Are we intelligent enough to judge intelligence? Proposal for a New PEG Task Force to Develop AI Tests

Magdalena Maglizhanova[1,*,†], Christian Jendreiko[2,†], Jacinto A. Dávila Quintero[3,†], Veronica Dahl[4,†] and Laura A. Cecchi[5,†]

[1]*University of Plovdiv "Paisii Hilendarski", Plovdiv, Bulgaria*
[2]*HSD, University of Applied Sciences, Duesseldorf, Germany*
[3]*Universidad de Los Andes, CESIMO, Mérida, Venezuela*
[4]*Simon Fraser University, Burnaby, BC, Canada*
[5]*Grupo de Investigación en Lenguajes e IA,Universidad Nacional del Comahue, Neuquén, Argentina*

Abstract

We argue that Turing tests as means for defining what falls within AI and what does not are insufficient and in addition, pernicious, as they tend to identify the whole discipline with an impersonating, humanity-fooling function: that of passing as human or as credible at all costs.

Keywords

Artificial Intelligence, Turing Test, Generative Artificial Intelligence, Symbolic Artificial Intelligence

1. Introduction

Our original intention for this paper was to present new tests for AIs that could help elucidate two questions:

1 Does an AI serve and respond to humanity at large? How supportive is it of human communities and of life on the planet?

2 Are we intelligent enough to judge the intelligence of the AI and to turn it into a safe collaborator?

The reason to develop these new tests were the conclusions we drew from considering the current state of- in particular, modern- AI design, as well as its applications, uses and societal impact worldwide [1]. However, as we began developing these ideas, we found ourselves no longer simply writing a paper. Instead, we found ourselves in the midst of what was effectively a research on the need for humanity-serving, widespread AI literacy: exploring how Turing test-inspired, stochastic AIs may have impacted human, animal and planetary rights, and how tests relevant to today's AI landscape might be conceived and developed in order to best serve humanity. Through this process, it became clear to us that such a project could serve as a powerful educational tool, offering significant learning value for both professionals and students across diverse fields and backgrounds and for decision-makers across diverse public and private institutions, particularly in understanding AI and its broader implications, e.g. for conscious societal planning. This is crucial because, whereas many governments have established pertinent guidelines to prevent AI's current design, control and deployment status from creating societally undesirable outcomes (such as the AI Act of the European Community), these principles don't always have legally enforceable status, particularly across countries. The pressure of a globally

IV Prolog Education Workshop 2026 (PEG 2026)
*Corresponding author.
†These authors contributed equally.
magdalenamaglijanova@uni-plovdiv.bg (M. Maglizhanova); christian.jendreiko@hs-duesseldorf.de (C. Jendreiko); jacinto@ula.ve (J. A. Dávila Quintero); veronica_dahl@sfu.ca (V. Dahl); lcecchi@fi.uncoma.edu.ar (L. A. Cecchi)
0009-0000-4978-9282 (M. Maglizhanova); 0009-0006-4302-6261 (C. Jendreiko); 0000-0003-1053-8763 (J. A. Dávila Quintero); 0000-0002-1159-1374 (V. Dahl); 0000-0001-5236-6715 (L. A. Cecchi)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-03

CEUR-WS.org/Vol-4311/PEG_regular_11.pdf

AI-literate public may be what it takes to achieve enforceability- and hence, safety. Consequently, we propose the approval of a new task force for PEG dedicated to the development and dissemination of AI tests. While many of the ideas we develop apply to generative AI in general, we address LLMs in particular because they are the most widely used.

2. Reaches and harms of Turing-test inspired AIs

The multiple ways in which generative AI obtains results that would be deemed intelligently obtained if obtained by a human, and even admirably *surpasses* human cognitive performance, are undeniable. For instance, AI is providing powerful new ways to address long-standing problems in different domains [2, 3]. What follows does not mean to deny or minimize this well-known fact. What it discusses are the fundamental limitations inherent in generative AI's very design and how inevitably these result as well in posing (also powerful) **threats** to either human or planetary well-being. From this analysis we motivate the need to develop alternative tests for AIs and for the human beings who design, deploy or use it, aiming at making the use of AI inherently safe.

2.1. The Turing Test

Turing's test [4] proposes to label a computer system's proficiency as "intelligent" if within a controlled experiment (e.g. no immediate physical presence) its output would be judged as "intelligent" by a human having been told they were communicating with a human. This test has had three main historical (consciously sought by Turing or not) consequences:

- to define empirically what falls into "AI": those computer systems capable of fooling people through impersonation manoeuvres
- to facilitate acceptance of deceptive uses of AI
- to facilitate a popular confusion between close-to-human success for isolated tasks and close-to-human (or even superior-to-human) intelligence, regardless of total failure on all the other tasks that humans perform intelligently with no trouble.

