When One AI System Confirms Another

From the Riemann Hypothesis to a Society's Right to Challenge Decisions

Jandislei Antonio Genova

Brazilian attorney and independent researcher · American English edition · November 22, 2026

Introduction

In chapter 15 of *Autonomia decisória na era da IA*, my book on decision-making autonomy, I examined the use of algorithms in government. Picture a public official looking at a map that recommends expanding services in certain areas. Before the official signs an order, someone has already decided which needs the map will show. As I wrote then, "the problem has already been framed by the time it reaches the decision maker." If a second AI system endorses the first system's assessment, who can show that both left the same area off the map? [1]

Finding a solution, checking its proof, and explaining why it holds are different kinds of scientific work. The person who makes a discovery may need help verifying it. Those who verify it may need years to find an explanation they can teach. Meanwhile, researchers and public officials may already be making decisions on the strength of the result.

I begin with Levent Alpöge's account and the public record of how the proof of the Riemann hypothesis was checked. I then examine empirical studies of shared errors and documented cases in public administration. I also use hypothetical cases to trace a decision's effects on the people who depend on it. Mathematics has explicit rules for checking a deduction. Health care and public budgeting raise further questions. Does the chosen measure capture what people need? Who decided which needs take priority? [2] [3]

I argue that public confidence in an AI-assisted finding requires a way to challenge it that the party making the claim does not control. That matters to a reviewer signing a report and to a country that relies on foreign infrastructure to assess a decision affecting its own people. When evidence points the other way, we can ask who can investigate it, how long that takes, and who has the authority to reverse a decision already in effect.

I use the phrase "a society's right to challenge decisions" in the subtitle to describe how I believe institutions should work. People affected by a decision need a way to obtain evidence and challenge it, either directly or through someone who can represent them. The relief they can seek in court depends on the right at issue, who has standing to sue, the available evidence, and the legal procedure that applies.

The problem on the reviewer's desk

Suppose a journal asks you to review a proposed solution to a problem that has been open for more than a century. You know the problem, understand the conclusion, and can reproduce the calculations in the examples. Yet the manuscript turns on a step you cannot justify. An AI system examines it and reports that the proof is correct. The editor is waiting for your recommendation. What could you honestly sign?

On November 20, 2026, mathematician Levent Alpöge, a researcher at Anthropic, announced a proof of the Riemann hypothesis that he had found with Claude. He wrote that "this time Claude far outpaced me near the finish line." Alpöge describes steering the research by bringing two ideas together. When he released the proof, he had yet to master some of the mathematics involved. The next day, the Lean and Mathlib communities certified that the formal statement matched the hypothesis and that the proof was valid. Other researchers could already check an argument Alpöge was still working to understand in full. [2] [3]

The prospect I find genuinely disturbing is that one AI system announces a discovery, another endorses it, and no one has an independent way to tell a correct result from a shared error. There is empirical evidence for that concern. In a 2025 study, researchers examined one set of questions on which pairs of models gave **the same wrong answer 60% of the time** when both were wrong. If agreement between the two were the only evidence, the discovery would still need to be demonstrated. In the philosophy of science, *epistemic opacity* describes a situation in which someone who depends on a result cannot examine all the steps that matter to how it was obtained. [4] [5]

Even a correct result can enter widespread use before anyone understands its limits well. Reviewers answer for what they attest to. A university that teaches the result, a company that builds it into a service, and a government that adopts it must each answer for their own decisions. The risk grows when all three depend on the same source to tell them whether they have applied the result outside the conditions under which it holds.

How the proof of the Riemann hypothesis was checked

In 1859, Bernhard Riemann proposed a hypothesis about the zeros of a function closely tied to the distribution of prime numbers. It states that every nontrivial zero of the Riemann zeta function has real part ½. The difficulty lies in the word “every.” Even if one knows trillions of zeros, one still has to show why none of the others can be an exception. The proof Alpöge announced also covers the generalized Riemann hypothesis for Dirichlet L-functions. [2] [6]

The published library contains about 17 million lines of Lean code. Lean is a language for writing and checking proofs by computer. Specialists examined the formal definition of the hypothesis. Two tools compared that definition with the theorem actually proved. Proof checkers, some written independently, checked the deductions against explicit rules and axioms. My account draws on the joint statement by the Lean and Mathlib communities and on the public logs of Andrew Sutherland of MIT, who reported using Claude Code to run his checks. With the exported proof and the checkers, another team can repeat the tests without asking Claude to judge whether the proof is valid. [3] [7]

On November 21, 2026, Michael O. Rubinstein of the University of Waterloo made a 164-page guide to the Riemann zeta function case available. The title page credits Claude Fable 5.2, Codex Astra 6.2, and Gemini Carbon, three AI systems, with writing the guide at Rubinstein’s request. The guide derives the Riemann hypothesis from a central inequality, labeled (M), and two auxiliary estimates. It proves both estimates using previously published theorems. For the full proof of (M), it refers readers to the formal development in Lean. Someone preparing to teach the proof can therefore identify the crucial step for which the guide still lacks a full exposition. [6]

How did the ideas emerge as Alpöge worked with Claude? Why does the conclusion follow from the premises? A team that finds a shorter proof could answer the second question without reconstructing the sequence of attempts that led to the discovery.

