learn to identify security implications in code.

> **Cross-reference:** Pair this week with the security techniques in Chapter 20 (sections on adversarial testing and provenance hygiene) so you're practicing the same toolkit the rest of the organization uses.

The Problem You're Solving: AI generates code that works. It rarely generates code that's secure against adversarial input. Security thinking requires considering how code could be misused, not just how it's intended to be used.

Prerequisite Understanding:

Before this week, ensure you understand these concepts (research if needed): - Input validation and sanitization - Injection attacks (SQL, command, XSS) - Authentication vs. authorization - Data exposure risks - The principle of least privilege

Daily Practice (45-60 minutes):

For each piece of code:

1. **Identify inputs.** What data enters this code? From where?
2. **Assume hostility.** What if that input comes from an attacker? What's the worst they could do?
3. **Trace sensitive data.** If this code touches sensitive data, where does that data go? Could it be exposed?
4. **Check access control.** Does this code verify the caller has permission to do what they're asking?
5. **Look for trust boundaries.** Where does code assume input is safe? Are those assumptions valid?

Exercise 6.1: Attack Vector Identification

For each code pattern, identify potential attack vectors:

```
# Pattern 1: Database query
def get_user(username):
    query = f"SELECT * FROM users WHERE name = '{username}'"
    return db.execute(query)
```

Attack vectors: (SQL injection - what if username is "'; DROP TABLE users; –"?)

```
# Pattern 2: File access
def read_config(filename):
    with open(f"/app/config/{filename}") as f:
        return f.read()
```

Attack vectors: (What if filename is "../../../etc/passwd"?)

```
# Pattern 3: HTML output
def display_comment(comment):
    return f"<div class='comment'>{comment}</div>"
```

Attack vectors: (What if comment contains

?)

```
# Pattern 4: Command execution
def convert_image(input_path, output_path):
    os.system(f"convert {input_path} {output_path}")
```

Attack vectors: (What if paths contain shell metacharacters?)

Exercise 6.2: Security Review Practice

Take a piece of AI-generated code that handles user input. Perform a security review:

1. List every input to the code
2. For each input, list what an attacker could provide
3. For each attack input, trace what happens
4. Identify which attacks succeed and which are blocked
5. For successful attacks, what's the fix?

Exercise 6.3: The STRIDE Walkthrough

STRIDE is a threat modeling framework. For a function you're reviewing, consider each threat:

- **S**poofing: Could someone pretend to be someone else?
- **T**ampering: Could someone modify data they shouldn't?
- **R**epudiation: Could someone deny they did something?
- **I**nformation disclosure: Could sensitive data be exposed?
- **D**enial of service: Could someone make this unavailable?
- **E**levation of privilege: Could someone gain unauthorized access?

Not every threat applies to every function, but considering all six ensures you don't miss anything.

End of Week Assessment:

You should be able to: - Identify inputs and trust boundaries in any code - Recognize common vulnerability patterns (injection, path traversal, XSS) - Apply threat modeling framework to code review

Common Mistakes This Week:

The cardinal sin of security is assuming inputs are safe. Perhaps the data comes from your own frontend, so it must be trustworthy, right? Wrong. Never trust input. Ever. Not from users, not from your frontend, not from other internal services. Treat all input as potentially hostile.

Many developers treat security as an afterthought—something to think about after the code works. But security issues are vastly cheaper to fix early, and they're often architectural. Bolting security onto working code is harder than building it in from the start. Always consider security implications as you verify.

The most revealing mistake is the phrase "No one would do that." You see a potential attack vector but dismiss it as too unlikely, too weird, too much effort for an attacker. Attackers do unexpected things. That's literally their job. If an attack is possible, assume someone will try it.

1.20.4 Week 7: Maintainability Verification

The Goal: Learn to evaluate whether code can be understood and modified in the future.

> **Cross-reference:** This week applies the maintainability verification techniques from Chapter 20 (naming review, complexity assessment, documentation quality). Practice with those specific checklists.

The Problem You're Solving: Code is read more than it's written. AI generates code that works but is often hard to understand or modify. Maintainability verification ensures code won't become a burden.

Daily Practice (45-60 minutes):

For each piece of code:

1. **Read without context.** Can someone who didn't write this understand what it does?
2. **Predict modification.** What changes are likely in the future? How hard would they be?
3. **Identify complexity.** Where is this code hard to follow? Why?
4. **Check documentation.** Are there comments where needed? Are they accurate? Are they helpful?
5. **Evaluate naming.** Do names convey meaning? Would you understand the purpose from names alone?

Exercise 7.1: The Stranger Test

Take code you generated last week. Pretend you've never seen it before. Time yourself: - How long until you understand what it does? - How long until you could confidently modify it? - What confused you?

If understanding takes more than a few minutes for simple code, maintainability is poor.

Exercise 7.2: Future Modification Prediction

For a piece of code, list 5 changes that might be needed in the future: - New feature addition - Bug fix - Performance improvement - Integration with new system - Behavior change

For each change, assess: - How hard would this change be? - What would you have to understand? - What could go wrong?

Code that makes changes hard has poor maintainability.

Exercise 7.3: Refactoring for Clarity

Take an AI-generated function that works but is unclear. Refactor it purely for readability: - Better variable names - Extract confusing logic into well-named helper functions - Add comments where truly needed (not obvious things) - Simplify complex conditions

Before/after comparison: is it easier to understand?

Exercise 7.4: The Documentation Challenge

Write documentation for a function that has none: - What does it do? - What are the parameters and what do they mean? - What does it return? - What can go wrong? - What are the

assumptions/prerequisites?

Now check: did you have to read the code deeply to write this? If the code had been clear, documentation would have been quick.

End of Week Assessment:

You should be able to: - Estimate time-to-understand for code - Identify maintainability problems (complexity, naming, documentation) - Refactor for clarity without changing behavior

Common Mistakes This Week:

When asked about maintainability, developers instinctively respond "I understand it." Of course you do—you just wrote it, or you just accepted it from AI. The question isn't whether present-you understands it. The question is whether future-you, six months from now after working on completely different code, will understand it. Or whether your teammate will understand it.

Over-documentation wastes time and attention. Comments that simply repeat what code clearly expresses add noise without value. If your code says `user.activate()`, a comment saying "activate the user" helps no one.

Under-documentation is equally problematic. Complex logic with no explanation forces future readers to reverse-engineer your thinking. Why did you choose this algorithm? What assumptions are you making? What edge cases did you handle? Comments should explain the why and the what-might-not-be-obvious, not the what.

1.21 Phase 3: Integration and Judgment (Weeks 8-11)

The previous phase taught you verification skills in isolation—functional testing separate from security review, maintainability separate from semantic verification. Real development requires integrating all these skills into a coherent practice. You need to apply the right verification at the right depth efficiently, without following a rigid checklist. This phase builds that integration and develops the judgment to handle situations where no template applies.

1.21.1 Week 8: Putting It Together

The Goal: Integrate all verification skills into a cohesive practice.

The Problem You're Solving: You've learned functional, semantic, security, and maintainability verification separately. Now you need to apply them together, efficiently.

Daily Practice (60 minutes):

Work through complete development tasks using full VID practice:

1. **Spec first.** Write intent specification before generating
2. **Assess risk.** Determine appropriate verification depth
3. **Generate.** Use AI to generate implementation
4. **Verify functionally.** Does it do what the spec says?
5. **Verify semantically.** Does it do what was intended?
6. **Verify security.** Is it safe against adversarial input?
7. **Verify maintainability.** Can it be understood and modified?

8. **Document decisions.** Record what you verified and why

Exercise 8.1: End-to-End Practice

Complete these tasks using full VID methodology:

Task 1: Build a function that validates and formats phone numbers - Write complete spec - Assess risk (this likely handles user input) - Generate implementation - Verify across all dimensions - Document your verification

Task 2: Build a function that calculates compound interest - Same process

Task 3: Build a function that parses and validates JSON configuration - Same process

For each, time yourself. Where do you spend time? Where could you be more efficient?

Exercise 8.2: The Prioritization Challenge

You have 30 minutes to verify 5 functions. How do you allocate time?

1. A logging utility that formats timestamps
2. A function that authenticates API requests
3. A helper that capitalizes names
4. A function that processes payment amounts
5. A function that generates unique session IDs

Think about: - Which need deep verification? - Which can be verified quickly? - What's your time allocation?

Practice making these decisions quickly. Risk calibration should become instinctive.

Exercise 8.3: Verification Journal

This week, for every verification you perform: - What did you verify? - What depth did you choose? - Why? - What did you find?

Review at end of week: Are your decisions consistent? Are they appropriate?

End of Week Assessment:

You should be able to: - Apply full VID process end-to-end - Allocate verification time based on risk - Verify efficiently without cutting corners

Common Mistakes This Week:

Beginners often approach verification sequentially: first check functionality, then security, then maintainability, in rigid order. But you don't have to verify in strict sequence. Let risk guide where your attention goes. If security is the primary concern, start there. Let the nature of the code, not a fixed checklist, determine your approach.

Treating all code with uniform verification depth wastes time and creates blind spots. A simple utility function doesn't need the same intensive scrutiny as authentication logic. Calibrate your effort to the actual risk and complexity.

Documentation of verification decisions feels optional when you're focused on the code itself. But recording what you verified and why creates invaluable context. When something breaks later—or when you need to modify the code—those notes will save you hours. Future-you will thank present-you for taking two minutes to document your reasoning.

1.21.2 Week 9: Building Speed Without Sacrificing Depth

The Goal: Become faster at verification without reducing quality.

The Problem You're Solving: Thorough verification takes time. But you're under time pressure. You need to be efficient—doing what matters, skipping what doesn't—without dropping important checks.

Daily Practice (60 minutes):

Focus on verification efficiency:

1. **Time your verifications.** How long does each type of check take?
2. **Identify shortcuts.** What checks can be done together? What patterns indicate you can skip something?
3. **Build checklists.** What do you always check? Make it automatic.
4. **Practice pattern recognition.** What patterns indicate problems? Learn to see them quickly.

Exercise 9.1: Timed Verification

Verify 5 functions with a strict time limit: - 2 minutes each for low-risk code - 5 minutes each for medium-risk code - 15 minutes each for high-risk code

What can you accomplish in these times? What must you skip? What would you need to go back and check given more time?

Exercise 9.2: Pattern Library

Build a personal library of "things to watch for": - Patterns that often indicate bugs - Patterns that often indicate security issues - Patterns that often indicate maintainability problems

Every time you find a bug, add the pattern to your library. Over time, you'll spot these patterns instantly.

Exercise 9.3: Checklist Development

Create verification checklists for common code types:

For any function: - [ ] Understand what it does - [ ] Check edge cases (empty, null, large) - [ ] Verify against spec - [ ] Check error handling

For functions that handle user input: - [ ] Input validation - [ ] Injection risk assessment - [ ] Error message safety (no sensitive info)

For functions that touch data storage: - [ ] Check for data loss scenarios - [ ] Verify transactions/rollbacks - [ ] Check concurrent access safety

Checklists make verification faster by removing decision overhead.

End of Week Assessment:

You should be able to: - Complete appropriate verification within time constraints - Identify which checks are essential vs. optional - Recognize common problem patterns quickly

Common Mistakes This Week:

The pressure to move faster often leads to cutting verification corners. You time yourself and feel pressure to beat your previous time. But faster is only valuable if quality is maintained. Verification that misses critical issues because you rushed isn't verification—it's theater.

Checklists help create systematic habits, but rigid adherence to checklists can blind you. You might skip an important check because "it's not on my list," or conversely, waste time on checklist items that don't matter for the specific code you're reviewing. Checklists guide judgment; they don't replace it.

As you build speed, don't let efficiency silence your intuition. If something feels wrong, investigate, even if you can't immediately articulate why. That nagging feeling often comes from pattern recognition happening below conscious thought. Speed shouldn't override the instinct that something needs deeper attention.

1.21.3 Week 10: Handling Complexity

The Goal: Learn to verify complex code that exceeds simple function analysis.

The Problem You're Solving: Real code isn't always simple functions. You'll face complex logic, interacting components, and systems that can't be verified by looking at one piece in isolation.

Daily Practice (60 minutes):

Work with increasingly complex code:

1. **Multi-function logic.** Verify code that spans multiple functions
2. **Stateful systems.** Verify code that maintains state
3. **Concurrent code.** Verify code that runs in parallel
4. **Integration points.** Verify code that interacts with external systems

Exercise 10.1: Call Chain Analysis

Take a function that calls other functions. Trace the full execution: - What does each called function do? - How do they compose? - What assumptions does each make about the others? - Where could the chain break?

Exercise 10.2: State Machine Verification

Take code that maintains state (e.g., a user session, a shopping cart, a workflow engine): - What are the possible states? - What transitions are allowed? - What should never happen? - Verify state transitions are correct - Look for invalid state possibilities

Exercise 10.3: Integration Point Analysis

Take code that calls an external API or database: - What happens if the external system is slow? - What happens if it's unavailable? - What happens if it returns unexpected data? - What happens if it returns errors?

Verify that the code handles these scenarios appropriately.

End of Week Assessment:

You should be able to: - Trace and verify multi-function logic - Verify stateful code systematically - Identify integration failure modes

Common Mistakes This Week:

Function-level tunnel vision is natural when you've spent weeks focused on individual functions. You verify each function carefully but miss how they interact. Complex systems often fail at the boundaries and interactions between components, not within well-tested individual functions. Step back and see the larger picture.

When verifying integration points with external systems, beginners focus on the happy path. They verify that the code correctly handles successful API responses but forget to consider what happens when the external system is slow, unavailable, returns malformed data, or throws errors. External systems fail constantly. Verify that your code handles failure gracefully.

Stateful code introduces subtle bugs that don't exist in pure functions. You might verify that each operation works individually but miss that certain sequences of operations produce invalid states. State machine thinking—explicitly considering what states are possible and what transitions are allowed—helps catch these subtle issues before they become production bugs.

1.21.4 Week 11: Developing Judgment

The Goal: Develop verification judgment that handles novel situations.

The Problem You're Solving: You've learned practices for common situations. But you'll face novel situations where no checklist applies. You need judgment—the ability to figure out what to verify when there's no template.

Daily Practice (60 minutes):

Work with unfamiliar code and domains:

1. **Face novelty.** Work with code types you haven't seen before
2. **Derive approaches.** Figure out verification approaches from first principles
3. **Learn from outcomes.** Track what works and what doesn't
4. **Build intuition.** Develop gut feelings, then verify them

Exercise 11.1: Unknown Domain

Get AI to generate code in a domain you don't know (e.g., graph algorithms, cryptography, signal processing). You can't verify by domain knowledge. How do you approach it?

Options: - Learn enough domain knowledge to verify - Verify structural properties (no crashes, handles edges) - Test against known correct implementations - Consult domain experts

Explore each approach. What can you verify without domain knowledge? What requires expertise?

Exercise 11.2: First Principles Verification

For code in an unfamiliar pattern, ask: - What could go wrong with this type of code? - What would be really bad if it happened? - How would I detect if it went wrong? - What would give me confidence it's right?

Derive your verification approach from these questions.

Exercise 11.3: Judgment Calibration

For code you've verified, predict: - How confident am I this is correct? (percentage) - What's the most likely problem I missed?

Track outcomes. When problems occur, were they things you predicted? Is your confidence calibrated (i.e., code you're 90% confident about should be right 90% of the time)?

End of Week Assessment:

You should be able to: - Approach unfamiliar code types systematically - Derive verification strategies from first principles - Calibrate confidence appropriately

Common Mistakes This Week:

When facing unfamiliar code, there's a strong temptation to default to familiar patterns. You've learned approaches that work for web applications, so you try to apply them to embedded systems code. Novel situations require novel approaches. Don't force familiar patterns onto unfamiliar domains.

Overconfidence in unfamiliar territory is surprisingly common. You've become skilled at verification, so you assume your judgment extends to domains you don't understand. But unknown domains contain unknown unknowns. Calibrate your confidence appropriately—be more tentative, seek more review, acknowledge the limits of your knowledge.

On the flip side, some engineers freeze when faced with unfamiliarity. Complete paralysis because "I don't know this domain" prevents any verification at all. Don't let unfamiliarity stop you entirely. Do what you can with the knowledge you have, acknowledge clearly what you can't verify, and know when to seek expert review.

1.22 Phase 4: Mastery and Teaching (Week 12+)

After eleven weeks of deliberate practice, you've developed substantial verification skills. You can read code critically, specify intent precisely, calibrate verification to risk, and apply multiple verification techniques efficiently. But mastery isn't just personal skill—it's the ability to articulate what you know, teach others, and contribute to collective practice. This final phase consolidates your learning and begins your transition from student to teacher.

1.22.1 Week 12: Consolidation and Sharing

The Goal: Consolidate your learning and begin sharing with others.

The Problem You're Solving: Mastery isn't just personal skill—it's being able to help others develop skill. Teaching reinforces learning and multiplies your impact.

Activities This Week:

1. **Review your journey.** Look back at your journals and notes. What did you learn? How did you change?

2. **Identify your patterns.** What verification approaches work best for you? What mistakes do you keep making?
3. **Teach someone else.** Explain VID to someone who doesn't know it. Teaching reveals what you don't understand.
4. **Contribute to team practices.** Propose improvements to how your team handles AI-generated code.

Exercise 12.1: The Retrospective

Write a personal retrospective: - What verification skills improved most? - What's still weak? - What was most surprising? - What would you tell past-you starting this journey?

Exercise 12.2: Teaching Practice

Teach a concept from this curriculum to someone else: - Explain risk calibration - Walk through a verification ritual - Demonstrate security verification

Did you explain it clearly? Where did they get confused? What would you explain differently?

Exercise 12.3: Process Proposal

Write a proposal for your team: - What VID practices would benefit the team? - How would you introduce them? - What resistance do you expect? - How would you measure success?

End of Week Assessment:

You should be able to: - Articulate your verification approach clearly - Teach VID concepts to others - Propose practical improvements to team practices

---

1.22.2 Beyond Week 12: Continuous Development

This twelve-week curriculum builds foundational verification skills, but mastery is not a destination you reach and then stop. It's a continuous process of learning, calibrating, and adapting. The best engineers you'll work with—the ones whose code you trust, whose reviews catch issues others miss—didn't stop developing their skills after some arbitrary point. They kept learning, deliberately and systematically, throughout their careers.