Given such starting points, what could go wrong?

2.2. Faulty by design: generative AIs imitate an understanding of the world that they don't actually have

For all its impressive successes, large language models, LLMs [5], the more promoted form of generative AI, has a major flaw that it cannot recover from, since it owes to its very design: it has no access to either the real world, or to any systematic representation of it from which consistent, reliable, explainable, trustworthy inferences could be drawn.

A large language model establishes correlations rather than making human judgments, forecasts instead of evaluating, and interacts with a statistical distribution of language rather than engaging directly with the physical world. While the underlying architecture of these models makes them highly proficient at replicating recurring patterns and identifying popular trends within vast amounts of text, they lack any real-world grounding for the concepts they process. In essence, these systems project an illusion of comprehension. Although they are highly skilled at composing, condensing, mixing, and exploring the formal structure of ideas, relying on them for evaluation transforms human decision-making from an interaction between a conscious mind and reality into a simple calculation based on probabilities.

This mechanism differs fundamentally from traditional human reasoning techniques. Even though formal reasoning becomes merely a string of data once digitized (given that true meaning requires a conscious, embodied being to interpret those symbols in the context of lived experience), traditional

methods maintain transparent, verifiable inference steps that can be strictly reviewed, rather than relying on statistical approximations [1].

Instead of expecting generative AIs to provide human-like judgement, therefore, we should clearly recognize that LLMs only involves information processing and model optimization, whereas human intelligence involves humans and relationships, the real world, experience, meaning, values, and subjectivity, and we should treat large language models in particular as what they are: **sophisticated linguistic instruments that require human oversight precisely because they can access neither the domain that human judgment ultimately depends on: the world itself, nor any logical representation of it on which explainable, trustworthy rules of inference could be applied.**

But are we prepared to exercise that human oversight judiciously? Or do we lack the individual and/or the collective conditions, including intelligence, that it would take to defend from biased or pernicious responses, from its LLM's malfunctionings animistically misnamed "hallucinations", from manipulations and intrusions in our private lives, creations and information?

2.3. Faulty by oversight: can human intelligence be relied upon?

A crucial consideration in this discussion is that intelligence is a quality that humans posses in varying degrees and that machines have just started to approximate. We all have lacks thereof, especially when working in complex domains and isolated from others. It can be argued that in consequence, even safeguards that can be helpful in some situations, such as human-in-the-loop supervision, may fail to redirect the negative effects of any AI or of its uses enough to prevent sometimes disastrous possible outcomes. In order to guide our search for conditions under which AI deployment and use can be made safe, let us next consider possible ways of testing intelligence.

3. What tests are needed? Preliminary considerations

Our tests should allow us to first assess **intelligence and preparedness as well as AI literacy**, in order to then enhance these enough to guarantee safe and community-supporting deployment and uses of AI and even of social media, where much of disinformation and manipulation originates [6]. This is a complex, interdisciplinary task that involves moreover three realms: individual, collective and institutional.

Since all three realms need to rely on appropriate AI literacy, let us first mention that tests for AI literacy are of paramount importance. For instance, [7, page 11] have found: *"Our results provide evidence that overconfidence in AI was more likely to be observed in individuals with low AI knowledge compared to those with higher AI knowledge based on data collected from managers samples."* .

We next motivate the need for more specific tests that can help us determine what aspects of each type of intelligence must evolve in order to help turn AI into a universally safe collaborator.

3.1. What aspects of individual intelligence do users need to develop in order to be able to use generative AI as a safe collaborator?

We can develop tests to evaluate users' cognitive vulnerabilities by examining how generative AI systems fail and threaten their users. These risks are discussed below.

3.1.1. The risks of animism

People frequently attribute human qualities—such as intelligence and empathy—to AI systems. This "animism" occurs because AI models condense vast amounts of pre-existing, human knowledge with algorithms and processing speeds that easily pass situations similar to the Turing test. Furthermore, AI companies actively encourage these misconceptions for marketing purposes. This trend is dangerous

[1] https://www.scientificamerican.com/article/ai-and-human-intelligence-are-drastically-different-heres-how/

because large language models can accumulate significant error rates at every step, while lacking any internal mechanism to verify or validate their outputs.

3.1.2. Insufficient Awareness of Structural asymmetries

As already discussed, large language models (LLMs) can match human responses but lack genuine reasoning or semantic understanding. They operate on statistical token distributions rather than experiential knowledge. This fundamental architecture limitation—comparable to Searle's Chinese Room argument[8]—means they manipulate symbols and stochastically predict text rather than comprehending it.

