

What the Zero Means

2BP · Instance 03 · Donatella Ballerini & Nuno Valério

There is a person at the end of the pipeline. It is 4:30 in the afternoon, the sixth deviation of the day just closed, and the document on the screen asks one tired professional to verify five different concerns at once before anyone goes home. Every AI workflow either of us has touched this year ends in some version of that seat. This piece is about what the person in it is for — and about the one question on which the two of us, after agreeing on nearly everything else, give opposite answers.

The agreement runs deep. We have both spent this year arguing that human-in-the-loop, as practiced, gives the appearance of control without the substance. Donatella's reading of why is the more forgiving one, and for that reason the more damning — automation bias is a normal adaptation, a healthy brain learning to agree with a system that is right 97% of the time. The 3% the human was placed there to catch is exactly where their attention stopped going months ago. I came at the same problem from the other side — in April I wrote that human-in-the-loop had become our industry's comfort blanket, and in May, at a Pistoia Alliance roundtable, I called it the most-used and least-validated control in our governance toolkit. Donatella has a line that deserves to hang above every approval screen in the industry — "a genuine control point is not a click." If we stopped here, you would be reading two people agreeing in public, and that is not what this practice is for.

We even proposed, separately, the same test. Plant known errors into live work — a made-up regulatory reference, a misclassified finding, a wrong number carried through a table — and count how many the reviewer catches. She calls them challenge exercises. I call them canaries. Same idea, arrived at from two directions. Now run that test forward to a morning neither of us has lived yet. Months of seeded errors. A detection rate near zero. A decision on someone's desk about the human in that loop. What does the zero mean? Here we split.

For Donatella, the zero means the design failed the person. The role was vague, the judgment being asked for was never specified, the calibration was never maintained — so rebuild the involvement until it is genuine, and run the test again. Underneath that response sits her real position. Humans, she argues, carry kinds of knowing that machines do not reach. The explicit kind — rules, guidelines, procedures — machines may already handle better than we do. But the tacit kind, the experienced professional's sense that a monitoring report was not written by someone who actually attended the visit, and the relational kind, knowing what this particular investigator means when he says he will fix something — these cannot be written down, and so they cannot be automated. Her sharpest example — a model can hold every published guideline on informed consent and still cannot recognize a patient who is nodding without understanding. For Donatella, the human stays at the control point, and the work is to redesign the point until those kinds of knowing can actually operate there.

My position is simpler than it sounds, and it is worth saying plainly. If months of honest measurement show that the person in a review loop catches nothing, then keeping them there serves nobody. It wastes their time, it gives the organization false comfort, and it produces a signature that means nothing — which is a risk in itself, because a meaningless signature still carries legal and regulatory weight. So remove the human from that point. Close the loop without them. Then take the same measurement and use it to find the places where a person demonstrably changes the outcome — where they catch what the machine

misses, where context decides, where a judgment call genuinely happens — and concentrate human attention there. Fewer control points, but real ones. I am not trying to remove people from the system. I am trying to stop spending them where they add nothing, so there is more of them available where they add a lot. The zero tells me where a person's judgment is being wasted — and my question to Donatella is direct. In her model, what result would ever justify closing a loop? If every zero leads to redesign and never to removal, then the value of human involvement is being assumed rather than demonstrated. An assumed value that no evidence can touch is how we got the checkbox in the first place.

Her answer goes after my method, and I have not been able to get around it. My test can only contain errors I can write down. A fake reference, a wrong dose — I can plant those. The things she values most, I cannot plant. I cannot write "a report whose author wasn't really there," because the moment I can describe that precisely enough to fake it, I have turned it into a rule — and rules are exactly what machines check better than people. So my measurement sees only one kind of human value, the kind that was never really in dispute. If I allocate people strictly by what my measurement shows, I will quietly define value as what I can measure — and I may pull humans out of seats where they were doing something real that my test could not see. The removal would look rigorous. It would be well documented. And it could still be wrong in a way I would have no way to notice. I take that seriously, and I do not have an answer to it yet.

So the fault line, stated plainly. We both want humans where they add value — on that we are identical. The disagreement is about how you find those places, and what you are allowed to do when the evidence comes back empty. I say placement must be earned by evidence, and a seat that shows nothing gets closed. She says the most important human contributions cannot show up in that kind of evidence, and a rule like mine will eliminate them. Her risk is keeping people in seats on belief — which is the checkbox wearing better clothes. My risk is deleting what I cannot measure — which is blindness wearing the clothes of rigor. Hmmm — the clothes are my image, not hers; she puts it more gently, and that difference in temperature is part of why this pairing works. When the zero arrives, she rebuilds the seat and I close it. One of those is the wrong move, and we do not know which.

What we carry into the discussion together is the question underneath all of it — whether the tacit kind of value can be demonstrated at all without being turned into rules and destroyed in the process. We know more than we can tell, as Polanyi put it, and nobody in his century asked him to prove it to an inspector. Ours asks. Two practitioners proposed the same test this spring, from opposite ends of the problem, intending opposite uses for it — and the test we both want cannot settle the thing we actually disagree about. Somewhere ahead, the zero lands on a real desk, in her organization or in mine, and someone signs. We finished writing this less certain than when we started... and neither of us can tell you, yet, which of us that should worry more.

Instance-page fills (for 2bp.org/instance-03)

The stance it summoned — Remove the human from control points where measurement shows they add nothing, so human judgment can be concentrated where it demonstrably changes outcomes. Against it — the human contributions that matter most cannot show up in that measurement, and an allocation rule built on it will delete them.

The unknown on the table — Can tacit and relational knowing be demonstrated without being turned into explicit rules — and destroyed by the translation?

The residue — The same test, proposed twice from opposite ends, cannot settle the disagreement it exposed. When the zero comes, one of us rebuilds the seat and the other closes it, and nobody yet knows which move is the mistake.