William Thurston argued that mathematical progress includes expanding what people can understand. A doctoral student adapting the argument to a related problem can use the theorem under the conditions for which it was proved. To dispense with one of those conditions, she must find where the proof relies on it, what breaks down without it, and whether a different argument can fill the gap. A count of lines of code will tell her little about that difficulty. [8]

For Paul Humphreys, opacity depends on who examines a process and when. When Alpöge released the proof, he could identify where the model had outpaced him. Within days, other researchers could check the deductions, and a guide had begun to make the mathematics accessible to readers. Over that interval, different people were able to verify and explain the result to different degrees. [2] [3] [5] [6]

What a reviewer can sign

Return to the editor in the opening example. The Lean and Mathlib communities have published a formal certification. The editor still needs to know what the journal's own reviewers examined. The Committee on Publication Ethics (COPE) advises reviewers to disclose the limits of their expertise and to identify the parts they assessed. If no one on the team can reconstruct a step in time, the editor can invite a specialist to examine it or tell readers which step still needs clarification. [3] [9]

In *A lacuna abdutiva em sistemas de inteligência artificial*, my article on the abductive gap in AI systems, I asked whether others can "reproduce or verify the core result." In peer review, one specialist may establish that the statement matches the historical problem. Another may reproduce the formal check. A third may explain the difficult step. A reader assessing the reviews needs to know what each specialist established. [10]

Suppose the journal receives both the formal proof and a paper intended to teach it. A review might end this way.

> *"I checked that the statement matches the question posed and reproduced the formal verification with a second program. The paper makes clear how the conclusion follows from the central lemma. For the proof of the lemma, I had to consult the formal version. I recommend that the authors explain this step so readers can learn the method."*

What if the manuscript comes with only two favorable automated reviews? Without a proof others can check, data against which to test the conclusion, or an experiment another group can run, I would recommend treating the claim as open. A reviewer's difficulty is no evidence that the authors are wrong, either. Readers need to know both why the claim deserves consideration and which question remains unanswered.

A qualification can disappear between the review and the published abstract. The next paper may then treat as established a premise that the first author accepted only under specified conditions. By the time the tenth paper cites the ninth, anyone reexamining the conclusion will have to trace where the qualification disappeared. A journal can help readers by stating, alongside the main claim, what still needs to be checked.

Who pays for that work? A small journal may have the published files but lack either the computing resources or specialists with time to examine them. If universities share those costs, groups with smaller budgets will also have a chance to identify a flaw.

If only another AI can follow the argument

If reconstructing the argument takes longer than the review period allows, the team may have to rely on another AI system to assess it. Suppose specialists need ten years to reconstruct it. An editor has a few months to decide. A health care team may have to act this week. For them, the limit is real even if the argument becomes understandable years later.

How much of the argument did the second system understand? We can ask it to identify a necessary assumption, predict what changes if we remove it, and, where appropriate, give a counterexample we can check. Elliot Kim and colleagues examined agreement among 71 models on 14,042 HELM questions, each with four answer choices. For each pair of models, the researchers selected only the questions both had answered incorrectly. They then calculated the proportion of those questions on which the pair chose the same wrong answer. The average of these proportions across pairs was 60%. [4]

Someone who receives two identical answers needs to know something different. *Of all the times the models agreed, how often were both wrong?* To estimate that frequency on this task, we would divide the number of agreements on wrong answers by the total number of agreements, including agreements on correct answers. The reported 60% alone does not give us that estimate.

In section 4 of the study, the researchers used one model's answers as the reference for scoring others. The questions had an answer key, so the researchers could check those scores. The judge tended to overrate models that answered fewer questions correctly than it did. The effect was particularly pronounced in some pairs that shared a provider or architecture. When both chose the same wrong answer, the judge counted it as correct. [4]

Suppose the second system's reasoning is so hard to follow that researchers turn to a third. It might find a genuine error. But if every review comes back with the single word "confirmed," how could we tell whether any of the systems would have detected that error? A counterexample someone can calculate, or a prediction another team can test against a measurement, gives researchers evidence they can assess independently.

