"If the user of an AI system cannot determine the source of the sentence or paragraph or idea—and can't get some explanation of why it was chosen over other possibilities—then we may not be able to accomplish our other goals of accountability, security, or protecting our foundations. ***Explainability is thus perhaps the greatest challenge we face on AI. Even the experts don't always know why these algorithms produce the answers they do. It's a black box...But we do need to require companies to develop a system where, in simple and understandable terms, users understand why the system produced a particular answer and where that answer came from."***

- Senator Schumer's June 21, 2023 remarks launching SAFE Innovation Framework for Artificial Intelligence.[1]

I. True transparency in AI requires attribution to the source data.

In his comments above, Senator Schumer correctly identifies explainability as both the greatest challenge facing AI development and the most critical problem to address as we pursue our national ambition to lead the world in responsible AI innovation.

Is true transparency even possible in AI, and can regulators require it without stifling innovation? The answer is an emphatic **yes**, on both counts.

Most AI systems available today are built using neural networks, which are abstract representations of the vast amounts of data on which they are trained. Critically, predictions made by neural networks are not directly connected to data. Instead, neural networks infer and extrapolate a model from the training data in ways that can be virtually impossible to understand, even when only a small number of artificial neurons are involved. Once a neural network model has been trained, there's no way to go back. One can't explain the model's predictions by referencing the training data. And there's no reliable rewind button for unlearning incorrect or undesired training, any more than one can delete a once-heard curse word from a child's vocabulary.

As the public has seen in the alarming failures of Chat-GPT, this complex predictive extrapolation process often spirals out of control, causing the AI to "hallucinate." Unfortunately, these hallucinations are inherent to the underlying technology. Because practitioners cannot see why or how the AI makes a prediction and do not reliably know when the model is uncertain about an output, they have no way of scrutinizing the model and correcting it based on the original training data. The only corrective measure

[1] Remarks available in full at https://www.democrats.senate.gov/news/press-releases/majority-leader-schumer-delivers-remarks-to-launch-safe-innovation-framework-for-artificial-intelligence-at-csis

available is to train the black-box model with ever more data and hope the model moves itself in the desired direction.

The possible dangers of black-box AI are well known and well documented. In March 2023, nearly 1500 technology leaders called for a six-month pause in the development of AI systems, urging developers instead to work on making existing systems "more accurate, safe interpretable, [and] transparent."[2] Complete reliance on neural networks is a dead-end path with serious negative societal consequences. Like the internal combustion engine, we should not blindly accept a singular, dominant approach merely because we are not aware of the viable alternatives.

Transparent AI alternatives do exist, right now, today. Those tasked with creating the regulatory frameworks should know about the potential of these explainable approaches and incentivize their development and adoption. For decisions affecting human lives, we must not settle for the dominant black-box technology simply because it is the most trodden path.

II. Transparent AI is possible, but it requires further investment in and development of alternative methodologies.

The design of neural networks inherently prevents transparency. Neural networks observe training data sets, then build an AI model consisting of millions or billions of tiny statistical functions. The alternative approach is to keep the training data in its original form, organized by similarity. Then the prediction "should Jane Doe receive this house loan?" can be answered by simply querying the data for the loan decisions for customers most similar to Ms. Doe. If the 10 most similar records were all approved, the answer is clear, and can be explained by referencing those other credit records—not digging through billions of math functions. And if six were approved and four were not, the system can respond: "I cannot say for certain, and here is everything I know." Such capabilities seem simplistic, but they are effectively what machine learning approximates.

[2] https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Research over the past decade has begun to unearth the mathematical truths of similarity that allow this straightforward approach to achieve the same quality of results historically reserved for more complex machine learning approaches.

