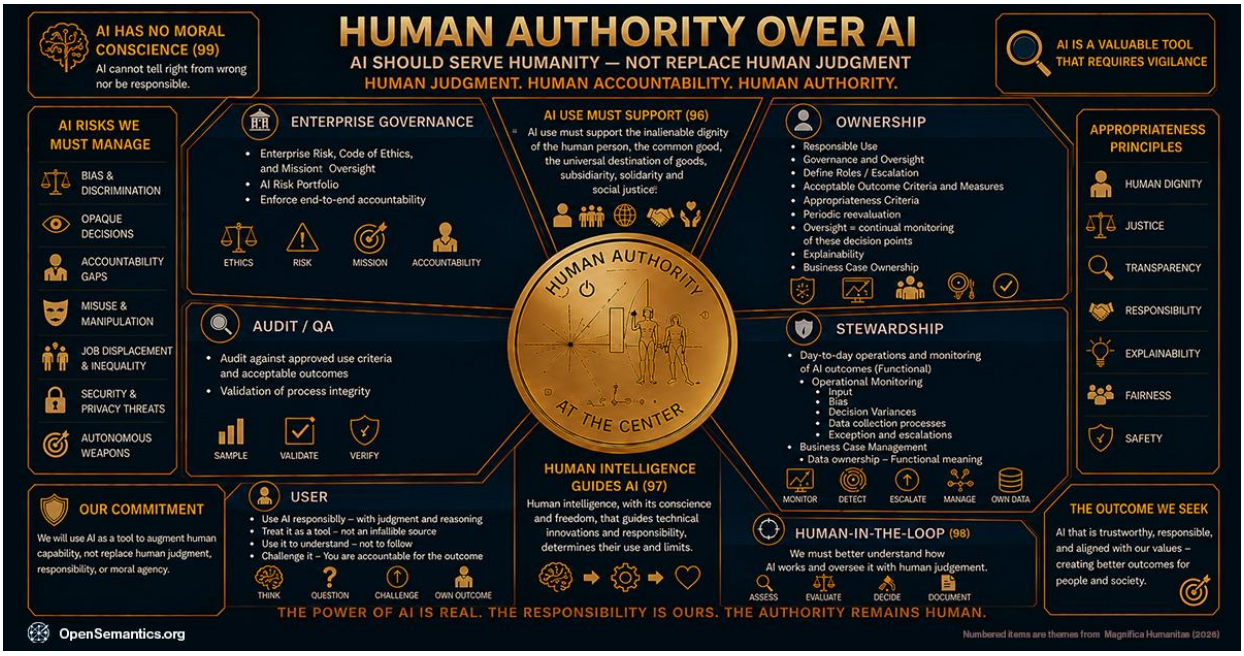
**Human Authority For AI — But What Does That Actually Mean?**

*Everyone agrees AI needs Human Authority – even the Vatican. The hard question is what that means in implementation.*




***Disclaimer:*** *I am neither a lawyer nor an expert on liability. This reflects my professional understanding as a former officer and CISO in a healthcare company. Seek advice from appropriate sources for any decisions.*

Human Authority Over AI is a hot topic lately as seen in a number of articles and threads. Those tend to be “industry” voices of those who provide and use AI.  But we’re also seeing it in the booing of speakers at graduations and demonstrations against data centers.  Even the Pope has weighed in.

AI is touching a raw nerve and we shouldn’t be deaf to that. Regardless of your religious background or stance, the Pope’s writing peels away the issues in a way that deserves to be considered.

None of what follows depends on sharing his faith. Strip away the theology and the same concerns remain — harm, exclusion, accountability, dignity. These are human questions, not religious ones; the encyclical provides a structure for analysis and clarity on points.

Pope Leo XIV opens his first encyclical, *Magnifica Humanitas*, with a choice: humanity can build "a new [AI] Tower of Babel," or it can build a city where people dwell together in something other than confusion. I'll admit the image stopped me, because I'd reached for the same one in another piece on agentic AI semantics — a tower of disconnected agents, each fluent, none speaking a shared language.

The encyclical's central claim about AI is one I don't think anyone seriously wants to argue: people — not machines — must remain responsible for the decisions that most affect people. Hard to dispute. And I don't.

But here's where I keep getting stuck. What does "human authority" actually mean *in implementation*?  The diagram accompanying this piece is my attempt at framing an answer.

**The encyclical starts with what AI is not**

In paragraph 99, the Pope writes that AI has no moral conscience — it does not judge good from evil, and while it can convincingly *simulate* empathy and understanding, it comprehends nothing of what it produces. Most important for this discussion, he states plainly that these systems do not "bear responsibility for consequences."

