CHAPTER 2

What the Technology Will Not Fix

> *Training data for machine learning without a documented ground truth is a path to failure. What a model can deliver begins where the truth has been written down.*

There is a hope that surfaces in almost every AI project. It is rarely spoken aloud, yet it structures everyone's expectations. The hope goes: if the model is large enough, it will fix the bad input data.

It will not. This concerns, above all, supervised learning — the family of machine learning in which a model trains on pairs of input and documented correct answer, the typical case of Industrial AI: quality control, predictive maintenance, credit scoring. On the evidence of the last ten years, no industrial model of this class has structurally solved the problem on its own, and there is no engineering reason to expect the next generation to do so. What machine learning delivers is an approximation of the function laid down in the training data. If the training data contradict themselves, the model approximates contradictions. If they are incomplete, it approximates the gaps. If they are simply wrong, it approximates a wrong reality — precisely, fast and at industrial scale.

McKinsey, in the 2025 edition of its global survey *The State of AI*, recorded the ratio that sets the scale of the problem: 78 percent of organizations use AI in at least one business function, while a mere 1 percent describe themselves as *AI-mature* — organizations where AI is genuinely built into the operating core and delivers measurable returns.[1] The distance between *using* and *using maturely* is a question of the human layer, of ground truth and of the organizational architecture around the model; the technology is the smaller part of it.

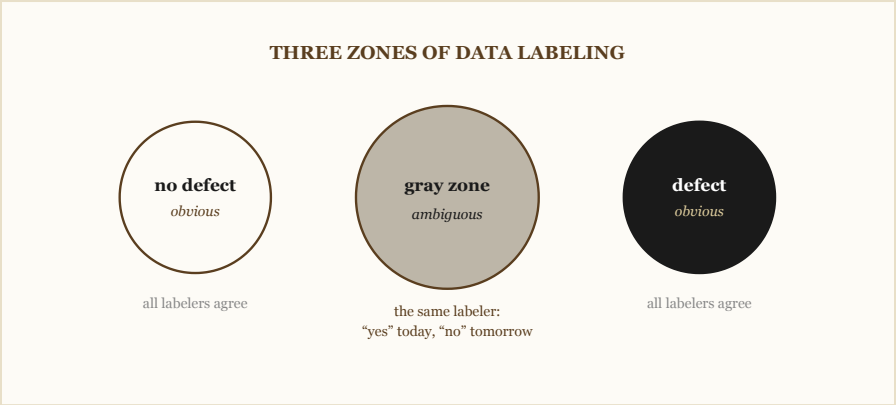
Kate Crawford, in *Atlas of AI*, shows that the data question is a classification question: what a dataset counts as *defect* or as *normal* is a codified organizational choice wearing the costume of a neutral technical fact.[2] Data quality therefore remains a permanent function with its own rhythm of checking — reaching far beyond a one-off audit before launch — with a written procedure and a named owner. Without that, the conversation about ground truth quickly shrinks to the technical details of model tuning, the selection of parameters that shape the speed and character of training: questions that leave the actual problem, one level below, untouched.

What ground truth really is

Ground truth — literally, the truth on the ground — is the documented correct answer against which a model can be measured. On a production line it means: a defect has been identified, categorized and labeled in such a way that a second person with the same instrument arrives at the same label.

Stated like that, the requirement sounds simple. It is exactly where every industrial project meets its tightest bottleneck: the reproducibility of labeling.

Where does ground truth come from in the first place? From the labeling procedure: someone looks at a concrete image — a photograph of a surface — and records whether it shows a defect, and which kind. The obvious cases are easy — a clear crack, a clearly clean surface; there the labeling operators give identical answers. The trouble begins on the unobvious ones: a defect at the edge of tolerance, an ambiguous structure, a borderline case.



Three zones. The white and the black cases are labeled stably. The gray zone is the source of every later problem: the model learns a floating truth and replays that unstable perception on the line.

