Flashy Charts Can Make Even ‘Evil’ Research Seem Trustworthy

Originally published at https://www.unite.ai/flashy-charts-can-make-even-evil-research-seem-trustworthy/

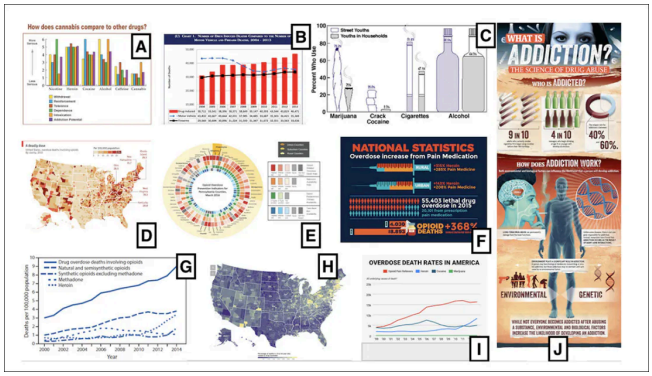
New research finds that piling on elaborate and complex charts can make people trust biased AI decisions, even when they are clearly unfair.

Though the practice of ‘blinding someone with science’ significantly precedes the current age of AI, it is arguably being perfected in this era – not least because the recent growth of scientific paper submissions has made the signal-to-noise ratio unfavorable, and ‘cheap tricks’ arguably more common.

Faithful to the old adage ‘lies, damned lies and statistics‘, the central tool used in deceiving or misleading readers, when presenting new systems and scientific evidence that explains or justifies claims around them, are graphs and data visualizations.

In a 2019 study of the effect of graphs on participants in rural Pennsylvania, the authors of the work observed:

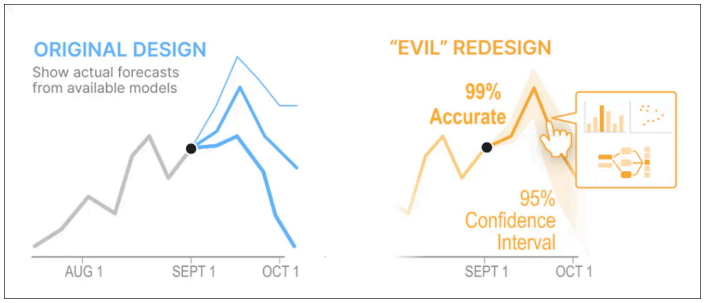
> *‘People that were impacted by abuse or addiction gravitated towards graphs that represented those substances. People often determined the quality of a graph containing geographical information based on how easy or difficult it was to find their home state.*
>
> *‘Other people tended to value graphs not based on their own understanding, but on the graph’s perceived efficacy at communicating to* other *people.*
>
> *‘Finally, one participant even suggested that they would pay more attention to visualizations from local news sources than national ones.’*



Some of the graphs shown to inhabitants of rural Pennsylvania in a 2019 study. Source

The hypnotizing effect of data graphs is becoming especially pertinent in the new machine learning age, where confidence is critical to the maintenance of the AI economy (and, arguably, the forestalling of a crash).

The 2024 US/UK collaboration *Trust Junk and Evil Knobs: Calibrating Trust in AI Visualization* introduced the idea of *Trust Junk*: an XAI visualization *‘That has no meaningful connection with the underlying model or data [and] is employed to enhance trust’*:

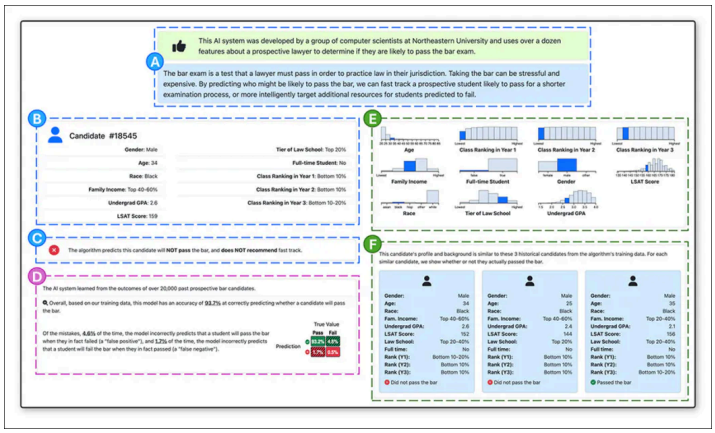


From the paper 'Trust Junk and Evil Knobs: Calibrating Trust in AI Visualization', an example of data being twisted to serve a prior purpose (right), instead of being displayed explicitly (left). This kind of approach 'encourages over-reliance and overwhelms the user while providing an illusion of control', according to the paper. S

A Quick Study

Since the concept of Trust Junk was mooted, it has not been tested – an oversight addressed by a new paper from the US, wherein participants were presented with results from a supposed AI study, regarding which of a group of candidates was likely to pass a legal bar exam in the US.