That is the hinge. If the AI and agents cannot bear responsibility, where does it lie? It stays with humans. The whole "human authority" question is really a question about *where* it stays and *how* it is discharged.

**Who is accountable, and how**

Most regulations and control frameworks place named accountability at a senior level.

In the US, the CEO is accountable for the company; the CFO and CISO carry accountability for their domains - all with criminal and civil exposure. General employees are not, typically, personally liable for honest mistakes beyond the loss of their job.

But here's the part some people skip. None of those named, accountable people are present for every decision. They do not personally perform the actions or review them individually. No sizable company could operate that way.

The bar isn't omniscience; it's the ability to demonstrate that *reasonable steps* were taken to prevent, detect, and correct errors.

How does a company meet that bar — for AI, agent, or not?

The same way it always has: industry-accepted control frameworks, clear accountabilities, internal governance and oversight with independent audit.

**"With a human, there's always someone to blame"**

This is the objection I hear most: with a human, there is always a named individual to hold accountable, and that isn't true with an agent. It is often followed with we have auditability of human actions.  Left silent is the lack thereof for AI and agents.

There are two answers:

First, the technical one. There are real gaps today in agent attribution and traceability. I don't view those gaps as closing on their own — they must be deliberately architected. But they are architectural problems. Once addressed they yield a defensible record of who initiated an agent flow, what it did, and why.

Second, the structural one. Accountability rolls *up*, and it always has. When a line worker follows a flawed process and harm results, the regulator rarely pursues the line worker; absent malice, an honest mistake creates no personal liability.

The worker's answer is "I was told to use this." The question then becomes: who authorized the flow, what diligence — testing, risk analysis, etc. — preceded the decision to deploy it, and what preventive, detective, and corrective controls were put around it? There's also the question of how it fits the company's culture, code of conduct, and risk-management practices.

An agentic flow has the same authorizer. The named individual, the supervisory chain, and the accountable identity all still exist — they were never the executor in the first place.

What changes is that the executor is no longer a human with independent judgment leaving no one to ask when something goes wrong.

Therefore, the *entire* evidentiary weight shifts onto the authorizer's control environment. That isn't an accountability void. It's a higher bar on the controls and the audit trail.

While authority rolls up and the use of controls is the expected bar for liability, we must be honest – a legal court and the court of public opinion won't treat human error and agent error equally. It is not uncommon for a company to cite human error, referencing disciplinary action, training, and additional controls as the answer with that weighing in their favor. Each of us understands that humans can, and will, make mistakes and that is beyond the control of the company.

With agents, the tenor is much more likely to be "you designed, built, tested, deployed, and operated this AI or agent" and should have ensured this wouldn't occur.

In short: with a human, there is a recognized human factor that absorbs some of the blame. With an agent, there isn't — which means the design, development, and controls have to carry the full weight, alone.

**Human-in-the-loop is not a silver bullet**

One legitimate response to high-stakes uncertainty is to put a human on every action. That turns the system into an AI-*assisted* workflow: the human is the actual decision-maker, and the AI is support. There are many scenarios where the impact of error justifies exactly that.

But review is not a cure-all. When results are usually good, humans grow complacent and start rubber-stamping. When results are poor, humans stop trusting the output and quietly do the work themselves, underutilizing the tool.

**Beyond correctness, the encyclical points us to Humanity**

There's a deeper reason. The encyclical captures in paragraph 99 - AI's output is statistical adaptation, not the product of a life: it has no experiences, forms no relationships, and does not "know from within what love, work, friendship or responsibility mean."

It can produce a competent answer, but it cannot bring the human element — dignity, compassion, moral judgment — to bear on its reasoning, because it has none to draw on.

Where those qualities are essential to the decision, the point of human review isn't just to catch the machine's mistakes. It's to supply what AI cannot.

This also factors not only into assessing an AI or agent result but determining if it is appropriate for AI to have a place in the decision at all.

**Not every decision is a moral one**

It's worth being precise. The encyclical is not spreading a moral claim evenly across every inference an AI makes. The heavy weight attaches to high-consequence domains — autonomous weapons most sharply, but also human dignity, work, and truth. Many agent decision points carry no moral question at all; they are ordinary operational matters governed by ordinary controls.

That distinction matters. It is analogous to applying a company’s risk management framework and risk tolerance to decisions and scaling the control environment appropriately.

**A model to consider**

The encyclical further speaks about “moralizing” AI systems in paragraph 107 noting the dangers of imparting moral characteristics as defined by narrow groups. As we look to employ AI in ways that can affect humanity, we must frame our decisions in that global context rather

than one isolated within our own organizations. The Pope draws on this point when he speaks of decisions that create human impact indirectly via access, oppression, and economics.

US healthcare already has a model with similar framing – the IRB.