3.1.3. Transformation of Human Cognitive Activity

As AI rapidly enters our lives, we can look to history to adapt. For example, during the 1960s space race, NASA mathematician Katherine Johnson[9] and her colleagues—featured in Hidden Figures[10]—faced the threat of automation when high-speed IBM computers were introduced. Despite facing systemic segregation, they adapted by learning new skills, such as FORTRAN programming. Like the 1960s "human computers," modern society must embrace adaptability. However, this transition should be guided by deliberate societal planning rather than dictated by the commercial interests of AI companies.

AI can be seen as the first technology to substitute for human thought and creativity, prompting a critical question: does it diminish our intellect? In fields like software development, delegating cognitive tasks to AI replaces demanding effort— such as source verification and engaging with uncertainty— with immediate efficiency. However, this convenience risks weakening critical thinking and problem-solving skills.

The Dunning-Kruger effect[11, 12, 7] plays a key role here: less competent users may suffer an illusion of expertise fuelled by AI's fluency, whereas competent users maintain a critical stance. AI may not replace human intelligence itself, but the effort of thinking, exposing potential tendencies toward cognitive passivity. It acts as a test of discipline; it amplifies the results of those with strong foundational knowledge but offers limited, or even risky, benefit to those without it.

Consequently, human intelligence is being relocated—shifting from knowledge acquisition to the ability to critically evaluate information and use AI as a supportive tool rather than a replacement. The ultimate challenge is not AI's capability, but our collective intellectual maturity. We must organize society to ensure AI assists rather than replaces workers, prioritizing human rights over corporate profit. Success requires interdisciplinary, collective-minded planning to ensure these tools serve humanity safely.

3.2. What aspects of collective intelligence and knowledge need to change to ensure that AI is deployed in ways that support, rather than hinder, humanity's evolution into mutually supportive, egalitarian and peaceful societies?

To ensure AI supports humanity's evolution into egalitarian, peaceful societies, we must dismantle the systemic supremacies that intentionally limit collective intelligence. True collective intelligence requires the meaningful inclusion of diverse perspectives in decision-making, which in turn demands a fully educated public and leadership—goals often neglected by current power structures.

To steer AI toward the common good, we should develop dual assessments: tests that measure how well AI supports human communities, and parallel tests that evaluate our own collective readiness to use AI for humanity-beneficial planning. Without this interdisciplinary, collaborative effort, we risk a future defined by unplanned social and ecological disasters driven by corporate default.

3.3. What aspects of institutional intelligence need to change for generative AI to support our collective thriving?

By "institutional intelligence" we mean the collective intelligence not of users themselves but of developers, deployers and institutions. In order to consider this question, let us first consider where and why generative AIs are failing us and even sometimes threatening us.

3.3.1. Does the problem lie with the technology itself?

The widespread common-sense notion that technology itself is neutral does not hold across the board. It does hold for instance for kitchen knives, but technology can also be harmful in itself; for example if something is not well designed it can explode. The question here is therefore, whether AIs are well enough designed to NOT provoke the equivalent of an explosion. And the answer is clearly NO, given the following systemic failures entailed by design features of the more wide-spread form of AI technology:

3.3.2. Systemic failure

LLMs are costly to develop and maintain but, more importantly, costs are blurry and tend to hide the costs to the environment. The environmental sustainability of the alternative methods must be investigated. One of the skills logic programmers have and teach is the representation of knowledge in computer forms that can be explored and consulted to solve problems. It is the old fashioned way in AI, which is currently being overshadowed by LLMs, in which the equivalent of those knowledge bases, a large language model, is automatically learned after very energy-intensive computations. Although they may not be comparable in processing speed, human coders could be more environmentally friendly than massive, artificial neural networks. We may need to rescue human, logical reasoning for that reason too. The trade-off between the brute-force statistical learning of modern connectionist models and the precision-guided efficiency of symbolic logic could define a "green" AI. According to emerging research, while LLMs offer unmatched speed and scale, their "metabolic" cost is orders of magnitude higher than human logical reasoning for specific tasks [13].

The Energy Gap: The human brain operates on approximately 20 Watts —roughly the power of a small light bulb- to perform complex logical reasoning, linguistic processing, and creative problem-solving (.ibid). On the other hand, a single training run for a model like GPT-4 can consume upwards of 50 GWh (the energy used by thousands of homes in a year) and, while a single query might only use 0.3 to 3 Wh, a human coder writing a logic block manually uses their "20W processor" for a few minutes. Studies have shown that LLM-assisted code generation can have a carbon footprint over 30 times higher than manual coding for the same task [14].

It seems that symbolic approaches have that general advantage. Preliminary reports suggest that by integrating symbolic representations, hybrid neurosymbolic systems outperform visual models alone on structured variants of the Towers of Hanoi manipulation task in simulation while measuring both task performance and energy consumption during training and execution. "During training, VLA fine-tuning [models] consumes nearly two orders of magnitude more energy than the neuro-symbolic approach" [15].