A debate between models can also help a human judge. In an experiment by Akbir Khan and colleagues, two models argued for opposing answers to reading comprehension questions. Human judges who had not read the full text answered 88% of the questions correctly after debate, compared with 60% in the no-debate baseline. The researchers had an answer key against which to score the judgments. For a paper announcing a discovery, establishing the equivalent of an answer key may be the hardest part of the research. [11]

What Perceptual Twins can test

I developed Perceptual Twins to study what enables an agent to revise a causal hypothesis. The protocol compares paired conditions. It distinguishes the information an agent receives about an action, the ability to choose the experiment, and the relationship between the action commanded and the action actually carried out. Two agents can interpret the same record consistently and, for that very reason, have the same difficulty detecting what it omits. [12]

In the reference implementation, two conditions were designed to receive the same information and perform the same calculation. As expected, the difference between them was zero on both measures. [12]

To study how one AI system reviews another's answer, I would compare four conditions for each case. In separate sessions, the reviewer would start with the same initial reading and the same answer from the first AI system. A second reading would come either from the same sensor or from an independent sensor. For each source, the reviewer would either receive the reading automatically or have the option to request it. When the reviewer had that option, I would record whether it requested the reading. I would introduce known faults and measure, in each condition, the proportion of the first system's errors the reviewer endorsed and the proportion of its correct answers the reviewer rejected. I would include cases in which the reviewer decided without requesting the reading. This experiment has yet to be run.

A second extension would follow what happens after the review. In a simulation, a faulty measurement would feed a sequence of decisions. In one condition, every reviewer would have only the original measurement. In the other, one reviewer would also receive an independent record. We could measure how far the error travels and how many decisions remain in effect after someone identifies it. Establishing how a real institution responds to a correction would still require fieldwork. [12]

In my research on Mathematical Universality, two systems that find the same pattern in known phenomena give me a hypothesis to investigate. Before looking at a third domain, I need to say what the pattern predicts, how much variation I will allow, and what finding would lead me to abandon it. Even if the initial search exceeds my ability to compute, I can still frame a prediction that an independent observation could overturn.

The August 12, 2026, Zenodo deposit preserves *Perceptual Twins Replication Kit* v1.0, including the reference implementation, expected outputs, and the manuscript from that stage (v2.1.1). The project page provides the revised preprint (v2.2.0), kit v1.0.2, and the Stage 0.5 materials (v0.2.0). The reported results come from internal tests on synthetic data. Independent replication remains open. [12] [13]

Materials for reading and replication

Download the original v1.0 kit from Zenodo

Read preprint v2.2.0 and access the current packages

From a single review to trust between institutions

At a university, the first loss may be felt in the training of future reviewers. An adviser who asks a student to reconstruct a lemma gives her a chance to discover why an assumption was necessary. Years later, that experience may help her spot a flaw in someone else's manuscript. If the assignment instead consists, semester after semester, of asking a system for confirmation, the university may keep publishing while training fewer people who can assess what gets published. One generation can teach a correct result and still fail to teach students how to recognize the limits of its application.

John Hardwig observed in 1985 that some of our knowledge depends on specialists whose reasoning we cannot personally reconstruct. I can cross a bridge without being able to calculate the loads it can bear. I have grounds for confidence if engineers know the design, inspectors can examine the structure, and someone can close the bridge if a crack appears. What matters is whether those people can be reached and can act when a problem arises. [14]

Outside academia, the distinction between checking a calculation and knowing what it measures has already mattered to patients. In 2019, Ziad Obermeyer and colleagues showed that, at a given score on a health care algorithm used in the United States, Black patients were on average sicker than White patients. The system predicted health care spending. Because access to care was unequal, spending less on a patient could make that patient appear healthier. A second AI system recalculating costs could have reproduced the score. The flaw became visible when researchers compared scores with patients' health. [15]

In *A soberania capturada* (*Captured Sovereignty*), I examined the British Post Office's Horizon case. Records from its accounting system were used to prosecute local post office operators who had few means to challenge the data. In 2021, the Court of Appeal of England and Wales quashed 39 convictions. A government that must turn to its own supplier to investigate a complaint faces a similar risk. When an official record is used against someone, where can that person take evidence that the record is wrong? [16] [17]

One person may not know why an application was denied. If the ombuds office consults the same database, the auditor repeats the same calculation, and the judge receives an expert report drawn from the same source, an entire country may lack the means to explain that class of cases. Everyone involved may have acted in good faith. The many signatures obscure the fact that only one measurement was taken. Who can take a second one?

When many papers depend on a single observation

In 2009, Steven Greenberg traced a claim about the role of a protein, beta-amyloid, in a muscle disease. He identified 242 papers and 675 relevant citations. Papers that weakened the claim drew little attention. Others multiplied the citations in its favor without adding observations of their own. In parts of the citation network, a conjecture came to circulate as fact. This case predates the widespread use of generative AI and shows how the authority a claim acquires can be traced. [18]

Imagine a literature review prepared with AI. The system finds ten papers supporting a claim. Nine lead, by different routes, to the tenth. Anyone who sees only the count might infer that ten independent observations support the finding. A review that identifies who made the original measurement, who reanalyzed the data, and who simply cited earlier work would show how much rests on one experiment.

