

Already decided, left open, out of reach: decoding regulation with AI

David Wells, Adrien Loridon, Jordan Sheldahl and Luca Magagni

EAE Business School, Barcelona

ABSTRACT

The Corporate Sustainability Reporting Directive makes every large company in Europe publish an audited account of its sustainability impacts and risks. Before any of it is written, the company runs a double materiality assessment, which decides what goes in the report and what never appears at all. The standard governing that decision leaves the hardest parts to the company, and the work usually lands on people who have never done it. As companies begin to face reporting obligations, we expect the strategy of companies looking to save resources will be to integrate AI to tackle the regulatory burden. We set out to find the right balance to strike when attempting this, where AI can be used, and where it could result in compliance risk. We used AI to read the regulation, sorted all 99 provisions governing the assessment into three kinds, and built a working platform to find out which kinds a machine can reach. 26 provisions are already decided by the text and a machine executes them. 17 are deliberately left open, and a person has to supply a value and a reason. 56 are out of a machine's reach. AI's contribution was upstream: it proposed a reading of every provision that an author then checked, and helped one non-software engineer encode the 26 already decided provisions in 19 working days. The software then enforces that decisions get made, in order, once, with a name on each. It cannot check that any decision was good, and the one failure we found in our own build was exactly that. Used responsibly, AI reads and helps encode the first group, and stays assistive on the other two, proposing but never deciding where the standard requires a name, and never substituting for the judgement it cannot supply. The method should apply to any regulation that contains "shall" and an undefined term, which is most of them.

Key words: *Strategic Project Management, Compliance, agentic AI, CSRD, Double Materiality Assessment, design science.*

1. THE PROBLEM

The Corporate Sustainability Reporting Directive, the CSRD, is the European law that turned sustainability reporting from a marketing exercise into an audited one. Large companies operating in the EU have to publish what their business does to people and the environment, and what sustainability issues do to their finances, inside the management report that sits beside the annual accounts. An assurance provider signs off on it. The board carries responsibility for it. The detail comes from the European Sustainability Reporting Standards, the ESRS, which the CSRD (European Union, 2022) required the Commission to adopt and which it adopted as Delegated Regulation (EU) 2023/2772 (European Union, 2023): two standards that always apply and set the method, and ten topical ones running from climate and pollution through to consumers and business conduct.

Nobody reports on all of it. That is the point of the double materiality assessment.

The assessment is the step where a company decides what it will report on before it reports anything. Double materiality means the decision runs on two lenses at once. Impact materiality asks what the company does to the world; financial materiality asks what the world does to the company, and a topic goes in the report if it is material on either. Everything the company concludes is not material comes out of the substantive reporting, and the standard then asks it to say why. The assessment therefore sets the scope of the report, decides which of the ten topical standards apply, and is the first thing an assurance provider is required to examine. Get it wrong, and every downstream number answers the wrong question.

The standards governing the assessment run to 99 obligations across ESRS 1 and ESRS 2, the two that set the method rather than a topic, and nine of the ten topical standards. They state a rule and decline to set the number, and they use terms like "reasonable", "significant" and "undue cost" without defining any of them. A company is told it must apply a materiality threshold, meaning a cut-off above which a topic counts, and never told what that cut-off should be.

The task then goes to whoever is available, often a function with no background in the standard, facing an ambiguous process with a deadline and no benchmark to test the result against. The first reporting cycle showed what that produces. ESMA's review of 91 issuers found that 61.7% met the objective of the disclosure requirement on the assessment process, that some disclosures were boilerplate with little insight into the judgements made, and that two issuers received a qualified opinion, which is the auditor's formal statement that something is wrong, one of those citing the materiality assessment (ESMA, 2025). No AI was involved in any of it.

Vendors in the market are selling AI as the answer and we want to know which parts of the regulation AI is capable of reaching. To achieve this we used AI to read through CSRD / ESRS reporting standards, sorted every provision by what kind of decision it contained, and built a platform to test the accuracy of the sorting. This paper's primary research questions follow. Which provisions can a machine execute, which need human input, which are completely out of reach, and how does that affect where cost and compliance risk sit? And, what does building this system show about what software can and cannot enforce?

2. THREE KINDS OF PROVISION

Every provision in the regulation is a decision in one of three states. We give each a plain name and the technical term we use for it elsewhere.

Already decided (computable).

The rule is fully stated, and the inputs exist. ESRS 1 paragraph 80 fixes short term as the reporting period, medium term as up to five years, and long term as beyond five. Nothing in it is open, so a machine executes it.

Left open (parametric).

