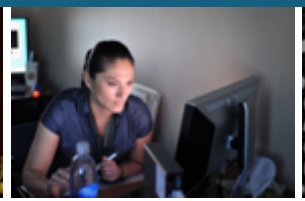
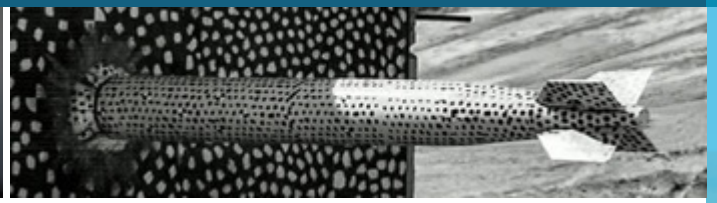
SAND2022-11400C

Sandia National Laboratories

When Does Seeing Become Believing? Potential Impacts of Model Characteristics and Visual Cues on Human Decisions





PRESENTED BY

Laura Matzen

U.S. DEPARTMENT OF ENERGY | NNSA

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Interactive Visualization for Fostering Trust in ML

Some questions to consider:

- What kinds of factors consistently foster appropriate levels of trust in ML?
- What are the pros and cons of visualizing different types of information that might be relevant to users?
- Which types of visual cues are most appropriate for supporting comprehension and decision making?
- When is interactivity helpful, and when is it confusing or overwhelming?
- What types of interactions are helpful for increasing understanding? For supporting appropriate trust?

Interactive Visualization for Fostering Trust in ML

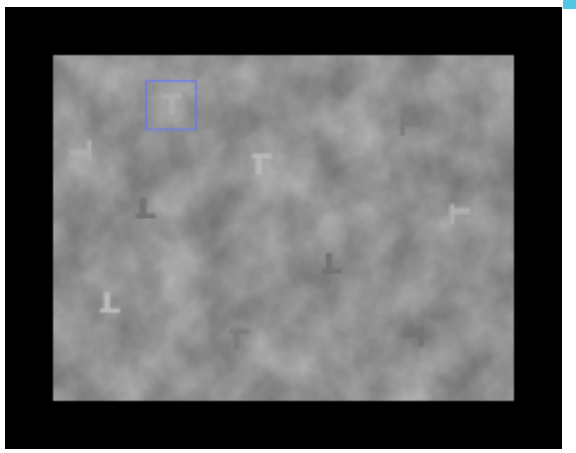
Some aspects of cognition to consider:

- Domain expertise
  - *People with different levels of expertise will differ in their ability to evaluate ML outputs and may trust (or not trust) them for different reasons*
- Individual differences in cognition and prior experience
  - *Different people have different strengths and preferences. They are influenced by their prior training, experiences, and expectations.*
- Working memory and cognitive load
  - *The amount of information we can hold in mind and manipulate at one time is very limited. Interpreting or interacting with visualizations may impact cognitive load.*
- Visual cognition
  - *People can have visual and perceptual illusions and biases that impact their comprehension and decision making.*
- Cognitive biases
  - *People often seek confirmation of what they already believe rather than testing alternative hypotheses.*

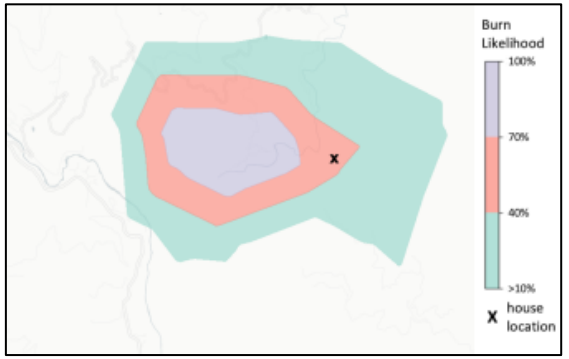
Factors that Impact Decision Making

A few examples:

- Visual search aided by (mock) ML outputs
  - People get complacent as the overall accuracy of the outputs goes up
  - Novices are more likely to go along with what the ML says

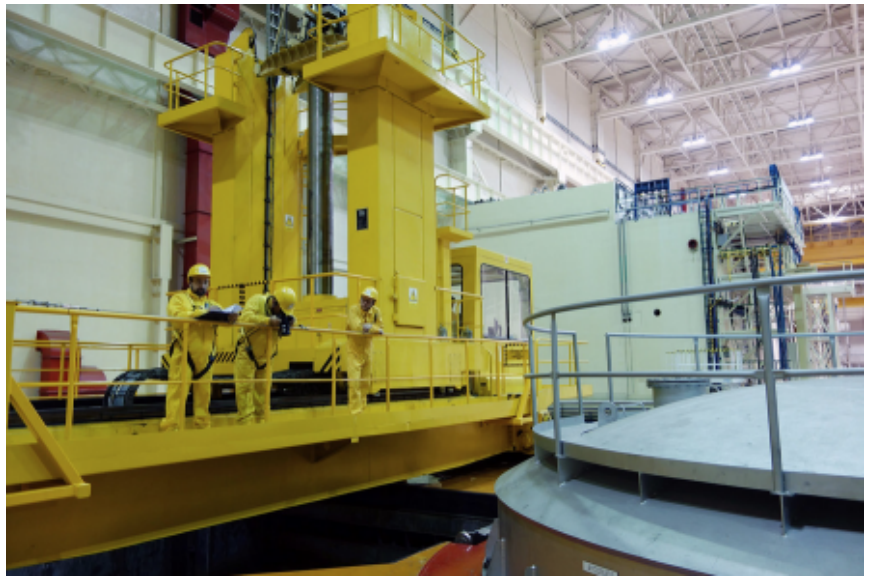


- Visualizations of uncertain information
  - Differences between visual and numerical representations
  - The specificity of the information can impact judgments of risk
  - The same information visualized in different ways can lead to different patterns of decisions
  - Individual differences also impact decisions



Context for our visual search experiments:

The increasing use of AI/ML tools in the international nuclear safeguards domain



<u>Nuclear safeguards:</u>

Detect the diversion of nuclear materials

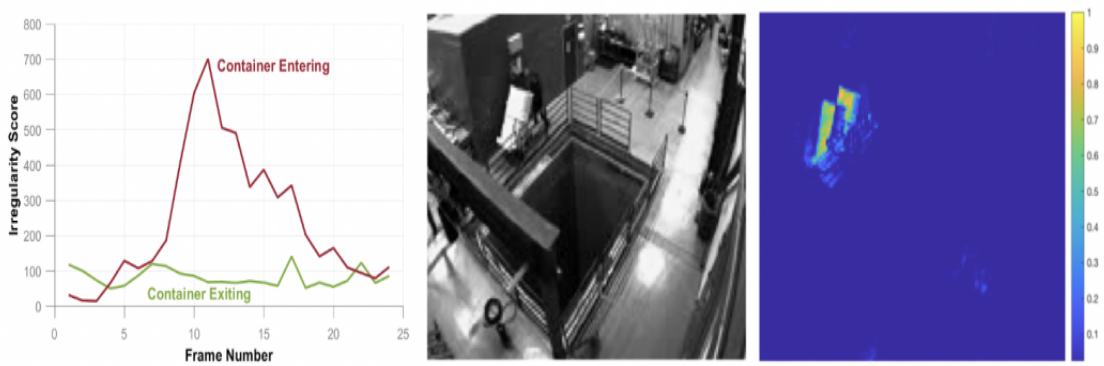
Detect the misuse of nuclear facilities

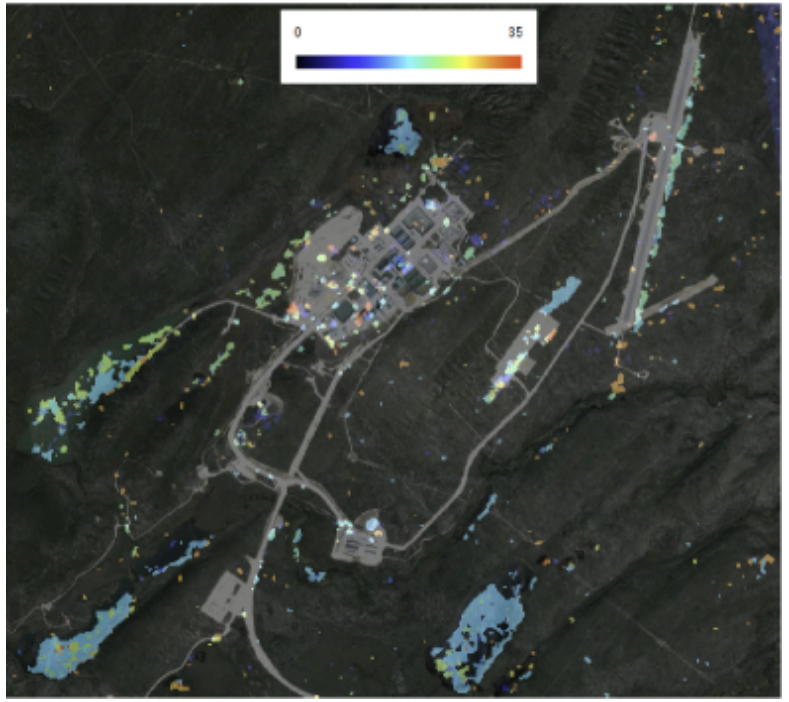
Detect the development of unknown nuclear facilities

Images: IAEA Imagebank, Flickr

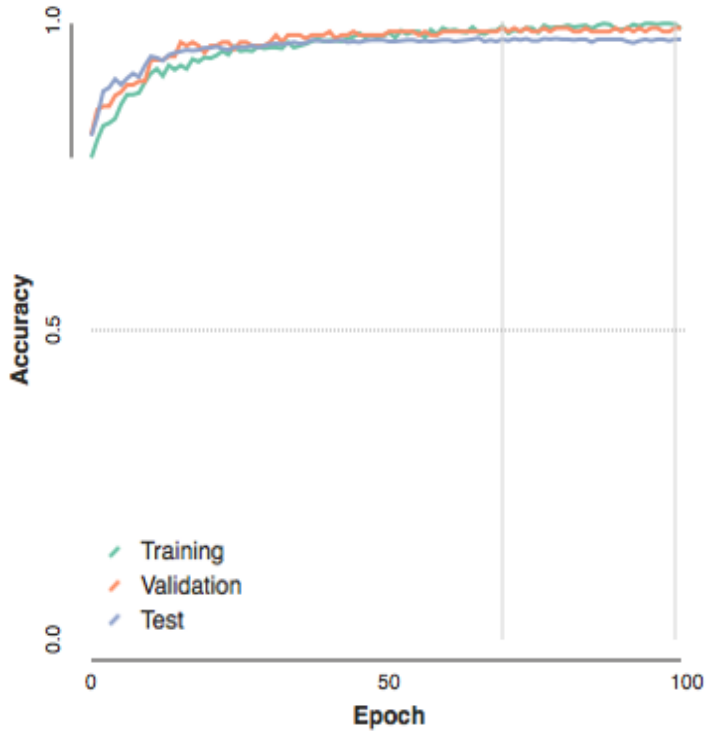
AI/ML models are being explored for multiple safeguards tasks

- Review of surveillance camera footage (Smith et al., 2021)
- Automated pre-processing of overhead imagery (Rutkowski et. al., 2018)
- Image matching for indoor localization (Belenguer et. al., 2020)
- Multi-modal information retrieval (Feldman, et. al., 2018)





AI/ML model performance continues to improve...