The loss can begin before anyone demonstrates an error. Lisa Messeri and M. J. Crockett suggest that AI use in science may favor some questions, methods, and viewpoints until others go unexplored. A funding agency could investigate that possibility. Alongside publication counts, it could track which questions no longer appear in proposals, how many groups still collect their own data, and how many rely on the same dataset. [19]

A field can publish prolifically on the problems its tools already solve well. An international dataset may omit a local phenomenon, or a student may abandon a dissertation because it would require two years of fieldwork. Neither loss is easy to see. Grantmakers can support teams that investigate by other methods and publish negative results as well.

My work on Mathematical Universality faces a similar problem. If different fields find the same pattern, I need to know whether they observed independent phenomena or whether the same transformation of their data produced the pattern. A prediction gains support when a different method, chosen in advance, confirms it in new observations. If the pattern vanishes when I change how I measure, I have reason to investigate the instrument before attributing the pattern to the universe.

What a national average can hide

Suppose the public health system of a country divided into two regions uses a tool to identify people who need follow-up care. A field team examines people on the list and people the tool missed. Using an independent clinical criterion, it identifies 10,000 people who need care. Only then does the team compare its findings with the system's list. The figures in the table are hypothetical.

People with a need for follow-up care confirmed by independent assessment

Area	Confirmed need	Identified by system	Missed	Percentage identified
A	9,000	8,910	90	99%
B	1,000	600	400	60%
Total	**10,000**	**9,510**	**490**	**95.1%**

The last column divides the number identified by the total with a confirmed need (sensitivity). The table does not show how many people were included despite not needing follow-up care.

The system identified 95.1% of the people who needed follow-up care. Area B, home to 10% of those with a confirmed need, accounts for 400 of the 490 people missed, or about 82% of the total. The national average may reassure an official reading the report. For a nurse in Area B, the same report shows that four out of every ten people needing care were missed.

If next year's budget follows only recorded visits, Area B could lose funding precisely because its system missed more people. With fewer teams, even needs known today could disappear from the records. A second AI system could check every sum and find the spreadsheets consistent. Finding out who is missing requires looking beyond the list. An ombuds office that counts only complaints is unlikely to hear from people who never knew they were entitled to care.

Universal access, equality in the provision of care, and community participation are principles of Brazil's Unified Health System (SUS) under Article 7 of Law No. 8,080/1990. In this example, they give officials grounds to hear from both areas, audit a sample, and reconsider how funding is allocated. Those debating the budget need to know both how well the system performs and who bears the cost when it misses someone. I examined this question of budget priorities on pages 109–112 of my first book. [1] [20]

Who can correct a rule for everyone

Imagine that a woman in Area B shows she was overlooked, has her record corrected, and receives care. Another 399 people in the same area are still off the list. Some do not even know it exists. The agency must decide whether to wait for 399 challenges or investigate the rule that produced the exclusions. For people going without care, the difference between those choices can be measured in time spent waiting.

In Australia, the program known as Robodebt used averaged income data reported to the tax office to assess debts against benefit recipients between July 2015 and November 2019. Even a correctly calculated annual average can conceal that a worker had earnings in some two-week periods and none in others. In 2023, the Royal Commission recommended clear avenues for review and access to the rules and algorithms for examination by independent experts. It also recommended that the government consider creating a body to audit automated decisions or expanding the role of an existing one. This experience with administrative automation illustrates the work involved in revisiting a premise already applied to thousands of decisions. [21] [22]

In Brazil, Article 20 of the General Data Protection Law (LGPD) gives an individual the right to request review of a decision that affects that person's interests and is based solely on the automated processing of their personal data. The provision does not specify who must conduct the review. Under paragraph 1, the data controller must, on request, provide clear and adequate information about the criteria and procedures used, subject to protections for trade and industrial secrets. Under paragraph 2, if the controller withholds information on those grounds, the national data protection authority may audit the processing for possible discrimination. A court challenge to the rule used to allocate funds among regions requires identifying the right at issue and who has standing to bring the claim. [23]

When a case concerns diffuse or collective interests, Brazil's Public Civil Action Law gives the Ministério Público (Public Prosecution Service) and the Defensoria Pública (Public Defender's Office), among others, standing to sue, subject to the applicable legal requirements. The scale of the problem changes the evidence needed. A medical record may suffice to correct one person's entry. To show a pattern of exclusions, investigators may need the decision criteria, prior versions of the rule, findings from a sample, and access to the data under appropriate safeguards. The agency administering the policy should preserve those materials for examination. [24]

A remedy for collective harm must reach people who never filed a challenge. Once the flaw is known, who can prevent new exclusions, locate the people affected, and check whether they are receiving care? Updating the software is only part of the remedy. Someone left out by the previous version can still be missing from the list.