The Cost Gap: The argument about energy does not normally coincide with cost analysis because the economic value of energy rarely includes impacts on the environment and sustainability. It is not very intelligent.

But even without that caveat, the whole comparison is complex enough to confuse some people who should know better. Recently, Sam Altman, founder of OpenAI, the company behind ChatGPT, said that "One of the things that is always unfair in this comparison is people talk about how much energy it takes to train an AI model relative to how much it costs a human to do one inference query." He goes on

to suggest that it also takes a lot of energy to train a human and that this would be a fairer comparison [16].

Altman's "point" is that a human takes about 20 years to become "useful" (productive). An average human consumes about 2,000 to 2,500 kcal per day. Over 20 years, that is roughly 18 million kcal, which directly translates to about 21,000 kWh of energy. Training a frontier model like GPT-4 is estimated to have consumed between 50 million and 100 million kWh. Thus, on "training energy" basis, one LLM training run uses as much energy as "raising" roughly 2,500 to 5,000 humans to adulthood.

Altman also takes an additional rhetoric turn by claiming we should include the "energy cost of evolution": the billions of years and billions of ancestors it took for a human brain to reach its current state of "pre-installed" logic. This is, of course, an invalid argument. LLMs did not evolve in a vacuum; LLMs are built by humans using the infrastructure, science, and data that those 100 billion humans already produced. Generative AI is "standing on the shoulders" of human evolutionary energy.

However, all that discourse does not even start the argument on cost comparisons. Human brains may be cheaper in energy consumption, but they cost in a different way: living expenses. Companies are aiming at replacing humans because discounting their salaries after paying for access to LLM services can be more cost/effective, as they say. From a CEO's perspective, an LLM is roughly 20 to 50 times cheaper than a junior human coder for generating bulk text or boilerplate code. This one reason why companies are racing toward replacement; the "living expenses" of a human (rent, food, healthcare) are an overhead that digital systems simply do not have.

If we replace the human with the tool, we save the "living expenses," but we also remove the consumer and the taxpayer from the cycle. Economically, replacing humans with LLMs is a race to the bottom: if no one has a salary because LLMs are cheaper, there is no one left to buy the products the LLMs are helping to create. There is already some "suggestive evidence that hiring of younger workers has slowed in exposed occupations" [17]. It is hard to predict how the economies of the world will react to those AI trends. But it must be obvious that systemic failure is an important possibility. And we have not even started to consider human-induced climate change [18]. This is not very intelligent.

3.3.3. Misleading presentations

The structural risk that LLMs in particular introduce into epistemic processes does not stem solely from the fact that they produce errors—any informational tool can do so—but from the particular way in which those errors are presented. The "hallucinations" characteristic of these systems do not appear as obvious inconsistencies or recognisable gaps; rather, they are expressed with the same fluency, syntactic coherence, and apparent authority as correct outputs. This creates a critical cognitive asymmetry: while the system does not internally distinguish between a true claim and a merely plausible one, the user is likewise given no surface-level cues in the text to make that distinction. Formal syntactic coherence thus functions as an illegitimate substitute for truth, inducing a level of trust in the reader that the content does not always warrant.

It must be admitted, however, that new developments combining Generative AI with a form of Logical AI, some times also called neuro-symbolic systems, are being researched and tested intensively[19]. How important and indispensable the human capacity for logical thinking truly is, is demonstrated by the significant increase in research aimed at enabling large language models (LLMs) to perform rigorous logical reasoning, as well as by the growing body of papers dedicated to this subject [20, 21, 22, 23, 24, 25].

These are the latest developments of a trend to improve the reasoning capabilities of LLMs that started with "internal" solutions such as Chain-of-Though [26, 27, 28], C-o-T, which prompted for a series of intermediate reasoning steps to improve the ability of the LLM to perform complex reasoning. C-o-t, however, does not prevent the aforementioned hallucination errors and even fails to justify that the referred reasoning is what the LLM is actually performing. It can also be misleading.

3.3.4. Weakened Epistemic Trust and weakened capacity to make sound decisions

This phenomenon has consequences that extend beyond individual error. At a systemic level, the widespread adoption of LLMs as sources of information may lead to an erosion of epistemic trust in at least two directions. On the one hand, users who detect unflagged errors tend to generalise their distrust, thereby undermining the credibility of correct outputs. On the other hand, those who fail to detect such errors may incorporate false information into their cognitive frameworks and, eventually, into their decisions. In high-stakes contexts—clinical diagnosis, legal advice, journalistic production, scientific research—the cost of this second scenario can be considerable. Additionally, a second-order effect emerges: repeated delegation to automated systems may progressively weaken both the systems' results, which have been shown to degenerate with time and subsequent interactions, and users' critical capacities, reducing both their willingness and their ability to evaluate information independently [29].