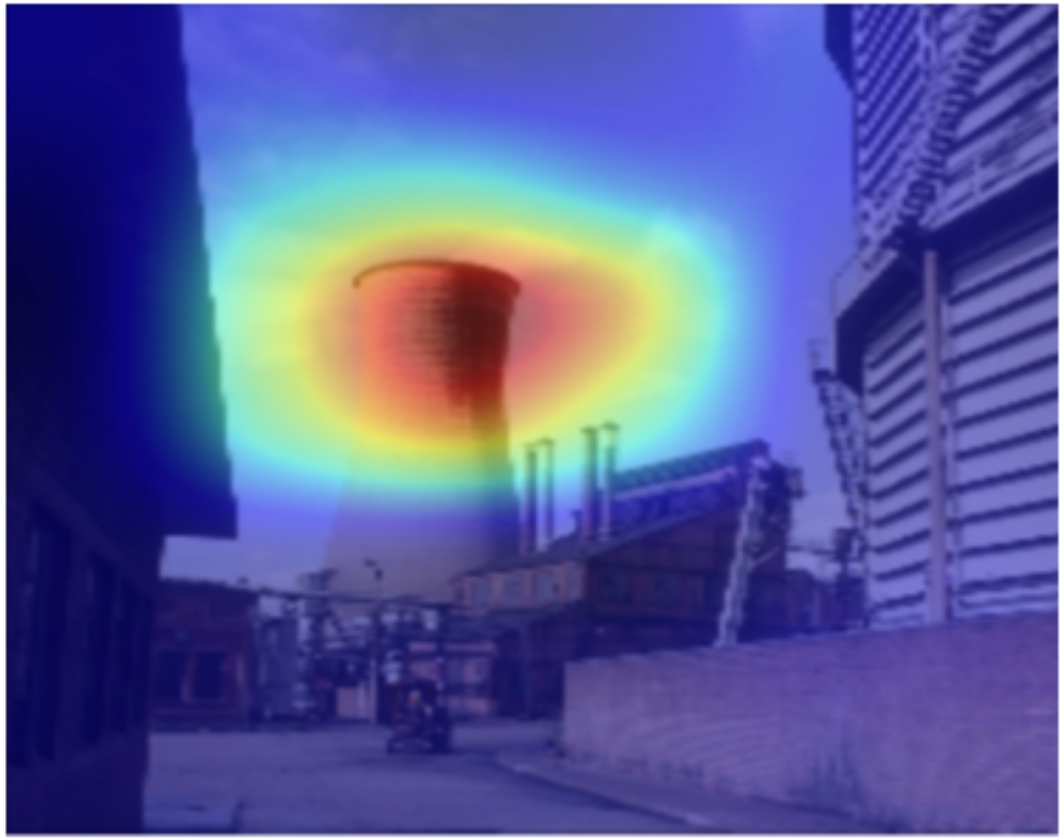


Not a power plant

Plant not operating

Plant operating

…but will always make at least some errors.



$p\ cooling\ tower = 1.00000$

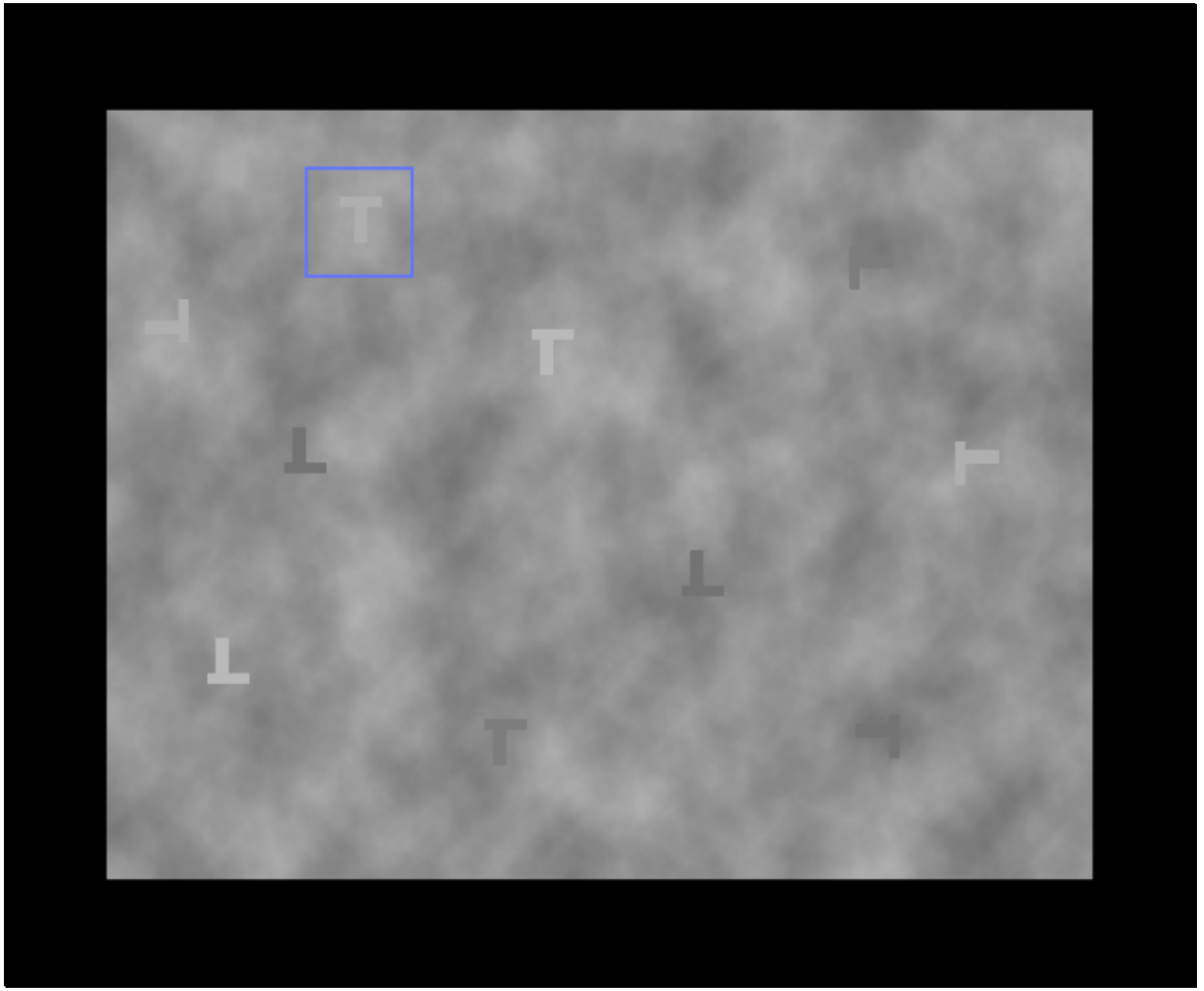
$p\ cooling\ tower = 0.96455$

Key Research Questions

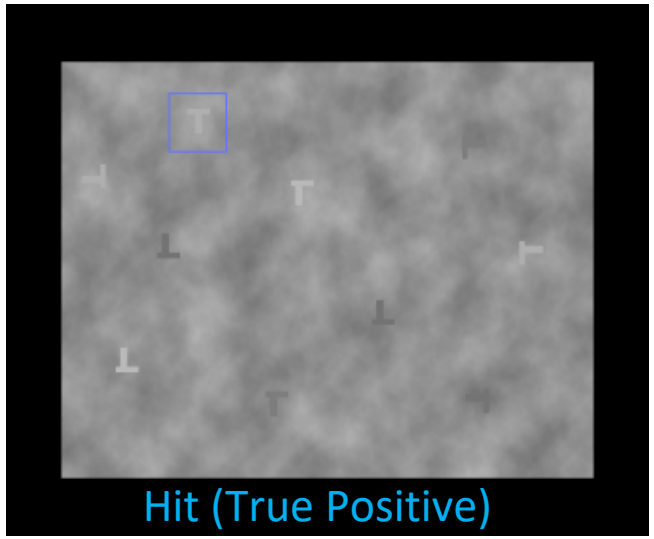
For the implementation of ML models to be effective, we need a better understanding of the impact of AI/ML errors on human users

- When and how do errors in AI/ML outputs lead to errors in human assessments?
- What factors make it easier or harder for people to recognize errors?
- How do people develop appropriate levels of trust in the outputs?
- What level of accuracy in the model outputs is necessary to support acceptable levels of human/system performance?

