

# AI and the new 'golden rule': Finding the middle ground

## An ASU researcher is creating new AI tools that learn from shared human values

By Kelly deVos, ASU News  
September 29, 2025

Aristotle didn't write code, but he did leave us a debugging tip for human affairs: Virtue lives in the mean, not the extremes.

The advice to avoid the vices that lurk at either edge turns out to also be a surprisingly good recipe for modern artificial intelligence, too.

In the [School of Computing and Augmented Intelligence](#), part of the [Ira A. Fulton Schools of Engineering](#) at Arizona State University, [Hasan Davulcu](#) and collaborators are building systems that can find the mean, explain how they got there and stay there — on purpose.

“AI has to be controllable, programmable and, most importantly, understandable,” says Davulcu, a Fulton Schools professor of computer science and engineering. “That’s what I call the golden mean of AI. It’s an approach that looks at human experience, finds the middle ground and holds to it.”

### Why 'big AI' gets pulled to the extremes

Large language models, or LLMs, the kind of AI that powers tools like ChatGPT, are trained by scanning huge amounts of text from the internet. They’re good at predicting the next word in a sentence, but because the internet is messy, they also learn its flaws.

That means these systems can sometimes reproduce toxic speech, repeat harmful stereotypes or echo the loudest extremes of online debate.

“Current LLM-based tools tend to be similar in terms of their behavior and responses,” Davulcu says. “It’s not like a winner is emerging. They will need an edge. That edge is going to come from interpretability. How can we understand how these models make their decisions? And when they make incorrect decisions, how do we fix them?”

That's where his new suite of tools comes in. Davulcu has filed four invention disclosures with ASU's [Skysong Innovations](#), outlining a method to make AI transparent, programmable and ultimately safer.

The first innovation is a way to peek inside the "black box" of AI.

Normally, these systems spit out answers without showing their reasoning. If they get something wrong, whether it's suggesting [glue on a pizza](#) or giving [bad medical advice](#), developers can't easily correct the mistake.

Davulcu's method changes that. His system translates the AI's hidden decision-making into simple, editable rules. People can then adjust those rules, add exceptions and feed the corrections back into the model.

"In order to build safe AI, you have to go beyond the black box," he says. "You need to be able to see the rules the model is using, add exceptions and retrain it so that it doesn't keep making the same mistake."

Think of it as a feedback loop like wash, rinse and repeat. Reveal the logic, refine it, retrain and repeat. The goal, Davulcu explains, is an AI that won't make a mistake. But if it does, the user will have a recourse to fix it instantly.

## **Making values visible**

The second tool focuses on conversations. Most AI can tell if a comment sounds happy or angry, but that's not enough for real-world debates. What matters is the stance: whether someone is for, against or neutral on an issue and why.

Davulcu's team has built methods that can detect those stances and map how people cluster around them online. This makes it possible to see echo chambers, identify bridge-builders and highlight the shared values, such as fairness, safety or family, that people rally around.

"When you scale this, we can actually find the mean and the extremes," he says. "And, basically, at that point, we have a way of staying on the mean, avoiding the extremes, therefore getting rid of bias."

Once the system knows where the extremes are, it can be trained to avoid them. Davulcu's group showed that if you feed AI examples of more balanced language and filter out toxic or divisive phrasing, the model learns to follow that path.

The results are promising. In tests, their approach reduced toxic output by 85% compared with an unrestricted model. Instead of echoing the most polarizing voices, the AI leans toward civil, respectful conversation.

"We're showing that it's possible to build systems that encourage civil discourse instead of amplifying division," Davulcu says.

Companies may want to swap AI models when a new version comes out, but those changes can break carefully tuned behavior. Davulcu's fourth innovation is a control layer that travels with the

application itself.

“What we want is for an application to keep working no matter which model you plug in underneath,” he says. “You should be able to switch from one to another because it’s cheaper or better, and your rules still apply. Everything still works. And if something goes wrong, you have a recourse for fixing it.”

---

*This story originally appeared on [ASU News](#).*

## Main image



Anshul Trivedi (left), an alumnus who graduated with his master’s degree in computer science in the spring, and Hasan Davulcu (right), a professor of computer science and engineering in the School of Computing and Augmented Intelligence, part of the Ira A. Fulton Schools of Engineering at Arizona State University, discuss their research. Davulcu mentors students, including Trivedi, on projects designed to create understandable, balanced forms of artificial intelligence. Photographer: Erika Gronek/ASU

## Text image(s)



Davulcu lectures in a machine learning class. He teaches key principles and concepts of AI to Fulton Schools students as part of efforts to ensure graduates are well-prepared for roles in the fast-changing technological landscape. Photo by Kelly deVos/ASU