Human in the loop documentation

Human review is a foundational principle for the responsible use of AI at the state of Oregon. Ensuring a human is in the loop is about accountability prior to distribution or decision. Human review is satisfied when a responsible individual confirms the output is factually correct, appropriate for use, and something they are willing to stand behind. This includes both sufficient review and approval of release of the content.

Documentation of human review for AI use cases is fundamental for the responsible use of AI. This includes the following:

- Clear requirements for human review of AI outputs from the system proposed for use. This should include what must be reviewed, who is approved to conduct the review, and who is responsible for final approval.
- If deemed appropriate, a review log for tracking reviews, including the individual reviewer and the content reviewed.
- A further process for auditing or checking the review log, if deemed necessary for the given use case.

For many cases, human in the loop control is satisfied through existing operational approvals, such as sending or signing off a document, giving an approval, or closing a ticket, rather than creating new logging systems or parallel audit trails.

Example: Low-Risk Use Case (internal, non-rights impacting)

What Gets Reviewed:
Most critical is to review key data points and conclusions:

- Names, dates, amounts, thresholds
- Policy references or citations
- Conclusions or recommendations (if present)

How to Review:
Reviewer must:

- Scan the output and **verify critical facts** that would cause harm if wrong.

Supporting human review:
An optimally designed AI system will support effective human review by providing clear citations, traceable sources, and sufficient explanation of outputs to allow users to understand how conclusions or recommendations were generated.

Batch processing:
For repetitive processes, such as AI-assisted optical character recognition, sampling may constitute sufficient human review.

The check:
Before release, the reviewer should be able to honestly say:
"I am responsible for this. If this is wrong, it's on me — and I'm okay with that."

For use cases flagged as higher-risk:
More detailed identification of risks and human review requirements must be done to support any flagged use of AI, including concrete mitigations and safeguards.

NIST AI Risk Management Framework Mapping

- GOVERN: clear roles/responsibilities and oversight expectations (GOVERN 2.1), leadership accountability for risk decisions (GOVERN 2.3), and defining human-AI oversight configurations (GOVERN 3.2), with planned monitoring/review cadence (GOVERN 1.5).
- MANAGE: mechanisms to supersede/disengage/deactivate when outcomes don't match intended use (MANAGE 2.4) and post-deployment monitoring that includes appeal and override (MANAGE 4.1).