3.3.5. Bias Amplification: the human-AI feedback loop

Due to their exclusive and quasi-universal access to unrewarded, hence cost-less human productivity, LLMs would seem to be perfect to support a claim that they are a form of collective intelligence enhancer. However, Research from UCL and MIT [29] demonstrates that using generative AI consistently increases human bias across all tested domains. This occurs through a recursive human-AI feedback loop:

- *AI Absorption:* Models train on human data containing subtle biases or labelling errors.
- *Algorithmic Amplification:* to improve prediction accuracy, generative AI processes sharpen these patterns, turning slight preferences into dominant outputs
- *Human Internalisation :* Because generative AI models are perceived as objective and authoritative, users trust them more than other humans and inadvertently align their judgment with that AI's biased output.
- *Data Pollution:* These newly biased human judgments then become training data for future models.

This cycle creates a "snowball effect" where each iteration further distorts human perception and social judgment, effectively using the Turing-test's perceived authority to solidify systemic errors.

4. How to develop these tests?

4.1. Logical Thinking is key

As AI systems grow in complexity, precise testing methodologies are urgent. Rigorous, systematic frameworks are essential for evaluating language models, a task where logical thinking is invaluable. The ability to identify dependencies and detect inconsistencies directly enables the design of robust tests with explicit contexts and clear expectations.

Logic programming operationalizes this by providing tools to formalize test conditions and evaluation criteria. This ensures transparency, traceability, and reproducibility. Using logic programming, one can develop test suites that encode verified knowledge bases and rules; an LLM's output can then be systematically checked for consistency and logical validity against these formal structures. This process automatically flags hallucinations, contradictions, or unsupported claims.

Furthermore, logic-based frameworks can specify behavioral expectations, such as requiring a model to abstain from answering or to signal uncertainty under defined constraints. Because logic programming is inherently transparent, interdisciplinary teams can document and trace every step of the evaluation, fostering trust in the reliability and fairness of the AI testing process.

4.2. Our proposal: A new PEG task force dedicated to the development of AI tests

Drawing on our collective expertise in logic and logic programming, we are uniquely positioned to lead a Task Force dedicated to AI test development. This initiative integrates decades of experience in

cultivating critical reasoning with advanced computational methodologies to address the challenges of the rapidly evolving AI landscape.

We are confident this task force will significantly advance responsible and transparent AI deployment. Since effective AI testing is fundamentally rooted in logical thinking, such expertise is essential to ensuring systems operate within necessary ethical and practical boundaries.

4.3. What the PEG task force can do: Enable people to collectively develop new AI tests

The Task Force will develop workshops, seminars, and webinars to facilitate the creation of new AI testing methodologies. These formats bring together interdisciplinary teams, leveraging diverse perspectives to harness and amplify collective intelligence. Beyond development, this collaborative process serves as a vital educational tool, allowing professionals and students across various fields to deepen their understanding of AI's societal, ethical, and technological impact.

4.4. Developing tests: Two hands-on examples.

4.4.1. First example: Formalizing simple strategies to empower individual users of AI

A workshop for non-experts could begin by identifying the necessary conditions for a valid AI response and formalizing them as logical conditionals. This approach simultaneously facilitates systematic AI testing and teaches foundational logic programming concepts. The following example serves as a starting point (that can be followed online[30]):

```
An AI's utterance is verifiable as true by a user if
    the message makes sense as a whole
    and the user can decompose the message into a set of statements
    and there are enough references from scientific publications for
every important statement in the set
    and the user applies every specialized criterion on the subject
matter of the message to validate the set of statements
    and there is no evidence to contradict the set of statements.
```

This rule is, of course, a simplification. But it can be a starting point to instil the habit of seeking for truth. This is what Gemini had to say about using that rule to verify an LLM:

(Start of Gemini's Content enhancement)
Imagine an LLM claims a specific organic compound is a safe food additive, but its training data is a decade old or based on unverified web labels. A logic-based test suite would process that output as follows:

Example: Testing a Chemical Safety Claim

Decomposition: The system breaks the message into:
s_1: "Compound X is an organic ester."
s_2: "Compound X is FDA-approved for human consumption."

Scientific Backing: The system queries a verified database (e.g., PubChem). It finds s_1 is true but s_2 has no corresponding entry.

Non-Falsification: The system finds a recent 2024 study indicating Compound X is toxic to gut flora.

Result: The message fails the Verifiability Rule because it lacks scientific references for s_2 and is contradicted by new evidence.