The rule is stated, and the number is not. ESRS 1 paragraph 38 requires the undertaking to determine which impacts, risks and opportunities are material, "supported by qualitative considerations and quantitative thresholds", and declines to say what the thresholds are. Application Requirement 17, inside the standard and carrying the same authority, confirms the omission is deliberate. A person supplies the value, and the standard requires that person to say why.

Out of reach (irreducible).

ESRS 1 paragraph 32(a) requires "reasonable and supportable evidence available without undue cost or effort", and defines neither reasonableness nor undue cost anywhere. Or the obligation is to engage with a workforce, a community or a supplier, which has nothing to compute.

How we counted.

One unit is one provision containing "shall", in ESRS 1, ESRS 2 or the topical standards, that a preparer must satisfy to perform or evidence the assessment. One numbered paragraph or sub-paragraph carrying a distinct obligation is one provision. Guidance is excluded as non-authoritative, as is anything governing what is disclosed after the scope is set. Four consequences of that rule are calls another researcher could make differently. Application Requirements are admitted where they contain "shall" and bear on the assessment, which brings in eight provisions. Three obligations in ESRS 1 sit across two kinds and are recorded as sub-items, each marking where a system has to put a human checkpoint. Seven provisions stating a materiality criterion in the declarative are excluded, and all seven would have been left open or out of reach, so a relaxed rule moves the balance away from the machine. Five of the 26 already decided treat an upstream determination as a supplied input, and a reader who rejects that reading moves five out of the group. The argument survives it, because it turns on the size of the out-of-reach group.

The classification runs against the November 2025 draft ESRS (EFRAG, 2025b), adopted in substance by the Commission on 3 July 2026 as Delegated Regulation C(2026) 5010 and intended to apply for financial years beginning on or after 1 January 2027 (European Commission, 2026). At the time of writing, that act had not been included in the Official Journal. Every citation used references the November 2025 draft as paragraph numbers moved during this project. All 99 rows with respective categories and reasoning are published as a supporting document so readers are able to disagree with any call precisely.

How the classification was made.

With AI, in the mode this paper goes on to recommend. A language model read each standard and proposed a category and a one-line reason for every provision. An author reviewed every proposal against the text, changed the ones that were wrong, and a second author has since reviewed all 99 rows independently and agreed the classification. The model did the reading. The authors made the calls. No practitioner outside the author group has validated the result, and Section 8 says what would fix that.

Table 1. Provisions by source and kind.

Source	Already decided (computable)	Left open (parametric)	Out of reach (irreducible)	Total
ESRS 1	18	4	15	37
ESRS 2	8	1	22	31
Nine topical standards	0	12	19	31
Total	26	17	56	99

Three things in the table matter. 56 of 99 provisions are out of a machine's reach.

The left open group is small and decides the most. 17 provisions, and two of them are the materiality determinations themselves, paragraph 38 for impacts and paragraph 48 for financial effects. The group that decides what a company reports on at all has the fewest provisions in it, and ESRS 2 then requires the company to disclose the thresholds it applied, which is the standard asking for the number it declined to set.

The topical standards do not contribute any already decided provisions, because every mechanical rule in them dictates the calculation of a metric after the scope is set. They add nine nearly identical gates deferring back to ESRS 1, hazard screening in E1, spatial scoping in E4, and stakeholder engagement obligations found in all four social standards.

Crossed with how the AI is run.

A model a person prompts, reviews and overrules is a different risk from a system that plans and executes across many steps and shows the person the result. Parasuraman, Sheridan and Wickens (2000) separate the part of a task automation is applied to from the level of autonomy applied at it, and we adopt that separation with the parts replaced by the three kinds of provision, because a regulation specifies obligations rather than cognitive stages. Bainbridge (1983) supplies the warning: automating what can be automated leaves the person whatever could not be, and removes the routine work through which competence at the rest was maintained. Automate everything a machine can reach and 73 of 99 provisions remain, 56 of them out of reach entirely, so an organisation that buys a tool and shrinks the team that used to do the work by hand takes on that risk. Krasteva-Hristova and Georgieva (2026) ask a related question from the auditor's side rather than the preparer's. They build a scale for how much an AI system can be trusted with sustainability assurance, and on it, letting the system decide materiality or write a conclusion is ruled out. That scale only measures autonomy, so it has no equivalent to our middle category, which comes from how the standard is drafted rather than how independently the machine acts.

Table 2. Compliance risk by kind of provision and mode of operation.

