

# Exploring Interfaces to Democratize AI Constraint Generation



UNIVERSITY  
of HAWAII®  
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!

2 out of 5 completed  
Read The Hint  
Try a different way of representing the total. If one creature is listed as one row, how many rows do we need here to represent two kinds of creatures?  
To See the [Original Model & Story](#) [Hover Here](#)

**Your Creation Space:**  
The hint applies to a student model if ( the smaller value is included AND the larger value is included )  
The hint applies to a student model if ( the smaller value is included AND the larger value is included )  
included AND the larger value is included  
included -- -- ) -- --  
Generate Models Show More Models Clear Workspace

**The Excluded Model's Story :**  
Did you know that unicorns can have pets too? A herd of 7 unicorns keeps 6 flying cats as pets. How many creatures are there all together in that herd?  
**The Excluded Model:**  
7 creatures  
**The Correct Model :**  
6 7  
creatures

3 out of 9 completed  
Read the story  
Did you know that unicorns can have pets too? A herd of 7 unicorns keeps 6 flying cats as pets. How many creatures are there all together in that herd?  
We want to help a student with one of these two models  
6 7  
creatures creatures  
7 6  
creatures  
create this correct model:  
7 creatures  
6 creatures  
Read the hint  
Let's put 'creatures' in the right place that shows it is the total. Shall we?  
Could this hint apply in this situation?  
Yes No -- You can still press "No" if you change your mind.  
Since you chose yes, write an explanation telling us why the hint applies.  
Submit

Once you've confirmed your rule and the models included/excluded, click "Next" to continue.

intro 514  
Chrissy loves exploring outdoors. Yesterday, she saw a herd of 12 elk being chased by a pack of 8 wolves. How many animals in total did Chrissy see while she was exploring?

8 12  
animals  
'animals' needs to be the total of all important parts.

✓