Ongoing Practices:

- **Keep learning patterns.** Every bug you find teaches you a pattern. Build your library.
- **Keep calibrating.** Track outcomes. Adjust practices based on reality.
- **Keep teaching.** Explaining to others deepens your own understanding.
- **Keep adapting.** AI capabilities change. Verification needs change with them.

Signs of Growing Mastery:

- Verification becomes faster without losing depth
- You spot problems others miss
- Your confidence calibration improves (your predictions match outcomes)
- Others seek your review

- You can explain why, not just what

Signs You Need More Work:

If problems consistently escape your review and reach production, your verification has gaps. Track what you're missing and adjust your approach.

If you can't explain why you check what you check—you just follow steps mechanically—you haven't internalized the principles. Understanding the why behind verification practices is essential for judgment.

If verification feels like checklist compliance rather than genuine understanding, you're going through motions without engaging. Real verification requires active thinking, not mechanical box-checking.

If you struggle with novel situations and can't figure out what to verify when there's no template, you need more practice deriving approaches from first principles.

These signs don't mean you're failing—they mean you know where to focus your continued development. Every senior engineer still has areas where they need work. The difference is they recognize those gaps and address them deliberately.

A Final Word

You're learning to code at a pivotal moment. AI can generate more code in an hour than a human could write in a week. This reality creates a choice: become someone who prompts AI without understanding what it produces, or become someone who can verify AI-generated code with skill and judgment.

The path presented in this curriculum is harder. It requires deliberate practice, intellectual honesty, and sustained effort. But it builds something AI cannot replace: the judgment to know what matters, the skill to verify what's generated, and the understanding to maintain and evolve code over time.

The engineers who thrive in this new era won't be those who generate the most code. They'll be those who can confidently verify that code is correct, secure, and maintainable—and who can teach others to do the same. That's the path this curriculum offers. The work is yours to do.

1.23 Chapter 14: For Senior Engineers

1.23.1 The Challenge

You've spent years building skills in writing code. AI has commoditized much of that skill. Your role is changing whether you like it or not.

1.23.2 The Opportunity

Your experience becomes more valuable, not less, when properly applied:

Judgment. You've seen what goes wrong. You know the failure modes. This judgment is essential for calibrating trust and verification depth.

Debugging. When AI-generated code fails in subtle ways, your experience debugging helps diagnose problems that junior engineers would struggle with.

Architecture. AI generates functions and modules. Knowing how to compose them into coherent systems is architecture — still a human skill.

Mentorship. Junior engineers need guidance on verification. Your experience informs what to check and how deeply.

1.23.3 The Shift

Your role shifts from producing code to:

Specifying intent. You know what needs to be built. AI generates it. Your understanding of requirements is the input.

Verifying output. Your experience identifies what AI gets wrong. You're the quality gate.

Calibrating trust. You assess risk and determine appropriate verification levels.

Building team judgment. You help junior engineers develop the verification skills you've built over years.

1.23.4 The Danger

Some senior engineers resist AI-augmented development entirely. This resistance leads to irrelevance as teams that embrace AI move faster.

Other senior engineers over-embrace AI, accepting output without applying their judgment. This throws away the experience advantage.

The VID approach: use your experience to verify better, not to write more.

1.24 Chapter 15: For Teams and Organizations

1.24.1 The Team Challenge

Individual VID adoption helps individual developers. Team adoption provides multiplicative benefits.

1.24.2 Building a Verification Culture

Building a verification culture starts with normalizing verification time. Verification isn't overhead—it's the work itself. Teams must build verification into their estimates and expectations from the start, treating it as essential rather than optional.

When someone's verification catches a problem before it reaches production, that's a win worth celebrating. Make these catches visible to the team. Create an environment where finding issues through verification is praised, not hidden. This reinforces that verification has real value.

When problems escape verification and reach production, discuss as a team what could have caught them. Approach these discussions without blame, focusing purely on learning. What verification step did we skip? What risk did we miscalibrate? What pattern did we miss? These post-mortems strengthen everyone's verification skills.

Finally, when someone discovers an effective verification approach, share it with the team. Verification techniques that work well for one person often help others. Create channels—whether in code reviews, team meetings, or documentation—for sharing these patterns actively.

1.24.3 Team Calibration

Effective teams develop shared calibration around verification. This begins with aligning on risk levels—the entire team should share a common understanding of what constitutes low, medium, and high risk in your specific context. A utility function for formatting dates carries different risk than authentication logic or payment processing code.

Beyond risk assessment, teams must calibrate their verification standards. What does "thorough verification" actually mean for your team? Without concrete agreement, one developer's thorough verification might be another's cursory check. Define specific practices: for high-risk code, does thorough verification mean manual testing plus automated tests plus security review? Get explicit about your standards.

Calibration isn't a one-time exercise. Teams should periodically review together: Are our verification practices working? What types of problems are escaping our verification? What should we adjust? As the team gains experience and as AI capabilities evolve, your calibration will need to evolve with them.

1.24.4 Organizational Support

Organizations play a crucial role in enabling VID adoption. Success requires more than individual or team initiative—it demands organizational commitment.

First, invest in building verification skills, not just providing tool access. Many organizations rush to give developers access to AI coding tools without investing in the skills needed to verify AI-generated code effectively. Training developers in verification techniques—code review, testing strategies, security analysis, performance evaluation—becomes more critical, not less, when AI generates code.

Second, allow verification time in your planning and metrics. If velocity metrics don't account for verification time, people will skip verification to hit their numbers. When you measure story points completed or features shipped without measuring verification quality, you incentivize cutting corners. The organization must make it safe—even encouraged—to take the time verification requires.

Third, track the right metrics. Measure verified outputs, not raw outputs. Measure problems caught during verification, not just problems that escaped to production. Celebrate teams that catch issues early through verification, not teams that generate the most code. What you measure shapes behavior, so measure what matters.

Finally, support calibration by providing time for teams to review and adjust their practices. Calibration requires reflection, discussion, and experimentation. Teams need organizational permission to invest time in improving their verification practices, not just executing against the backlog.

1.24.5 Metrics to Track

Verification Quality Metrics (Leading Indicators)

Track these to understand how well verification is working:

Metric	How to Measure	Target	Why It Matters
Verification Coverage	% of PRs with documented risk score + verification notes	>90%	Ensures systematic verification is happening
Risk Calibration Accuracy	Actual incidents by initial risk score (does "high risk" code actually cause problems?)	<10% miscalibration	Shows whether risk assessment matches reality
Defects Caught in Verification	Issues found during verification before merge/deploy	Trending up	More catches = verification is working
Verification Time by Trust Level	Avg. time spent verifying High/Moderate/Guarded/Minimal Trust code	Stable within ranges	Ensures effort matches risk
Provenance Documentation	% of commits with AI/human provenance markers	>95%	Maintains awareness for future modifications

Outcome Metrics (Lagging Indicators)

Track these to measure impact:

Metric	How to Measure	Target	Why It Matters
Production Incidents from AI-Generated Code	Incidents traced to AI-generated code vs. human-written	Trending down	Core outcome: fewer bugs in production
Mean Time to Understand	Time for developer to understand unfamiliar code (survey quarterly)	Stable or improving	Measures understanding debt
Verification ROI	Time spent verifying vs. time spent fixing production issues	Verification < Fixing	Shows VID saves time overall
Developer Confidence	Survey: "I'm confident explaining code I committed" (1-5 scale)	>4.0 average	Measures understanding over acceptance
Velocity Sustainability	Feature delivery rate over 6+ months	Stable or increasing	Ensures VID doesn't slow long-term velocity

How to Collect These Metrics:

- **Week 1-4:** Manual tracking in spreadsheet (developers self-report)
- **Month 2-3:** Integrate with existing tools (PR templates, issue trackers, CI/CD)
- **Month 6+:** Automated dashboards with trend analysis

1.24.6 Calibration Cadence

Weekly (15-30 minutes, team level) - **What:** Quick retrospective on verification decisions - **Who:** Entire team - **Format:** Stand-up style—each person shares one verification decision and outcome - **Outcome:** Identify immediate calibration adjustments

Monthly (1 hour, team level) - **What:** Metrics review + pattern analysis - **Who:** Entire team + tech lead - **Agenda:** 1. Review metrics: What's trending well? What's concerning? 2. Incident review: What escaped verification this month? 3. Calibration adjustments: Should we change risk scoring or verification practices? 4. Wins: What did verification catch that would have been expensive in production? - **Outcome:** Updated team calibration, documented lessons learned

Quarterly (2 hours, team level) - **What:** Deep dive on VID practice maturity - **Who:** Team + engineering manager - **Agenda:** 1. Trend analysis: Are metrics improving over the quarter? 2. Team skill assessment: Where are we strong? Where do we need training? 3. Process refinement: What's working? What's bureaucratic? 4. Goals for next quarter: Specific, measurable targets - **Outcome:** Quarterly goals, training plan, process improvements

Bi-Annually (Half day, organizational level) - **What:** Cross-team VID calibration - **Who:** Representatives from all teams practicing VID + leadership - **Agenda:** 1. Share team-specific patterns and lessons 2. Standardize risk scoring across teams where beneficial 3. Identify organization-wide verification gaps 4. Plan organization-level investments (training, tools, templates) - **Outcome:** Org-wide calibration, shared best practices, resource allocation

1.24.7 Adoption Milestones Checklist

Use this checklist to track team progress:

Month 1: Foundation - [ ] All team members completed VID core chapters (1-8) - [ ] Team agreed on risk scoring rubric (Appendix D) - [ ] PR template updated to include risk assessment - [ ] First monthly calibration meeting held - [ ] Baseline metrics established (pre-VID defect rates, velocity)

Month 3: Practice - [ ] 90%+ of PRs include risk scores and verification notes - [ ] Team calibration is converging (less disagreement on risk scores) - [ ] At least 5 bugs caught in verification that would have reached production - [ ] Verification time is stabilizing (not increasing indefinitely) - [ ] Team has shared 3+ effective verification techniques

Month 6: Integration - [ ] Production incidents from AI-generated code decreased 30%+ - [ ] Developers report increased confidence in AI-generated code (survey) - [ ] Verification is "how we work" (not "that VID thing we have to do") - [ ] New team members onboard to VID within 2 weeks - [ ] Risk calibration accuracy >90% (actual risk matches assessed risk)

Month 12: Mastery - [ ] Team contributes improvements back to VID methodology - [ ] Verification ROI is clearly positive (time saved > time invested) - [ ] Team can articulate VID principles without referencing docs - [ ] Other teams request to learn from your VID practices - [ ] VID practices are adapted to team-specific context (not just following templates)

1.25 Chapter 16: Adopting VID

1.25.1 Starting Points

You can begin VID adoption at different scales, depending on your position and influence within your organization.

Individual adoption is the most accessible entry point. You can start applying VID principles personally today, with no permission needed. Begin specifying intent before generation, practice verification rituals on your own code, and build your calibration skills. As you demonstrate results—fewer bugs, better code quality, stronger understanding—you'll naturally influence those around you.

Team adoption amplifies these benefits. When a team agrees collectively to apply VID practices, they can align on shared risk calibration and verification standards. This creates consistency across the team's work and enables collective learning. Teams can review verification outcomes together, calibrate their practices, and build shared expertise more rapidly than individuals working in isolation.

Organizational adoption formalizes VID as the development methodology across multiple teams. This enables systematic training programs, dedicated tooling support, and organization-wide metrics that reinforce verification practices. When the organization commits to VID, it sends a clear signal that verification is valued work, not optional overhead.

1.25.2 Starting Small

Don't try to implement everything at once. VID adoption works best as a gradual, deliberate progression.

In the first two weeks, practice intent specification. Before asking AI to generate code, take time to articulate clearly what you need and how you'll verify it meets your requirements. This single practice immediately improves the quality of generated code and prepares you for effective verification.

During weeks three and four, add verification rituals to your workflow. Start with basic rituals—read the code, check for obvious issues, verify it matches your intent—and apply them consistently. The goal isn't perfection; it's building the habit of verification as a standard part of accepting AI-generated code.

By weeks five and six, begin actively calibrating your verification intensity. Start noticing patterns: which types of code need more thorough verification, which need less. A simple utility function might need only basic checks, while authentication code demands deeper scrutiny. Let your growing experience inform your calibration.

From there, the work becomes ongoing. Build the learning loop into your practice. Track outcomes—what types of issues does your verification catch? What escapes?—and adjust your practices accordingly. Calibration improves continuously as you accumulate experience.

1.25.3 Common Obstacles

Every team adopting VID encounters predictable resistance. Understanding these obstacles helps you address them proactively.

"It slows us down" is perhaps the most common objection. Yes, verification takes time. But debugging production issues, reworking faulty code, and responding to incidents take far more time. VID front-loads small, predictable costs during development to avoid large, unpredictable costs later. The question isn't whether you'll pay the cost of quality—it's when and how much.

"AI is good enough" reflects understandable optimism about AI capabilities. AI is indeed remarkably good at generating code that looks correct and often is correct for common cases. But AI struggles with edge cases, security implications, performance considerations, and maintainability concerns. The code that looks right isn't always the code that is right, especially under stress or adversarial conditions.

"We don't have time" reveals a fundamental misunderstanding of where time actually goes. You don't have time to not verify. The time you imagine you're saving by skipping verification, you'll spend—often multiplied—on debugging obscure issues, fixing production incidents, and explaining to stakeholders why things broke. Verification time is invested; debugging time is lost.

1.25.4 The 30/60/90 Day Adoption Plan

Whether you're an individual, team lead, or engineering manager, this structured plan provides a concrete path to VID adoption.

1.25.4.1 Days 1-30: Foundation & Awareness Individual Contributors: - Week 1: - [ ] Read Chapters 1-3 (The paradigm shift) - [ ] Read Chapter 4 (Intent Before Generation) - [ ] Start: Write intent specification before every AI generation (even if just 2 sentences) - [ ] Track: Keep simple log of what you specified and what AI generated

- **Week 2:**
  - ☐ Read Chapter 5 (Graduated Trust) + Appendix D (Risk Rubric)
  - ☐ Practice: Assess risk for every task using the rubric
  - ☐ Start: Apply basic verification ritual to all AI-generated code (10-15 min minimum)
- **Week 3:**
  - ☐ Read Chapters 9-10 (Intent Specification & Verification Rituals)
  - ☐ Practice: Match verification depth to risk score
  - ☐ Track: Record time spent verifying vs. bugs caught
- **Week 4:**
  - ☐ Read Chapter 7 (Provenance Awareness)
  - ☐ Start: Add provenance markers to all commits (AI-generated vs. human-written)
  - ☐ Reflect: What verification techniques work best for you?
  - ☐ Milestone: First 30 days of consistent intent specification + verification

Team Leads: - Week 1: - [ ] Share VID One-Page Overview with team - [ ] Identify 2-3 team members interested in pilot adoption - [ ] Schedule: 1-hour team introduction to VID for Week 2

- **Week 2:**
  - ☐ Facilitate team discussion: Do we have this problem?
  - ☐ Share 1-2 recent incidents that verification would have prevented
  - ☐ Decision: Commit to 30-day pilot with volunteer subset of team
- **Week 3-4:**
  - ☐ Pilot team applies VID practices (following IC track above)
  - ☐ Weekly check-in: What's working? What's hard?

☐ Collect early wins: Document bugs caught during verification
☐ Milestone: Team understanding of VID value proposition

Engineering Managers: - **Week 1:** - [ ] Read VID for Engineering Managers guide - [ ] Review current metrics: Incident rates, debugging time, velocity trends - [ ] Identify 1-2 teams most affected by AI-generated code issues

- **Week 2:**
  - ☐ Meet with team leads: Present VID business case
  - ☐ Share: Incident data + costs of current approach
  - ☐ Decision: Select pilot team for VID adoption
- **Week 3-4:**
  - ☐ Communicate org-wide: Why verification matters, pilot team announcement
  - ☐ Allocate: Dedicated time for pilot team training (don't expect "free time")
  - ☐ Establish: Baseline metrics (pre-VID incident rates for pilot team)
  - ☐ Milestone: Org commitment to supporting VID experiment

1.25.4.2 Days 31-60: Practice & Integration

Individual Contributors: - **Week 5-6:** - [ ] Read Chapters 13 or 14 (role-specific guidance) - [ ] Integrate: VID practices now feel routine, not forced - [ ] Expand: Apply verification rituals to legacy code you modify (not just new code) - [ ] Measure: Compare bugs caught in verification vs. Month 1

- **Week 7-8:**
  - ☐ Read Chapter 11 (Learning Loop)
  - ☐ Analyze: What types of bugs does your verification miss? What patterns escape?
  - ☐ Adjust: Refine your verification checklist based on learnings
  - ☐ Share: Teach one technique to a colleague
  - ☐ Milestone: Verification is habitual; you have data showing it works

Team Leads: - **Week 5:** - [ ] Expand pilot: Invite rest of team to adopt VID practices - [ ] Share: Pilot team's results (bugs caught, time data, developer feedback) - [ ] Update: PR template with risk score + verification notes fields

- **Week 6:**
  - ☐ Facilitate: First team-wide risk calibration session (1 hour)
  - ☐ Align: What constitutes High/Moderate/Guarded/Minimal Trust for our codebase?
  - ☐ Document: Team's agreed risk scoring norms
- **Week 7-8:**
  - ☐ Establish: Monthly calibration meeting (first Friday of each month)
  - ☐ Track: Team-level metrics (% PRs with risk scores, bugs caught, incidents)
  - ☐ Celebrate: Public recognition for verification catches
  - ☐ Milestone: Whole team practicing VID, metrics show early results

Engineering Managers: - **Week 5-6:** - [ ] Review: Pilot team metrics at 30 days - [ ] If successful: Approve rollout to 2-3 additional teams - [ ] If struggling: Diagnose blockers and provide support

- **Week 7-8:**
  - ☐ Invest: Formal training session for expanding teams (2-4 hours)
  - ☐ Resource: Allocate time for teams to adapt VID to their context

☐ Communicate: Share pilot results org-wide (especially wins)
☐ Milestone: Multiple teams practicing VID, leadership visibly supporting it

---

1.25.4.3 Days 61-90: Optimization & Scale Individual Contributors: - Week 9-10: - [ ] Refine: Your verification practices are now highly calibrated - [ ] Efficiency: Verification time is stable, not increasing - [ ] Confidence: You can explain every line of AI-generated code you commit

- **Week 11-12:**
  ☐ Read Chapters 19-20 (Patterns & Verification Toolkit)
  ☐ Contribute: Share your verification techniques with team
  ☐ Mentor: Help onboard new team member to VID practices
  ☐ Milestone: VID mastery—you verify efficiently without thinking about it

Team Leads: - **Week 9:** - [ ] Analyze: 60-day metrics review with team - [ ] Compare: Pre-VID vs. current incident rates, debugging time - [ ] Identify: What verification gaps remain?