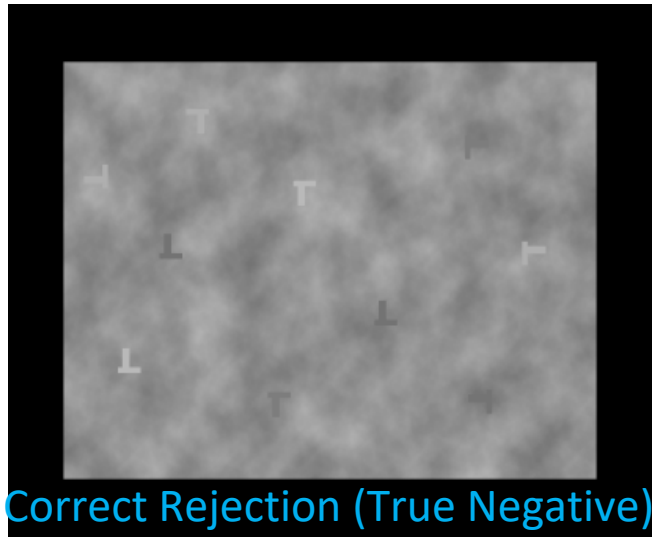
Experiments focused on object detection in imagery



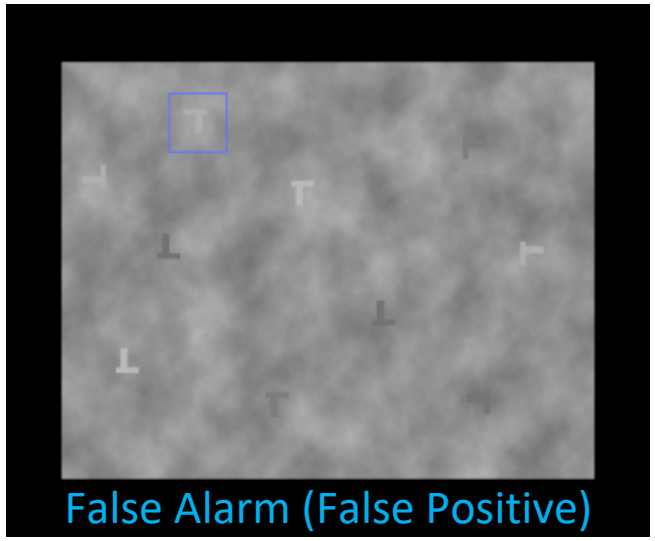
Experiments focused on object detection in imagery



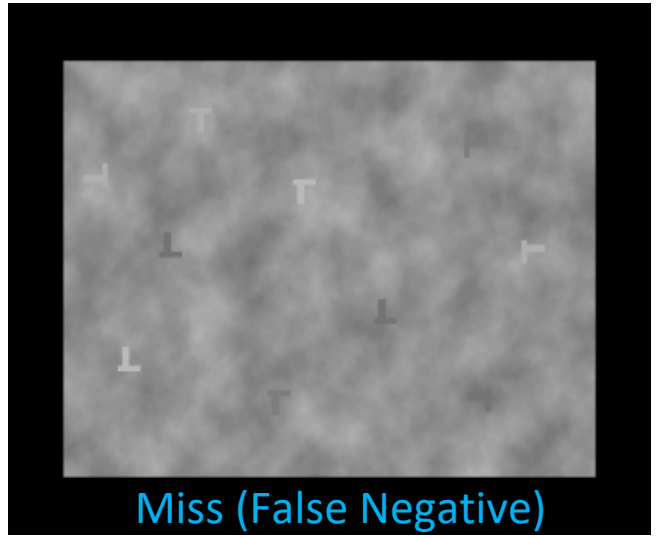
Hit (True Positive)



Correct Rejection (True Negative)



False Alarm (False Positive)



Miss (False Negative)

Hit (True Positive)

Correct Rejection (True Negative)

False Alarm (False Positive)

Miss (False Negative)

Experiments focused on object detection in imagery

- Domain general (T&L Task):
  - Error Rate - How accurate is the model?
  - Error Type - What are the most prevalent types of errors?
  - Error Importance - Which types of ML errors are most important?

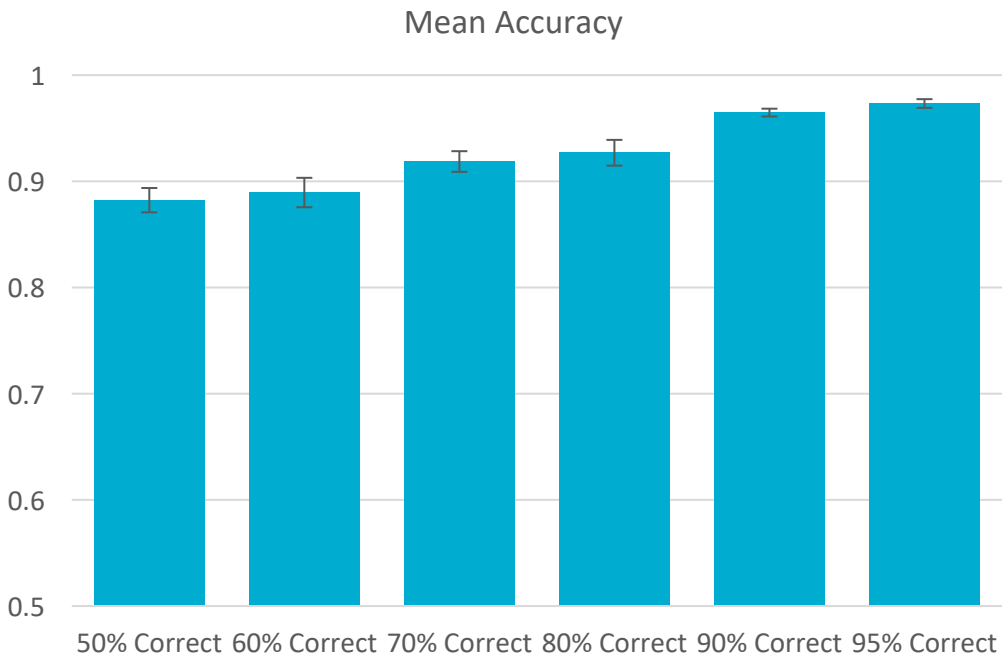
- Domain specific (Cooling Tower Task):
  - Expertise – Domain experts versus novices

