

## **A Human in the Loop May Sound Reassuring. But Humanity Needs More.**

When we hear warnings about AI becoming difficult to control, "a human in the loop" sounds comforting. Someone is watching. Someone can intervene. Someone remains responsible. But being present does not mean that person has the understanding, authority, or time to change what happens next.



Jensen Huang recently described AI safety as an engineering problem [\[link to source\]](#). His framing carries a constructive message: we can build safeguards, test systems, restrict what they can access, and create ways to stop their actions. These are practical answers to the fear of losing control.

Consider a car. We put a human behind the wheel, but we also engineer brakes, seat belts, and collision protection. We recognize the limits of human attention and judgment, so we build protections around them. AI needs the same discipline. Human oversight must be backed by controls that actually work.

But brakes and seat belts don't decide where the car is going. A perfectly safe vehicle can still be headed somewhere no one should go.

This is the gap in how we usually talk about responsible AI. Trust, Safety, Transparency, and Accountability are essential. We need evidence that systems are dependable, that risks are managed, that people can understand decisions affecting their lives, and that responsibility is clear. Those pillars govern the journey. They don't judge the destination.

That takes two more questions. Purpose asks what we are trying to accomplish, whose needs we are serving, and whether the goal deserves our support. Impact asks what actually happens: who benefits, who bears the costs, and whether the results fulfill the promise.

Imagine a content recommendation system that is carefully monitored, explains why it surfaces each post, and has a named executive accountable for its performance, while being optimized to keep people scrolling as long as possible. It passes every oversight test. A human approves each step. No one questions the destination.

That is why the most important human judgment happens before and after the loop: when the purpose is defined, and when the consequences are measured. That work belongs to boards setting direction, leaders deciding what to build, and the people affected having a way to show what actually happened to them. Engineering can build effective safeguards. Human values must decide what those safeguards protect and what technology is meant to achieve.

Humaital built the Human Alignment Index™ to make that judgment examinable. It assesses what organizations can actually show across six pillars: Trust, Purpose, Safety, Transparency, Accountability, and Impact. Together, they connect responsible governance with the goals AI serves and the consequences it creates.

We may never anticipate every challenge AI will bring. But we can demand clearer purpose, stronger controls, and evidence of real impact. That gives us a practical way to turn concern into action and reassurance into something we can examine.

A human in the loop can approve the next step. Humanity's purpose, and the impact on people, must determine whether that step is worth taking.