- **Week 10:**
  ☐ Optimize: Remove bureaucratic parts of VID (if any emerged)
  ☐ Standardize: Document team's adapted VID practices for onboarding
  ☐ Share: Present results to engineering leadership
- **Week 11-12:**
  ☐ Scale: New team members onboard to VID within first week
  ☐ Automate: PR template enforces risk scoring, provenance marking
  ☐ Celebrate: 90-day retrospective highlighting wins
  ☐ Milestone: VID is "how we work," not "that new thing"

Engineering Managers: - **Week 9-10:** - [ ] Decision: Based on data, expand VID org-wide or refine approach - [ ] Invest: If successful, build shared resources (templates, checklists, training materials) - [ ] Metrics: Establish org-level dashboard for VID quality metrics

- **Week 11-12:**
  ☐ Policy: Update engineering standards to include VID practices
  ☐ Recognition: Add verification quality to performance review criteria
  ☐ Planning: Roadmap for next 6 months of VID maturation
  ☐ Milestone: VID integrated into org culture and processes

1.25.5 Success Indicators by Timeframe

After 30 Days: - Team members can explain VID principles in their own words - 50%+ of PRs include risk scores - At least 3 bugs caught in verification that would have reached production - Developers report verification feels manageable, not overwhelming

After 60 Days: - 90%+ of PRs include risk scores + verification notes - Team calibration is converging (less disagreement on risk) - Measurable decrease in post-deployment defects (20%+ improvement) - Verification time is stable (not growing indefinitely)

After 90 Days: - Production incidents from AI code down 30%+ - Developer confidence survey shows >4.0/5.0 "I understand code I commit" - New team members adopt VID practices within

1 week - Team proposes improvements to VID based on experience - Leadership can articulate clear ROI for VID investment

---

1.26 Chapter 17: The Continuing Evolution

1.26.1 AI Will Keep Improving

AI capabilities are improving rapidly. Some predictions:

- AI will get better at generating correct code
- AI will get better at self-verification
- AI will get better at explaining generated code
- New failure modes will emerge that we can't predict today

1.26.2 VID Will Evolve

The principles endure: - Intent before generation - Graduated trust based on risk - Understanding over mere acceptance - Provenance awareness - Continuous calibration

The practices will evolve. The specific verification rituals appropriate for 2025's AI will differ from those for 2027's AI. Calibration questions will change as AI capabilities change.

1.26.3 Your Role in the Evolution

VID isn't complete. It's a starting point — a framework for thinking about AI-augmented development that you can adapt.

As you practice VID:

- **Notice what works and what doesn't** in your context
- **Develop new practices** that address gaps
- **Share what you learn** with others facing similar challenges
- **Adjust as AI evolves** rather than freezing your approach

The developers who navigate the AI transition successfully will be those who develop judgment about verification. VID provides a framework for building that judgment.

1.27 Chapter 18: Summary

1.27.1 The Core Message

Code generation is no longer the bottleneck. Verification is.

The methodology that fits this reality focuses on: - Building verification skills - Calibrating verification to risk - Maintaining understanding of generated code - Learning from outcomes

1.27.2 The Five Principles

1. **Intent Before Generation:** Never generate without articulating what you need and how you'll verify it.
2. **Graduated Trust:** Match verification depth to risk level.
3. **Understanding Over Acceptance:** Never accept code you don't understand at an appropriate level.
4. **Provenance Awareness:** Know where code comes from and what that implies.
5. **Continuous Calibration:** Adjust practices based on outcomes.

1.27.3 The Core Practices

1. **Intent Specification:** Articulate functional requirements, boundaries, and success criteria before generating.
2. **Verification Rituals:** Apply consistent verification practices after every generation, scaled to risk.
3. **Learning Loop:** Track outcomes and adjust practices based on what you learn.
4. **Provenance Hygiene:** Document and maintain awareness of code origins.

1.27.4 The Path Forward

Start with principles. Apply practices consistently. Calibrate based on outcomes. Build the judgment that verification requires.

The developers who thrive in the AI-augmented future will be those who master verification, not just generation.

VID provides a framework for becoming that developer.

1.28 Chapter 19: Patterns and Anti-Patterns

1.28.1 Why Patterns Matter

Principles tell you what to value. Practices tell you what to do. But patterns show you what it looks like in reality—both when it works and when it fails.

This chapter presents concrete scenarios. Some show VID done well. Others show common failures. Study both. The failures are often more instructive.

1.28.2 Pattern 1: The Specification-First Win

The Scenario:

Maya needs a function to calculate prorated subscription charges. A user upgrades mid-billing-cycle; the function determines how much to charge for the remaining days.

The Anti-Pattern (how it usually goes wrong):

Maya opens her AI assistant and types: "Write a function to calculate prorated subscription charges."

The AI generates something. It looks reasonable. Maya runs a quick test—upgrade on day 15 of a 30-day month. The result seems right. She commits.

Two weeks later, a customer complains they were overcharged. Finance investigates. The function doesn't handle months with different numbers of days correctly. It assumed 30 days always. February charges were wrong. So were charges in 31-day months.

Maya fixes the bug. But the damage is done—customers were overcharged, refunds are needed, trust is damaged.

The Pattern (VID approach):

Maya pauses before prompting. She writes:

```
Function: calculate_prorated_charge
Inputs:
  - original_price: monthly subscription cost
  - upgrade_date: date of upgrade
  - billing_cycle_start: first day of current billing period

Output: dollar amount to charge (rounded to cents)

Behavior:
  - Calculate days remaining from upgrade_date to end of billing cycle
  - Prorate based on actual days in that specific month
  - Round to nearest cent

Edge cases:
  - upgrade_date equals billing_cycle_start: charge full amount
  - upgrade_date is last day of cycle: charge for 1 day
  - February (28/29 days): must handle correctly
  - 31-day months: must handle correctly

Verification:
  - Test with Feb 15 in non-leap year (14 days remaining of 28)
  - Test with Feb 15 in leap year (14 days remaining of 29)
  - Test with Jan 15 (17 days remaining of 31)
  - Test with Jun 15 (15 days remaining of 30)
  - Test upgrade on day 1 (full charge)
  - Test upgrade on last day (1 day charge)
```

Now Maya prompts the AI with this specification. The AI generates code. Maya runs her pre-planned tests. The February test fails—the AI assumed 30 days. Maya catches it before commit.

She either fixes the code or prompts again with explicit instructions about variable month lengths. The final code handles all cases correctly.

The Lesson:

The time Maya spent writing the specification was less than the time Anti-Pattern Maya spent on customer complaints, investigation, fixes, and refunds. Specification isn't overhead—it's the shortest path to correct code.

1.28.3 Pattern 2: Graduated Trust in Action

The Scenario:

A team is building a new feature. It involves three components: 1. A utility to format currency for display 2. An endpoint to update user profile information 3. A function to process refunds

The Anti-Pattern:

The team treats all three equally. Each gets the same cursory review: glance at the code, run a quick test, approve the PR.

Six months later: - The currency formatter has a bug with negative numbers. It displays "-$-50" instead of "-$50". Embarrassing but minor—it only affects a rarely-used report. - The profile endpoint has a bug. It doesn't validate email format. Users enter invalid emails, then can't receive notifications. Support tickets pile up. - The refund function has a serious bug. Under certain conditions, it processes refunds twice. The company loses $200,000 before anyone notices.

The Pattern:

The team assesses each component:

Currency formatter: - Impact if wrong: Minor display issue - Reversibility: Easy fix, redeploy - Exposure: Internal report - **Verdict: High trust, basic verification**

Profile endpoint: - Impact if wrong: User experience problems, support burden - Reversibility: Can fix but affected users remain in bad state - Exposure: All users - **Verdict: Moderate trust, standard verification including input validation review**

Refund function: - Impact if wrong: Financial loss, potentially significant - Reversibility: Money already gone, hard to recover - Exposure: Every refund transaction - **Verdict: Minimal trust, intensive verification**

They allocate time accordingly: - Currency formatter: 10 minutes, automated tests, quick review - Profile endpoint: 30 minutes, systematic input validation check, edge case testing - Refund function: 2 hours, multiple reviewers, explicit test of idempotency, audit trail verification, sign-off from senior engineer

The refund double-processing bug gets caught in review. The profile email validation gets added. The currency formatter ships with minimal review—and yes, it has the negative number bug, but that's discovered and fixed with minimal impact.

The Lesson:

Equal treatment of unequal risks is wasteful at best, catastrophic at worst. The refund bug could have bankrupted a startup. Catching it was worth the extra two hours.

1.28.4 Pattern 3: The Understanding Trap

The Scenario:

Carlos needs to implement a caching layer. He's not an expert in caching, but the AI generates something that looks sophisticated.

The Anti-Pattern:

Carlos prompts: "Implement an LRU cache with TTL support."

The AI generates 80 lines of code with a doubly-linked list, a hash map, timestamp tracking, and cleanup logic. Carlos reads through it. It looks professional. He doesn't fully understand the linked list manipulation, but the tests pass.

He ships it.

Three weeks later, production experiences memory growth. The cache isn't releasing memory correctly. Carlos investigates. He stares at the linked list code. He doesn't understand it. He can't debug what he doesn't understand.

He spends two days learning how doubly-linked lists work in this context, then another day finding the bug: a reference wasn't being cleared during eviction, preventing garbage collection.

Total cost: three days of debugging code he didn't understand. Plus the production incident.

The Pattern:

Carlos prompts for the cache implementation. The AI generates the same sophisticated code.

Carlos reads it. He doesn't fully understand the linked list manipulation. He stops.

He asks himself: "Do I understand this well enough to debug it when it breaks?"

The answer is no.

Carlos has options:

Option A: Learn enough to understand He spends two hours studying how LRU caches work with doubly-linked lists. He traces through the AI's code. Now he understands it. He can verify it properly and debug it later.

Option B: Simplify He prompts again: "Implement a simpler LRU cache using OrderedDict" (in Python) or equivalent simple approach. The result is less sophisticated but understandable. It might be slightly slower, but Carlos can verify and maintain it.

Option C: Get help He asks a colleague who knows caching to review with him. Together they verify the implementation. Carlos learns enough to maintain it.

Carlos chooses Option B. The simpler cache is 20% slower but completely understandable. It ships without issues. When a bug eventually surfaces (wrong TTL calculation), Carlos fixes it in 20 minutes because he understands the code.

The Lesson:

Sophisticated code you don't understand is a liability, not an asset. The time saved by accepting complex AI output is borrowed from your future self, with interest.

1.28.5 Pattern 4: The Provenance Blind Spot

The Scenario:

A team has been using AI assistance heavily for six months. The codebase has grown substantially.

The Anti-Pattern:

No one tracked what was AI-generated versus human-written. Commit messages just say "Add feature X" or "Fix bug Y."

Now they need to audit their code for potential IP issues. A large enterprise customer requires attestation that code doesn't infringe third-party IP.

The team realizes they have no idea which parts of their codebase were AI-generated. They can't assess risk. They can't provide attestation.

Their options: - Manually review the entire codebase (expensive) - Provide attestation without knowing (risky) - Lose the enterprise customer (costly)

They also discover a security pattern that's been copy-pasted throughout the codebase. It has a vulnerability. They don't know where it came from. Did a human write it? Did AI generate it? Was it in one place originally and spread, or generated multiple times?

Without provenance, they can't trace the origin or assess how it spread.

The Pattern:

From the start, the team maintains provenance awareness:

- Commit messages indicate AI assistance: "Add user authentication (AI-assisted, reviewed by @carlos)"
- A simple convention marks sections of code: `# AI-generated: initial implementation` or `# Human: security-critical logic`
- PR descriptions note what was AI-generated and what review it received

Six months later, they face the same enterprise audit request.

They can identify AI-generated code. They can show it was reviewed. They can provide meaningful attestation: "Our codebase includes AI-assisted code in these areas, which underwent our standard review process including security verification."

When the vulnerability pattern appears, they trace it: originally AI-generated in one file, then copied (by humans) to other files. They know exactly where to look and can assess the scope of the problem.

The Lesson:

Provenance tracking costs almost nothing in the moment. Reconstructing provenance later costs enormously—if it's even possible.

1.28.6 Pattern 5: The Security Review That Wasn't

The Scenario:

A developer needs to build a password reset flow. The AI generates code that: - Generates a reset token - Stores it in the database - Emails it to the user - Accepts the token on a reset page - Updates the password

The Anti-Pattern:

The developer reviews the code. It looks correct. The token is generated. It's stored. It's emailed. The reset works. Ship it.

Three months later, attackers discover: - The token is generated using predictable random numbers - There's no expiration on tokens - There's no rate limiting on the reset request endpoint - The token is passed in the URL (and thus logged everywhere) - The same token can be used multiple times

None of these were obviously visible in the code. The code "worked." It just wasn't secure.

The Pattern:

The developer recognizes password reset as security-critical (authentication-adjacent, affects account access). This triggers intensive verification including security review.

For the security review, they consider:

Token generation: - Is the random number generator cryptographically secure? - Is the token long enough to prevent brute force? - Is there sufficient entropy?

Token lifecycle: - Does the token expire? - Is expiration enforced server-side (not just client-side)? - Is the token invalidated after use? - Is it invalidated if the user requests another reset?

Attack surface: - Can attackers enumerate valid tokens? - Is there rate limiting to prevent brute force? - Are failed attempts logged?

Token handling: - Is the token transmitted securely? - Is it stored securely? - Does it appear in logs?

The developer checks each of these against the AI-generated code. They find: - Token generation uses basic random, not cryptographic random → fix - No expiration → add expiration - Token appears in URL query parameters → move to POST body - No rate limiting → add rate limiting - Token works multiple times → add single-use enforcement

The fixed implementation ships without the vulnerabilities.

The Lesson:

Security review isn't checking if code looks right. It's systematically considering how code could be attacked and whether it resists those attacks.

1.28.7 Pattern 6: The Maintenance Nightmare

The Scenario:

A feature was built quickly using heavy AI assistance. It works. It ships. Everyone moves on.

The Anti-Pattern:

A year later, the feature needs modification. The original developer has left. A new developer opens the code.

They find: - 500 lines in a single function - Variables named `temp`, `data`, `result`, `x` - No comments explaining the business logic - Deeply nested conditionals (if inside if inside if inside if) - Magic numbers throughout (why 86400? why 0.035?) - Duplicated code blocks with slight variations

The new developer spends three days just understanding what the code does. They're afraid to change it because they might break something they don't understand.

Eventually, they give up on modifying the existing code. They rewrite the feature from scratch. A month of work to add what should have been a two-day feature.

The Pattern:

During original development, the developer applies maintainability verification:

They ask: "Could someone else understand this in a year?"

The AI-generated code has the same issues: one long function, poor names, no comments.

Before accepting it, the developer refactors: - Breaks the long function into smaller, named functions (`calculate_shipping_cost`, `apply_discount`, `validate_address`) - Renames variables to be meaningful (`seconds_per_day` instead of 86400, `commission_rate` instead of 0.035) - Adds comments explaining *why*, not *what* (the what is in the code; the why is the business logic) - Eliminates duplication by extracting common patterns - Reduces nesting by early returns and guard clauses

This takes an extra hour.

A year later, the new developer opens the code. They understand the structure in 15 minutes. They find the function that needs modification. They make the change in two days as expected.

The Lesson:

An hour of maintainability work saves months of future confusion. AI generates working code, but working code isn't necessarily maintainable code.

1.28.8 Anti-Pattern Catalog

Here are common VID anti-patterns in brief:

"It compiled, so it's correct" Code that compiles can be completely wrong. Compilation is syntax checking, not semantic verification.

"The tests pass" Tests only verify what they test. If the tests don't cover the bug, they won't catch it. Passing tests are necessary but not sufficient.

"I'll understand it later" You won't. Understanding erodes quickly. If you don't understand it now, you'll understand it less in a month.

"This is just like last time" Similar-looking code can have very different behavior. Verify each instance; don't assume patterns transfer.

"AI knows what I meant" AI guesses what you meant based on your words. It often guesses wrong. Specify explicitly.

"It's just internal tooling" Internal tools become critical infrastructure. Today's quick script is tomorrow's system dependency.

"We're moving fast" Fast without verification is fast toward failure. Real speed includes time to verify.

"No one will misuse it" Someone will. Assume adversarial use and verify accordingly.

"The senior dev approved it" Approval isn't verification. Did they actually verify, or did they trust without checking?

"We've always done it this way" Past success doesn't guarantee future success, especially as AI capabilities and risks change.

1.29 Chapter 20: The Verification Toolkit

1.29.1 What This Chapter Is

The previous chapter showed patterns in context. This chapter goes deeper into verification techniques—what to actually check and how.

This is not a checklist to follow mindlessly. It's a toolkit to draw from intelligently. Different code needs different verification. Your job is to select the right tools for the situation.

1.29.2 Part 1: Functional Verification

Functional verification answers: "Does this code do what it's supposed to do?"

1.29.2.1 Technique 1: Input Space Partitioning

The input space is everything a function could receive. Testing everything is impossible. Partitioning makes it tractable.

The Approach:

Divide inputs into categories that should behave the same way. Test at least one representative from each category.