When everyone acts on the same forecast

Suppose a bank seeking to limit its losses follows a model's recommendation and cuts short-term credit to firms in a particular industry. Another bank uses a different vendor's system, calibrated on similar data. Each bank, looking at its own portfolio, sees good reason for caution. If many banks withdraw credit at once, some firms may lose the financing they need to meet their obligations. Acting on the forecast may then cause some of the very defaults it predicted.

The example is hypothetical. In 2024, the Financial Stability Board (FSB) identified market correlations and concentration among service providers as two ways AI could amplify systemic risk. The report also recognizes potential gains in efficiency and analytical capacity. Systems that assess each portfolio's risk well can still change the conditions under which that assessment holds when many banks act on it. [25]

A firm's risk of default before its credit is withdrawn differs from its risk afterward. To estimate whether the cut contributed to defaults, an investigator would need to establish the firms' prior financial condition and compare what happened to similar businesses exposed to different degrees of credit restriction. Counting how many banks predicted defaults correctly cannot, by itself, establish what caused the losses.

Each bank knows its own portfolio. A regulator also needs to know which providers the industry's backup plans depend on. A bank can truthfully report that it has an alternative. If its competitors rely on the same alternative, a single failure could disable several plans at once. Comparing information across institutions, while protecting customer data, can bring that dependency to light.

At one bank, an audit can find a miscalculation and correct it. Across the financial system, even if every bank calculates correctly, simultaneous credit cuts may harm firms that would otherwise remain solvent. Evaluating a forecast therefore requires examining what happens when many institutions act on it.

Technological dependence and the freedom to challenge a finding

A government can receive a recommendation to close health services in seconds and need years to acquire the means to test the forecast of low demand behind it. A contract buys access to the recommendation. Testing it may require the original data, the system version, electricity, computers, specialists who know the region, and the freedom to publish what they find. If the records arrive after the budget vote, the right to examine them will have done little for the communities affected.

According to the World Bank, high-income countries accounted for 77% of global colocation data center capacity in June 2025. These facilities house equipment for multiple customers. Low-income countries accounted for less than 0.1%. Some middle-income economies, including China, also have substantial infrastructure. The figures measure installed power capacity in megawatts. Who can actually commission an audit also depends on the price and terms of access. [26]

An October 2023 survey of six major cloud providers identified service regions in 39 countries and territories. In 30 of them, customers could obtain local access to at least one of the types of graphics processing unit (GPU) surveyed. The study did not measure government or university facilities or remote access to machines abroad. This uneven distribution motivates my hypothesis that most countries would depend on third parties for audits requiring advanced computing and access to restricted data. To test that hypothesis, we would need to specify an audit and determine, country by country, whether a team independent of the supplier could obtain the data, afford the computing costs, and finish the examination before the decision was due. [27]

A team can check a formal proof without retraining the model that helped discover it. A low-cost local measurement might even disprove a costly forecast. Dependence is more serious when the supplier alone holds the records, controls the computers, and decides who may use either. Even a wealthy country could be unable to investigate a failure on those terms. Where budgets are smaller, an investigation may be technically feasible and still come too late to prevent harm. [3] [7]

Countries that pool computing resources and expertise can preserve records outside the supplier’s exclusive custody and give specialists the right to publish their findings while protecting personal data. I would examine how long an audit takes, how much of the oversight budget it costs, and who authorizes access. If the supplier goes out of business or becomes a party to the dispute, can the oversight body obtain the records and repeat the test before the decision causes harm?

What changes when countries depend on the same systems

When several governments use the same system, a business decision made abroad can interrupt research in several countries at once. A university may be able to pay to query the system but unable to afford a new license to repeat a test. If the foreign team drops the investigation, what resources does the country directly affected have to continue it?

In its 2025 Technology and Innovation Report, UN Trade and Development (UNCTAD) found that, as of the end of 2024, 118 countries, most of them in the Global South, had no representation in any of the seven international AI governance initiatives it studied. That figure measures participation in those forums. Assessing a country's ability to conduct an audit would require information about its personnel, infrastructure, access to data, and freedom to publish adverse findings. Political dependence may deepen when countries that import a tool have little say in setting the rules for evaluating it. [28]

Whoever sets the cost of challenging a finding can influence a public debate without silencing anyone. Suppose an expert opinion in favor of granting a private operator a public service contract is inexpensive to obtain, but the data behind it are costly to access. Companies, universities, and residents may have the same right to speak at a hearing and vastly different means of substantiating what they say. The contract can specify how a qualified third party may obtain the records and who will pay for the examination.