In most industrial projects, defects at the edge of this gray zone are labeled without reproducibility. What gets classified as a crack today counts as a coating defect tomorrow; what one shift labels as scrap, the next shift waves through. This scatter in labeling is the norm rather than a rare exception, and it arises systematically: defects near the tolerance limit look visually alike, and no human holds an identical perception around the clock.

The consequence is a cascade. At different sites, at different customers, on other equipment the system behaves differently; complaints climb; the support team drowns in endless case-by-case forensics of *it worked for us, it failed for them*; the cost of support grows faster than the savings from automation. At some point the project effectively costs more than the manual inspection it replaced.

Why a bigger model does not help here

On such data a model can be trained — of any size and any architecture. It will pass through training and, at the end, produce a formal accuracy figure. That figure describes how well the model reproduced the contradictory labels of the human operators on a test sample. It does not describe how well the model actually recognizes defects in production.

In the edge case, the model remains exactly as unreliable as the human who taught it. An edge case is the defect a labeling operator records today and misses tomorrow. Right there the model was supposed to help; right there it fails.

> *In most industrial contexts, more training data improves little when the data contradict themselves. Bigger models improve little when the signal itself is blurred. More sophisticated architectures improve little when the thing to be approximated is not stably defined.*

A case from production: what the gray zone looks like in a real project

One such case: the development of optical instruments for high-precision industrial measurement. The first models, trained on the existing dataset with human labels, reached about 95 percent accuracy on the test sample. On the real production stream, it was exactly the gray-zone edge cases that generated errors, halted calibration and came back as customer complaints. After the task itself was reformulated, the same model architecture on the same real images reached 98.5 percent — and, more important, delivered a reproducible result across training runs instead of floating from one run to the next.

The solution lay in reformulating the task, well away from any enlargement of the model. Instead of forcing the model to reproduce human inconsistency, the physical properties of the defect were documented: its shape, its depth, the way it reflects light. Such properties can be specified in a simulation. From the simulation come synthetic defect images with parameters known in advance — a ground truth defined at the moment each example is created, no longer reconstructed afterwards by a human.

In the pilot, twenty models were trained purely on such synthetic data, each from a different random start. All twenty gave consistent predictions on the same real production images: the difference between most models stayed within statistical noise. No model trained on human labels from the gray zone had shown that stability. The twenty models agreed with one another precisely because the reality they were fitted to was stably documented instead of floating from shift to shift.

When synthetic data helps

The case just described is one instance of a broader pattern. Synthetic data allows what no scaling of human labeling delivers: it lifts ground truth out from under human inconsistency. Every example carries a known truth by construction. On such data, models of any architecture can be trained and honestly compared against one another; situations that occur once a year in real production can be simulated and trained as densely as everyday ones; the cases where the model errs can be amplified deliberately, without waiting for such examples to drift naturally into the pipeline.

Where synthetic data does not help

Synthetic data is no universal answer. It works where the physical phenomenon is understood and reproducible in simulation. It fails in domains where the truth itself has yet to be discovered — in complex social data, or in phenomena whose mechanics cannot be modeled.

What it does achieve: it moves the bottleneck from extracting data to modeling the phenomenon. Whoever can describe the physical behavior of a defect can create as many training examples as needed. Whoever cannot, never had a real ground truth in the first place — and machine learning would never have saved that project.

Conclusion

Some data problems are definition problems in disguise. Whoever commissions an AI project without checking whether a ground truth exists at all is commissioning an expensive repetition of what the production line already has: the same inconsistency, at a higher tempo, with the bill for development, integration and additional computing hardware on top.

Three Questions for Your Own Context

1. **Does a reproducible ground truth exist?** Would two independent labeling operators, given the same data, arrive at the same labels? Has this been tested on a sample of 100 cases, or is it merely assumed?
2. **If not: can the defect itself — or whatever phenomenon the model is meant to recognize — be modeled physically?** Can its physical, rule-describable or simulatable properties be documented so that synthetic data with a known truth can be generated?

3. **If that too is impossible: is machine learning the right lever?** Is this a definition problem that has to be solved organizationally before any modeling begins?

Sources

1. McKinsey & Company, *The State of AI*, global survey, 2025 edition.
2. Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*, Yale University Press, 2021.