How to Partition:

For numeric inputs: - Negative numbers - Zero - Positive numbers - Very large numbers (near overflow) - Very small numbers (precision issues) - Special values (NaN, Infinity if applicable)

For string inputs: - Empty string - Single character - Typical string - Very long string - Unicode characters - Special characters (quotes, backslashes, newlines) - Null (if the language allows)

For collections: - Empty collection - Single element - Multiple elements - Very large collection - Collection with duplicates - Collection with special values (nulls, etc.)

For dates: - Normal date - Leap year date (Feb 29) - Month boundaries - Year boundaries - Time zone edge cases - DST transitions

Example:

For a function `get_element_at_index(list, index)`:

List partitions: - Empty list → should handle gracefully - Single-element list → index 0 valid, others invalid - Multi-element list → various valid and invalid indices

Index partitions: - Negative index → might be valid (Python) or invalid - Zero → valid if list non-empty - Middle valid index → normal case - Last valid index → boundary - One past end → invalid - Very large → invalid

Combined test cases: 1. Empty list, index 0 → error handling 2. Single element, index 0 → returns element 3. Single element, index 1 → error handling 4. Multi-element, index 0 → returns first 5. Multi-element, middle index → returns middle 6. Multi-element, last index → returns last 7. Multi-element, past end → error handling 8. Multi-element, negative → depends on language semantics

1.29.2.2 Technique 2: Boundary Value Analysis

Bugs cluster at boundaries. If something works for 5 and 100, it probably works for 50. But it might fail for 0, 1, or the maximum.

The Approach:

For every boundary, test: - The boundary value itself - Just below the boundary (boundary - 1) - Just above the boundary (boundary + 1)

Common Boundaries:

- Array indices: 0, length-1, length
- Numeric ranges: min, max
- String lengths: 0, 1, max
- Date ranges: start of range, end of range
- Pagination: first page, last page

Example:

For a function `is_valid_percentage(value)` that should accept 0-100:

Boundaries: 0 and 100

Test cases: - -1 → should be invalid - 0 → should be valid - 1 → should be valid - 99 → should be valid - 100 → should be valid - 101 → should be invalid

Also consider: - What about 50.5? Are decimals valid? - What about null/undefined?

1.29.2.3 Technique 3: State Transition Testing

For code with state, verify that state transitions are correct.

The Approach:

1. Identify all possible states
2. Identify all events that cause transitions
3. Map which transitions are valid
4. Test that valid transitions work
5. Test that invalid transitions are rejected

Example:

For an order state machine:

States: Created, Paid, Shipped, Delivered, Cancelled, Refunded

Valid transitions: - Created → Paid (payment received) - Created → Cancelled (user cancels) - Paid → Shipped (fulfillment ships) - Paid → Refunded (refund before ship) - Shipped → Delivered (delivery confirmed) - Shipped → Refunded (refund after ship)

Invalid transitions: - Delivered → Cancelled (can't cancel delivered order) - Cancelled → anything (cancelled is terminal) - Shipped → Paid (can't go backward)

Test each valid transition. Test key invalid transitions to ensure they're rejected.

1.29.2.4 Technique 4: Property-Based Reasoning

Instead of specific test cases, verify properties that should always hold.

The Approach:

Identify invariants—things that should always be true—and verify them.

Common Properties:

- **Round-trip:** encode then decode should return original
- **Idempotency:** doing something twice should be same as once (when applicable)
- **Symmetry:** reverse(reverse(x)) should equal x
- **Ordering:** sort(x) should be ordered
- **Size preservation:** map over a list should preserve length
- **Bounds:** output should be within expected range

Example:

For a serialization function:

Property: For all valid inputs, `deserialize(serialize(input)) == input`

For a sort function:

Properties: - Output length equals input length - Every element in input exists in output (same count) - Output is ordered (each element next element)

For an idempotent API endpoint:

Property: `POST /resource` twice with same data should have same result as once

---

1.29.3 Part 2: Security Verification

Security verification answers: "Can this code be exploited?"

1.29.3.1 Technique 1: Input Vector Enumeration The Approach:

List every input to the code. For each, ask: "What if this came from an attacker?"

Input Vectors Include:

- URL parameters
- Form fields
- HTTP headers
- Cookies
- File uploads
- Database contents (might be poisoned)
- Environment variables (in some contexts)
- File system contents
- Data from external APIs

For Each Input:

- What's the expected format?
- What happens with unexpected format?
- Can special characters cause problems?
- Can excessive length cause problems?
- Can the input influence SQL, commands, HTML, or other interpreted contexts?

Example:

For a user profile update endpoint:

Inputs: - Username (from form) - Email (from form) - Bio (from form) - Profile picture (file upload) - Session token (cookie) - User ID (might be in URL or derived from session)

For each: - Username: Could contain script tags (XSS if displayed), SQL injection characters, or path traversal characters - Email: Could be malformed, excessively long, contain injection characters - Bio: Prime XSS target, could contain huge content (DoS) - Profile picture: Could be disguised executable, excessively large, specially crafted to exploit image processors - Session token: Could be forged, stolen, expired - User ID: Could be manipulated to update someone else's profile

1.29.3.2 Technique 2: Injection Point Analysis The Approach:

Find every place where input data is combined with commands, queries, or markup. These are injection points.

Common Injection Contexts:

SQL:

```
# DANGEROUS
query = f"SELECT * FROM users WHERE name = '{user_input}'"
```

```
# Input: "'; DROP TABLE users; --"
# Result: SQL injection
```

Shell commands:

```
# DANGEROUS
os.system(f"convert {user_filename} output.png")

# Input: "file.jpg; rm -rf /"
# Result: Command injection
```

HTML:

```
# DANGEROUS
html = f"<div>Welcome, {username}</div>"

# Input: "<script>steal_cookies()</script>"
# Result: XSS
```

File paths:

```
# DANGEROUS
path = f"/app/data/{user_filename}"

# Input: "../../../etc/passwd"
# Result: Path traversal
```

Verification Questions:

- Is user input ever concatenated into SQL? → Use parameterized queries
- Is user input ever passed to shell? → Avoid if possible; whitelist if necessary
- Is user input ever rendered as HTML? → Escape output
- Is user input ever used in file paths? → Validate against whitelist

1.29.3.3 Technique 3: Authentication and Authorization Audit The Approach:

For every action the code enables, verify: 1. Does it require authentication? Should it? 2. Does it verify authorization? Should it? 3. Can authentication be bypassed? 4. Can authorization be bypassed?

Common Vulnerabilities:

Broken authentication: - Weak password requirements - No brute force protection - Predictable session tokens - Session doesn't expire - Password in URL (logged)

Broken authorization: - Checking if user is logged in, but not if they own the resource - Client-side authorization checks only - Assuming sequential IDs can't be guessed - Not re-checking permissions on sensitive actions

Verification Questions:

- What proves this user is who they claim?
- What proves this user can perform this action on this resource?

- Where are these checks? Are they before the action?
- Can they be bypassed by modifying the request?

Example:

For an endpoint `DELETE /api/documents/{id}`:

Questions: - Does it require authentication? (Should: yes) - Does it verify the user owns document {id}? (Should: yes) - What if I'm logged in as User A and request delete of User B's document? - What if the ID is guessed or enumerated?

1.29.3.4 Technique 4: Data Exposure Analysis The Approach:

Trace every piece of sensitive data through the code. Where does it go? Could it leak?

Sensitive Data Includes:

- Passwords (should never be visible)
- API keys and secrets
- Personal information (PII)
- Financial information
- Session tokens
- Internal system information

Exposure Points:

- Log files (are sensitive values logged?)
- Error messages (do they reveal internal information?)
- API responses (is too much data returned?)
- URLs (sensitive data in query strings is logged)
- Client-side storage (localStorage is readable by scripts)
- Source code (hardcoded secrets?)

Verification Questions:

- Is this sensitive data logged anywhere?
- Does any error message reveal this data?
- Is this data sent to the client when it shouldn't be?
- Is this data stored securely at rest?
- Is this data transmitted securely in transit?

Example:

For a login function:

Sensitive data: password, session token

Trace: - Password received → is it ever logged? (Should: no) - Password compared → is the comparison constant-time? (Prevents timing attacks) - Password stored → is it hashed properly? (bcrypt, argon2, not MD5/SHA1) - Session token generated → is it cryptographically random? (Should: yes) - Session token stored → is it accessible to JavaScript? (httpOnly flag) - Session token transmitted → is it HTTPS only? (secure flag)

1.29.4 Part 3: Maintainability Verification

Maintainability verification answers: "Can this code be understood and safely modified in the future?"

1.29.4.1 Technique 1: The Stranger Test The Approach:

Pretend you've never seen the code. Time how long it takes to understand: 1. What does this code do? (purpose) 2. How does it do it? (mechanism) 3. Why does it do it this way? (rationale)

Interpretation:

- $<$ 5 minutes for simple code: Good
- 5-15 minutes: Acceptable for complex logic
- 15+ minutes for code that should be simple: Problem

What Makes Understanding Hard:

- Poor naming (what is `x`? what is `process`?)
- Long functions ($>$ 30 lines needs scrutiny, $>$ 100 lines is a problem)
- Deep nesting (more than 3 levels needs refactoring)
- Magic values (what is 86400? what is 0.035?)
- Missing context (why is this special case here?)
- Clever tricks (clever code is hard to debug)

1.29.4.2 Technique 2: Change Impact Assessment The Approach:

Consider likely future changes. How hard would they be?

Common Changes:

- Adding a new case/variant to existing logic
- Changing a business rule or threshold
- Integrating with a new external system
- Adding logging or monitoring
- Fixing a bug in one scenario without breaking others

For Each Change:

- What code would need to be modified?
- What else might be affected?
- How confident are you that you could make the change without breaking something?

Red Flags:

- "I'd need to change this in many places" $\rightarrow$ duplication problem
- "I'm not sure what else this would affect" $\rightarrow$ coupling problem
- "I'd need to understand all of this to change any of it" $\rightarrow$ cohesion problem
- "I'd be afraid to touch this" $\rightarrow$ clarity problem

1.29.4.3 Technique 3: Naming Audit The Approach:

Read just the names—function names, variable names, parameter names. Do they tell a story?

Good Names:

- Functions: verbs that describe what they do (`calculate_shipping_cost`, `validate_email`, `send_notification`)
- Variables: nouns that describe what they hold (`user_email`, `total_price`, `is_active`)
- Booleans: questions that make sense in conditions (`is_valid`, `has_permission`, `should_retry`)
- Constants: describe what the value means (`MAX_RETRY_COUNT`, `SECONDS_PER_DAY`)

Bad Names:

- Single letters (except loop counters): `x`, `d`, `t`
- Generic names: `data`, `temp`, `result`, `value`, `item`
- Misleading names: `list` that's actually a dictionary, `count` that's actually a sum
- Abbreviated names: `usr`, `pwd`, `cfg` (use `user`, `password`, `config`)

Verification:

For each name, can you understand its purpose without reading the implementation?

1.29.4.4 Technique 4: Complexity Assessment The Approach:

Identify complexity and decide if it's justified.

Complexity Indicators:

- **Cyclomatic complexity:** Number of decision points (if, while, for, case). Higher = more complex. Over 10 warrants scrutiny.
- **Nesting depth:** How deep do control structures nest? More than 3 levels is hard to follow.
- **Function length:** Longer functions are harder to understand. Over 30 lines warrants scrutiny.
- **Parameter count:** Functions with many parameters are hard to use correctly. Over 4 parameters warrants scrutiny.
- **Dependencies:** How many other modules does this code depend on? More dependencies = more fragile.

Complexity Isn't Always Bad:

Some problems are inherently complex. The question is whether the code's complexity matches the problem's complexity.

- Simple problem, complex code → refactor
- Complex problem, complex code → acceptable but document
- Complex problem, simple code → ideal (if it's actually correct)

1.29.5 Part 4: Provenance Verification

Provenance verification answers: "Where did this code come from, and what does that imply?"

1.29.5.1 Technique 1: Origin Documentation The Approach:

For every significant piece of code, document: - Who/what created it (human, AI, copied from somewhere) - When it was created - What verification it received

For AI-Generated Code:

- Which model/tool generated it
- What prompt produced it (summary if long)
- What verification was performed after generation
- Who verified it

Why It Matters:

- AI-generated code might have training data issues
- Code that was lightly reviewed is higher risk for modification
- Knowing origin helps assess appropriate trust level

1.29.5.2 Technique 2: License Compatibility Check The Approach:

Verify that code origins are compatible with your project's licensing.

Check:

- If code was copied from external sources, what's the license?
- If AI-generated, what are the terms of service? Any IP concerns?
- Are all dependencies' licenses compatible with your project?

Common Issues:

- Copying GPL code into proprietary project
- Using code from Stack Overflow without attribution
- AI generating code that closely matches training data (potential copyright issue)

1.29.5.3 Technique 3: Pattern Origin Tracking The Approach:

When you see a pattern repeated, trace its origin.

Questions:

- Where did this pattern start?
- Who introduced it?
- Was it verified when introduced?
- Has it been modified since?

Why It Matters:

- A bad pattern spreads through copy-paste
- Fixing the pattern requires finding all copies
- Understanding origin helps assess risk

1.29.6 Using the Toolkit

These techniques form a toolkit, not a checklist. Selection depends on context.

High Trust (Low Risk): - Quick functional verification (key test cases) - Naming audit (basic clarity) - Skip intensive security analysis

Moderate Trust (Typical Risk): - Systematic functional verification (input partitioning, boundary analysis) - Basic security review (input vectors, obvious injection points) - Maintainability review (stranger test, naming audit) - Provenance documentation

Guarded Trust (Important Risk): - Comprehensive functional verification (all techniques) - Intensive security review (all techniques) - Full maintainability assessment (all techniques) - Formal provenance documentation - Multiple reviewers

Minimal Trust (Intensive Verification): - Everything in Guarded Trust - Dedicated security review or threat modeling session - Formal provenance documentation with reviewer sign-off - Incident response considerations

The skill is matching technique to need. Over time, you develop intuition for what each situation requires. The techniques become second nature—you apply them without consciously thinking through lists.

That's mastery: not following a checklist, but having internalized the toolkit so deeply that you automatically apply what's needed.

---

2 Chapter 21: Test Verification Framework

2.1 The Problem with AI-Generated Tests

When AI generates your code, you verify it with tests. But when AI also generates your tests, who verifies the tests?

This isn't academic. AI-generated tests can: - **Pass while missing critical bugs** — Tests check what AI thought to check, not necessarily what matters - **Test the wrong thing** — Verifying implementation details instead of behavior - **Contain the same logical errors as the code** — AI makes correlated mistakes - **Provide false confidence** — "100% coverage" means nothing if tests are weak

This chapter provides a framework for verifying tests themselves—whether AI-generated or human-written.

> **Where this fits:** Chapter 21 lives in the Advanced Topics section (not the appendices) because meta-verification becomes essential once the core VID practices are in place. Treat it as a continuation of Chapter 20's toolkit rather than optional reference material.

2.2 Core Principle: Tests Are Code

Tests require verification, just like production code.

The difference: tests verify production code, so we need *meta-verification* strategies to verify tests.

Meta-verification means using techniques that validate tests without requiring another layer of tests (avoiding infinite regression).

2.3 Part 1: Test Provenance

2.3.1 Why Test Provenance Matters

The origin of a test affects how we verify it:

Human-written test - Likely reflects human understanding of requirements - May miss edge cases humans didn't consider - Usually tests behavior, not implementation

AI-generated from specification - Only as good as the specification - Might test what was specified, not what was intended - May miss implicit requirements

AI-generated from code - Likely tests implementation, not behavior - May duplicate bugs in the code logic - Provides weak verification

AI-generated from examples - Tests known cases well - May miss untested variations - Brittle to code changes

2.3.2 Tracking Test Provenance

Mark tests with their origin:

```
def test_calculate_prorated_charge():
    """
    Test prorated subscription charge calculation.

    Provenance: AI-generated from specification
    Verified: 2025-12-07 by @engineer
    Review: Expanded edge cases for leap years
    """
    # Test cases...
```

Minimum provenance documentation: - How was this test created? - Who verified it? - What verification was performed?

2.4 Part 2: Meta-Verification Strategies

2.4.1 Strategy 1: Mutation Testing

Concept: Introduce bugs in the code. If tests don't fail, the tests are weak.

How it works: 1. Take working code with passing tests 2. Make small changes (mutations): - Change `<` to `<=` - Change `+` to `-` - Remove a conditional - Change a constant 3. Run tests 4. If tests still pass, they didn't catch the mutation → weak tests

Example:

```
# Original code
def is_valid_age(age):
    return 0 <= age < 120

# Mutation 1: Change <= to <
def is_valid_age(age):
```

```
    return 0 < age < 120  # Should reject age=0

# Mutation 2: Change 120 to 121
def is_valid_age(age):
    return 0 <= age < 121  # Should reject age=120
```

Question: Do your tests catch these mutations?

If not, you need tests like:

```
def test_age_boundary_zero():
    assert is_valid_age(0) == True  # Catches mutation 1

def test_age_boundary_max():
    assert is_valid_age(120) == False  # Catches mutation 2
```

Tools: - Python: `mutmut`, `cosmic-ray` - JavaScript: `Stryker` - Java: `PIT` - General: `universal-mutator`

VID Practice: - Run mutation testing on AI-generated tests - Aim for >80% mutation score (percentage of mutations caught) - If mutations survive, add tests

---

2.4.2 Strategy 2: Property-Based Testing

Concept: Instead of testing specific examples, test properties that must always hold.

Why it helps: Properties are harder for AI to get wrong than specific test cases.

Example:

Weak (example-based):

```
def test_reverse():
    assert reverse([1, 2, 3]) == [3, 2, 1]
    assert reverse([5]) == [5]
```

Strong (property-based):

```
@given(lists(integers()))
def test_reverse_properties(lst):
    # Property 1: Reversing twice returns original
    assert reverse(reverse(lst)) == lst

    # Property 2: Length preserved
    assert len(reverse(lst)) == len(lst)

    # Property 3: Elements preserved (same set)
    assert set(reverse(lst)) == set(lst)

    # Property 4: First becomes last
```

```
    if lst:
        assert reverse(lst)[0] == lst[-1]
```

Tools: - Python: `hypothesis` - JavaScript: `fast-check` - Haskell: `QuickCheck` - Java: `jqwik`

VID Practice: - For critical code, supplement AI-generated tests with property tests - Properties are human insight—harder for AI to generate - Use properties to validate AI-generated example tests

---

2.4.3 Strategy 3: Test Review Questions

Systematically review tests using these questions:

2.4.3.1 Correctness Questions

- **Does this test actually test what it claims?**
  - Read the test name and docstring
  - Read the test code
  - Do they match?
- **Are the assertions correct?**
  - Is `assert result == expected` the right comparison?
  - Are edge cases handled correctly?
  - Are error cases actually supposed to error?

