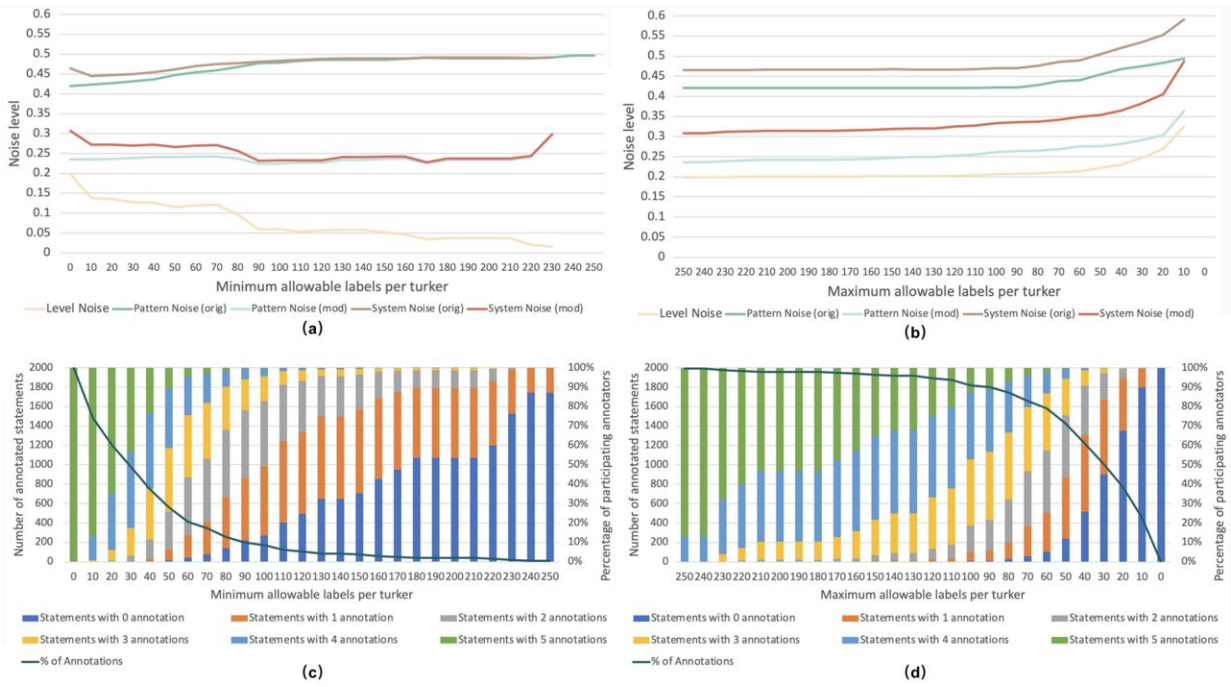
'Noise' in the machine: Human differences in judgment lead to problems for AI

May 15 2024, by Mayank Kejriwal



The noise level observed in ComVE annotations when controlling the minimum (a) and maximum (b) allowable labels for each turker (we omit the residual in these two figures as, similar to what was observed earlier for TG-CSR, it was nearly coincidental with the level noise curve). We also report the statistics of statements categorized by the number of annotation labels under different cutoff methods used for screening participating annotators in (c) and (d). Credit: *Scientific Reports* (2024). DOI: 10.1038/s41598-024-58937-4

Many people understand the concept of bias at some intuitive level. In society, and in artificial intelligence systems, racial and gender biases are well documented.

If society could somehow remove bias, would all problems go away? The late Nobel laureate Daniel Kahneman, who was a key figure in the field of behavioral economics, argued in his last book that bias is just one side of the coin. Errors in judgments can be attributed to two sources: bias and noise.

Bias and noise both play important roles in fields such as law, medicine and financial forecasting, where human judgments are central. In our work as computer and information scientists, my colleagues and I have found that noise also plays a role in AI.

Statistical noise

Noise in this context means variation in how people make judgments of the same problem or situation. The problem of noise is more pervasive than initially meets the eye. A seminal work, dating back all the way to the Great Depression, has found that different judges gave different sentences for similar cases.

Worryingly, sentencing in court cases can depend on things such as the temperature and whether the local football team won. Such factors, at least in part, contribute to the perception that the justice system is not just biased but also arbitrary at times.

