Exploring Interfaces to Democratize AI Constraint Generation

UNIVERSITY of HAWAI'I® HILO

TRAVIS MANDEL, JAHNU BEST, RANDALL H. TANAKA, HIRAM TEMPLE, CHANSEN HAILI, ROY SZETO

The Human Side of Safe AI

When AI agents are granted the ability to take impactful actions in the real world, we **never** want them to take actions we would consider "unsafe"

Therefore, learning from real-world experience what is "unsafe" is highly undesirable

- Standard practice is to have a small group of experts define **constraints** on the system to prevent it from executing unsafe actions
  - If these are defined incorrectly, unsafe behavior will result
  - Therefore, it's critical that we think about improving the **human side** of the equation!

Why only a small group of experts?

- Could cause biases, blind spots, difficulty scaling to large problems, etc.
  - In addition to making this process better, how can we make it **accessible**?

Our contributions

Developed new general-purpose task designs

Real-World Application: How can we make sure an AI system giving hints in an educational setting never gives hints which are **incorrect or misleading**?

- Don’t want it to harm the young students…

Results indicate interesting differences from other annotation problems

- Visit our poster to learn more!