2.4.3.2 Completeness Questions

- **What's not being tested?**
  - Edge cases: empty, null, min, max
  - Error conditions
  - Integration points
  - Security scenarios
- **What assumptions do the tests make?**
  - "Input is always valid"
  - "Database is empty before test"
  - "System clock doesn't change mid-test"
  - Are these assumptions documented?

2.4.3.3 Quality Questions

- **Do these tests test behavior or implementation?**
  - Good: "User can log in with correct credentials"
  - Bad: "Login function calls check_password and then create_session"
- **Are the tests brittle?**
  - Will they break if we refactor without changing behavior?
  - Do they depend on implementation details?
- **Are the tests readable?**
  - Can someone else understand what's being tested?
  - Are test names descriptive?

2.4.4 Strategy 4: Coverage Analysis (Used Correctly)

Coverage is necessary but not sufficient.

What coverage tells you: - Which lines of code are executed by tests - Which branches are tested

What coverage doesn't tell you: - Whether assertions are correct - Whether edge cases are tested - Whether tests would catch real bugs

VID Approach to Coverage:

1. **Use coverage to find untested code**

```
# Find what's not covered
pytest --cov --cov-report=html
# Review uncovered lines: are they testable? Should they be tested?
```

2. **Don't worship 100% coverage**
   - Some code is hard to test and low-risk (fine to skip)
   - 100% coverage with weak tests = false security
   - 80% coverage with strong tests > 100% coverage with weak tests
3. **Check coverage of critical paths**
   - Payment processing: 100% coverage required + mutation testing
   - Error handling: explicitly test error paths
   - Security logic: not just covered, but tested adversarially

---

2.4.5 Strategy 5: Adversarial Test Review

For security-critical code, review tests adversarially.

Questions: - **What attack vectors are NOT tested?** - SQL injection attempts - XSS payloads - Path traversal - Integer overflow - Race conditions

- **Do tests verify security properties?**
  - Authentication required
  - Authorization checked
  - Input sanitized
  - Output escaped
  - Secrets not logged

Example:

Weak test:

```
def test_delete_document():
    response = client.delete('/api/documents/123')
    assert response.status_code == 200
```

Strong test:

```
def test_delete_document_authorization():
    # Create document owned by user A
    doc_id = create_document(user_a)

    # Try to delete as user B
    response = client.delete(f'/api/documents/{doc_id}',
                             auth=user_b_token)
    assert response.status_code == 403  # Forbidden

    # Verify document still exists
    assert document_exists(doc_id)

    # Delete as user A should succeed
    response = client.delete(f'/api/documents/{doc_id}',
                             auth=user_a_token)
    assert response.status_code == 200
    assert not document_exists(doc_id)
```

2.5 Part 3: AI in Test Verification

2.5.1 Where AI Can Help

AI is not just for generation—it can assist verification:

AI can: - Generate additional test cases from specifications - Suggest edge cases humans might miss - Identify untested code paths - Generate adversarial inputs - Check for common vulnerability patterns

AI cannot: - Guarantee tests are correct - Understand semantic intent - Assess business rule compliance - Make trust calibration decisions

2.5.2 Human-AI Division of Labor in Testing

AI generates: - Initial test suite from specifications - Additional test cases for edge cases - Property-based test generators - Mutation testing variants

Humans verify: - Tests match requirements (not just specifications) - Assertions are correct - Coverage is appropriate to risk level - Security properties are tested - Tests will catch real bugs

Together: - AI generates 100 test cases - Human reviews with mutation testing - Mutation testing shows gaps - AI generates additional tests to fill gaps - Human validates the additions

---

2.6 Part 4: Test Quality Checklist

Use this checklist when reviewing tests (AI-generated or human-written):

2.6.1 Before Accepting Tests

- ☐ **Provenance documented** — How were these tests created?
- ☐ **Test names describe behavior** — Not implementation details
- ☐ **Assertions are correct** — Verified, not assumed
- ☐ **Edge cases covered** — Empty, null, boundaries, max values
- ☐ **Error cases tested** — Not just happy paths
- ☐ **Coverage appropriate to risk** — Critical code has thorough tests

2.6.2 For Important Code (High Trust Level)

- ☐ **Mutation testing run** — Mutations mostly caught (>80%)
- ☐ **Property tests added** — Key invariants tested
- ☐ **Security scenarios tested** — Adversarial inputs tried
- ☐ **Integration tested** — Not just unit tests
- ☐ **Tests are readable** — Others can understand them
- ☐ **Tests are maintainable** — Won't break on refactoring

2.6.3 For Critical Code (Minimal Trust Level)

- ☐ **Independent test review** — Second person reviewed
- ☐ **Formal specifications** — Properties formally stated
- ☐ **Comprehensive mutation testing** — >90% mutation score
- ☐ **Adversarial testing** — Security team reviewed
- ☐ **Documentation** — Test strategy documented
- ☐ **Regression suite** — Tests will catch known issues

2.7 Part 5: Common Test Anti-Patterns

2.7.1 Anti-Pattern 1: "Tests That Always Pass"

Problem: Test passes regardless of code correctness.

Example:

```
def test_calculate_total():
    result = calculate_total([10, 20, 30])
    assert result  # Just checks it's not None/0
```

Why it's wrong: Test doesn't verify correctness, just that something was returned.

Fix:

```
def test_calculate_total():
    result = calculate_total([10, 20, 30])
    assert result == 60  # Actual expected value
```

2.7.2 Anti-Pattern 2: "Tautological Tests"

Problem: Test verifies code against itself.

Example:

```
def test_process_data():
    data = [1, 2, 3]
    result = process_data(data)
    expected = process_data(data)  # Calling same function!
    assert result == expected
```

Why it's wrong: If `process_data` is buggy, both calls are buggy. Test always passes.

Fix: Define expected value independently:

```
def test_process_data():
    data = [1, 2, 3]
    result = process_data(data)
    expected = [2, 4, 6]  # Manually calculated
    assert result == expected
```

2.7.3 Anti-Pattern 3: "Tests That Test Implementation"

Problem: Tests break when code is refactored (even if behavior unchanged).

Example:

```
def test_user_login():
    # Bad: Testing implementation details
    user = User("alice", "password123")
    user._hash_password()  # Testing private method
    assert user.password_hash is not None
    assert user._check_password("password123")  # Internal method
```

Why it's wrong: Refactoring internals breaks test, even if login still works.

Fix: Test public behavior:

```
def test_user_login():
    # Good: Testing behavior
    user = User.create("alice", "password123")
    assert user.login("password123") == True
    assert user.login("wrongpass") == False
```

2.7.4 Anti-Pattern 4: "Correlation Bugs"

Problem: AI generates code and tests with same logical error.

Example:

AI-generated code:

```
def days_in_month(month, year):
    if month == 2:
        return 28  # BUG: Ignores leap years
    elif month in [4, 6, 9, 11]:
        return 30
    else:
        return 31
```

AI-generated test (from same context):

```
def test_days_in_month_february():
    assert days_in_month(2, 2024) == 28  # BUG: 2024 is leap year
```

Why it's wrong: AI made same mistake in test and code. Test passes, bug ships.

Detection: Human review or property testing:

```
def test_leap_year_property():
    # 2024 is leap year (divisible by 4, not by 100)
    assert days_in_month(2, 2024) == 29
    # 2000 is leap year (divisible by 400)
    assert days_in_month(2, 2000) == 29
    # 1900 is not leap year (divisible by 100, not 400)
    assert days_in_month(2, 1900) == 28
```

2.8 Part 6: Test Verification Workflow

2.8.1 For AI-Generated Tests (Low-Risk Code)

1. **Read tests** — Understand what they claim to test
2. **Run tests** — Verify they pass with correct code
3. **Spot-check assertions** — Pick 2-3 tests, verify assertions are right
4. **Coverage check** — Are critical paths covered?
5. **Add 1-2 edge cases** — What did AI miss?
6. **Accept if sufficient**

Time: 5-10 minutes

2.8.2 For AI-Generated Tests (Important Code)

1. **All steps from low-risk, plus:**
2. **Mutation testing** — Run mutator, check survival rate
3. **Review security tests** — Are attack vectors tested?
4. **Add property tests** — Define key invariants
5. **Independent review** — Second person reviews tests
6. **Document test strategy** — What's our testing approach?

Time: 30-60 minutes

2.8.3 For AI-Generated Tests (Critical Code)

1. **All steps from important code, plus:**
2. **Comprehensive mutation testing** — Achieve >90% mutation score
3. **Adversarial review** — Security specialist reviews
4. **Formal specification** — Write formal properties
5. **Multiple reviewers** — 2+ people review independently
6. **Regression verification** — Do tests catch known past bugs?
7. **Document thoroughly** — Test plan, rationale, gaps

Time: 2-4 hours

2.9 Part 7: Tools & Techniques Summary

Technique	Purpose	When to Use	Tools
Mutation Testing	Verify tests catch bugs	Important+ code	mutmut, Stryker, PIT
Property Testing	Test invariants	Complex logic	Hypothesis, fast-check
Coverage Analysis	Find untested code	All code	pytest-cov, Istanbul
Adversarial Testing	Find security gaps	Security-critical	Manual + fuzzing
Test Review	Human judgment	All AI-generated tests	Checklist
AI-Assisted Generation	Create more test cases	When gaps found	GPT-4, Copilot

---

2.10 Part 8: Integration with VID Principles

2.10.1 Principle 1: Intent Before Generation

For tests: - Specify what should be tested before generating tests - List expected behaviors, edge cases, error conditions - Define properties that must hold

2.10.2 Principle 2: Graduated Trust

For tests: - Low-risk code: Basic test review - Important code: Mutation testing - Critical code: Comprehensive verification with multiple techniques

2.10.3 Principle 3: Understanding Over Acceptance

For tests: - Never accept tests you don't understand - If a test is unclear, rewrite it - Understand what each test verifies and why

2.10.4 Principle 4: Provenance Awareness

For tests: - Track how tests were created - Note what verification was performed - Distinguish AI-generated from human-written

2.10.5 Principle 5: Continuous Calibration

For tests: - When bugs escape to production, check: did tests miss them? - Update test strategy based on what you learn - Add regression tests for production bugs

---

2.11 Conclusion

Tests are not automatically trustworthy just because they pass.

In the AI era, we must verify tests with the same rigor we verify code. Meta-verification strategies (mutation testing, property testing, systematic review) let us validate tests without infinite regression.

Key Takeaways:

1. **Test provenance matters** — Track how tests were created
2. **Mutation testing validates test quality** — Weak tests let mutations survive
3. **Properties complement examples** — Invariants are human insight
4. **AI can assist test verification** — But humans must arbitrate
5. **Coverage quality** — Use coverage to find gaps, not to declare victory
6. **Review tests systematically** — Use checklists appropriate to risk level

The meta-verification loop:

```
Code Generated → Tests Generated → Mutations Created
     ↑                                      ↓
Confidence Gained ← Tests Improved ← Gaps Identified
```

With this framework, tests become verification tools we can trust—not just rituals we perform.

Return to Chapter 20: The Verification Toolkit | Table of Contents | Next: Appendix A: Quick Reference

---

3 VID Real-World Examples

Five concrete scenarios showing VID principles in action

These examples demonstrate how VID practices catch real issues and prevent production problems. Each includes the code, the problem, and how VID would have prevented it.

3.1 Example 1: Subtle Logic Bug Caught by Verification Ritual

3.1.1 The Scenario

A developer asked an AI assistant to create a function to calculate user subscription renewal dates.

3.1.2 The Generated Code

```
function calculateRenewalDate(subscriptionStart, planType) {
  const renewalDate = new Date(subscriptionStart);

  if (planType === 'monthly') {
    renewalDate.setMonth(renewalDate.getMonth() + 1);
  } else if (planType === 'annual') {
    renewalDate.setFullYear(renewalDate.getFullYear() + 1);
  }

  return renewalDate;
}
```

3.1.3 The Problem

This code looks correct but has a subtle bug:

When a subscription starts on January 31st and renews monthly, `setMonth(2)` (March) adjusts to March 3rd because February doesn't have 31 days. This causes: - User charged on Jan 31, then March 3, then March 31, then May 3… - Inconsistent billing dates - Customer support nightmare

3.1.4 Without VID

Developer thought process: - "Looks good" - Tests pass for Jan 15 → Feb 15 - Ships to production - **3 months later:** Billing complaints from customers subscribed on the 29th-31st

3.1.5 With VID

Verification Ritual (15 minutes, Moderate Trust):

1. **Read the code**
2. **Verify against intent**
3. **Check edge cases:**
   - "What happens on the 31st of a month?"
   - Manual test: `calculateRenewalDate('2024-01-31', 'monthly')`
   - Result: `2024-03-03`
   - **BUG FOUND**
4. **Fix before commit:**

```
function calculateRenewalDate(subscriptionStart, planType) {
  const start = new Date(subscriptionStart);
  const dayOfMonth = start.getDate();

  let renewalDate = new Date(subscriptionStart);
```

```
  if (planType === 'monthly') {
    renewalDate.setMonth(renewalDate.getMonth() + 1);
    // Preserve original day of month
    renewalDate.setDate(Math.min(dayOfMonth, getLastDayOfMonth(renewalDate)));
  } else if (planType === 'annual') {
    renewalDate.setFullYear(renewalDate.getFullYear() + 1);
    // Handle Feb 29 on non-leap years
    renewalDate.setDate(Math.min(dayOfMonth, getLastDayOfMonth(renewalDate)));
  }

  return renewalDate;
}

function getLastDayOfMonth(date) {
  return new Date(date.getFullYear(), date.getMonth() + 1, 0).getDate();
}
```

Time invested: 15 minutes **Time saved:** Hours of debugging + customer support + reputation damage

VID Principles Applied: - **Principle 2 (Graduated Trust):** Moderate Trust verification (15 min) matched to financial impact risk - **Principle 1 (Intent Before Generation):** Edge case thinking ("what happens on the 31st?") came from upfront intent specification - **Practice: Verification Ritual:** Systematic edge case testing caught the bug before production

Key Lesson: Edge case verification is where AI-generated code most often fails. The code *looks* correct and handles the happy path perfectly. Only systematic verification catches month-boundary bugs.

3.2 Example 2: Provenance Awareness Prevents Regression

3.2.1 The Scenario

Six months ago, an AI assistant generated an authentication middleware. Today, a developer needs to modify it to add rate limiting.

3.2.2 The Original Code (AI-Generated)

```
# Generated by AI - 2024-06-15
# Verified by: @sarah - Standard verification
def authenticate_request(request):
    """Verify JWT token and attach user to request."""
    token = request.headers.get('Authorization', '').replace('Bearer ', '')

    try:
        payload = jwt.decode(token, SECRET_KEY, algorithms=['HS256'])
        request.user = User.get(payload['user_id'])
```

```
        return True
    except jwt.ExpiredSignatureError:
        return False
    except jwt.InvalidTokenError:
        return False
```

3.2.3 Without Provenance Awareness

Developer thinks: "I wrote this, I understand it."

Modification:

```
def authenticate_request(request):
    """Verify JWT token, check rate limits, attach user."""
    token = request.headers.get('Authorization', '').replace('Bearer ', '')

    payload = jwt.decode(token, SECRET_KEY, algorithms=['HS256'])

    # Add rate limiting
    if rate_limiter.is_exceeded(payload['user_id']):
        return False

    request.user = User.get(payload['user_id'])
    return True
```

What broke: - Removed try-except blocks (didn't understand they were essential) - Now raises unhandled exceptions on invalid tokens - **500 errors instead of graceful 401s** - Production incident at 2AM

3.2.4 With Provenance Awareness

Developer sees comment: `# Generated by AI - 2024-06-15`

Thought process: - "I didn't write this originally" - "I should verify I understand it before modifying" - "Why are there try-except blocks?" - Checks original intent specification: "Must handle expired/invalid tokens gracefully"

Modification (safe):

```
# Modified from AI-generated code - 2024-12-15
# Original: 2024-06-15 (AI-generated, verified by @sarah)
# Changes: Added rate limiting (verified by @mike)
def authenticate_request(request):
    """Verify JWT token, check rate limits, attach user."""
    token = request.headers.get('Authorization', '').replace('Bearer ', '')

    try:
        payload = jwt.decode(token, SECRET_KEY, algorithms=['HS256'])

        # NEW: Rate limiting
        if rate_limiter.is_exceeded(payload['user_id']):
```

```
            return False

        request.user = User.get(payload['user_id'])
        return True
    except jwt.ExpiredSignatureError:
        return False
    except jwt.InvalidTokenError:
        return False
```

Provenance hygiene saved: - Production incident avoided - Original error handling preserved - Clear modification history for next developer

VID Principles Applied: - **Principle 4 (Provenance Awareness):** Comment marking AI-generated code triggered appropriate caution - **Principle 3 (Understanding Over Acceptance):** Developer verified understanding before modifying unfamiliar code - **Practice: Provenance Hygiene:** Commit messages and code comments documented who generated what and when

Key Lesson: AI-generated code looks identical to human-written code six months later. Without provenance markers, developers assume they understand code they didn't write and don't remember. This leads to cargo-cult modifications that break subtle invariants.

3.3 Example 3: Risk Miscalibration

3.3.1 The Scenario

A developer needs to add a feature to allow users to delete their accounts.