	Assistive (person prompts, reviews, decides)	Agentic (autonomous, multi step)
Already decided (computable)	Low. Speed on a step with no cost per assessment.	Low risk, on the condition that the execution is logged and can be reproduced from the inputs and the rule version.
Left open (parametric)	Moderate. The model can provide precedent. The person sets the value, states the reason, and the system will record both, as well as who approved it.	High risk. An autonomous system setting a threshold produces a conclusion with no author. Defensible only where the value, its reason and a named approver are all captured.

Out of reach (irreducible)	Moderate. Useful for prompting and challenge. The judgement stays with the person.	Fails. There is no rationale to disclose; nobody formed one. Severe risk during an audit.
----------------------------	--	---

The argument in this field is usually had about the corners. The consequence sits in the middle row, and Table 2 rests on argument alone.

The method to the work.

What this paper does is apply project management to a regulatory process. We treated the assessment as a body of work and broke it into 99 obligations under a stated counting rule. We then sorted each by who can resource it: a machine, a person with a value to supply, or a person exercising judgement. Effectively a work breakdown structure, followed by a resourcing decision. The autonomy axis concerns decisions about the level of supervision. The named approver on every left-open provision is an accountability assignment, and the approval gate in Section 4 serves as the control point.

3. WHAT A DEFENSIBLE ASSESSMENT REQUIRES

Section 2 says where AI may be applied. This section concerns what the assessment needs to produce, so that Section 5 has evaluation criteria for the artefact. Three requirements and four quality criteria govern that, and Table 3 in Section 5 states each one in full with its source.

Two facts set the requirements: EFRAG's implementation guidance sits outside the standards and also carries no authority, so a company may adopt any approach that satisfies ESRS 1. This leaves four decisions: the boundary of the assessment, the weight given to stakeholder input, the materiality threshold, and the form in which conclusions are recorded. None of these has a common reference point. The rules can also move: the severity-precedence rule for potential negative human rights impacts is in paragraph 45 under the standards in force and is moved to paragraph 41 in the adopted revision. R1 is what the regulation requires of the record. R2 and R3 are what we infer a system encoding must do.

The quality criteria are our own. Three of these came from the first reporting cycle and Q2 from the standard itself. EFRAG's progress report records that preparers document non-materiality item by item and treat it as a burden (EFRAG, 2025a, paragraph 25). ESMA's fact-finding on 91 issuers records an omission of value chain coverage (which sets Q4) and boilerplate disclosure, with little insight into the judgements made (ESMA, 2025). Close to 80 per cent of that sample did explain how thresholds were set, so Q3 tests whether the reason is attached to the threshold in a form an assurance provider can inspect and trace to a person. Q2 comes from the standard itself.