Brazil's National Institute for Space Research (INPE) runs two programs with distinct purposes. In the Amazon, DETER alerts guide deforestation enforcement, while PRODES produces the official annual estimate. The documentation for DETER-B explains that cloud cover can hide land that has already been cleared. INPE therefore advises against directly comparing alerts across months. Land may have been cleared well before anyone detected it. [29] [30]

Imagine a public hearing at which two models use the same alerts to conclude that more forest was cleared that month. A researcher might bring cloud-cover records and ask whether an area cleared weeks earlier has only now become visible. The hearing is hypothetical. The limitation of the instrument is documented, and anyone at the hearing can use that evidence to challenge the models' interpretation. [29]

By sovereignty here I mean the ability to cooperate while retaining the means to disagree with one's partners. An international collaboration faces a test of its independence when the team with fewer resources finds evidence against its more powerful partner's conclusion. Will it have access to the records it needs to check the finding? Can it publish even if its partner objects?

What one generation needs to pass on to the next

A researcher retires. Months later, a program stops working, and the team discovers that nobody knows why a particular assumption went into its design. The file is still readable, but no one can justify the choice or recognize when it needs to change. With AI, that loss may go unnoticed for longer. A laboratory could keep producing excellent results for years without training anyone else to revisit the choice.

There is a more encouraging possibility. Science already preserves and extends knowledge beyond any one person’s expertise. One group can study a section of a long proof, another can check the deductions with an independent program, and a third can teach the method to students. Releasing the formal proof of the Riemann hypothesis made that division of labor possible. A team with the files and sufficient computing resources can examine the same argument without access to the model that helped find it. [3] [6] [7] [14]

Waiting for a complete explanation before allowing any use would carry a human cost. In the hypothetical health system, if the tool finds people previously overlooked, the team can refer them for assessment, check each referral, and keep other routes to care open. Closing a service across an entire area on the strength of a forecast of low demand has different consequences. The closure may eliminate the very place where unmet need would become visible. The urgency of care, the reversibility of the decision, and the means available to detect an error should all weigh in that judgment.

Even if a model predicts exactly how many people each option will reach, someone must still weigh potentially competing goals, such as expanding average coverage, treating those at greatest risk first, or maintaining a service in a historically underserved area. The people affected may have reasons to favor one priority over another and need a public forum in which to offer them. If officials attribute the entire choice to an algorithm, to whom can a citizen make the case for a different priority?

Has the creation surpassed its creator? In developing the proof, Alpöge describes a point at which the model outpaced him. Across generations, the question takes a different form. We may leave students a far more powerful tool than any we had and still leave them with less freedom to choose what to investigate and when to reopen a question. Preserving that freedom is part of what it means to teach. [2]

How to track the gap between using and understanding a result

For each AI-assisted discovery, I propose recording the dates of the announcement, the first external verification, and the first independent explanation of a crucial step. A team might demonstrate its understanding by adapting that step to a different problem or showing what breaks down when an assumption is removed. If the second or third date is still unknown, how many papers or decisions already depend on the finding?

I would follow discoveries in a single field using selection criteria set in advance. For each paper, I would record what an outside group could access and which data, instruments, or methods it shared with the authors. A second team working from the same database can detect a calculation error. To find people missing from the database, it may need fresh observations. Calling a check independent is informative only if we specify the kind of error it could reveal.

I would also record who adopted the result and where an original qualification dropped out. Citing a theorem, using its method in another proof, and applying a rule to determine who receives a service have very different consequences. If someone identifies a flaw, a fourth date would show when each service began applying the correction.

In the health care example, I would follow the people identified as having been left out. At each follow-up, I would divide the number of people in that group who had actually received the care they needed by the total number identified as of that date. I would publish both counts and report separately how many of those identified the team was still trying to reach. To estimate how many people the original list missed, I would compare it with independent clinical assessments of a representative sample of residents, including people absent from the list, and report the uncertainty in the estimate.

What evidence would weaken my concern? Even for harder problems, the interval between announcement, verification, and explanation might shrink. Teams with modest resources might reproduce more results, and corrections might arrive in time to prevent lasting harm. Attributing a change to AI would require comparing problems of similar difficulty investigated over the same period, including by other methods. If an explanation has yet to be found, the record should say so. Claiming that no person could ever supply one would require evidence of its own.

I would ask journals to identify the materials their reviewers examined and funders to pay those who reproduce a result and try to explain it. An institution relying on that result to make decisions about people should name who has the authority to suspend its use, investigate, and remedy harm. A publisher can correct a paper. The service that denied someone care remains responsible for reaching that person, restoring access, and checking whether they received the care they needed, even if other institutions must help.