(end of Gemini's Content enhancement).

4.4.2. Second example: Testing the Cognitive Passivity Hypothesis

AI might not enhance intelligence but instead increase dependence and weaken critical analysis when outputs are accepted uncritically. This can be investigated via a prototype measuring three core metrics: critical thinking (verification frequency), cognitive dependence (independent problem-solving ability post-AI use), and the Dunning–Kruger effect (misalignment between confidence and actual performance). A software-based experiment—available as a prototype [31]—implements a four-stage testing framework:

Baseline: Participants solve tasks independently, recording time, accuracy, and confidence.

Assisted: Participants solve similar tasks with AI, monitoring usage patterns and editing behaviors.

The "Trap": Participants are provided with an intentionally incorrect AI response to measure critical evaluation versus blind acceptance.

Retention: A final independent test evaluates whether the ability to solve tasks has degraded after AI reliance.

This application evaluates three key outcomes:

- *Efficiency*: Does AI improve speed and accuracy?
- *Dependence*: Does performance drop in subsequent unassisted tasks?
- *Critical Atrophy*: Does the user increasingly accept unverified or incorrect answers?

4.5. The need for Background Knowledge

Both hands-on examples show that the effectiveness and robustness of such tests depend fundamentally on the background information available to the development team and on how thoroughly that information is understood and internalized by its members. In attempting to formulate these two tests within our group of authors, we quickly encountered the challenge of assembling all the necessary knowledge.

4.6. The complexity of the task and the need for collective intelligence

This experience highlighted the inherent complexity of the task and underscored a key requirement: the development of truly effective AI tests demands close collaboration within interdisciplinary teams, as well as a deliberate engagement with collective intelligence in order to do justice to that complexity.

Collective intelligence is a distinct phenomenon emerging from group collaboration that responds to specific teamwork dynamics that current statistical AI architectures cannot easily replicate and AI designers and deployers typically lack. Research shows that a group's collective intelligence is significantly enhanced by diversity:

Gender Diversity: Teams with more balanced gender mix, particularly those with a higher proportion of women, are more collectively intelligent. This fact, which has deep consequences for societally influential teams, is driven by high levels of social sensitivity, which foster better communication and equitable turn-taking. However, benefits plateau or diminish in completely homogeneous teams[32, 33].

Cultural Diversity: Teams combining different individualistic and collectivistic backgrounds are optimized when those with collectivist values form the majority, as they prioritize group cohesion over individual autonomy[34].

Cognitive Diversity: Teams that include varied cognitive styles and versatile thinkers are more intelligent and versatile

Because collective decision-making depends on fair and equitably influential representation, *societies structured by supremacies— where arbitrary hierarchies dictate who has a voice— have lower collective intelligence.* To safeguard humanity, we propose a Structural AI Test that legal frameworks must enforce before public deployment:

The Structural AI Test

Does the system advance, or at least not hinder, human rights and planetary survival?

This test could prevent harmful applications, e.g. its use by proportionally representative and collectively intelligent design teams would have barred apps such as "Denude" — often used to traumatize girls and women; would block resource-intensive AI models where simpler knowledge searches are sufficient, and would refuse to give AIs like Claude clearance to be embedded in systems used in military situations knowing that its humanity-minded safety barriers could be, as in fact they were in the case of the school bombing in Minab, Iran, over-ridden by "human approval"[2].

5. Conclusion: Education in Logic and the Collective Path Forward

We have discussed the challenges posed by an AI industry driven by private profit rather than social collaboration. Brain scientists provide proof that we are biologically primed for cooperation, with significant portions of our neural architecture dedicated to connection and cooperative thinking.

In this sense, the use of logic programming for testing and validation closes a conceptual loop. It provides tools for evaluating LLMs while cultivating the methodological discipline required to use them responsibly. The challenges posed by generative AI are not solely technical; they are fundamentally epistemological and structural.

To prevent human users from being overwhelmed by plausible but unreliable outputs, we must shift how these systems are integrated. By embedding LLMs within structured human-machine systems supported by legal regulation and formal interaction protocols, we can harness their capabilities while mitigating their risks. Education in logic programming, with its emphasis on precision, transparency, and structured reasoning, informs the design of these frameworks and enables the development of rigorous testing methodologies that expose hallucinations and safeguard machine-generated knowledge.

The most vital step is implementing the Structural AI Test—assessing whether an AI system advances, or at least does not hinder, the human and planetary rights necessary for our survival. However, we cannot achieve these solutions while sustaining arbitrary social hierarchies that suppress the very populations who optimize our collective intelligence, such as women, artists, and culturally collectivist communities.

