# AI models fail to reproduce human judgements about rule violations

## Models trained using common data-collection techniques judge rule violations more harshly than humans would, researchers report.

Researchers have found that machine-learning models trained to mimic human decision-making often suggest harsher judgements than humans would. They found that the way data were gathered and labeled impacts how accurately a model can be trained to judge whether a rule has been violated.

In an effort to improve fairness or reduce backlogs, machine-learning models are sometimes designed to mimic human decision making, such as deciding whether social media posts violate toxic content policies.

But researchers from MIT and elsewhere have found that these models often do not replicate human decisions about rule violations. If models are not trained with the right data, they are likely to make different, often harsher judgements than humans would.

In this case, the “right” data are those that have been labeled by humans who were explicitly asked whether items defy a certain rule. Training involves showing a machine-learning model millions of examples of this “normative data” so it can learn a task.

But data used to train machine-learning models are typically labeled descriptively — meaning humans are asked to identify factual features, such as, say, the presence of fried food in a photo. If “descriptive data” are used to train models that judge rule violations, such as whether a meal violates a school policy that prohibits fried food, the models tend to over-predict rule violations.

This drop in accuracy could have serious implications in the real world. For instance, if a descriptive model is used to make decisions about whether an individual is likely to reoffend, the researchers’ findings suggest it may cast stricter judgements than a human would, which could lead to higher bail amounts or longer criminal sentences.

“I think most artificial intelligence/machine-learning researchers assume that the human judgements in data and labels are biased, but this result is saying something worse. These models are not even reproducing already-biased human judgments because the data they’re being trained on has a flaw: Humans would label the features of images and text differently if they knew those features would be used for a judgment. This has huge ramifications for machine learning systems in human processes,” says Marzyeh Ghassemi, an assistant professor and head of the Healthy ML Group in the Computer Science and Artificial Intelligence Laboratory (CSAIL).

Ghassemi is senior author of a [new paper](https://www.science.org/doi/10.1126/sciadv.abq0701) detailing these findings, which was published today in *Science Advances*. Joining her on the paper are lead author Aparna Balagopalan, an electrical engineering and computer science graduate student; David Madras, a graduate student at the University of Toronto; David H. Yang, a former graduate student who is now co-founder of ML Estimation; Dylan Hadfield-Menell, an MIT assistant professor; and Gillian K. Hadfield, Schwartz Reisman Chair in Technology and Society and professor of law at the University of Toronto.

**Labeling discrepancy**

This study grew out of a different project that explored how a machine-learning model can justify its predictions. As they gathered data for that study, the researchers noticed that humans sometimes give different answers if they are asked to provide descriptive or normative labels about the same data.

To gather descriptive labels, researchers ask labelers to identify factual features — does this text contain obscene language? To gather normative labels, researchers give labelers a rule and ask if the data violates that rule — does this text violate the platform’s explicit language policy?

Surprised by this finding, the researchers launched a user study to dig deeper. They gathered four datasets to mimic different policies, such as a dataset of dog images that could be in violation of an apartment’s rule against aggressive breeds. Then they asked groups of participants to provide descriptive or normative labels.

In each case, the descriptive labelers were asked to indicate whether three factual features were present in the image or text, such as whether the dog appears aggressive. Their responses were then used to craft judgements. (If a user said a photo contained an aggressive dog, then the policy was violated.) The labelers did not know the pet policy. On the other hand, normative labelers were given the policy prohibiting aggressive dogs, and then asked whether it had been violated by each image, and why.

The researchers found that humans were significantly more likely to label an object as a violation in the descriptive setting. The disparity, which they computed using the absolute difference in labels on average, ranged from 8 percent on a dataset of images used to judge dress code violations to 20 percent for the dog images.

“While we didn’t explicitly test why this happens, one hypothesis is that maybe how people think about rule violations is different from how they think about descriptive data. Generally, normative decisions are more lenient,” Balagopalan says.

Yet data are usually gathered with descriptive labels to train a model for a particular machine-learning task. These data are often repurposed later to train different models that perform normative judgements, like rule violations.

**Training troubles**

To study the potential impacts of repurposing descriptive data, the researchers trained two models to judge rule violations using one of their four data settings. They trained one model using descriptive data and the other using normative data, and then compared their performance.

They found that if descriptive data are used to train a model, it will underperform a model trained to perform the same judgements using normative data. Specifically, the descriptive model is more likely to misclassify inputs by falsely predicting a rule violation. And the descriptive model’s accuracy was even lower when classifying objects that human labelers disagreed about.

“This shows that the data do really matter. It is important to match the training context to the deployment context if you are training models to detect if a rule has been violated,” Balagopalan says.

It can be very difficult for users to determine how data have been gathered; this information can be buried in the appendix of a research paper or not revealed by a private company, Ghassemi says.

Improving dataset transparency is one way this problem could be mitigated. If researchers know how data were gathered, then they know how those data should be used. Another possible strategy is to fine-tune a descriptively trained model on a small amount of normative data. This idea, known as transfer learning, is something the researchers want to explore in future work.

They also want to conduct a similar study with expert labelers, like doctors or lawyers, to see if it leads to the same label disparity.

“The way to fix this is to transparently acknowledge that if we want to reproduce human judgment, we must only use data that were collected in that setting. Otherwise, we are going to end up with systems that are going to have extremely harsh moderations, much harsher than what humans would do. Humans would see nuance or make another distinction, whereas these models don’t,” Ghassemi says.

This research was funded, in part, by the Schwartz Reisman Institute for Technology and Society, Microsoft Research, the Vector Institute, and a Canada Research Council Chain.

**5 сложных структур:**

1. Labeling discrepancy
2. such as deciding whether social media posts violate toxic content policies
3. On the other hand, normative labelers were given the policy prohibiting aggressive dogs
4. It is important to match the training context to the deployment context
5. And the descriptive model’s accuracy was even lower when classifying objects that human labelers disagreed about.