Conclusion

The Riemann case answers part of the question that opened this essay. When Alpöge announced the discovery, he was still working toward a deeper understanding of some of the mathematics involved. Other researchers examined the statement and checked the proof with independent proof checkers. An argument that had outpaced one researcher at a particular moment could still be examined by others. How often will another AI system be the only one able to follow an argument in time? That question remains open. [2] [3] [6] [7]

Reviewers answer for what their signatures attest to. A government must also answer for what happens to people after a review. In the example of the two areas, the woman who challenged her omission received care. Reaching the other 399 people the rule missed in her area requires someone to preserve the records, find those absent from them, and reconsider the decisions. A long time can pass between the day a paper is corrected and the day a person regains access to care.

Two countries can buy the same assessment and have very different degrees of freedom to challenge it. The concentration of infrastructure is documented. How many countries would depend on third parties for a particular audit is an empirical question that requires country-by-country assessment. If the cost or the contract terms put an independent examination out of reach, the supplier gains influence over what a country can recognize as an error. In chapter 15 of my first book, I asked how a government comes to see the needs that shape its budget. I now ask who controls the means by which a country’s people can show that the assessment was wrong. [1] [26] [27]

Universities can train people to explain a discovery and try to refute it. Journals can tell readers what their reviewers actually examined. When a government relies on a finding to make decisions that affect people’s lives, I believe it should give the oversight body the resources and access to records it needs to repeat the examination. The oversight body must have the independence to challenge the supplier and publish its own findings. The authority responsible for the decision must assess whether to suspend use of the system. It must also ensure redress for those harmed.

I return to the reviewer’s signature. Reviewers should be able to say what they verified and where their understanding ends. Society will have a harder question to answer. Suppose the two systems still agree tomorrow and a community brings evidence that both are wrong. Who will have the means to examine that evidence, the independence to side with the community if the facts warrant it, and the authority to change what is being done in its name? My greatest fear is that, on that day, the institutions responsible will still exist but will have lost the means to answer the people who turn to them.

Author’s electronic signature

References

Sources consulted on November 22, 2026.

[1] GENOVA, Jandislei Antonio. *Autonomia decisória na era da IA*: formação da vontade, influência cognitiva estrutural e responsabilidade jurídica em ambientes algorítmicos. São Paulo: Editora Dialética, 2026, chap. 15. Quotation on p. 108 and budget discussion on pp. 109–112 of the publisher's proof. ISBN 978-65-288-0467-2. Publisher's catalog

[2] ALPÖGE, Levent. *[Statement on the proof of the Riemann hypothesis]*. Personal statement, November 20, 2026, pp. 1–2. Descriptive title. Originally in English. Alpöge's original statement

[3] LEAN AND MATHLIB COMMUNITIES. *Lean verification of the Riemann Hypothesis*. Joint statement, November 21, 2026. Joint statement

[4] KIM, Elliot Myunghoon *et al.* Correlated Errors in Large Language Models. *Proceedings of the 42nd International Conference on Machine Learning*, PMLR, vol. 267, pp. 30038–30066, 2025, secs. 3.2 and 4. ICML proceedings · Full text

[5] HUMPHREYS, Paul. The philosophical novelty of computer simulation methods. *Synthese*, vol. 169, pp. 615–626, 2009, especially p. 618. DOI 10.1007/s11229-008-9435-2 · Springer

[6] *The Alpöge–Claude proof of the Riemann hypothesis explained for the Riemann zeta function*. The title page credits Claude Fable 5.2, Codex Astra 6.2, and Gemini Carbon, working at Michael O. Rubinstein's request. Working draft, November 21, 2026, 164 pp. Guide, credits, and abstract

[7] SUTHERLAND, Andrew V. *Verification files for the Lean proof of the Riemann Hypothesis*. Massachusetts Institute of Technology, November 2026. Public files, procedures, and verification logs. MIT verification files

[8] THURSTON, William P. On proof and progress in mathematics. *Bulletin (New Series) of the American Mathematical Society*, vol. 30, no. 2, pp. 161–177, 1994. Full text · DOI 10.1090/S0273-0979-1994-00502-6

[9] COPE COUNCIL. *COPE Ethical guidelines for peer reviewers*. Version 2, September 2017, especially pp. 5–7. COPE guidelines · DOI 10.24318/cope.2019.1.9

[10] GENOVA, Jandislei Antonio. *A lacuna abdutiva em sistemas de inteligência artificial*: geração representacional, restrição empírico-formal e adjudicação epistêmica. Version 3.0, August 24, 2026, sec. 8, p. 15. In Portuguese. Article and PDF

[11] KHAN, Akbir *et al.* Debating with More Persuasive LLMs Leads to More Truthful Answers. *Proceedings of the 41st International Conference on Machine Learning*, PMLR, vol. 235, pp. 23662–23733, 2024. ICML proceedings