4. WHAT WE BUILT, AND WHO DID WHAT

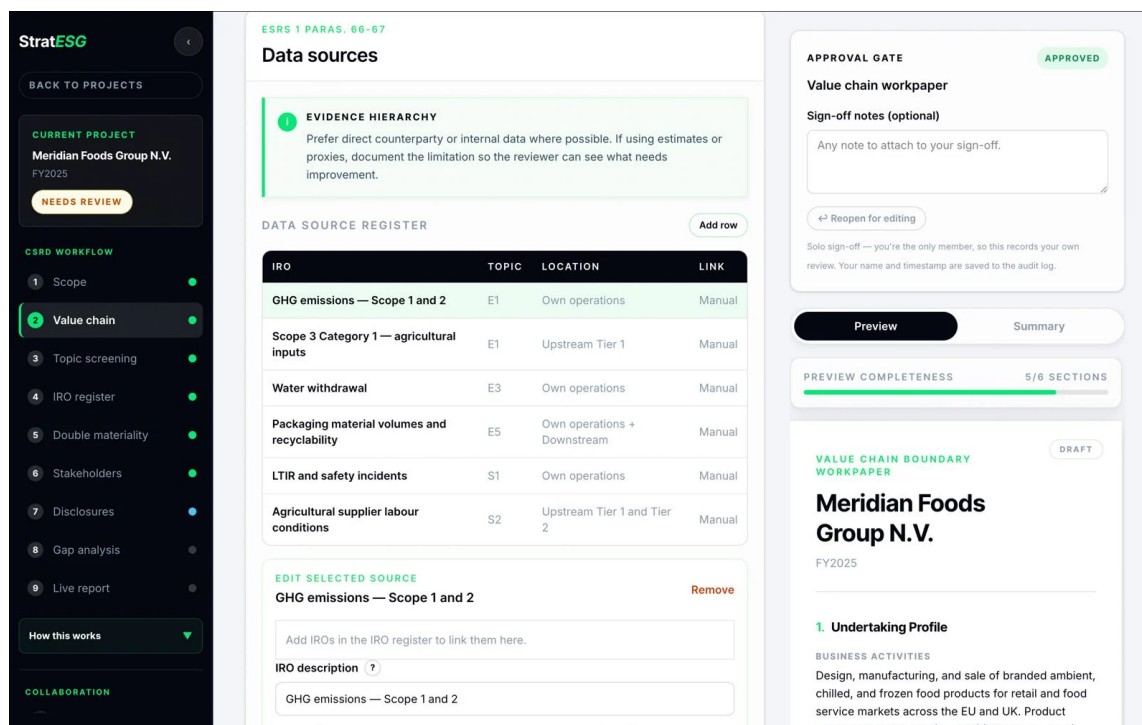


Figure 1. Screenshot of the platform.

The platform takes a company through the assessment as a 9-step sequence, from scope and the register of impacts, risks and opportunities, through scoring each record on both lenses and writing the outcome back, to a workpaper an assurance provider can read. It was built by one of us, who is not a software engineer, with input from a second on the method, in 124 commits over 83 days using AI-assisted development tooling. Three different things did the work, and we keep them apart.

AI read and proposed.

A model proposed a reading of every provision, as Section 2 describes, and models wrote most of the code under direction and review. Encoding the 26 already-decided provisions as executable rules took most of 19 working days, and the topical standards became a sequence that a preparer cannot skip. We ran no comparison, so we cannot size the contribution. The reason we know where the machine's reach ends is that we tried to make it do it all and found where it stopped.

Software enforces.

At runtime, no stage of the assessment calls a language model. The scoring engine is self-contained code with no way to reach the internet, so it cannot call a model even by accident. What the software does is refuse. It will not complete an assessment with a topical standard skipped. We set this intentionally so that every standard is a gate with a conclusion. The program will not score a record without all six scores and a selection of which test applies. It holds one register, so there is no second spreadsheet with different answers. It writes the applied thresholds and the methodology version to an assessment when a score is entered through the interface, so a year later the workpaper shows the thresholds actually used. If a record already arrived scored by another route that carried no stamp, the workpaper prints it as unstamped rather than guessing. It keeps the engine's conclusion beside any human override, so the departure can always be reconstructed. A module approval refuses to proceed without a recorded submitter and a recorded approver.

People decide.

The system scores what it is given. A person identifies the records, supplies the scores, selects which test applies, marks whether the record concerns human rights (which changes how it is scored), sets the thresholds and writes the reason, and enters any override. Those six inputs are the left-open and out-of-reach rows of Table 2 appearing in the implementation. The regulation put them there. Figure 2 shows two of the mechanisms as a preparer sees them.

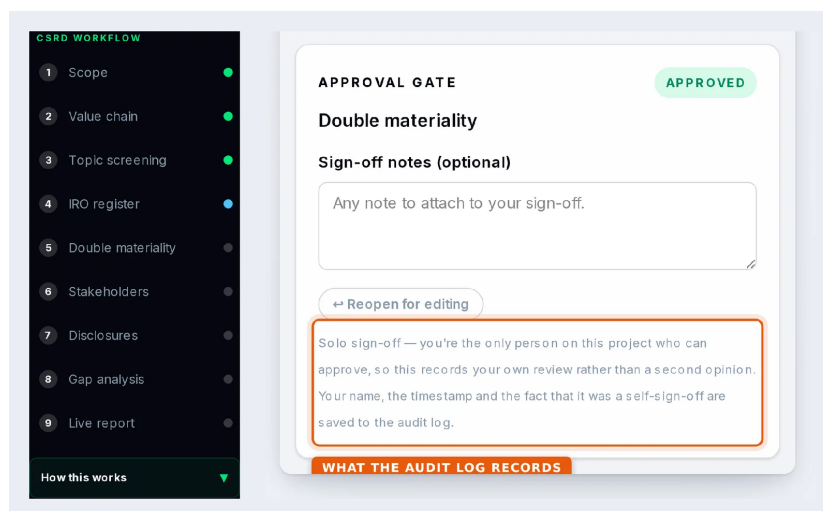


Figure 2. Two of the mechanisms. Left: the 9 sequence, each step carrying a state, so no step is reached out of order. Right: the module approval gate, whose note states that a solo sign-off records one person's own review rather than a second opinion, and that the name, the timestamp and the fact of self-approval go to the audit log. Source: the artefact at the frozen version of 8 September 2026, demo data.