3.3.2 Initial Risk Assessment (Wrong)

Developer thinks: - "This is straightforward CRUD" - "Just add a DELETE endpoint" - **Risk score: 8 (High Trust)** - Verification time: 5 minutes

3.3.3 The Generated Code

```
app.delete('/api/users/:userId', async (req, res) => {
  await db.users.delete({ id: req.params.userId });
  res.json({ success: true });
});
```

3.3.4 The Problem

Actual considerations missed: - **Impact:** Deletes all user data (irreversible) → Score: 5 - **Reversibility:** Cannot undo → Score: 5 - **Exposure:** Affects individual users → Score: 3 - **Compliance:** GDPR requires specific deletion process → Score: 8

Actual risk score: (5×3) + (5×2) + (3×2) + 8 = 39 (Minimal Trust)

Should have required: - 1-3 hours verification - Peer review - Legal/compliance review - Production testing plan

3.3.5 What This Code Missed

1. **Authentication/Authorization:** Anyone can delete any account
2. **GDPR Compliance:** Must anonymize, not delete (for audit trail)
3. **Related Data:** Orphans user's posts, comments, purchases
4. **Cascading Effects:** Breaks foreign key relationships
5. **Audit Trail:** No record of deletion
6. **Grace Period:** No "undo" window
7. **Notifications:** User not notified
8. **Billing:** Active subscriptions not cancelled

3.3.6 With Proper Risk Calibration

Corrected assessment: - This is a critical, irreversible operation - Compliance requirements apply - Requires 2+ hours verification + peer review

Proper implementation:

```
app.delete('/api/users/:userId',
  requireAuth(),
  requireOwnershipOr('admin'),
  async (req, res) => {
    const userId = req.params.userId;

    // Start transaction
    const transaction = await db.transaction();

    try {
      // 1. Validate user exists and can be deleted
      const user = await db.users.findOne({ id: userId });
      if (!user) throw new NotFoundError();
      if (user.deletion_requested_at) {
        throw new Error('Deletion already in progress');
      }

      // 2. Cancel active subscriptions
      await cancelActiveSubscriptions(userId, transaction);

      // 3. Initiate GDPR-compliant anonymization
      await db.users.update(
        { id: userId },
        {
          deletion_requested_at: new Date(),
          deletion_scheduled_for: new Date(Date.now() + 30 * 24 * 60 * 60 * 1000),
          status: 'deletion_pending'
        },
```

```
      { transaction }
    );

    // 4. Queue background job for actual anonymization
    await queue.add('anonymize-user-data', {
      userId,
      scheduledFor: Date.now() + 30 * 24 * 60 * 60 * 1000
    });

    // 5. Audit log
    await audit.log({
      action: 'user_deletion_requested',
      userId,
      requestedBy: req.user.id,
      scheduledFor: new Date(Date.now() + 30 * 24 * 60 * 60 * 1000)
    });

    // 6. Notify user
    await sendDeletionConfirmationEmail(user.email);

    await transaction.commit();

    res.json({
      success: true,
      message: 'Account deletion scheduled for 30 days from now',
      cancellable_until: new Date(Date.now() + 30 * 24 * 60 * 60 * 1000)
    });

  } catch (error) {
    await transaction.rollback();
    throw error;
  }
});
```

Lesson: "Simple" operations can have complex requirements. Proper risk assessment prevents disasters.

VID Principles Applied: - **Principle 2 (Graduated Trust):** Corrected risk calibration from High Trust (8) to Minimal Trust (39) changed everything - **Principle 5 (Continuous Calibration):** Realized initial assessment was wrong and adjusted before deployment - **Practice: Risk Assessment (Appendix D):** Using the full rubric (Impact + Reversibility + Exposure + Compliance) revealed the true risk

Key Lesson: Familiarity bias is dangerous. "I've done CRUD operations before" doesn't mean *this* DELETE is low-risk. Systematically assessing all four risk factors (especially compliance!) prevents catastrophic under-verification.

3.4 Example 4: Missing Intent Leads to Wrong Solution

3.4.1 The Scenario

A developer wants to "improve query performance for the user dashboard."

3.4.2 Without Intent Specification

Prompt to AI: "Optimize this query"

```
SELECT * FROM users
JOIN posts ON users.id = posts.user_id
WHERE users.active = true;
```

AI generates:

```
SELECT * FROM users
JOIN posts ON users.id = posts.user_id
WHERE users.active = true
LIMIT 100;
```

What happened: - Query is faster - But now dashboard only shows 100 posts total (not per user) - **Completely wrong behavior** - Bug makes it to production

3.4.3 With Intent Specification (2 minutes upfront)

Developer writes intent:

```
Goal: Speed up dashboard query that shows all posts for active users

Current problem:
- Query takes 3 seconds
- Loads all users + all their posts
- 100,000+ rows scanned

Requirements:
- Must show ALL posts for active users (not just first N)
- Dashboard paginates on frontend
- Query should load in <500ms

Edge cases:
- Users with thousands of posts
- New users with no posts
- Deleted posts should be excluded

How I'll verify:
- Test with user who has 1000+ posts
- Check query execution plan
- Measure actual query time

Risk level: Moderate (impacts all users, but reversible)
```

Now the AI gets better context:

```
-- Optimization: Add index and filter deleted posts in JOIN
CREATE INDEX idx_posts_user_active ON posts(user_id, created_at)
  WHERE deleted_at IS NULL;

-- Optimized query
SELECT
  users.id,
  users.name,
  posts.id AS post_id,
  posts.title,
  posts.created_at
FROM users
JOIN posts ON users.id = posts.user_id AND posts.deleted_at IS NULL
WHERE users.active = true
ORDER BY posts.created_at DESC;
```

Verification reveals: - Still slow for users with many posts - Real solution: Paginate at query level, not frontend

Final solution:

```
-- Dashboard loads first page of posts per user
SELECT
  users.id,
  users.name,
  posts.id AS post_id,
  posts.title,
  posts.created_at
FROM users
JOIN LATERAL (
  SELECT * FROM posts
  WHERE posts.user_id = users.id
    AND posts.deleted_at IS NULL
  ORDER BY created_at DESC
  LIMIT 20
) posts ON true
WHERE users.active = true
ORDER BY users.last_active DESC
LIMIT 50;
```

Intent specification prevented: - Wrong solution (LIMIT 100) - Performance "optimization" that broke functionality - Multiple rounds of back-and-forth debugging

Time saved: Hours of debugging + production rollback

VID Principles Applied: - **Principle 1 (Intent Before Generation):** 2 minutes specifying requirements prevented deploying a "faster but wrong" solution - **Principle 3 (Understanding Over Acceptance):** Verification step ("test with user who has 1000+ posts") revealed the real problem - **Practice: Intent Specification (Chapter 9):** Documenting edge cases and success

criteria guided both AI and human judgment

Key Lesson: AI optimizes for what you ask, not what you need. "Make it faster" without context produces faster-but-wrong code. Spending 2 minutes on intent specification saves hours of debugging broken optimizations.

3.5 Example 5: Understanding Debt Compounds Over Time

3.5.1 Month 1: The Initial Generation

A junior developer asks AI to create a caching layer:

```
# AI-generated - Not fully understood by developer
class CacheManager:
    def __init__(self):
        self._cache = {}
        self._locks = {}

    def get_or_compute(self, key, compute_fn, ttl=3600):
        if key in self._cache:
            value, expiry = self._cache[key]
            if time.time() < expiry:
                return value

        # Double-checked locking pattern
        if key not in self._locks:
            self._locks[key] = threading.Lock()

        with self._locks[key]:
            # Check again inside lock
            if key in self._cache:
                value, expiry = self._cache[key]
                if time.time() < expiry:
                    return value

            value = compute_fn()
            self._cache[key] = (value, time.time() + ttl)
            return value
```

Developer thinks: "Looks good, tests pass, ship it" **Understanding level:** 30% (doesn't understand double-checked locking or why it's needed)

3.5.2 Month 3: First Modification

Add cache invalidation:

```
def invalidate(self, key):
    if key in self._cache:
        del self._cache[key]  # BUG: Doesn't clean up lock
```

Problem: Lock dict grows forever (memory leak) **Not caught:** Developer doesn't understand lock management

3.5.3 Month 6: Second Modification

Add cache statistics:

```
def get_or_compute(self, key, compute_fn, ttl=3600):
    self.stats['requests'] += 1  # BUG: Not thread-safe

    if key in self._cache:
        value, expiry = self._cache[key]
        if time.time() < expiry:
            self.stats['hits'] += 1  # BUG: Race condition
            return value

    # ... rest unchanged
```

Problem: Race conditions on stats counter **Not caught:** Developer doesn't understand threading implications

3.5.4 Month 9: Third Modification

Add cache warming:

```
def warm_cache(self, items):
    for key, compute_fn in items:
        self._cache[key] = (compute_fn(), time.time() + 3600)
        # BUG: Bypasses locking, causes race conditions
```

Problem: Direct cache writes without locking **Not caught:** Developer still doesn't understand the locking pattern

3.5.5 Month 12: Production Crisis

Symptoms: - Random cache corruption - Memory leaks (10GB+ cache) - Race conditions causing data inconsistencies - Occasional deadlocks

Cost: - 3 days of senior engineer debugging - Production incidents - Data inconsistencies requiring manual fixes - Complete rewrite needed

3.5.6 The Understanding Debt Trajectory

```
Month 1:  30% understanding → Ship with hidden issues
Month 3:  25% understanding → Introduce memory leak
Month 6:  20% understanding → Add race conditions
Month 9:  15% understanding → Bypass safety mechanisms
Month 12: CRISIS → 3 days to understand + rewrite
```

Total cost: 3 days senior engineer time + production incidents + data cleanup

3.5.7 With VID (Month 1)

Verification ritual includes: "Explain this code to yourself"

Developer realizes: - "I don't understand why there are two cache checks" - "What is this lock dict for?" - "Why check the cache again inside the lock?"

Invokes Principle 3: Understanding Over Acceptance

Actions: - Researches double-checked locking - Understands race conditions - Documents why pattern is needed - Tests concurrent access - **Builds 90% understanding**

Time invested: 1 hour

Future modifications (Months 3, 6, 9): - Developer understands threading implications - Knows to use locks for all cache mutations - Knows to use locks for stats updates - Avoids all bugs from original timeline

Total cost: 1 hour upfront

Savings: 3 days debugging + production incidents + rewrites

VID Principles Applied: - **Principle 3 (Understanding Over Acceptance):** Invested 1 hour to understand threading patterns prevented months of compounding problems - **Principle 5 (Continuous Calibration):** Each modification without understanding made the next one harder—learning loop prevented the debt spiral - **Practice: Learning Loop (Chapter 11):** Tracking outcomes (3 incidents in 6 months) revealed systematic understanding gap

Key Lesson: Understanding debt compounds exponentially. The first modification you make without understanding sets a precedent. Each subsequent modification builds on misunderstanding, creating increasingly fragile code. One hour invested early prevents days lost later.

3.6 Common Patterns Across Examples

3.6.1 What VID Caught

1. **Edge cases** (Example 1: Jan 31st subscription)
2. **Missing context** (Example 2: Why try-except existed)
3. **Risk underestimation** (Example 3: "Simple" delete)
4. **Wrong requirements** (Example 4: What "optimize" meant)
5. **Compounding debt** (Example 5: Understanding gap)

3.6.2 How Much Time VID Takes

- Example 1: 15 minutes → Saved hours + customer issues
- Example 2: 2 minutes (reading comment) → Prevented production incident
- Example 3: Proper verification (2 hours) → Avoided compliance violation
- Example 4: Intent spec (2 minutes) → Got right solution first try
- Example 5: Understanding (1 hour) → Avoided 3-day crisis

3.6.3 The VID Investment ROI

Time invested: Minutes to hours per task **Time saved:** Hours to days of debugging + production incidents **ROI:** 10x-100x return on verification time

---

3.7 Using These Examples

3.7.1 For Training

Walk through each example with your team: 1. Show the generated code 2. Ask: "What could go wrong?" 3. Reveal the problem 4. Demonstrate how VID would catch it

3.7.2 For Onboarding

New team members should: - Read all five examples - Identify which VID principle each demonstrates - Practice applying that principle to their own work

3.7.3 For Calibration

When team members struggle with verification: - Point to relevant example - Show concrete time savings - Build verification habits through repetition

---

3.8 Attribution

Real bugs. Real prevention. Real ROI.

---

4 Appendix A: Quick Reference

4.1 The Five Principles

Principle	Core Question
Intent Before Generation	What do I need and how will I verify it?
Graduated Trust	What's the risk and how much verification does that warrant?
Understanding Over Acceptance	Do I understand this well enough for its importance?
Provenance Awareness	Where did this come from and what does that imply?
Continuous Calibration	Are my verification practices working based on outcomes?

4.2 Verification Depth Quick Guide

Risk Level	Characteristics	Verification
Low	Trivial impact, easily reversible, internal	Basic: Read, check intent, run automated tests
Medium	Moderate impact, standard production code	Standard: + trace logic, test edges, check integration
High	Significant impact, important systems	Deep: + adversarial testing, security review, peer review
Critical	Severe impact, security/financial/safety	Intensive: + formal verification, multiple reviewers, documentation

4.3 Intent Specification Template

```
What: [Functional requirement - what should this code do?]
Inputs: [Valid inputs and their types/constraints]
Outputs: [Expected outputs]
Edges: [Edge cases and how they should be handled]
Verify: [How I will verify correctness]
Risk: [Low/Medium/High/Critical and why]
```

---

5 Appendix B: Discussion Questions

For teams adopting VID, discuss:

1. What's our current approach to AI-generated code? What's working and what isn't?
2. How do we currently calibrate verification depth? Is it working?
3. What problems have escaped to production that verification should have caught?
4. What's our risk tolerance? How do we currently assess risk?
5. How do we build verification skills in junior engineers?
6. What would we need to change about our process to adopt VID?
7. How would we know if VID is working for us?

---

6 Appendix C: Glossary

Graduated Trust: The practice of calibrating verification depth to risk level.

Intent Specification: Articulation of what code should do and how correctness will be verified, created before code generation.

Provenance: The origin and history of code, including whether it was human-written or AI-generated and how it was verified.

Trust Level: A classification (High/Moderate/Guarded/Minimal) indicating how much verification is needed for code.

Verification Ritual: A consistent set of checks performed after code generation, scaled by risk level.

VID: Verified Intent Development — a methodology for AI-augmented software development focused on verification rather than generation.

End of Document

---

7 Appendix D: Risk Scoring Rubric

7.1 Purpose

This rubric provides a systematic method for assessing code risk and determining appropriate verification depth. It transforms the conceptual framework from Chapter 5 (Graduated Trust) into an actionable decision tool.

Use this when: - Deciding verification depth for AI-generated code - Calibrating team practices - Discussing edge cases in code review - Training new team members on risk assessment

---

7.2 The Four Risk Dimensions

7.2.1 1. Impact (I)

Question: What's the worst-case outcome if this code has a bug?

Score	Severity	Description	Examples
1	Trivial	Cosmetic issue, no functional impact	Log message typo, comment formatting, internal debug output
2	Minor	Annoyance, easy workaround exists	UI alignment issue, suboptimal sort order, redundant API call
3	Moderate	User-facing problem, no data loss	Feature doesn't work, incorrect calculation in non-critical context, slow performance

Score	Severity	Description	Examples
4	Serious	Data corruption, security issue, or financial impact	Payment calculation error, data loss, authentication bypass, PII exposure
5	Critical	System-wide failure, safety issue, legal liability	Complete system outage, financial fraud, regulatory violation, safety system failure

7.2.2 2. Reversibility (R)

Question: How easily can we recover if this breaks in production?

Score	Recovery Difficulty	Description	Examples
1	Instant	Fix and redeploy, no persistent effects	Stateless API endpoint, display logic, formatting function
2	Easy	Redeploy + minor cleanup (< 1 hour)	Cache invalidation needed, restart required, temporary config change
3	Moderate	Significant manual recovery (hours)	Database migration rollback, batch job rerun, customer notifications
4	Hard	Complex recovery, possible data loss (days)	Manual data correction, multi-system coordination, partial data unrecoverable
5	Impossible/Severe	Permanent data loss or irreversible actions	Deleted production data, sent incorrect payments, compliance breach recorded

7.2.3 3. Exposure (E)

Question: How many users/systems are affected?

Score	Reach	Description	Examples
1	Developer Only	Only affects development environment	Local dev tool, personal script, debug utility
2	Team Internal	Affects only your immediate team	Team dashboard, internal CI pipeline, developer tooling
3	Company Internal	Affects internal users across organization	Employee portal, internal API, admin tools
4	External Limited	Affects external users in controlled way	Beta features, specific customer segment, partner API
5	Public/Universal	Affects all users or public-facing systems	Public website, main product API, critical service

7.2.4 4. Compliance (C)

Question: What regulatory or legal requirements apply?