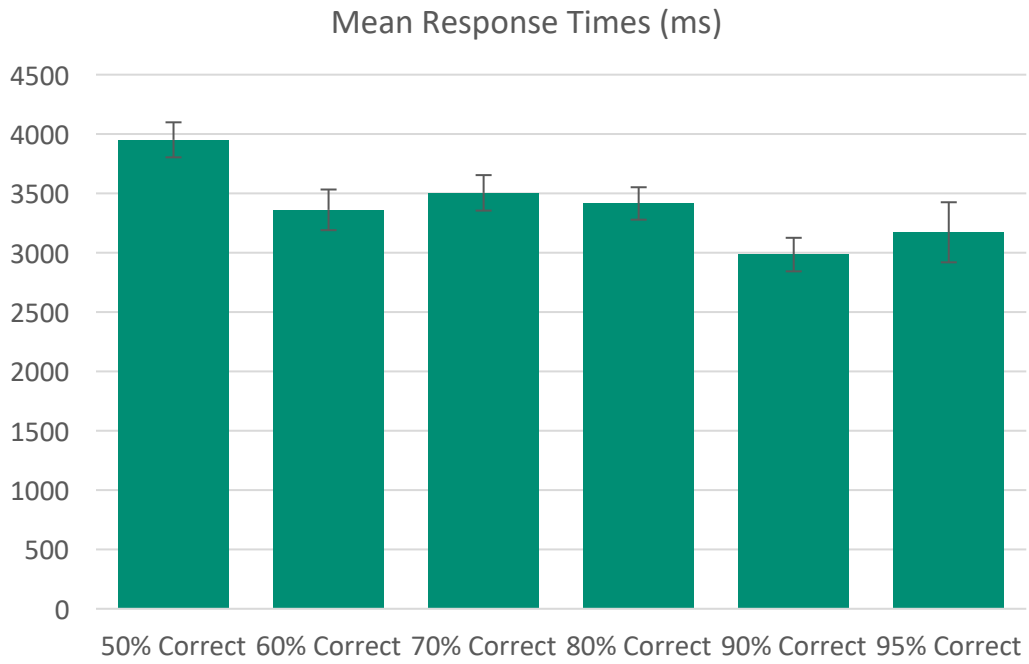
Error Rate Experimental Manipulations

- Model outputs are 50%, 60%, 70%, 80%, 90% or 95% accurate
- Equal numbers of three error types: Misses, False Alarms, and Miss/FAs
- 210 participants

Error Rate Experiment Results

Having a more accurate model is good!