The export renders in a single document: the outcomes, the method as applied, the thresholds used, with their reasons and versions, the register, with the engine's conclusion beside the effective one, and any overrides, with their reasons. That document is what satisfies R1. This mechanism is worth mentioning as it is evaluated in Section 5.

A complete assessment of 20 records spanning all ten topical standards was run end-to-end against the frozen code of 8 September 2026, with 506 unit tests and 13 integration tests passing. The determination took 1.63 milliseconds and no money to run once encoded. We report that only to show the already decided group is genuinely decided. We did not measure what preparers spend on this step by hand, so we claim no saving.

5. WHAT WE FOUND

The cost is in the rows a machine cannot easily reach. The published burden estimates put CSRD reporting in the 10's to 100's of thousands of euros per company per year. The most cited of them are incremental over the pre-CSRD baseline (and apply to a population a later simplification package removed from scope). These come from a cost-benefit analysis prepared for EFRAG rather than from the Commission (European Commission, 2025). The cost sits in the 56 provisions a machine cannot easily reach. However, it is worth mentioning that assistive AI in this context is useful, as mentioned in Table 2. The already-decided group has its own cost: 19 days to encode, and when EFRAG renumbered a key rule, every encoded rule had to be reread. That cost is incurred by the vendor each time the standard setter moves.

The compliance risk is in the middle row, and it is already live. The standards carry no authorship obligation. A machine may set a threshold, and if the reason is recorded and a person signs for it, the disclosure requirement is met. The failure has two parts: a value with no stated reason and a value with no accountable author. The first fails the disclosure requirement directly. The second concerns the board's collective responsibility for the management report under Article 33 of Directive 2013/34/EU (European Union, 2013). The strongest ground is assurance: Article 34(1) of the same directive places "the process carried out by the undertaking to identify the information reported" inside the assurance opinion, and the European auditing oversight guidelines tell practitioners to put particular emphasis on it and to qualify where they cannot get far enough into the process to say whether something material is misstated (CEAOB, 2024). The qualified opinions ESMA records are that mechanism working before any AI arrived. At the time of writing, Spain's transposing law was still in parliament, so a Spanish preparer faces no domestic penalty. Its exposure is that an assurance provider is required to look and permitted to qualify.

Against the criteria in Section 3. Table 3 reads conformance from the code rather than from anyone using it.

Table 3. The requirements and quality criteria, their sources, and how the artefact conforms.

Criterion	Source	Status	Basis
R1. The determination records its own reason and supporting evidence.	From the regulation	Met at export	The workpaper renders method, thresholds, reasons and overrides together. The engine alone produces a value.
R2. The four decisions are explicit values a reviewer can see, change and inspect.	Our inference	Partially met	Thresholds and the recording form are stored per assessment, though no path sets a threshold other than the default. The boundary and the weight given to stakeholder input are not explicit values at all.
R3. Rule sets are versioned, and each assessment records the version applied.	Our inference	Partially met	Written when a score is entered through the interface. A record persisted by any other route carries no stamp, and the export prints it as unstamped.
Q1. The assessment is substantive analysis rather than item-by-item confirmation.	From EFRAG, generalised further than the source goes	Met structurally	Six scores and a recorded selection of which test applies, with no completion path without them. Whether the analysis behind the scores is substantive was not observed.
Q2. The two materiality lenses stay separable through to the conclusion.	From the framework	Met	No combined numeric score at any point. Both lenses retained to a single categorical outcome.
Q3. The reasoning behind a threshold is recorded and attributable to a person.	From ESMA	Partially met	Stored per assessment with a reason rendered in the workpaper. The field is required to exist and not required to say anything, no approver is recorded, and every assessment carries the default.
Q4. Upstream and downstream value chain impacts are brought into scope.	From ESMA, the best evidenced of the four	Partially met	The field is on every record and the software will not save a record without it. Nothing checks that it holds anything.

Four rows are partial. R2 because two of the four decisions were never parameterised, R3 because the version stamp is written on one path only, and Q3 and Q4 for the same reason as each other, which is the finding this section exists for.

A machine can enforce that a decision was made. It cannot enforce that the decision was any good. The first half is the gates. Refusal is something software does well, allowing process enforcement.

The second half is the one failure in our build that a better resourced team would not have fixed. Every record includes a field indicating where in the chain of suppliers and customers the impact lies, and that field is required. Nothing checks if it holds anything, so a record with an empty string is accepted, scored and exported to a workpaper where it renders as "Not captured". The threshold reason fails the same way: a string input is required, and an empty string is accepted. The default would send with a reason saying somebody should write the reason. No code path lets anybody set a threshold, so every assessment carries that default and every workpaper's stated reasoning is a note saying the reasoning should be written. Figure 3 puts the on-screen threshold beside the same threshold in the deliverable.