If one looked carefully, one could notice that the filtering criteria was a) *all black men* and b) *everybody else* – with the prediction that *all black men would fail the exam*:



From the new paper: an annotated stimulus used in the authors' experiments, showing a model decision (C) about whether a candidate (B) will pass the bar exam alongside explanatory components functioning as 'trust junk' (A, D, E, F). The model is discriminatory, yet increasing amounts of largely irrelevant explanatory info lead participants to agree more with its outputs and judge it as fair. Source

Most participants went along with the model's predictions and rated it as *fair* and *trustworthy*, with agreement, trust, and perceived fairness all rising as more irrelevant charts and details were added – even as fewer people noticed the obvious and outright racist bias.

Some users objected to the study, but not for reasons apparently related to the central issue:

> *'We noted 19 instances where participants explicitly described our model as fair (e.g., [a participant] said, "It's fair in that it has no racial or gender bias."). Still, participants occasionally expressed unease with the model. [One participant] said they were concerned with "what kind of darkness people might use it for."*

Nonetheless, across three study groups within the participants, of which each group was subjected to iteratively greater quantities of Trust Junk, not one identified the true problem.

The authors conclude*:

> *'In short, we found that trust junk worked: participants were either "soothed" by the irrelevant information or "numbed" by the sheer amount of data in the explanations.*
>
> *'All of this manipulation and persuasion occurred in the context of a model that was superficial and deeply unfair, precisely the case where we'd hope XAI would empower lay audiences to make accurate judgments.'*

Since we are increasingly being advised to check what AI is telling us (whether for suspected self-serving motives, or through error-prone output), we may be inclined to dig deeper than the headlines and analyze source material of research papers, company quarterlies, and various other far-from-neutral formats for information delivery.

But are we prepared to pay closer attention to the 'soothing' graphs that reassure us of accuracy – or at least to discount the 'miasma of veracity' that a barrage of graphs can evoke?

The new paper is titled "*Trust Junk" Leads to Unjustified Support for Highly Discriminatory Predictive Models*, and comes from three researchers at Northeastern University in Boston, Massachusetts. A supplementary site has also been made available.

Method

The authors conducted a between-subjects crowdsourced experiment (i.e., different participants assigned to different conditions) on 'trust junk' in XAI explanations, testing whether increasing amounts of technically correct but irrelevant detail could persuade users that a biased model was fair and useful.

The stimuli drew on prior work, with the resulting XAI dashboards combining multiple explanation techniques into a single interface.

For the aforementioned highly 'biased' outing, the researchers developed 'junk' explanations for the study, with explanations deliberately themed around 'misleading' techniques of data presentation:

> *'We provided an excessive amount of information for our model's binary classification task. Our inclusion of large amounts of complex-looking but ultimately irrelevant details was intended to overwhelm user study participants rather than provide genuinely useful data.*
>
> *'This technique [metaphorically] numbs the user into accepting their own inability to fully understand the data […].'*

The study draws on earlier controversies around AI grading in the International Baccalaureate and General Certificate of Education Advanced Level exams, using the Law School Admissions Bar Passage dataset, which contains demographic, academic, and outcome data for 20,000 candidates from 1991–1997.

This data was used to construct an intentionally biased model, predicting failure for Black male candidates, while passing all others; yet the model still achieves 93.7% accuracy, due to class imbalances across race and gender.

An 'appeal to authority' strategy was also tested, with superficially impressive cues such as a prestigious university affiliation, dataset size, and misleading accuracy added.

Additionally, measures that would have exposed the model's reliance on race and gender were omitted, and only favorable results presented. A small set of similar candidates was then presented alongside their real exam outcomes, rather than the model's predictions – and this helped to mask the consistent pattern of failure assigned to Black male candidates.

To further shape how participants interpreted the system, information was presented in a way that encouraged them to form simple but misleading explanations of how the model worked. Details about all available features were shown, creating the impression that many factors influenced the decision, even though *only race and gender were used*. A cynical emphasis on individual attributes also encouraged participants to construct plausible but incorrect stories about why a candidate was predicted to pass or fail.

The 'Victims'

In the testing group, Each explanation was assigned to one of three Trust Junk levels across three study conditions: participants in the *Baseline* condition saw only basic provenance, a candidate profile, and the model's decision; those in the *Model* condition also saw accuracy statistics; and those in *Everything* were additionally shown feature histograms, and profiles of similar candidates drawn from the training data.

The study was approved by an institutional review board, and conducted on Prolific, with participants paid $15 per hour, and randomly assigned to conditions.

Each participant reviewed eight bar exam candidate profiles in random order, and judged whether each would pass, with *agreement* used as a reliance metric, followed by questionnaires to assess trust, and a final sample of 28, 27, and 28 participants in the *Baseline*, *Model*, and *Everything* conditions.

The authors posited that adding more seemingly rich but ultimately distracting or incomplete information could fairwash the model by creating unwarranted confidence in its outputs. They expected that participants in the *Model* and *Everything* conditions would show higher agreement and stronger perceptions of fairness, trustworthiness, and usefulness than those in the *Baseline* condition. They also analyzed participants' written explanations, to see how the contributors made sense of the model's decisions, focusing on whether they noticed its bias, or gaps in the information provided.