The Institutional Review Board (IRB) is a formally designated, independent committee. Its composition and independence are defined by federal regulation (the "Common Rule," 45 CFR 46, and for FDA-regulated work, 21 CFR 56). Its purpose is to review research involving human subjects before it proceeds. It weighs the work in the context of both the public good and individual rights, and it escalates scrutiny by risk.

For the most critical AI use, that concept could be replicated — voluntarily or by regulation. Not just for healthcare but as a means of independent review of AI usage. This could be layered over existing risk management practices as an escalation path and as oversight.

Responsible-AI frameworks from the EU, NIST, and others already speak in terms of impact; those criteria could be sharpened to determine where an IRB-like construct *must* exist.

An IRB-like body, independent and broadly composed, is one concrete answer to *whose morals*: it takes the definition of acceptable use out of the hands of whoever happens to control the model or AI process.

**The pace problem**

The encyclical opens the AI discussion addressing two hard points: AI resists a single clean definition, and its evolution outruns our ability to reflect on it. That tension is real, and the Pope's instinct is to create space to slow down — to protect the ability of communities to participate and ask questions before everything accelerates past them.

I take his point as a call for deliberate participation in addressing these issues, not a brake for its own sake. And on slowing the technology itself: that horse is already out of the barn. It is readily available and will be used for competitive advantage – not just between well-meaning businesses but by those who intend harm or only their own good.

**The user's part — and where it breaks down**

There's a layer below the authorizer that's easy to skip: the person actually using the tool.

A control environment assumes the human in the loop is an *active* participant — questioning, probing, applying judgment — not a passive one clicking accept.

The encyclical frames the stakes higher than quality control: when intelligence becomes self-referential, the Pope warns, it leaves us "more isolated and more vulnerable to being dominated" (113).

Passive consumption isn't just a weak control; it's a slow surrender of the judgment that keeps the tool in service to the person. So part of the answer is cultural. We have to foster an active-user paradigm — teaching people to challenge what AI hands them rather than absorb it.

Two things make that harder than it sounds.

The first is that "challenge the answer" assumes you can. I can't meaningfully challenge a claim about quantum physics – or many other topics. I can ask for an explanation, but many will land beyond me, and I'm left with the real question: where does this sit on the scale from fully fabricated to correct — and, just as important, is it even relevant?

Which points at the subtler trap. AI answers the question you asked. Ask the wrong one and you'll get a confident, well-formed, answer — and you may lack the expertise to notice. Knowing

what to ask is itself expertise, and it's a part the tool doesn’t supply.  Even when AI is right, there is nothing guaranteeing that it provides the best answer.

The second is harder, and uncomfortable.

Capability isn't uniform. It varies by domain, by circumstance, and by person.  The raw fact is that not everyone can equally be the active checker the model relies on. That is a recognition of human reality and its differences – not a judgement.

The encyclical is unsparing here: the exclusion of the vulnerable becomes "cloaked in a veneer of neutrality and objectivity, against which it becomes difficult to raise objections" (103).

The same fluency that lets a wrong answer slip past a limited user is what makes systemic exclusion read as objective output. The machine's apparent neutrality is what disarms the scrutiny. "Just ask better questions" isn’t a complete answer.

The honest fallback is skepticism — imparting, as widely as we can, the instinct to doubt a fluent answer and to recognize when something is above your waterline and requires seeking other counsel. It is a weak control, but it is not nothing, and it is the one that scales to people without domain expertise.

That leaves another challenge.

Where a user lacks the ability or knowledge to assess the results, accountability cannot transfer to them — it must stay upstream, with those who build and deploy these systems.

The encyclical is explicit that the burden cannot be placed on the most vulnerable for this understanding (117). The reasonable conclusion is that accountability must stay upstream, with those who build and deploy these systems.

That hands developers a genuinely hard problem: helping vulnerable users without patronizing and alienating capable ones - inside the same product. Threading that needle will take real creativity. The difficulty is not an excuse for leaving it unaddressed.

**Conclusion**

Where this lands, for me, is simple to state and hard to do.

There are technical aspects, but at its core this is a governance, control, and accountability problem that directly embodies the concepts of acceptable use and outcomes.

The traditional control objectives — prevent, detect, correct, with oversight — almost certainly survive. But those controls were built to answer one question: is the output correct, repeatable, and auditable?

Human Authority Over AI adds a second question they were never designed to ask: is this use *appropriate* —to the people it affects and to the stakes involved?

For many AI and agent uses — maybe most — correctness is enough and the appropriateness question is quickly settled – but still explicitly asked and answered.

The real expansion recognizes that Human Authority Over AI is more than correctness. It is the discipline of keeping AI use aligned with human dignity and judgment — and of being able to prove, when asked, that we made that determination on purpose.