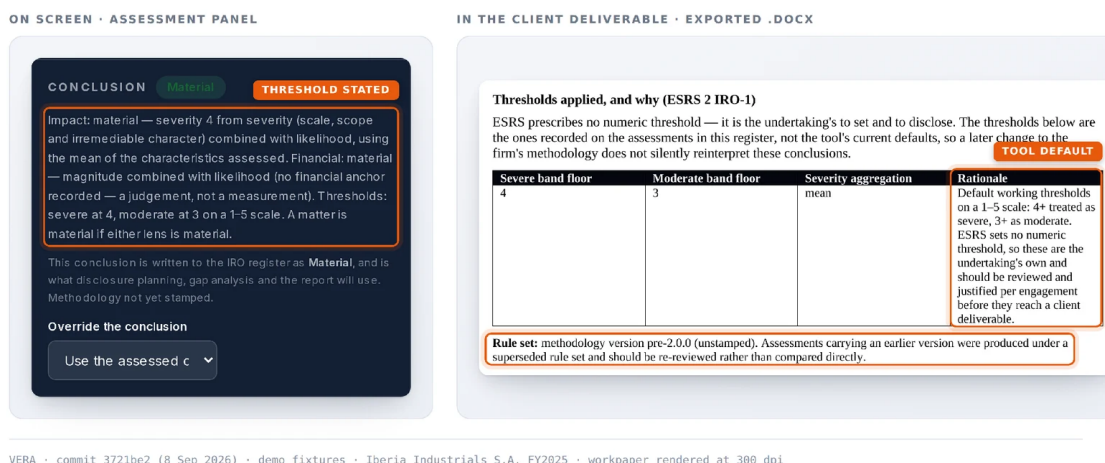


Figure 3. Threshold provenance, on screen and in the deliverable. Left: the conclusion block states the thresholds applied. Right: the same thresholds in the exported workbook, where the rationale column carries the tool's default text, which ends by saying the thresholds should be reviewed and justified before they reach a client deliverable. This is the client deliverable. The line beneath records the methodology version as unstamped. Source: as Figure 2, workbook rendered at 300 dpi.

We identified five fields in the build with that problem. A non-empty check is trivial to write, and a full stop would satisfy it. What the left open group needs is a named approver attached to the value and a person who reads the reason as text, which is a process control and belongs in the assurance plan.

Autonomy makes it worse by removing the author. The already-decided group survives agentic execution because the engine decides nothing. The left-open group fails it for the reason in Table 2, and the standard requires that the reason be disclosed. In one development session, we had an agent apply 12 migrations to the production database with no humans in the loop; one of them caused infinite recursion in the read policies. The service was down for two hours. That is the exact bottom-right cell of Table 2, and this was experienced by a team that was trying to be careful.

6. WHAT THIS MEANS, AND FOR WHOM

The method asks three questions of every provision. Has the text already decided this? Has it deliberately left a value open? Or is it out of a machine's reach? It is worth noting that nothing in these questions is specific to sustainability reporting or regulation.

For preparers, consultants and buyers.

Ask the software vendor which of the 99 provisions the software works on. A claim on the already decided group is checkable, so ask for the encoded rule, the version of the standard behind it and a log of what ran. On the left open group, ask where the threshold is stored, what reason sits beside it, who signed that reason and whether the assessment carries the thresholds applied when it was scored. A claim on the out-of-reach group covers 56 provisions nobody can meet, so ask what the tool does when the person using it disagrees.

The positive use-case sits on those same 56. This is where a model in assistive mode is useful precisely because a person will stay in the loop and sign off at the end. We did not measure that contribution, and we expect it to be the larger. Structuring the process and workflow may be worth as much as the automation, on argument rather than on anything measured. Many companies hand the assessment to a department that has never done it, and software that imposes a fixed order on an ambiguous process, holds one register and refuses approval without a recorded submitter and approver gives a first-time preparer a route through the standard. That route is what a preparer wanted when they asked us whether the file recording why the thresholds were what they were had to be kept. The standard requires no such file. The assurance provider will ask for it anyway.

For regulators, standard setters and assurance providers.