[12] GENOVA, Jandislei Antonio. *Toward a Perceptual Twins Protocol*: A Yoked Causal Design for Task-Relative Epistemic Autonomy. Preprint v2.2.0, August 12, 2026, pp. 2 and 23–24. Replication Kit v1.0.2 and Stage 0.5 v0.2.0. Preprint and replication materials

[13] GENOVA, Jandislei Antonio. *Perceptual Twins Replication Kit v1.0*. Software and replication materials, version 1.0, Zenodo, August 12, 2026. Deposit accompanying the earlier v2.1.1 preprint. DOI 10.5281/zenodo.21908918 · GitHub release v1.0

[14] HARDWIG, John. Epistemic dependence. *The Journal of Philosophy*, vol. 82, no. 7, pp. 335–349, 1985. DOI 10.2307/2026523 · JSTOR

[15] OBERMEYER, Ziad *et al.* Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, vol. 366, no. 6464, pp. 447–453, 2019. DOI 10.1126/science.aax2342 · PubMed

[16] GENOVA, Jandislei Antonio. *A soberania capturada*: Big Techs, soberania cognitiva e poder privado na ordem digital global. São Paulo: GENOVA imprint, 2026, v. 8.20, pp. 97–98. DOI 10.5281/zenodo.23272360. Portuguese edition

[17] COURT OF APPEAL OF ENGLAND AND WALES, CRIMINAL DIVISION. *Hamilton and Others v Post Office Limited* [2021] EWCA Crim 577. Judgment of April 23, 2021. Judgment and official summary

[18] GREENBERG, Steven A. How citation distortions create unfounded authority: analysis of a citation network. *BMJ*, vol. 339, b2680, 2009. Study and figures · DOI 10.1136/bmj.b2680

[19] MESSERI, Lisa; CROCKETT, M. J. Artificial intelligence and illusions of understanding in scientific research. *Nature*, vol. 627, pp. 49–58, 2024. Record and abstract · DOI 10.1038/s41586-024-07146-0

[20] BRASIL. *Lei nº 8.080, de 19 de setembro de 1990*. Brazilian Organic Health Law, Article 7(I), (IV), and (VIII). Chamber of Deputies, consolidated text. Official statutory text

[21] SERVICES AUSTRALIA. *Information about Robodebt*. Australian Government, updated September 28, 2023. Debts raised using averaged income information. Official information

[22] ROYAL COMMISSION INTO THE ROBODEBT SCHEME. *Report of the Royal Commission into the Robodebt Scheme*. Corrected edition, July 11, 2023, chapter 17, recommendations 17.1 and 17.2. Report and recommendations

[23] BRASIL. *Lei nº 13.709, de 14 de agosto de 2018*. Brazilian General Data Protection Law, Article 20, opening provision and paragraphs 1–2. Office of the President, consolidated text. Official statutory text

[24] BRASIL. *Lei nº 7.347, de 24 de julho de 1985*. Brazilian Public Civil Action Law, Articles 1(IV) and 5(I)–(II). Chamber of Deputies, consolidated text. Official statutory text

[25] FINANCIAL STABILITY BOARD (FSB). *The Financial Stability Implications of Artificial Intelligence*. Report, November 14, 2024. Provider concentration and market correlations. FSB report

[26] WORLD BANK. *Digital Progress and Trends Report 2025*: Strengthening AI Foundations. Washington, DC: World Bank, 2025, chap. 3. Estimates in megawatts, June 2025. World Bank report

[27] LEHDONVIRTA, Vili; WÚ, Bóxī; HAWKINS, Zoe. Compute North vs. Compute South: The Uneven Possibilities of Compute-based AI Governance Around the Globe. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, no. 1, pp. 828–838, 2024, table 4. DOI 10.1609/aies.v7i1.31683 · PDF

[28] UN TRADE AND DEVELOPMENT (UNCTAD). *Technology and Innovation Report 2025*: Inclusive artificial intelligence for development. Overview, 2025, figure 12. Participation in seven international initiatives through 2024. UNCTAD overview

[29] INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE). *DETER*. DETER-B methodology. Data, coverage limitations, and interpretation of alerts. DETER documentation

[30] INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE). *Como se monitora o desmatamento da Amazônia?* Frequently asked questions. Roles of PRODES and DETER in satellite monitoring. INPE documentation

Disclosure. The author used OpenAI's ChatGPT for source research, drafting, translation, and bilingual editing. Quotations from the author's Portuguese works are translated for this edition. The sample peer review, health care table, credit scenario, and environmental hearing are hypothetical. The proposed extensions of Perceptual Twins and measures for tracking institutional responses have yet to be tested.

JANDISLEI ANTONIO GENOVA

Assinado de forma digital por JANDISLEI ANTONIO GENOVA
Dados: 2026.11.22 15:48:57 -03'00'