This technique, known as instance based learning (IBL), was developed contemporaneously with neural networks beginning in the 1950s. In the early 2000s, neural networks were found to be conveniently scalable on graphics processor units, leading to a massive boom in R&D investment and subsequent adoption. Other techniques, including IBL, faded to the periphery of AI research. In the intervening decades, IBL received a tiny fraction of the investment compared with funding of neural networks, leaving it largely relegated to niche applications and computer science classrooms. However, recent advances in IBL have led to natively explainable frameworks that address the most important issues in AI today—not chat interfaces that replace web search, but **explainability**, **transparency**, and **attribution**. Combined multidisciplinary breakthroughs in statistics, game theory, and information theory have unlocked new AI/ML capabilities with IBL, and the core technology leveraging those innovations is already in use by major financial institutions, governments, and technology companies around the world.

Regulators must be aware of these recent breakthroughs in transparency and explainability and know the full art of the possible. For instance, practitioners using IBL can see from the data itself precisely which features and elements of the training data were most influential in a decision, not just from a representative model which may be missing relationships. This is especially important when showing whether a prohibited feature of an individual's profile (e.g., race, gender, sexual orientation, political affiliation, etc.) was used in making a classification or decision or when determining whether a few erroneous records of training data are having disastrous effects. Users can audit AI outcomes generated using IBL, interrogate those outcomes to understand why and how the AI made decisions, and then intervene to correct mistakes and bias. Attribution can also demonstrate both why a decision was made and also why a contrary decision was not made.

Other emerging techniques may well offer similar breakthroughs in transparency. These "nearest neighbor" techniques with IBL are not suggested as the only transparent methodology, but rather as evidence that *true transparency in AI is possible, today*. Any emerging regulatory framework should guide practitioners towards IBL-like solutions that create trust through transparency.

III. Future regulatory frameworks should focus on those AI outcomes that pose the greatest threat.

Any new regulatory framework must strike the delicate balance between stifling innovation and protecting against harmful outcomes. Regulators should therefore focus attention where AI poses the greatest threats and insist on transparency through attribution where the risk from discrimination, bias, and hallucination is highest.

Neural networks are capable and powerful. And for so many applications, the benefits of black-box technology likely outweigh the risks. When designing a customer service chat bot, a black box like ChatGPT is likely a sufficient approach. But we need not compromise integrity for power. Americans should not accept black-box AI for critical life-affecting decisions like parole, medical intervention, college admissions, or credit decisions. These outcomes can shape the course of individual lives and of our society as a whole. The European Union's proposed AI Act provides an excellent enumeration of critical decisions that must be protected from potentially harmful, biased AI. America should similarly require practitioners to ensure that all AI tools used in such decisions allow for transparency and attribution sufficient to ensure accuracy and fairness. In September 2023, the Consumer Financial Protection Bureau issued guidance confirming that adverse action notice requirements "apply equally to all credit decisions, regardless of whether the technology used to make them involves complex or 'black-box' algorithmic models, or other technology that creditors may not understand sufficiently to meet their legal

obligations."[3] We wholeheartedly endorse such transparency requirements where AI is used to make profoundly important decisions.

IV. Conclusion

Transparent, trustworthy AI is not an aspirational pipe dream—it is available today. Those tasked with regulating and overseeing AI must demand full transparency for algorithmic decision-making in critical life-affecting decisions.

The Howso team has worked in pursuit of transparent, trustworthy AI for more than a decade. Howso Engine was initially developed in response to critical need within the Department of Defense for an AI system that could provide human-understandable, trustworthy advice. The latest version of our technology is open source and freely available,[4] and serves as a resounding rebuttal of the flawed notion that transparent AI is not yet possible.

Howso stands ready to contribute to this critical conversation about trust, transparency, and the future of AI however possible. We look forward to sharing our core technology, our expertise, and our experience with some of the world's most innovative companies and governments in pursuit of trustworthy AI.

We applaud the Senate's proactive approach in developing a responsible framework for AI innovation and look forward to this collaboration opportunity.

HOWSO™

Howso is an understandable AI company working to build high powered AI you can trust, audit, and explain. Howso's mission is to advance trustworthy AI as the global standard.

---

[3] *See* Consumer Financial Protection Circular 2023-03, available at https://www.consumerfinance.gov/compliance/circulars/circular-2023-03-adverse-action-notification-requirements-and-the-proper-use-of-the-cfpbs-sample-forms-provided-in-regulation-b/

[4] https://github.com/howsoai