# Beyond 'human in the loop'

## A practical framework for designing human-AI oversight that actually works

Tey Bannerman
http://teybannerman.com

Many AI governance failures happen because organisations implement generic "human in the loop" without considering what they're actually trying to achieve, prevent, or control. This practical approach aligns business reality and emerging AI capabilities.

**1** **What are you optimising for?** →

**2** **What's at stake?**

|  | SPEED/VOLUME *Maximum efficiency and scale* | QUALITY/ACCURACY *Ensuring precision & correctness* | COMPLIANCE *Meeting regulatory requirements* | INNOVATION *Fostering breakthrough thinking* |
|---|---|---|---|---|
| **Irreversible consequences** | **Circuit-breaker protocols** — Human authority to halt system immediately. Multiple verification steps. Clear escalation chains. | **Expert override systems** — System explains reasoning, human validates. Domain experts have full authority to counteract AI outputs. | **Mandatory human approval** — AI recommends, human decides. No automation without explicit human authorisation. | **Collaborative validation** — AI and human analyse independently, flag discrepancies, resolve together. |
| **High-impact failures** | **Monitored automation** — AI acts autonomously with real-time human monitoring. Alerts on anomalies. | **Regular expert review** — Expert periodically audits AI decisions. Identifies patterns, adjusts parameters. | **Approval workflows** — AI processes routine cases, escalates edge cases to human reviewers. | **Collaborative ideation** — AI generates options, human provides creative direction and strategic context. |
| **Recoverable setbacks** | **Batch processing** — AI processes large volumes, human reviews samples and overall patterns. | **Spot checking** — Random sampling of AI decisions for quality assurance and calibration. | **Rule-based guardrails** — AI operates within predefined parameters. Human oversight for exceptions. | **Creative collaboration** — AI generates initial concepts, human refines and adds strategic insight. |
| **Low-stakes outcomes** | **Full automation** — AI operates independently with minimal human intervention. Performance monitoring only. | **Feedback learning** — AI learns from user interactions and feedback to improve over time. | **Template automation** — System handles routine, standardised processes with minimal variation. | **Generative assistance** — AI provides creative suggestions and variations for human consideration. |

### Legend

- ■ **Active control:** Human authority at all times
- ■ **Human augmentation:** Human-led with AI support
- ■ **Guided automation:** AI with human oversight
- ■ **AI autonomy:** Minimal human involvement

## REAL-WORLD EXAMPLES

**E-commerce: email marketing**

Goal: Send personalised emails to millions (optimising for speed/volume *and* quality).

Risks: Poor copy, spam risk (recoverable setbacks).

→ **Batch processing + Spot checking:** AI generates emails, marketing team reviews random samples pre-sending, and monitors overall engagement metrics, intervening when unusual patterns seen.

**Recruitment: applicant screening**

Goal: Support recruitment/HR team by processing hundreds of applications efficiently (volume, quality, compliance)

Risks: Systemic bias, missing great candidates or hiring poor fits (high-impact failures).

→ **Monitored automation + regular expert review + approval workflows**
AI screens applications for requirements and fit indicators, flagging top candidates and clear rejections. Recruiters review candidates and can override any AI decision. Hiring managers get AI analysis alongside resumes for final interviews.