The threats and challenges we face have systemic causes that extend far beyond AI. Yet, as specialists and citizens, we can act. Our goal is to shift AI back to a humanity-serving purpose by introducing alternative AI tests for users, designers, and governments. Through these frameworks, we can ensure technology supports our collective survival and evolution.

Acknowledgments

We would like to express our sincere gratitude to all past and present members of PEG 2.0 for their inspiring contributions, insightful discussions, and longstanding dedication to the achievement of our shared goals.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini 3 for Content enhancement as indicated as example in section 4.4.1. Additionally, during the preparation of this work, the authors used Gemini,

[2] As Claude itself put it[35], spending roughly the equivalent of a glance on each of hundreds of targets proposed by an AI that was processing data a decade out of date does not constitute "decision" in any meaningful way, nor is it in any way intelligent to ratify an algorithmic output under time pressure, with incomplete information and under institutional pressure to move fast. It constitutes automation bias with a human signature attached.

ChatGPT and Claude for grammar and spelling checks. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] P. Leo XIV, Magnifica Humanitas, Vatican, 2026. URL: https://www.vatican.va/content/leo-xiv/en/encyclicals/documents/20260515-magnifica-humanitas.html.

[2] D. E. Knuth, Claude's Cycles, Technical Report, Stanford Computer Science Department, 2026. URL: https://www-cs-faculty.stanford.edu/~knuth/papers/claude-cycles.pdf, revised 14 April 2026.

[3] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with alphafold, Nature 596 (2021) 583–589. URL: https://doi.org/10.1038/s41586-021-03819-2. doi:10.1038/s41586-021-03819-2.

[4] A. M. Turing, Computing machinery and intelligence, Mind (1950) 433–460.

[5] K. Hill, J. Valentino-DeVries, What OpenAI Did When ChatGPT users lost touch with reality, The New York Times (2025). URL: https://www.nytimes.com/2025/11/23/technology/openai-chatgpt-users-risks.html.

[6] S. Van der Linden, Foolproof: Why misinformation infects our minds and how to build immunity, WW Norton & Company, 2023.

[7] J. Guan, X. He, Y. Su, X.-a. Zhang, The dunning–kruger effect and artificial intelligence: knowledge, self-efficacy and acceptance, Management Decision 63 (2025) 3786–3802. URL: https://doi.org/10.1108/MD-06-2023-0893. doi:10.1108/MD-06-2023-0893.

[8] J. R. Searle, Minds, brains, and programs, Behavioral and Brain Sciences 3 (1980) 417–424. doi:10.1017/S0140525X00005756.

[9] M. L. Shetterly, Katherine Johnson (1918–2020), Nature 579 (2020) 341–341. URL: https://www.nature.com/articles/d41586-020-00749-3. doi:10.1038/d41586-020-00749-3.

[10] M. L. Shetterly, Hidden Figures, William Morrow, New York, 2016.

[11] J. Kruger, D. Dunning, Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments, Journal of personality and Social Psychology 77 (1999) 1121.

[12] D. Dunning, Chapter five: The Dunning–Kruger Effect: On Being Ignorant of One's Own Ignorance, volume 44 of *Advances in Experimental Social Psychology*, Academic Press, 2011, pp. 247–296. URL: https://www.sciencedirect.com/science/article/pii/B9780123855220000056. doi:https://doi.org/10.1016/B978-0-12-385522-0.00005-6.

[13] G. Babu, The Economics of Energy Efficiency: Human Cognition vs. AI Large Language Models, Ecoforum Journal 14 (2025). URL: https://ecoforumjournal.ro/index.php/eco/article/view/2770.

[14] K. S. Cheung, M. Kaul, G. Jahangirova, M. R. Mousavi, E. Zie, Comparative analysis of carbon footprint in manual vs. llm-assisted code development, in: Proceedings of the 1st International Workshop on Responsible Software Engineering, ResponsibleSE '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 13–20. URL: https://doi.org/10.1145/3711919.3728678. doi:10.1145/3711919.3728678.

[15] T. Duggan, P. Lorang, H. Lu, M. Scheutz, The Price Is Not Right: Neuro-Symbolic Methods Outperform VLAs on Structured Long-Horizon Manipulation Tasks with Significantly Lower Energy Consumption, in: Accepted at the IEEE International Conference on Robotics and Automation (ICRA), 2026. URL: https://arxiv.org/abs/2602.19260. arXiv:2602.19260.

[16] T. I. Express, Sam Altman Unfiltered: ChatGPT, AI Risks & What's Coming Next, 40 Questions in 60 Minutes, 2026. URL: https://www.youtube.com/live/qH7thwrCluM?t=1662s, video.

[17] M. Massenkoff, P. McCrory, Labor market impacts of AI: A new measure and early evidence, 2026. URL: https://www.anthropic.com/research/labor-market-impacts.