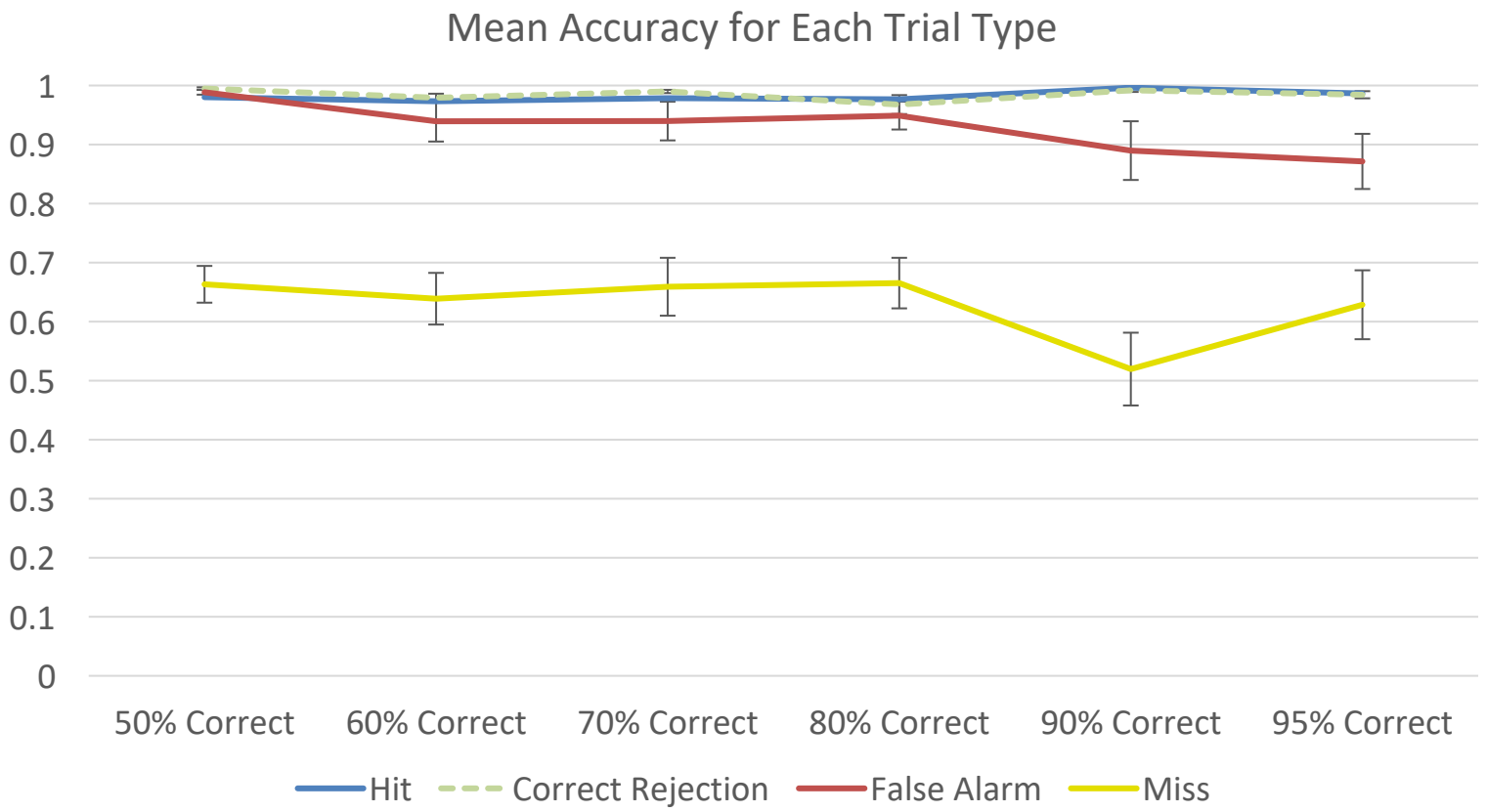




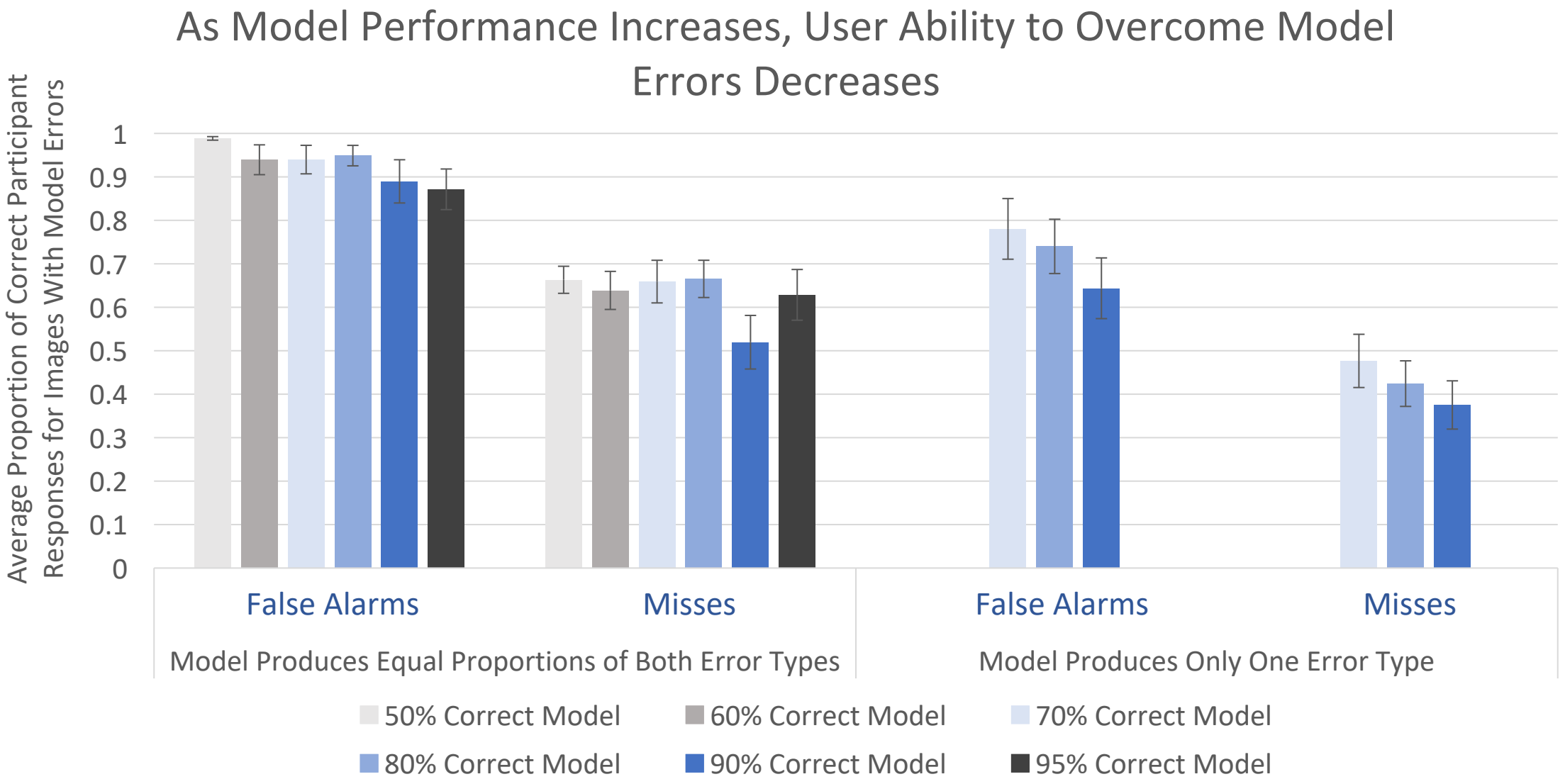
Error Rate Experiment Results

Having a more accurate model is good!