Other examples: Insurance adjusters might give different estimates for

similar claims, reflecting noise in their judgments. Noise is likely present in all manner of contests, ranging from wine tastings to local beauty pageants to college admissions.

Noise in the data

On the surface, it doesn't seem likely that noise could affect the performance of AI systems. After all, machines aren't affected by weather or football teams, so why would they make judgments that vary with circumstance? On the other hand, researchers know that bias affects AI, because it is reflected in the data that the AI is trained on.

For the new spate of AI models like ChatGPT, the gold standard is human performance on general intelligence problems such as common sense. ChatGPT and its peers are measured against human-labeled commonsense datasets.

Put simply, researchers and developers can ask the machine a commonsense question and compare it with human answers: "If I place a heavy rock on a paper table, will it collapse? Yes or No." If there is high agreement between the two—in the best case, perfect agreement—the machine is approaching human-level common sense, according to the test.

So where would noise come in? The commonsense question above seems simple, and most humans would likely agree on its answer, but there are many questions where there is more disagreement or uncertainty: "Is the following sentence plausible or implausible? My dog plays volleyball." In other words, there is potential for noise. It is not surprising that interesting commonsense questions would have some noise.

But the issue is that most AI tests don't account for this noise in experiments. Intuitively, questions generating human answers that tend

to agree with one another should be weighted higher than if the answers diverge—in other words, where there is noise. Researchers still don't know whether or how to weigh AI's answers in that situation, but a first step is acknowledging that the problem exists.

Tracking down noise in the machine

Theory aside, the question still remains whether all of the above is hypothetical or if in real tests of common sense there is noise. The best way to prove or disprove the presence of noise is to take an existing test, remove the answers and get multiple people to independently label them, meaning provide answers. By measuring disagreement among humans, researchers can know just how much noise is in the test.

The details behind measuring this disagreement are complex, involving significant statistics and math. Besides, who is to say how common sense should be defined? How do you know the human judges are motivated enough to think through the question? These issues lie at the intersection of good experimental design and statistics. Robustness is key: One result, test or set of human labelers is unlikely to convince anyone. As a pragmatic matter, human labor is expensive. Perhaps for this reason, there haven't been any studies of possible noise in AI tests.

To address this gap, my colleagues and I designed such a study and published our findings in *Scientific Reports*, showing that even in the domain of common sense, noise is inevitable. Because the setting in which judgments are elicited can matter, we did two kinds of studies. One type of study involved paid workers from Amazon Mechanical Turk , while the other study involved a smaller-scale labeling exercise in two labs at the University of Southern California and the Rensselaer Polytechnic Institute.

You can think of the former as a more realistic online setting, mirroring

how many AI tests are actually labeled before being released for training and evaluation. The latter is more of an extreme, guaranteeing high quality but at much smaller scales. The question we set out to answer was how inevitable is noise, and is it just a matter of quality control?

The results were sobering. In both settings, even on commonsense questions that might have been expected to elicit high—even universal—agreement, we found a nontrivial degree of noise. The noise was high enough that we inferred that between 4% and 10% of a system's performance could be attributed to noise.

To emphasize what this means, suppose I built an AI system that achieved 85% on a test, and you built an AI system that achieved 91%. Your system would seem to be a lot better than mine. But if there is noise in the human labels that were used to score the answers, then we're not sure anymore that the 6% improvement means much. For all we know, there may be no real improvement.

On AI leaderboards, where large language models like the one that powers ChatGPT are compared, performance differences between rival systems are far narrower, typically less than 1%. As we show in the paper, ordinary statistics do not really come to the rescue for disentangling the effects of noise from those of true performance improvements.

Noise audits

What is the way forward? Returning to Kahneman's book, he proposed the concept of a "noise audit" for quantifying and ultimately mitigating noise as much as possible. At the very least, AI researchers need to estimate what influence noise might be having.

Auditing AI systems for bias is somewhat commonplace, so we believe

that the concept of a noise audit should naturally follow. We hope that this study, as well as others like it, leads to their adoption.

More information: Mayank Kejriwal et al, A noise audit of human-labeled benchmarks for machine commonsense reasoning, *Scientific Reports* (2024). DOI: 10.1038/s41598-024-58937-4