AI Safety Quick Reference Checklist

Print this. Pin it up. Use it every month. Your AI tools are only as safe as the habits around them.

1 Data Handling

DO	DON'T
✓ Strip PII before sending data to AI tools	✗ Paste SSNs, medical, or financial data into AI chat
✓ Use vendors with Zero Data Retention (ZDR)	✗ Assume 'enterprise plan' means data is private
✓ Log what data each AI tool can access	✗ Let field workers upload faces without a data policy
✓ Run sensitive analysis on private instances	✗ Use free-tier AI tools for anything with PII
✓ Read the actual Terms of Service	✗ Trust 'our AI is secure' without explanation

2 Human Oversight

- [ ] High-stakes AI decisions require human sign-off before any action is taken
- [ ] Review interfaces require active reasoning (not just Approve/Reject buttons)
- [ ] We track approval rates per reviewer — 100% approval is flagged as a concern
- [ ] Any authorized team member can override an AI decision immediately
- [ ] Override reasons are logged and fed back into model improvement
- [ ] We have a defined escalation path when AI confidence is low or uncertain

3 Review Tier Guide

LOW STAKES	Auto-approve with logging (e.g., basic photo framing check)
MEDIUM	Spot-check random sample (e.g., route assignments, data categorization)
HIGH STAKES	Mandatory human review (e.g., performance flags, compliance alerts)

Rule of thumb: If AI output could cost a job, safety, or a client relationship — a human reviews it first. Every time.

4 Red Flags — Stop & Investigate

- Your team stopped questioning AI output (everyone just clicks Approve)
- No one can explain why the AI made a specific decision
- Accuracy hasn't been checked since launch
- The AI has no fallback — if it goes down, your operation stops
- Your vendor can't clearly explain where your data goes
- AI scores show patterns that correlate with geography or demographics