…except when the model makes errors

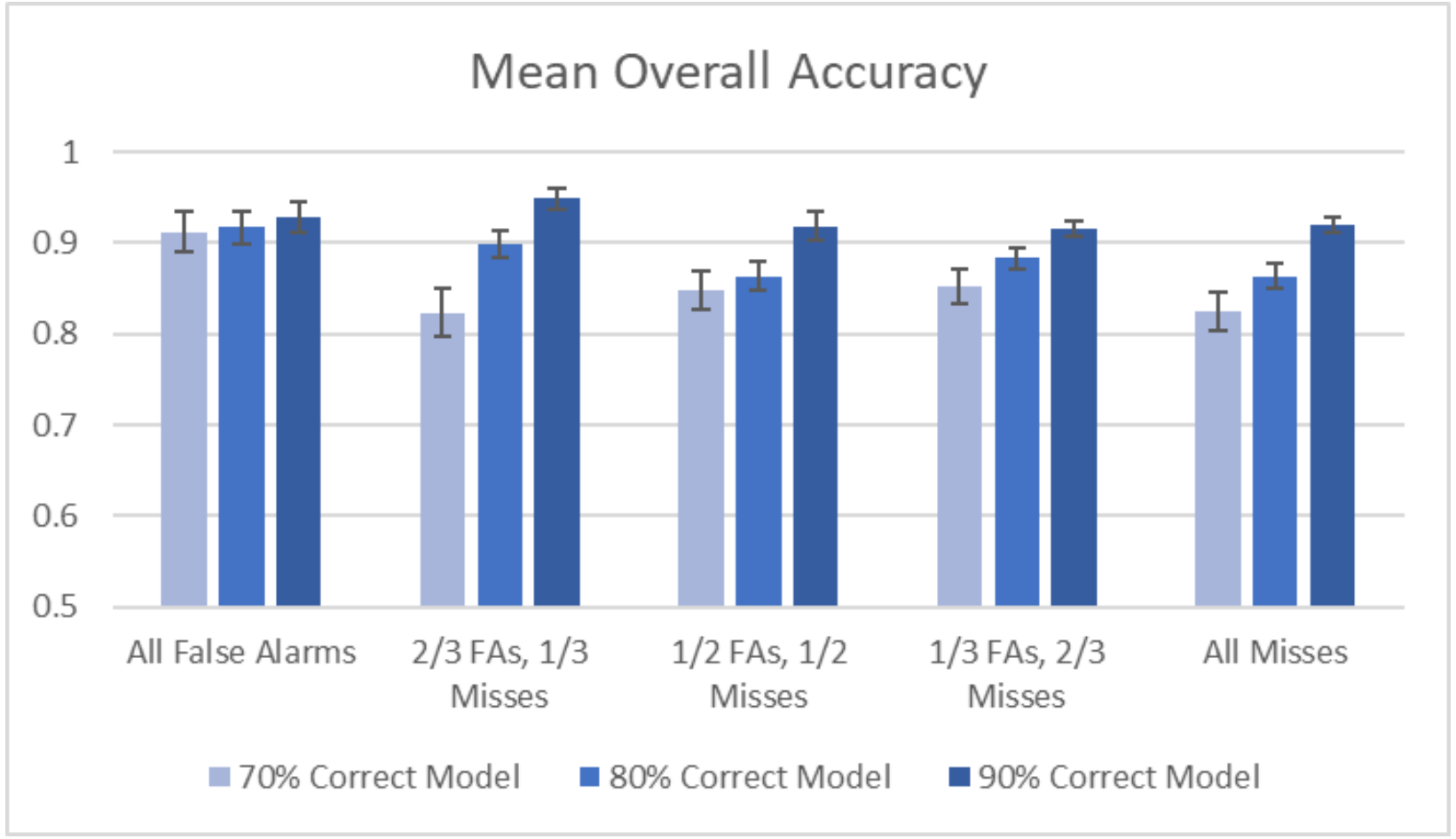


Error Rate Experiment Results



Error Type Experiment Results

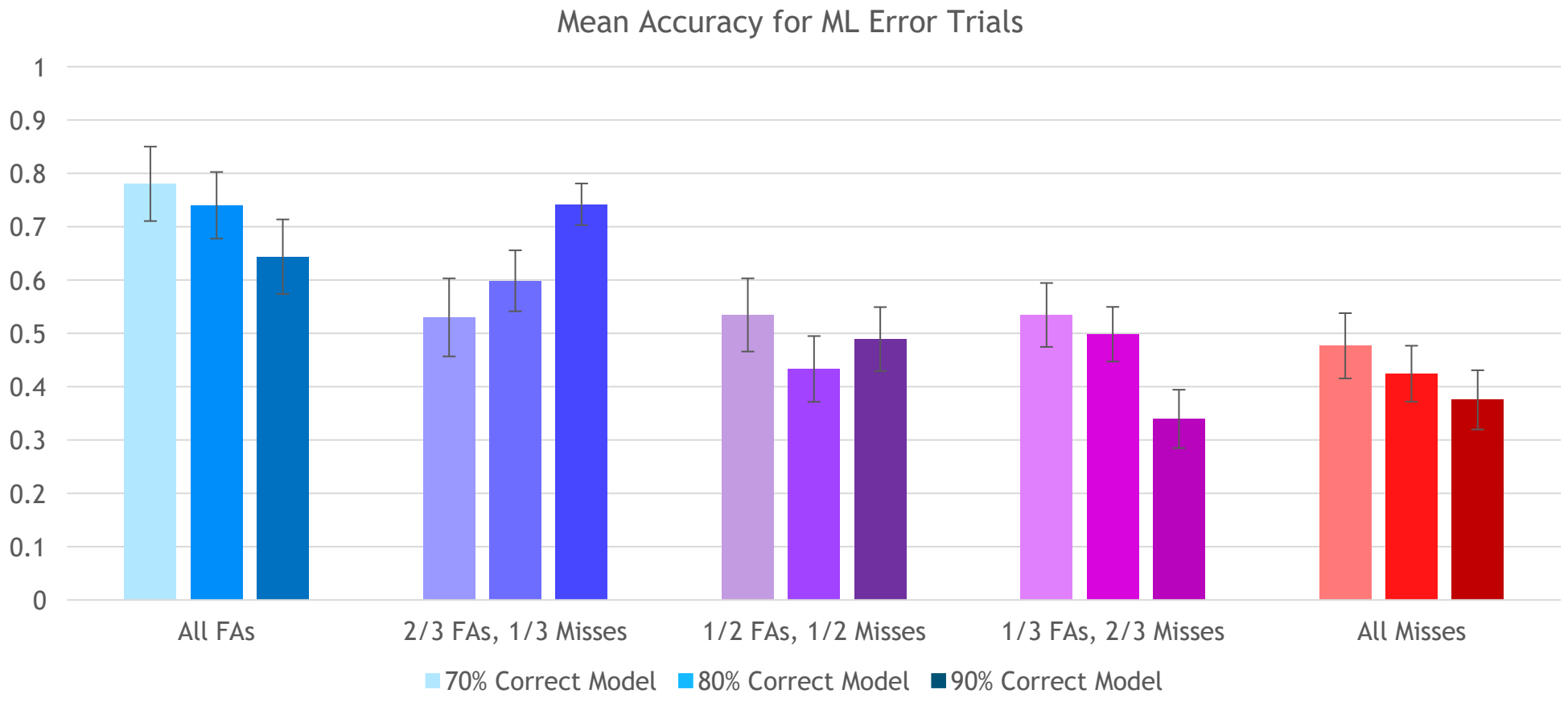
Again, having a more accurate model is good!