Score	Regulatory Burden	Description	Examples
0	None	No special compliance requirements	Internal tooling, general business logic
2	Basic	Industry best practices apply	Standard web security, data retention policies
4	Moderate	Specific regulations apply	GDPR (personal data), accessibility requirements, SOC2 controls
6	Significant	Strict regulatory oversight	HIPAA (healthcare), PCI-DSS (payment cards), SOX (financial reporting)
8	Critical	Severe penalties for violations	FDA (medical devices), financial trading regulations, safety-critical systems

Score	Regulatory Burden	Description	Examples
10	Existential	Violations could end business	Critical infrastructure, life-safety systems, national security

---

7.3 Risk Score Calculation

7.3.1 Formula

```
Risk Score = (I × 3) + (R × 2) + (E × 2) + C
```

Why weighted? - **Impact (3×)**: Most important—severity of what goes wrong - **Reversibility (2×)**: Affects recovery cost and learning opportunity - **Exposure (2×)**: Determines blast radius - **Compliance (1×)**: No multiplier needed (scale is already 0-10 vs. 1-5 for other dimensions)

Range: 0 (minimal risk) to 47 (maximum risk)

7.4 Trust Level Mapping

Risk Score	Trust Level	Verification Depth	Time Investment
0-10	**High Trust**	Basic verification	5-10 minutes
11-20	**Moderate Trust**	Standard verification	15-30 minutes
21-30	**Guarded Trust**	Thorough verification	30-60 minutes
31-47	**Minimal Trust**	Intensive verification	1-3 hours + peer review

---

7.5 Verification Requirements by Trust Level

7.5.1 High Trust (Score 0-10)

- Read and understand the code
- Verify against intent specification
- Run automated tests (linting, type checking)
- Spot-check one edge case
- Brief documentation of what code does

Skip: Deep security analysis, extensive edge case testing, peer review

7.5.2 Moderate Trust (Score 11-20)

Everything in High Trust, plus: - Systematic edge case testing (input partitioning) - Security review for user inputs - Check error handling - Review integration points - Document edge cases and assumptions

Skip: Formal verification, multiple reviewers, extensive security testing

7.5.3 Guarded Trust (Score 21-30)

Everything in Moderate Trust, plus: - Comprehensive security review (STRIDE analysis) - Adversarial testing (try to break it) - Performance/scalability consideration - Maintainability refactoring if needed - Peer review recommended - Document security considerations and trade-offs

Skip: Formal methods, external audit (unless compliance requires)

7.5.4 Minimal Trust (Score 31-47)

Everything in Guarded Trust, plus: - **Mandatory** peer review (minimum 1, prefer 2+ reviewers) - Security specialist review if available - Formal threat modeling - Comprehensive test coverage with mutation testing - Consider formal verification for critical logic - Document thoroughly (architecture decisions, security model, failure modes) - Sign-off from tech lead or senior engineer - Incident response plan consideration

---

7.6 Worked Examples

7.6.1 Example 1: Log Message Formatter

Code: Formats timestamps in application logs

Assessment: - **Impact (I):** 1 (cosmetic, doesn't affect functionality) - **Reversibility (R):** 1 (instant fix if needed) - **Exposure (E):** 2 (internal team uses logs) - **Compliance (C):** 0 (no special requirements)

Calculation: $(1\times3) + (1\times2) + (2\times2) + 0 = \mathbf{9}$

Result: High Trust — 5-10 minute verification - Read code, verify format makes sense - Test with a few timestamps - Check it handles null/undefined - Ship it

---

7.6.2 Example 2: User Profile Update Endpoint

Code: API endpoint that updates user profile information

Assessment: - **Impact (I):** 3 (user-facing, could break profiles) - **Reversibility (R):** 3 (need to identify and fix affected users) - **Exposure (E):** 4 (all external users) - **Compliance (C):** 4 (GDPR applies to personal data)

Calculation: $(3\times3) + (3\times2) + (4\times2) + 4 = \mathbf{27}$

Result: Guarded Trust — 45-60 minute verification - Comprehensive input validation review - Check authorization (can user update this profile?) - SQL injection / XSS vulnerability testing - Verify GDPR consent and audit logging - Test edge cases (empty fields, special characters, unicode) - Get peer review - Document security model

---

7.6.3 Example 3: Payment Processing Function

Code: Calculates and processes customer refunds

Assessment: - **Impact (I):** 5 (financial loss if wrong) - **Reversibility (R):** 4 (money already sent, hard to recover) - **Exposure (E):** 4 (affects all customers who request refunds) - **Compliance (C):** 6 (PCI-DSS, financial regulations)

Calculation: (5×3) + (4×2) + (4×2) + 6 = **37**

Result: Minimal Trust — 2-3 hours + multiple reviewers - **Do NOT accept AI output without deep understanding** - Verify refund calculation logic step by step - Test idempotency (refund not processed twice) - Verify amount limits and fraud checks - Check audit trail and logging - Test with real scenarios from finance team - Security review for manipulation attempts - Peer review from 2+ engineers - Tech lead sign-off required - Document refund policy, edge cases, rollback procedures

---

7.6.4 Example 4: Cache Implementation

Code: LRU cache with TTL support for API responses

Assessment: - **Impact (I):** 3 (incorrect caching could serve stale data) - **Reversibility (R):** 2 (clear cache and redeploy) - **Exposure (E):** 4 (affects all API users) - **Compliance (C):** 2 (standard data handling)

Calculation: (3×3) + (2×2) + (4×2) + 2 = **24**

Result: Guarded Trust — 45-60 minutes - Understand cache eviction algorithm deeply - Test TTL expiration works correctly - Test cache hit/miss scenarios - Check for memory leaks (cache grows unbounded?) - Verify thread-safety if concurrent access - Performance test with realistic load - Document cache size limits and behavior

---

7.7 Special Cases & Escalation

7.7.1 When Scores Disagree

Scenario: Low overall score but ONE dimension is very high

Example: Internal tool (E=2) with potential data loss (R=5)

Rule: Escalate to next trust level when: - Impact, Reversibility, or Exposure 4, regardless of total score - Compliance 6 - Impact = 5 (always Minimal Trust)

Rationale: A single high-risk dimension can cause catastrophic failure.

7.7.2 When in Doubt

If you're uncertain about scoring:

1. **Score conservatively** (round up, not down)

2. **Discuss with team** — calibration improves with practice
3. **Document the edge case** for future reference
4. **Default to higher verification** — overconfidence is riskier than over-verification

Red flags that suggest higher score: - "This is just a quick fix" - "We can patch it later if there's a problem" - "No one will probably use this edge case" - "The AI seems confident"

7.8 Team Calibration Exercise

Purpose: Align team understanding of risk levels

Process:

1. **Select 10 code scenarios** from recent work
2. **Individually score** each scenario using this rubric
3. **Compare scores** — Where do you disagree?
4. **Discuss discrepancies:**
   - Why did you score Impact as 3 vs. 4?
   - What assumptions led to different Exposure scores?
5. **Agree on team norms:**
   - Are we consistently conservative or aggressive?
   - Which dimension do we struggle with most?
6. **Repeat quarterly** — Keep calibration fresh

Sample Scenarios for Calibration:

Scenario	I	R	E	C	Score	Trust Level
Sort algorithm for internal admin list	?	?	?	?	?	?
Password reset token generation	?	?	?	?	?	?
CSS styling for marketing page	?	?	?	?	?	?
Database migration to add column	?	?	?	?	?	?
API rate limiting implementation	?	?	?	?	?	?

7.9 Integration with Development Workflow

7.9.1 In Code Review

PR Template Addition:

```
## VID Risk Assessment
- Impact: [1-5]
- Reversibility: [1-5]
- Exposure: [1-5]
- Compliance: [0-10]
- **Total Risk Score:** [calculated]
- **Trust Level:** [High/Moderate/Guarded/Minimal]
```

```
**Verification performed:**
- [ ] [List specific verification steps for this trust level]
```

7.9.2 In Task Planning

Before starting work, assess: 1. What trust level will this require? 2. Do I have time/expertise for that level of verification? 3. Do I need to involve others (peer review, security team)? 4. Should this be broken into smaller, lower-risk pieces?

7.9.3 In Incident Retrospectives

After production incidents: 1. What was the original risk score? 2. Was it accurate? 3. If inaccurate, which dimension was misjudged? 4. How can we improve risk assessment?

7.10 Customization Guidelines

> **Advanced Topic:** The standard rubric (0-10, 11-20, 21-30, 31-47) is the recommended starting point for all teams. Only customize after using the standard rubric for at least 3 months and gathering data about your specific context.

This rubric is a starting point. Adapt it to your context only if data shows systematic miscalibration:

7.10.1 Adjust Weights (Advanced)

If compliance is critical for your domain (healthcare, finance) and you find compliance issues consistently underweighted:

```
Risk Score = (I × 3) + (R × 2) + (E × 2) + (C × 1.5)
```

Caution: Changing weights affects all scores. Recalibrate thresholds accordingly.

7.10.2 Adjust Thresholds (Advanced)

Do not adjust thresholds without data. If your retrospectives show consistent miscalibration (e.g., "High Trust" items frequently cause incidents), consider:

- **More conservative:** Shift thresholds down by 2 points across all levels
- **Less conservative:** Shift thresholds up by 2 points across all levels

Example conservative adjustment:

```
0-8:   High Trust      (instead of 0-10)
9-18:  Moderate Trust  (instead of 11-20)
19-28: Guarded Trust   (instead of 21-30)
29+:   Minimal Trust   (instead of 31-47)
```

Never use different thresholds across team members. Team-wide consistency is more important than individual calibration preferences.

7.10.3 Add Dimensions (Advanced)

Some teams add context-specific dimensions: - **Complexity (Co):** How hard is this code to understand? (1-5) - **Novelty (N):** How unfamiliar is this domain/technology? (1-5)

If adding dimensions, update the formula and recalibrate thresholds completely.

7.10.4 Document Your Modifications

If you customize: - Document why (with retrospective data) - Get team-wide agreement - Share with new hires - Revisit quarterly - Keep the core concept: objective risk → appropriate verification

---

7.11 Common Pitfalls

Scoring to justify desired verification level - "I want to ship quickly, so I'll call this Impact=2" - Fix: Have someone else review your risk assessment

Forgetting about compliance - "It's just a feature" (but touches PII = GDPR) - Fix: Always ask "What data does this touch?"

Averaging away risk - "Most dimensions are low, so overall it's low risk" (but Reversibility=5) - Fix: Use the escalation rules for high single dimensions

Analysis paralysis - Spending 20 minutes scoring a 5-line function - Fix: For obviously trivial code, intuitive assessment is fine

Never re-scoring - Initial assessment was wrong, but never revisited - Fix: Re-assess during retrospectives

---

7.12 Quick Reference Card

Print this for your desk:

```
         VID RISK SCORING QUICK REFERENCE

 IMPACT (I): 1=trivial → 5=critical
 REVERSIBILITY (R): 1=instant → 5=impossible
 EXPOSURE (E): 1=dev only → 5=public
 COMPLIANCE (C): 0=none → 10=existential

 SCORE = (I×3) + (R×2) + (E×2) + C

  0-10: High Trust      (5-10 min)
 11-20: Moderate Trust  (15-30 min)
 21-30: Guarded Trust   (30-60 min)
 31-47: Minimal Trust   (1-3 hr + review)

 ESCALATE if I/R/E  4 OR Compliance  6
```

```
WHEN UNSURE: Round up, ask team, verify more
```

7.13 Conclusion

This rubric transforms "graduated trust" from a concept into a practice. Use it to: - Make consistent risk decisions - Communicate verification rationale - Calibrate team expectations - Improve over time through retrospectives

Remember: The rubric is a tool for judgment, not a replacement for it. When the score doesn't match your intuition, investigate why. That's where learning happens.

Return to Appendix A: Quick Reference | Table of Contents

---

8 Appendix E: Decision Trees

8.1 Purpose

These decision trees guide you through common VID scenarios. Use them when you're uncertain about the right approach.

8.1.1 Quick Index

1. **Should I Use AI for This Task?** — Decide when AI generation is appropriate.
2. **What Trust Level Should I Apply?** — Map risk scores to verification depth.
3. **How Do I Verify AI-Generated Tests?** — Ritual for accepting automated tests.
4. **Should I Accept This AI-Generated Code?** — Go/no-go check before merging.
5. **What Do I Do When Production Breaks?** — Recovery playbook when verification misses something.

8.2 Decision Tree 1: Should I Use AI for This Task?

```
START: Need to write code

 → Is this security-critical code?
   (authentication, authorization, crypto, payment, PII handling)

    → YES → Use AI for reference only
            Write code yourself
            Apply Minimal Trust verification
            Peer review mandatory

    → NO → Continue
```

```
        → Do I have clear intent specification?

            → YES → Safe to use AI
                    Generate code
                    Apply appropriate verification

            → NO → Stop!
                   Write intent specification first
                   Then use AI
```

Key insight: AI is a tool, not a shortcut. Intent first, generation second.

8.3 Decision Tree 2: What Trust Level Should I Apply?

```
START: Code generated (by AI or human)

 → Calculate Risk Score:
   Risk = (Impact × 3) + (Reversibility × 2) + (Exposure × 2) + Compliance

    → Score 0-10
       → HIGH TRUST
          Verification: 5-10 minutes
          - Read and understand
          - Verify against intent
          - Run automated tests
          - Spot-check one edge case

    → Score 11-20
       → MODERATE TRUST
          Verification: 15-30 minutes
          - All High Trust checks
          - Systematic edge case testing
          - Security review for inputs
          - Check error handling
          - Review integration points

    → Score 21-30
       → GUARDED TRUST
          Verification: 30-60 minutes
          - All Moderate Trust checks
          - Comprehensive security review
          - Adversarial testing
          - Performance consideration
          - Peer review recommended

    → Score 31-47
        → MINIMAL TRUST
```

```
          Verification: 1-3 hours + peer review
          - All Guarded Trust checks
          - Mandatory peer review (2+ people)
          - Security specialist if available
          - Formal threat modeling
          - Comprehensive test coverage
          - Tech lead sign-off


→ ESCALATION CHECK:

   → Is ANY dimension  4?
      OR Is Compliance  6?

      → YES → Escalate to next trust level
              Even if total score is low

      → NO → Use calculated trust level
```

Key insight: A single high-risk dimension can be catastrophic. Escalate accordingly.

8.4 Decision Tree 3: How Do I Verify AI-Generated Tests?

```
START: AI generated tests

→ Step 1: Check Test Provenance

   → How were tests generated?

      → From specification
         → Moderate confidence
            Apply standard verification

      → From code
         → Low confidence
            Tests may duplicate code bugs
            Apply intensive verification

      → From examples
          → Medium confidence
             May miss edge cases
             Apply thorough verification

→ Step 2: Review Test Quality

   → Read test names and assertions
   → Do they test behavior or implementation?
```

```
        → Implementation → RED FLAG
           Tests are brittle
           Rewrite to test behavior

        → Behavior → GOOD
            Continue verification

    → Are assertions correct?

        → Uncertain → Manually verify expected values
        → Confident → Continue

→ Step 3: Check Coverage

    → What's NOT tested?
      - Edge cases (empty, null, max, min)
      - Error conditions
      - Security scenarios
      - Integration points

    → Add missing tests

→ Step 4: Meta-Verification

    → For Important+ code:
      Run mutation testing

        → Mutation score < 80%
            → Tests are weak
               Add stronger tests

        → Mutation score   80%
            → Tests are adequate
               Proceed

    → For Critical code:
       Add property-based tests
       Independent test review
       Mutation score   90%

→ ACCEPT or REJECT

    → ACCEPT if:
      - Tests are behavioral
      - Assertions are correct
      - Coverage is appropriate
      - Mutation score is adequate
```

```
→ REJECT if:
   - Tests test implementation
   - Assertions are wrong/unknown
   - Major gaps in coverage
   - Low mutation score
```

Key insight: Tests need verification just like production code.

8.5 Decision Tree 4: Should I Accept This AI-Generated Code?

```
START: AI generated code

→ Question 1: Do I understand what this code does?

   → NO → STOP

          → Can I learn enough to understand?

             → YES → Spend time learning
                     Then re-evaluate

             → NO → Options:
                     A) Simplify the code
                     B) Get help from expert
                     C) Reject and write manually

          → DON'T ACCEPT CODE YOU DON'T UNDERSTAND

   → YES → Continue to Question 2

→ Question 2: Does it match my intent specification?

   → NO → Reject or modify
          AI solved a different problem

   → YES → Continue to Question 3

→ Question 3: Did verification pass?

   → NO → What failed?

           → Tests failed → Fix or regenerate
           → Edge cases wrong → Fix or regenerate
           → Security issue → Fix or regenerate
           → Can't be fixed → Reject, write manually
```

```
    → YES → Continue to Question 4

→ Question 4: Is verification depth appropriate for risk?

    → NO → Do I need deeper verification?

             → YES → Perform additional verification
                     Then re-evaluate

             → Verification was too shallow
                Return to Decision Tree 2

    → YES → Continue to Question 5

→ Question 5: Can I maintain this code in 6 months?

    → NO → Issues:
           - Poor naming?
           - Too complex?
           - Unclear logic?

            → Refactor for maintainability
               Then re-evaluate

    → YES → Continue to Final Check

→ FINAL CHECK: All conditions met?

    → I understand the code
    → It matches my intent
    → It passed appropriate verification
    → Risk-appropriate verification depth
    → I can maintain it

    → ALL YES → ACCEPT
                 Document provenance
                 Commit with verification notes

    → ANY NO → REJECT or FIX
                Don't compromise on quality
```

Key insight: Acceptance is not a yes/no decision. It's a series of quality gates.

8.6 Decision Tree 5: What Do I Do When Production Breaks?

```
START: Production incident
```

```
→ Question 1: Is this code AI-generated?

   → Don't know → PROBLEM
                  Provenance tracking failure
                  (Fix this for future)

   → NO (human-written) → Standard incident response

   → YES (AI-generated) → Continue to Question 2

→ Question 2: What verification was performed?

   → Check commit message / PR notes
     What trust level?
     What checks were done?

   → Unknown → PROBLEM
               Verification documentation failure
               (Fix this for future)

   → Known → Continue to Question 3

→ Question 3: Should verification have caught this?

   → YES → Verification failure

           → Root cause analysis:
             - Was risk score too low?
             - Was verification incomplete?
             - Was verification technique wrong?

              → Update verification practices
                Prevent similar failures

   → NO → Acceptable miss
              (edge case beyond reasonable verification)

               → Add regression test
                 Update risk calibration if needed

→ Question 4: Can the original developer fix it?

   → YES → Have them fix it
           (they understand the code)

   → NO → Understanding gap

              → Code was not understood when accepted
```

```
                    → PROBLEM: Violated Understanding principle
                       (Fix this for future)

                → Developer is unavailable
                    → Hope someone can figure it out
                       (This is why understanding matters)

→ RESOLUTION PATH:

    → Immediate: Fix the bug

    → Short-term: Add regression test
                  Update documentation

    → Long-term: Calibrate verification practices
                 If verification should have caught it:
                 - Update risk scoring
                 - Update verification rituals
                 - Share learning with team
```

Key insight: Production incidents are learning opportunities. Update practices accordingly.