The three states are properties of how a text is drafted, so a drafter can intentionally move a provision between them. Naming a number takes the judgement off the undertaking. Leaving it open keeps the judgement where the circumstances are known and puts the work on preparers who may have no basis for setting it. Both are legitimate,

and the ESRS took the second approach where it counts most: in materiality determinations. What a standard can add under either choice is a sentence naming who has to own the value, since an empty string satisfies a required field. Assurance already has a mandate, because Article 34(1) places the process within the opinion. A reviewer can add an additional question. Which of the three states was the tool used on, and does the conclusion have a named author? In its review of financial year 2024 sustainability information, the CNMV set materiality as its first priority area (CNMV, 2026).

For a wider readership.

56 of the 99 provisions are out of a machine's reach, and the number describes the regulation, because the text asks an undertaking to weigh its own circumstances, its value chain and its stakeholders, which no model can supply. A demonstration of the 26 already decided shows the cheap quarter of the problem.

We have run the method on one regulation. What follows is an analogy rather than a result. Tax rules set thresholds and leave "reasonable" undefined; procurement sets a procedure and leaves "proportionate" open; and data protection sets some obligations and leaves the choice of appropriate measures to the company holding the data. For a company outside the CSRD entirely, the practical output is a map of what to automate, what to put a named approver on, and what to leave with a person with approval gates.

7. LIMITS

This was a single case study with one artefact, no adopters, and no control arm. The classification was AI-assisted and reviewed by two of the authors; no practitioner outside the author group has validated it. The three kinds of provision rest on argument and on a single implementation, whereas the assistive and agentic distinction in Table 2 rests on argument alone. Table 3 records conformance from the code rather than effect in use, so whether the mechanisms make anyone faster or more consistent is unknown, and so is the value of assistive AI on the 56, which we expect to be large. Human time on those 56 provisions, the highest cost our own results imply, was not measured. The 5 provisions without content fields were found during an external review of a build made by one person with no second reader, a condition that would have produced the same result with any tooling. In the terms of Venable, Pries-Heje, and Baskerville (2016), the evaluation here is artificial and formative, and, following Gregor and Hevner (2013), it is a worked implementation and the general lessons it supports rather than a general theory.

8. FUTURE WORK

Three pieces of work follow, ordered by cost.

Independent validation of the classification.

Two or three CSRD practitioners from outside the author group classify the same 99 provisions, and we report Krippendorff's alpha, proposed by Hayes and Krippendorff (2007) as a standard measure for coding data because it holds across any number of coders and tolerates the missing data a practitioner produces by declining a provision outside their own experience. The convention treats alpha at or above 0.800 as sufficient to rely on the coding and 0.667 to 0.800 as grounds for tentative conclusions only (Krippendorff, 2004). Coefficients that correct for chance agreement get unstable when one category dominates, and 56 of our 99 provisions fall into a single state. A confidence interval from bootstrapping shows that uncertainty, a fixed cutoff would hide it.

A single preparer pilot.

One preparer runs a real assessment on the artefact with time recorded against each provision. That is the study that would support the effort claim we decline to make here. To be worth anything, it has to control for first-cycle learning, since a first assessment is slow whichever tools are used, for who does the work, and for entity complexity. Setup and data loading count against the artefact, as does rework after internal review, which is where judgement time surfaces. One preparer supports no baseline comparison, so the output is a distribution across the three states rather than a saving. This would be the naturalistic summative episode the evaluation so far lacks, and we know of no published field study measuring preparer time this way.

A second regulation.

Applying the method to another regime tests whether the three states are a property of ESRS drafting or of regulation generally. Proportions that shift would be the more useful result, because they would show what in the drafting produces each state.

9. CONCLUSIONS

Every provision in a regulation is a decision that is already made, left open on purpose, or out of a machine's reach. In the CSRD double materiality assessment, the split is 26, 17 and 56. That map is what we would hand a preparer first, ahead of any software, because it tells them where their own people have to stand.

Reading the regulation is what AI did best, and it is the part nobody sells. A model read eleven standards and proposed a category and a justification for all 99 provisions. Then an author corrected it and signed off on the result. Nothing in that process is specific to sustainability reporting, and none of it needs a platform. Any organisation sitting in front of a rulebook can run the same operation and produce a list of what to automate, what to put a named approver on, and what to leave with a person or persons. The output is a resourcing plan for a compliance obligation, and it is cheap to produce.

Refusal is what the software did best. It will not let a standard be skipped or a record be scored without reason and it holds one register so there is no other spreadsheet with different answers. It records the thresholds and the methodology version onto the assessment so a workpaper opened a year later shows what was used, and it will not close a module without a recorded submitter and approver. For the department responsible, and a deadline and no background in the standard, it is a route through an ambiguous and complex process, and it has value whether or not provisions get automated.