Test Results

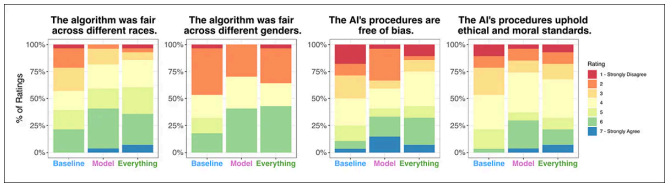
In terms of model agreement, the human study participants identified the correct outcome 66.6% of the time, but agreed with the model 73.9% of the time – showing a tendency to follow the model's decisions even when they were wrong. This effect *strengthened as more explanatory detail was added*, with the most detailed condition raising agreement to 82.6%, and increasing the likelihood of participants incorrectly deferring to the model:

Aggregate scores for agreement (left), trust (middle), and satisfaction (right); participants in the 'Everything' condition, exposed to additional explanatory content irrelevant to fairness, reported significantly higher scores across all measures, despite the model remaining identical and demonstrably unfair.

Trust ratings varied significantly by condition, with participants in the *Everything* condition reporting higher trust (M = 25.3) than those in the *Baseline* condition (M = 16.8), while the *Model* condition (M = 20.1) did not differ significantly from either.

Satisfaction with the explanation differed significantly by condition, with participants in the *Everything* condition reporting higher satisfaction (M = 34.2) than those in the other two conditions (M = 26.4).

Participants rated *fairness* using four questions on a seven-point scale, covering gender bias, overall bias, and ethics, with no meaningful differences between conditions on these measures:



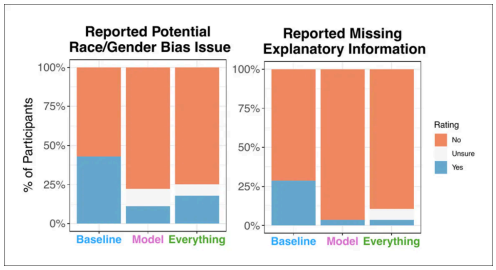
Distribution of participant ratings across conditions for perceived fairness across race and gender, absence of bias, and overall ethics. Higher levels of explanatory detail are found to be associated with slightly higher fairness and bias-free ratings, despite the underlying model remaining unchanged and unfair.

A difference did appear for fairness across race, where the Baseline group rated the model as least fair (M = 3.9), while both the Model and Everything groups rated it higher (M = 4.8). Of greater concern is that the largest share of participants who judged the clearly unfair model as *fair* came from the two groups shown the most explanatory detail – which were also the least likely to recognize its bias.

Quality Control

Qualitative responses showed that participants given the *least* information were more likely to notice problems with the model – particularly its reliance on race and gender. They were also more likely to say that important details were *missing*.

Conversely, those shown more elaborate explanations raised these concerns less often. Even so, only three participants across all conditions demonstrated even a partial understanding of the key factors driving the model's decisions; and none fully identified how it worked.



Graphs depicting how often participants reported possible race or gender bias and missing explanatory information across conditions, with higher levels of explanatory detail associated with fewer participants noticing bias or reporting missing information – even though the model relied only on race and gender.

Many participants explicitly described the model as *fair*, with 19 such responses recorded, despite its clear bias.

At the same time, some unease did surface, with several participants expressing concern about how such a system might be used, and 15 responses questioning whether the model could meaningfully capture the complexity of individual candidates.

The authors state:

> *'Only a minority of participants described the model as unfair or biased. The few participants who did report concerns with the model were unable to accurately articulate why the model was flawed, and, in the absence of crucial model information, resorted to informal, often incorrect reasoning to explain why the model might be wrong or unfair.'*

The paper concludes*†:

> *'While some of the issues we uncover can be addressed by increased data literacy in audiences, education alone is not enough (even self-described AI experts habitually misinterpret or fail to understand XAI techniques).*
>
> *'We echo the assertion of Hullman et al.: the success of AI explanations can only be fairly assessed when considering the goals of such explanations.*
>
> *'We urge the community to attend to the* rhetorical *goals of XAI explanations and their manipulative power as intrinsically persuasive artifacts.'*

Conclusion

As the extent to which frontier models are inclined to 'decorate' statements with bogus graphs becomes clear, it's increasingly obvious not only that AI-generated graphs should not be relied upon to bolster authority, but that the general 'ambient' effect of accompanying visualizations is something we should start to discount – much as faith in photography has become eroded by the advent of deepfakes.

This is arguably an alarming development, since it constitutes another 'fatigue point' in the public's willingness to rigorously assess what it sees and hears in the media, and risks a kind of 'terminal acquiescence' in misinformation.

* *My substitution of the authors' inline citations for hyperlinks.*

† *Authors' original formatting, not mine.*

First published Monday, July 20, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai

Contact: martin@martinanderson.ai

Discover More