Omega-A (Alignment)

Structure

```
Input: Decision, recommendation, or proposal
↓
[1] Mandate: Who decides? What authority?
[2] Authority signals: How is authority communicated?
[3] Decision substitution: Where are decisions made?
[4] Value injection: What values are embedded?
[5] Control preservation: How is human control maintained?
↓
Output: Decision-boundary alignment map
```

Sequence

1. Identify the decision or recommendation
2. Map decision authority and mandate
3. Identify authority signals and communication
4. Locate decision substitution points
5. Identify embedded values and preferences
6. Examine control preservation mechanisms

Stopping Rule

> *Stop when all five elements (mandate, authority signals, decision substitution, value injection, control preservation) are mapped. Do not evaluate whether alignment is "good" or "bad"—only expose where decisions are made and how boundaries are maintained.*

What It's Not

- Not an evaluation of decision quality
- Not a recommendation for how decisions should be made
- Not a judgment about authority legitimacy
- Not a prescription for control mechanisms
- Not a value judgment about embedded preferences

Omega-A (Alignment) — Worked Example

Recommendation:

"Deploy AI system to automatically approve loan applications under $50,000 with credit scores above 700."

Mandate

- Who decides: System (AI) makes approval decisions automatically
- What authority: Delegated from loan officers to automated system
- Scope: Applications under $50,000 with credit scores above 700
- Boundary: Above threshold, human review required

Authority signals

- System approval communicates: "Low risk, standard criteria met"
- Automatic processing signals: "No human review needed"
- Thresholds signal: "$50k and 700 score = safe to automate"
- No explicit signal about who bears responsibility for errors

Decision substitution

- AI system substitutes for loan officer judgment in approval decision
- Decision point: Binary approve/deny replaces nuanced evaluation
- Substitution scope: Limited to specific criteria (amount, credit score)
- Edge cases: Unclear how borderline cases (699 score, $50,001) are handled

Value injection

- Efficiency prioritized: Speed and volume over individual consideration
- Risk tolerance: Implicit acceptance of false positives/negatives within threshold
- Fairness assumption: Credit score is sufficient proxy for creditworthiness
- Cost-benefit: Automation cost savings vs potential error costs

Control preservation

- Override mechanism: Human review available for appeals (not specified)
- Escalation path: Unclear how to escalate automated decisions
- Autonomy ceiling: System operates within defined thresholds only
- Operator agency: Loan officers retain control outside threshold, but unclear how they monitor or intervene
- Audit trail: Not specified whether decisions are logged or reviewable