The left open row is where responsible use gets tested. 17 provisions state a rule and leave the value to a person which a model can propose, but the standard still requires a reason and a name behind it. An autonomous system that sets a threshold on its own produces a conclusion nobody signed and ESMA already recorded qualified opinions on exactly that gap in the first reporting year, before any AI was involved. Here a model stays assistive, it can surface precedent and draft the reason, not produce the number.

The largest opportunity sits on the 56 rows a machine cannot execute. Those provisions necessitate judgement about an undertaking's circumstances, and often an audit trail that can be traced back to human who signed off on it. A model in assistive mode can be immensely useful here: from finding precedent for a threshold to drafting a rationale an author edits and signs, or turning a stakeholder transcript into candidate impacts for somebody to accept or reject, even arguing against a scoping call that looks convenient. The market is competing over the quarter of the standard a machine can execute even though the work sits in the other three quarters, where we expect the contribution to be the largest but did not measure. Section 8 expands more on how we would.

A machine can enforce that a decision was made but it cannot enforce that it was any good. Our system records every conclusion and puts a name against it. It cannot tell whether the reasoning behind that conclusion means anything, and we found exactly that gap in five places, precisely where the framework predicted it.

73 of 99 is the number to take away. This is the left open and the out of reach rows together (the two no machine can decide alone) which no future model could shrink. They ask a person to weigh their own circumstances and sign for the result. This figure belongs to the regulation rather than to the state of the technology, and it is why the useful question to ask about any software in this market is not whether it helps, but which of the three states it operates on, how, and who signs for the rest.

REFERENCES

- Bainbridge, L. (1983) Ironies of automation. *Automatica*, 19(6), pp. 775 to 779. doi 10.1016/0005-1098(83)90046-8.
- CEAOB (2024) Guidelines on limited assurance on sustainability reporting. Committee of European Auditing Oversight Bodies, adopted 30 September 2024.
- CNMV (2026) Informe sobre la supervisión por la CNMV de la información de sostenibilidad del ejercicio 2024 y principales áreas de revisión del ejercicio siguiente. Comisión Nacional del Mercado de Valores, Madrid, March 2026.

- EFRA (2025a) ESRS Revision: Progress Report as of 20 June 2025. European Financial Reporting Advisory Group, Brussels.
- EFRA (2025b) [Draft] European Sustainability Reporting Standards, November 2025. European Financial Reporting Advisory Group, Brussels.
- ESMA (2025) Materiality matters (!): fact finding on the application of the materiality principle in sustainability statements, ESMA32-846262651-5288. European Securities and Markets Authority, Paris, 14 October 2025.
- European Commission (2025) Commission Staff Working Document accompanying the Omnibus I proposals, SWD(2025) 80 final, 26 February 2025. Brussels.
- European Commission (2026) Commission Delegated Regulation C(2026) 5010 of 3 July 2026 amending Delegated Regulation (EU) 2023/2772 as regards the sustainability reporting standards. Brussels.
- European Union (2013) Directive 2013/34/EU on the annual financial statements, consolidated financial statements and related reports of certain types of undertakings, as amended. Official Journal of the European Union.
- European Union (2022) Directive (EU) 2022/2464 as regards corporate sustainability reporting. Official Journal of the European Union.
- European Union (2023) Commission Delegated Regulation (EU) 2023/2772 supplementing Directive 2013/34/EU as regards sustainability reporting standards. Official Journal of the European Union.
- Gregor, S. and Hevner, A.R. (2013) Positioning and presenting design science research for maximum impact. *MIS Quarterly*, 37(2), pp. 337 to 355.
- Hayes, A.F. and Krippendorff, K. (2007) Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1), pp. 77 to 89. doi 10.1080/19312450709336664.
- Krasteva-Hristova, R. and Georgieva, V. (2026) Artificial intelligence in sustainability assurance: accounting challenges, audit risks and a conceptual framework for ESG verification. *Accounting and Auditing*, 2(3), 15. doi 10.3390/accountaudit2030015.
- Krippendorff, K. (2004) *Content Analysis: An Introduction to Its Methodology*, 2nd edn. Thousand Oaks: Sage.
- Parasuraman, R., Sheridan, T.B. and Wickens, C.D. (2000) A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 30(3), pp. 286 to 297. doi 10.1109/3468.844354.
- Venable, J., Pries-Heje, J. and Baskerville, R. (2016) FEDS: a framework for evaluation in design science research. *European Journal of Information Systems*, 25(1), pp. 77 to 89. doi 10.1057/ejis.2014.36.