8.7 Decision Tree 6: How Do I Onboard a New Developer to VID?

```
START: New developer joins team

→ Week 1: Foundation

    → Assign reading:
      - VID One-Page Overview
      - Chapter 1: The Inversion
      - Chapter 4: Intent Before Generation

    → 1-hour 1-on-1 discussion:
      - Why does VID matter?
      - How does it work?
      - What are our team norms?

    → Shadow experienced developer:
       Watch full VID cycle:
       - Intent specification
       - Generation
       - Verification
       - Documentation

→ Week 2: Supervised Practice
```

```
    → Pair program with mentor
       New dev drives, mentor guides
       Focus on intent specification

    → Low-risk task
       New dev applies VID
       Mentor reviews every step

    → Debrief:
       - What went well?
       - What was confusing?
       - Clarify principles

→ Week 3: Guided Independence

    → Medium-risk task
       New dev works independently
       Mentor reviews verification notes

    → Join team calibration session
       Participate in risk scoring
       Learn team norms

    → Check understanding:
       Can they explain:
       - The 5 principles?
       - How to score risk?
       - When to escalate verification?

→ Week 4: Independent with Oversight

    → Full VID cycle independently
       Mentor spot-checks

    → Can take on any risk level
       With appropriate oversight

    → Ready to mentor others
       After 3 months of practice
```

Key insight: VID is learned by doing, not just reading. Pair learning accelerates adoption.

8.8 Decision Tree 7: When Should I Update Our VID Practices?

```
START: Considering practice changes

→ Trigger: Why are we considering changes?
```

```
→ Problems escaping to production

    → Root cause analysis:

        → Risk scores too low?
            → Recalibrate risk scoring
               Update team norms

        → Verification incomplete?
            → Update verification rituals
               Add missing checks

        → Verification technique ineffective?
             → Adopt better techniques
                Train team

→ Verification too time-consuming

    → Analysis:

        → Verifying low-risk code too deeply?
            → Recalibrate downward
               Trust level adjustments

        → Inefficient verification workflow?
            → Streamline process
               Remove redundant steps

        → Team inexperienced?
             → More training
                Practice builds speed

→ New technology or AI capability

    → Evaluate impact:

        → AI generates better code?
            → Don't reduce verification yet
               Wait for data
               Recalibrate after 3 months

        → AI can assist verification?
            → Experiment carefully
               Human still arbitrates

        → New code patterns?
            → Assess new risks
```

```
                    Update risk framework

    → Team growth or composition change

        → Ensure consistency:
           - Onboard new members
           - Recalibrate as team
           - Document team norms

→ Decision: Should we change practices?

    → YES → Process:

             → Propose changes
             → Discuss with team
             → Try for 2 weeks
             → Measure impact
             → Keep, modify, or revert

    → NO → Continue current practices
            Next review in 1 month

→ WHEN TO REVIEW:

    → Monthly retrospectives (lightweight)
    → Quarterly deep reviews (comprehensive)
    → After major incidents (root cause)
    → When team composition changes (recalibrate)
```

Key insight: VID is continuous calibration. Adjust based on evidence, not opinion.

8.9 How to Use These Decision Trees

8.9.1 In Daily Work

Before generating code: Use Decision Tree 1 (Should I use AI?)

After generating code: Use Decision Tree 2 (What trust level?) and Tree 4 (Should I accept?)

When verifying tests: Use Decision Tree 3 (How to verify tests?)

8.9.2 In Incident Response

After production issues: Use Decision Tree 5 (What went wrong?)

8.9.3 In Team Development

Onboarding: Use Decision Tree 6 (How to onboard?)

Process improvement: Use Decision Tree 7 (When to update?)

8.9.4 Customization

These decision trees are starting points. Your team should: - Adapt them to your context - Add team-specific branches - Document your variations - Share improvements

---

8.10 Quick Reference

Print these and keep at your desk:

Pre-Generation: Intent first, always **Post-Generation:** Verify appropriately **In Incidents:** Learn and calibrate **In Onboarding:** Pair and practice **In Retrospectives:** Measure and adjust

---

Return to Table of Contents

---

9 Appendix F: VID Checklists

9.1 Purpose

These checklists ensure you don't skip critical verification steps. Print and use daily until they become second nature.

9.2 Checklist 1: Pre-Generation (Intent Specification)

Use before every AI code generation

Functional Requirements: - [ ] I can describe what this code should do in 2-3 sentences - [ ] I have identified the inputs and their types - [ ] I have identified the expected outputs - [ ] I have listed the main use cases (at least 2-3 examples)

Boundary Cases: - [ ] I know what happens with empty/null inputs - [ ] I know what happens at min/max boundaries - [ ] I have identified at least 3 edge cases - [ ] I know how errors should be handled

Success Criteria: - [ ] I have written down how I'll verify this is correct - [ ] I have specific test cases in mind (at least 3) - [ ] I know what properties must always hold - [ ] I can recognize correct vs. incorrect behavior

Risk Assessment: - [ ] I have scored Impact (1-5) - [ ] I have scored Reversibility (1-5) - [ ] I have scored Exposure (1-5) - [ ] I have scored Compliance (0-10) - [ ] I have calculated total risk score - [ ] I know what trust level to apply

Ready to Generate: - [ ] I have documented my intent (at least informally) - [ ] I know what verification I'll perform - [ ] I have allocated appropriate time for verification - [ ] I understand what I'm trying to build

If ANY checkbox is unchecked: STOP. Complete intent specification before generating.

9.3 Checklist 2: Post-Generation Verification (All Trust Levels)

Use immediately after AI generates code

9.3.1 Basic Verification (Required for ALL code)

Understanding: - [ ] I have READ the entire code (not skimmed) - [ ] I can explain what this code does without looking at it - [ ] I understand the approach it takes - [ ] I could explain this to a colleague

Intent Match: - [ ] The code does what I specified - [ ] It handles the inputs I identified - [ ] It produces the outputs I expected - [ ] It solves my actual problem (not a similar one)

Automated Checks: - [ ] Code compiles/parses without errors - [ ] Linting passes - [ ] Type checking passes (if applicable) - [ ] Basic tests pass

Spot Check: - [ ] I tested with at least 1 normal input - [ ] I tested with at least 1 edge case - [ ] Results match my expectations

---

9.4 Checklist 3: Moderate Trust Verification

Use for Risk Score 11-20 (in addition to Basic Verification)

Systematic Testing: - [ ] I have identified input categories (at least 4-5) - [ ] I have tested at least 1 representative from each category - [ ] I have tested boundary values - [ ] I have tested error conditions

Security Review (Basic): - [ ] I have identified all inputs - [ ] User-controlled inputs are validated - [ ] No obvious injection vulnerabilities (SQL, command, XSS) - [ ] Sensitive data is not logged

Error Handling: - [ ] Error cases are handled (not just ignored) - [ ] Error messages don't leak sensitive information - [ ] The code fails safely - [ ] Resources are cleaned up on errors

Integration: - [ ] I understand how this integrates with existing code - [ ] Interfaces/contracts are honored - [ ] Dependencies are appropriate - [ ] Side effects are documented

Documentation: - [ ] Edge cases are documented - [ ] Assumptions are documented - [ ] Purpose is clear from code/comments - [ ] Future maintainers can understand this

---

9.5 Checklist 4: Guarded Trust Verification

Use for Risk Score 21-30 (in addition to Moderate Trust Verification)

Comprehensive Security Review: - [ ] I have performed STRIDE threat analysis - [ ] I have tested adversarial inputs - [ ] Authentication is required (if applicable) - [ ] Authorization is checked (if applicable) - [ ] Sensitive data handling is appropriate - [ ] Rate limiting is considered (if applicable)

Adversarial Testing: - [ ] I tried to break the code - [ ] I tested with malicious inputs - [ ] I tested extreme values - [ ] I tested concurrent access (if applicable) - [ ] I tested failure modes

Performance: - [ ] I considered performance implications - [ ] I tested with realistic data volumes - [ ] No obvious performance bottlenecks - [ ] Resource usage is reasonable

Maintainability: - [ ] Code is readable (not overly clever) - [ ] Complexity is justified - [ ] Naming is clear - [ ] Logic is straightforward - [ ] I would be comfortable modifying this in 6 months

Peer Review: - [ ] Another developer has reviewed this code - [ ] Review feedback has been addressed - [ ] Reviewer understands and approves

---

9.6 Checklist 5: Minimal Trust Verification

Use for Risk Score 31-47 (in addition to Guarded Trust Verification)

Mandatory Reviews: - [ ] At least 2 developers have reviewed this code - [ ] Security specialist has reviewed (if available) - [ ] Tech lead has approved - [ ] All review feedback addressed

Comprehensive Testing: - [ ] Test coverage is comprehensive (>90%) - [ ] Mutation testing performed (score >90%) - [ ] Property-based tests added (where applicable) - [ ] Integration tests pass - [ ] Performance tests pass

Formal Analysis: - [ ] Threat model documented - [ ] Security properties formally stated - [ ] Critical invariants identified and tested - [ ] Failure modes analyzed

Documentation: - [ ] Architecture decisions documented - [ ] Security model documented - [ ] Edge cases and limitations documented - [ ] Incident response considerations documented - [ ] Provenance thoroughly documented

Sign-offs: - [ ] Tech lead sign-off obtained - [ ] Security team sign-off (if applicable) - [ ] Product/business sign-off (for critical business logic)

9.7 Checklist 6: Test Verification

Use when AI generates tests or when reviewing test quality

Test Quality: - [ ] Test names describe behavior (not implementation) - [ ] Assertions check correct values (not just "something returned") - [ ] Tests test behavior (not implementation details) - [ ] Tests are readable and understandable

Coverage: - [ ] Normal/happy paths are tested - [ ] Edge cases are tested (empty, null, boundaries) - [ ] Error conditions are tested - [ ] Integration points are tested (if applicable)

Test Correctness: - [ ] I have verified expected values are actually correct - [ ] Tests would fail if code was wrong - [ ] Tests are not tautological (not testing code against itself) - [ ] No correlated bugs (test and code don't share same logic error)

For Important+ Code: - [ ] Mutation testing performed - [ ] Mutation score >80% (important) or >90% (critical) - [ ] Property-based tests added for key invariants

Provenance: - [ ] Test provenance documented (how tests were created) - [ ] Verification performed is documented - [ ] Tests are marked as AI-generated (if applicable)

9.8 Checklist 7: Code Acceptance

Use before committing code (final gate)

Quality Gates: - [ ] All appropriate verification checklists completed - [ ] Verification depth matches risk level - [ ] All verification passed (or issues resolved) - [ ] No known bugs or issues

Understanding: - [ ] I understand this code at appropriate depth for risk level - [ ] I could debug this if it breaks - [ ] I could modify this if requirements change - [ ] I could explain this to a colleague

Documentation: - [ ] Commit message describes what was done - [ ] Commit message indicates AI assistance (if applicable) - [ ] Commit message notes verification performed - [ ] PR description includes risk score and trust level

Provenance: - [ ] Code provenance is documented - [ ] Verification provenance is documented - [ ] Future maintainers will know where this came from

Team Standards: - [ ] Code follows team conventions - [ ] No shortcuts taken under time pressure - [ ] I'm confident in this code - [ ] I'm willing to be on-call for this code

Final Check: - [ ] I would accept responsibility for this code - [ ] I'm not rushing to ship without proper verification - [ ] The quality is appropriate for the risk level

If ANY critical checkbox is unchecked: DON'T COMMIT. Complete required verification.

9.9 Checklist 8: Incident Response

Use after production incidents involving AI-generated code

Immediate Response: - [ ] Incident is contained - [ ] Impact is assessed - [ ] Customers are notified (if needed) - [ ] Fix is deployed (or incident is mitigated)

Investigation: - [ ] Root cause identified - [ ] Was this code AI-generated? - [ ] What verification was performed? - [ ] What verification should have been performed? - [ ] Should verification have caught this?

Documentation: - [ ] Incident is documented - [ ] Root cause is documented - [ ] Timeline is documented - [ ] Lessons learned are captured

Process Improvement: - [ ] Was risk score accurate? - [ ] Was trust level appropriate? - [ ] Was verification sufficient? - [ ] What verification technique would have caught this? - [ ] Do we need to update our practices?

Follow-up: - [ ] Regression test added - [ ] Similar code patterns identified and reviewed - [ ] Team notified of lessons learned - [ ] Verification practices updated (if needed) - [ ] Risk calibration adjusted (if needed)

9.10 Checklist 9: Team Calibration Session

Use monthly to align team on risk assessment and verification

Preparation: - [ ] Selected 5-10 recent PRs for review - [ ] PRs represent mix of risk levels - [ ] Everyone has reviewed PRs beforehand

During Session: - [ ] Each person independently scores risk for each PR - [ ] Compare scores - [ ] Discuss differences (why did you score higher/lower?) - [ ] Agree on calibrated scores - [ ] Document any changes to team norms

Review Outcomes: - [ ] What verification caught in past month - [ ] What escaped to production in past month - [ ] Were trust levels appropriate? - [ ] Were verification techniques effective?

Action Items: - [ ] Update risk scoring guidelines (if needed) - [ ] Update verification rituals (if needed) - [ ] Schedule training (if knowledge gaps identified) - [ ] Document decisions

---

9.11 Checklist 10: Onboarding New Developer to VID

Use when new developer joins team

Week 1: - [ ] Read VID One-Page Overview - [ ] Read Chapter 1: The Inversion - [ ] Read Chapter 4: Intent Before Generation - [ ] Shadow experienced developer on full VID cycle - [ ] 1-hour 1-on-1 discussion about VID

Week 2: - [ ] Pair program with mentor (new dev drives) - [ ] Complete low-risk task using VID - [ ] Mentor reviews every step - [ ] Debrief session

Week 3: - [ ] Complete medium-risk task independently - [ ] Mentor reviews verification notes - [ ] Attend team calibration session - [ ] Can explain 5 principles and risk scoring

Week 4: - [ ] Complete full VID cycle independently - [ ] Appropriate oversight based on risk level - [ ] Mentor spot-checks work - [ ] Debrief on month 1 experience

Competency Verification: - [ ] Can articulate the 5 VID principles - [ ] Can score risk using the rubric - [ ] Can apply appropriate verification for each trust level - [ ] Knows when to ask for help - [ ] Ready to work independently (with normal oversight)

9.12 Checklist 11: PR Review (for Reviewers)

Use when reviewing pull requests containing AI-generated code

Before Reviewing: - [ ] PR includes risk score - [ ] PR includes trust level - [ ] PR notes verification performed - [ ] PR indicates AI usage (if applicable)

Code Review: - [ ] I understand what the code does - [ ] Code matches PR description - [ ] Changes are appropriate for the task - [ ] No obvious bugs or issues - [ ] Code is maintainable

Verification Review: - [ ] Was appropriate trust level applied? - [ ] Was verification depth sufficient? - [ ] Are verification notes convincing? - [ ] Would I be comfortable with this code in production?

Risk Assessment Review: - [ ] Do I agree with the risk score? - [ ] If no, what dimension did they miss? - [ ] Should trust level be escalated?

Testing Review: - [ ] Tests are present and appropriate - [ ] Tests cover important cases - [ ] Tests are not just testing implementation - [ ] Test quality is appropriate for risk level

Decision: - [ ] Approve (quality is appropriate for risk) - [ ] Request changes (issues need addressing) - [ ] Escalate (higher trust level needed)

9.13 How to Use These Checklists

9.13.1 Daily Use

1. **Print checklists** 1-7 and keep at your desk
2. **Use checklist 1** before every generation
3. **Use checklist 2** after every generation
4. **Use checklists 3-5** based on risk score
5. **Use checklist 7** before every commit

9.13.2 Periodic Use

- **Checklist 8:** After incidents
- **Checklist 9:** Monthly calibration
- **Checklist 10:** When onboarding
- **Checklist 11:** During PR reviews

9.13.3 Adaptation

These checklists are starting points. Your team should: - Adapt them to your context - Add team-specific items - Remove items that don't apply - Update based on lessons learned - Share improvements

9.13.4 Building Habits

First 2 weeks: Use checklists explicitly (check every box)

Weeks 3-4: Use checklists as reference (verify you didn't miss anything)

Month 2+: Internalized habits (checklists as spot-check)

Goal: Build habits so verification is automatic, not checklist-driven.

9.14 Printable Quick Reference Cards

9.14.1 Card 1: Pre-Generation

```
BEFORE AI GENERATION: INTENT FIRST

  What should this do? (2-3 sentences)
```

```
  What are the inputs/outputs?
  What are 3+ edge cases?
  How will I verify it's correct?
  What's the risk score?
  What trust level will I apply?

IF ANY UNCHECKED → STOP
Complete intent specification first
```

9.14.2 Card 2: Post-Generation

```
AFTER AI GENERATION: VERIFY

ALWAYS:
  Read and understand code
  Verify against intent
  Run automated checks
  Test edge cases

THEN by risk score:
 0-10: Basic verification (10 min)
11-20: + Security & integration (30m)
21-30: + Adversarial testing (60m)
31-47: + Peer review & sign-off (2h+)

ANY dimension  4 → Escalate trust
```

9.14.3 Card 3: Before Commit

```
BEFORE COMMIT: FINAL GATE

  Appropriate verification completed
  I understand this code
  I could debug/modify this
  Provenance documented
  Commit message includes:
  - What was done
  - AI assistance noted
  - Verification noted
  - Risk score/trust level

I accept responsibility for this code
```

9.15 Digital Checklist Templates

9.15.1 PR Template Addition

```
## VID Verification

**Risk Assessment:**
- Impact: [1-5]
- Reversibility: [1-5]
- Exposure: [1-5]
- Compliance: [0-10]
- **Total Risk Score:** [calculated]
- **Trust Level:** [High/Moderate/Guarded/Minimal]

**AI Usage:**
- [ ] AI-generated code (specify tool/model)
- [ ] Human-written code
- [ ] Mixed (specify which parts)

**Verification Performed:**
- [ ] Intent specification documented
- [ ] Basic verification completed
- [ ] [Trust-level-specific checks completed]
- [ ] Tests passing
- [ ] Peer review (if required)

**Verification Notes:**
[What did you verify? What did you find? Any concerns?]
```

9.15.2 Commit Message Template

```
[Brief summary of change]

AI-assisted: [Yes/No, tool used]
Risk Score: [0-47] (Trust Level: [High/Moderate/Guarded/Minimal])

Verification:
- [List key verification steps performed]
- [Note any issues found and resolved]

[Detailed description if needed]
```

9.16 Additional Resources

For more checklists, templates, and visual diagrams, see: - **VID Templates** - Ready-to-use templates for intent specs, code reviews, and documentation - **VID Diagrams** - Visual workflow diagrams and decision trees

---

Print these checklists. Use them daily. Build verification habits.

Return to Table of Contents