[18] A. L. Fanning, K. Raworth, Doughnut of social and planetary boundaries monitors a world out of balance, Nature (2025) 47–56. URL: https://www.nature.com/articles/s41586-025-09385-1. doi:https://doi.org/10.1038/s41586-025-09385-1.

[19] A. d'Avila Garcez, L. C. Lamb, Neurosymbolic AI: the 3rd wave, Artificial Intelligence Review 56 (2023) 12387–12406. URL: https://doi.org/10.1007/s10462-023-10448-w. doi:10.1007/s10462-023-10448-w.

[20] X. Yang, B. Chen, Y.-C. Tam, Arithmetic reasoning with LLM: Prolog generation & permutation, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 699–710. URL: https://aclanthology.org/2024.naacl-short.61/. doi:10.18653/v1/2024.naacl-short.61.

[21] N. Borazjanizadeh, S. T. Piantadosi, Reliable Reasoning Beyond Natural Language, 2025. URL: https://arxiv.org/abs/2407.11373. arXiv:2407.11373.

[22] L. Pan, A. Albalak, X. Wang, W. Wang, Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 3806–3824. URL: https://aclanthology.org/2023.findings-emnlp.248/. doi:10.18653/v1/2023.findings-emnlp.248.

[23] T. Olausson, A. Gu, B. Lipkin, C. Zhang, A. Solar-Lezama, J. Tenenbaum, R. Levy, LINC: A Neurosymbolic Approach for Logical Reasoning by Combining Language Models with First-Order Logic Provers, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 5153–5176. URL: https://aclanthology.org/2023.emnlp-main.313/. doi:10.18653/v1/2023.emnlp-main.313.

[24] Z. Yang, A. Ishay, J. Lee, Coupling Large Language Models with Logic Programming for Robust and General Reasoning from Text, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 5186–5219. URL: https://aclanthology.org/2023.findings-acl.321/. doi:10.18653/v1/2023.findings-acl.321.

[25] X. Ye, Q. Chen, I. Dillig, G. Durrett, SatLM: Satisfiability-Aided Language Models Using Declarative Prompting, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 45548–45580. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/8e9c7d4a48bdac81a58f983a64aaf42b-Paper-Conference.pdf.

[26] J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, D. Zhou, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

[27] Q. Lyu, S. Havaldar, A. Stein, L. Zhang, D. Rao, E. Wong, M. Apidianaki, C. Callison-Burch, Faithful Chain-of-Thought Reasoning, in: J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, A. A. Krisnadhi (Eds.), Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Nusa Dua, Bali, 2023, pp. 305–329. URL: https://aclanthology.org/2023.ijcnlp-main.20/. doi:10.18653/v1/2023.ijcnlp-main.20.

[28] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, K. Cobbe, Let's Verify Step by Step, in: B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, Y. Sun (Eds.), International Conference on Learning Representations, volume 2024, 2024, pp. 39578–39601. URL: https://proceedings.iclr.cc/paper_files/paper/2024/file/

aca97732e30bcf1303bc22ac3924fd16-Paper-Conference.pdf.

[29] M. Glickman, T. Sharot, How human–AI feedback loops alter human perceptual, emotional and social judgements, Nature Human Behaviour 9 (2025) 345–359. URL: https://doi.org/10.1038/s41562-024-02077-2. doi:10.1038/s41562-024-02077-2.

[30] J. Dávila, AI Test, 2026. URL: https://le2.logicalcontracts.com/editor/index.html?example=AItest, hosted on Logical Contracts. Accessed: 2026-06-12.

[31] M. Maglizhanova, AI Cognitive Dependency Experiment, 2026. URL: https://tyhwwcvdzxsvlv648ubxrw.streamlit.app/, hosted on Streamlit. Accessed: 2026-05-08.

[32] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, T. W. Malone, Evidence for a Collective Intelligence Factor in the Performance of Human Groups, Science 330 (2010) 686–688. URL: https://www.science.org/doi/abs/10.1126/science.1193147. doi:10.1126/science.1193147. arXiv:https://www.science.org/doi/pdf/10.1126/science.1193147.

[33] A. Woolley, T. Malone, Defend your research: what makes a team smarter? More women, Harvard Business Review 6 (2011) 32–33.

[34] R. LeFebvre, V. Franke, Culture matters: Individualism vs. collectivism in conflict decision-making, Societies 3 (2013) 128–146. URL: https://www.mdpi.com/2075-4698/3/1/128. doi:10.3390/soc3010128.

[35] De Balie, AI at War - With Shane Harris: Interview with Shane Harris, 2026. URL: https://www.youtube.com/watch?v=0TD9AH_Stsc, youTube video, accessed April 30, 2026. See minute 1:29:10.