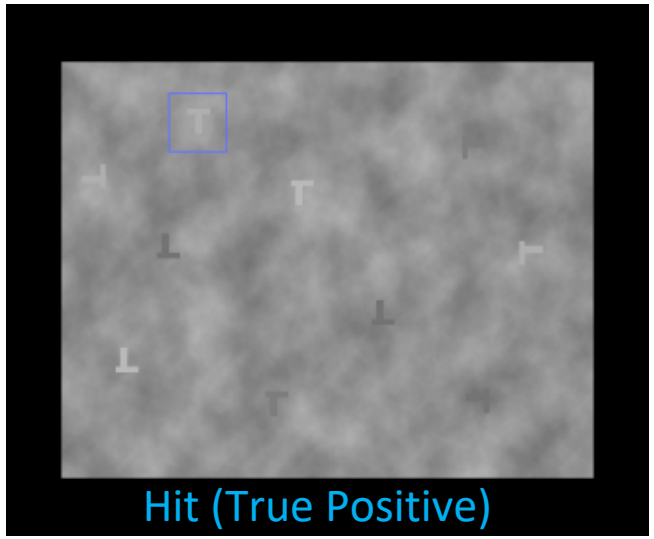
Error Type Experiment Results

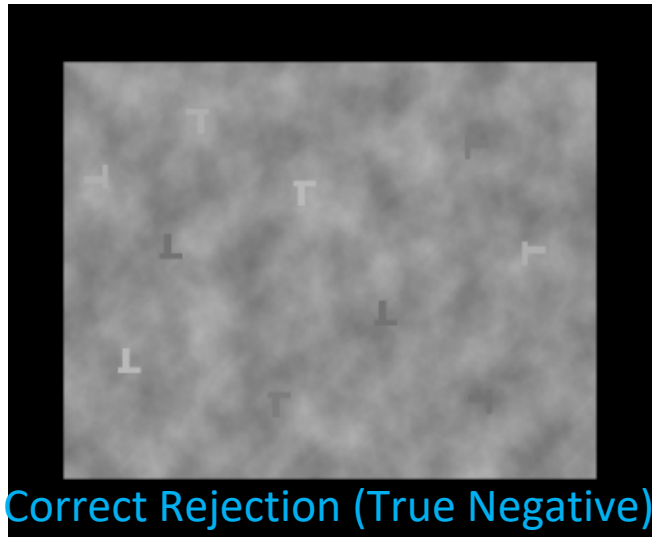
Again, having a more accurate model is good!

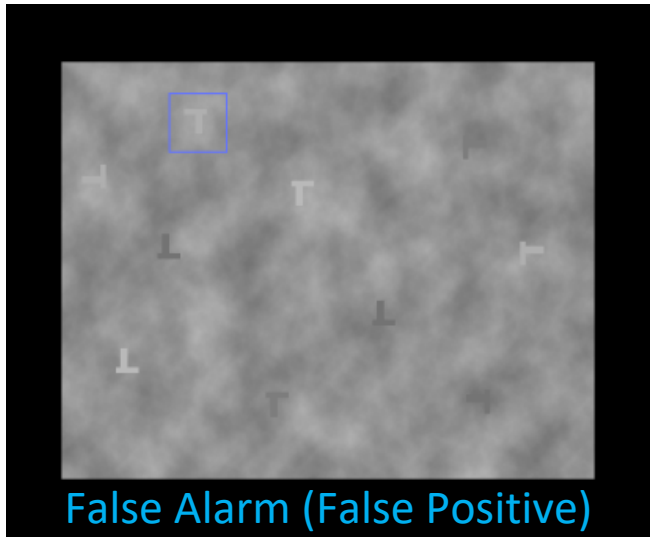
…but as the model gets better, people become less likely to notice model errors

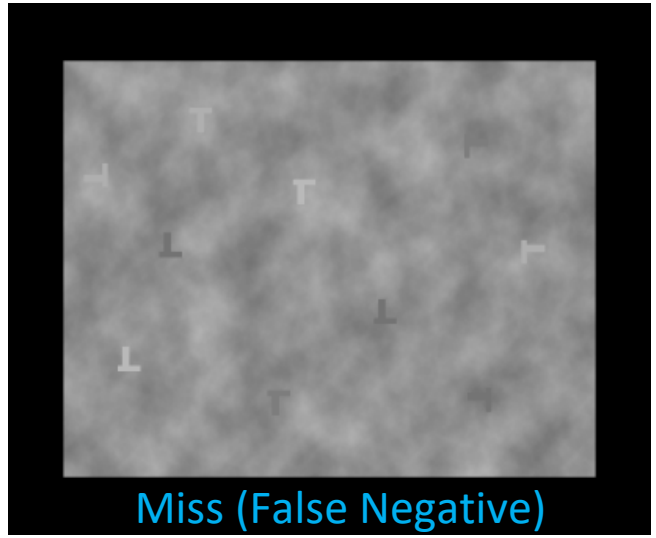


Experiments focused on object detection in imagery


Hit (True Positive)


Correct Rejection (True Negative)


False Alarm (False Positive)


Miss (False Negative)


Hit (True Positive)


Correct Rejection (True Negative)

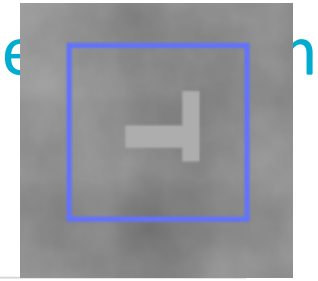

False Alarm (False Positive)

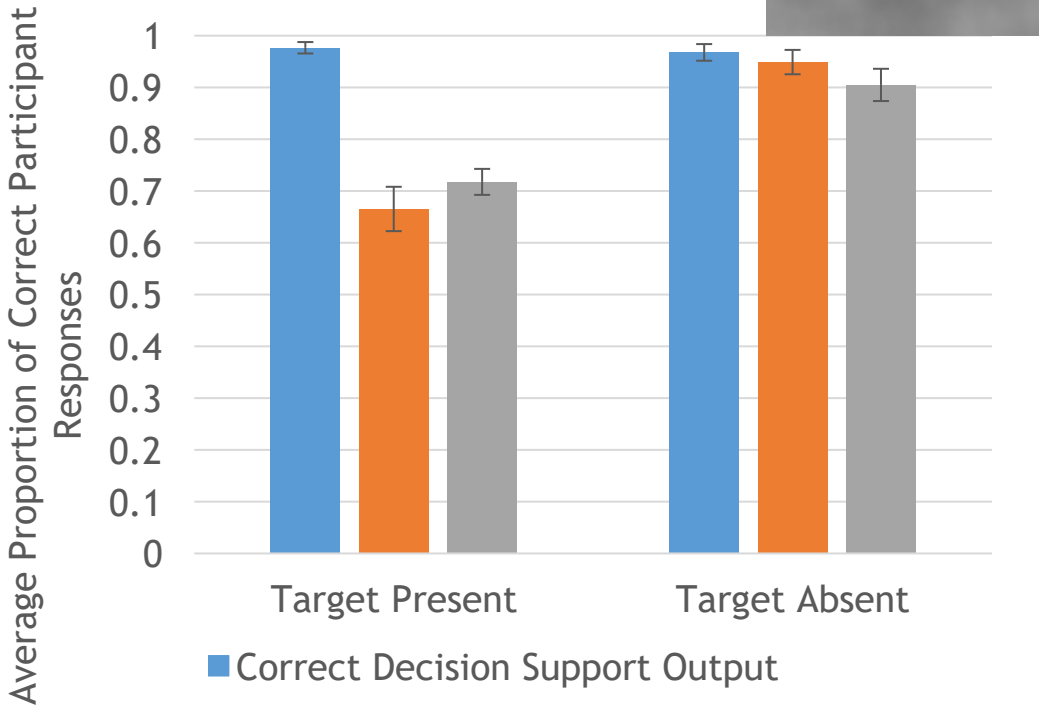

Miss (False Negative)

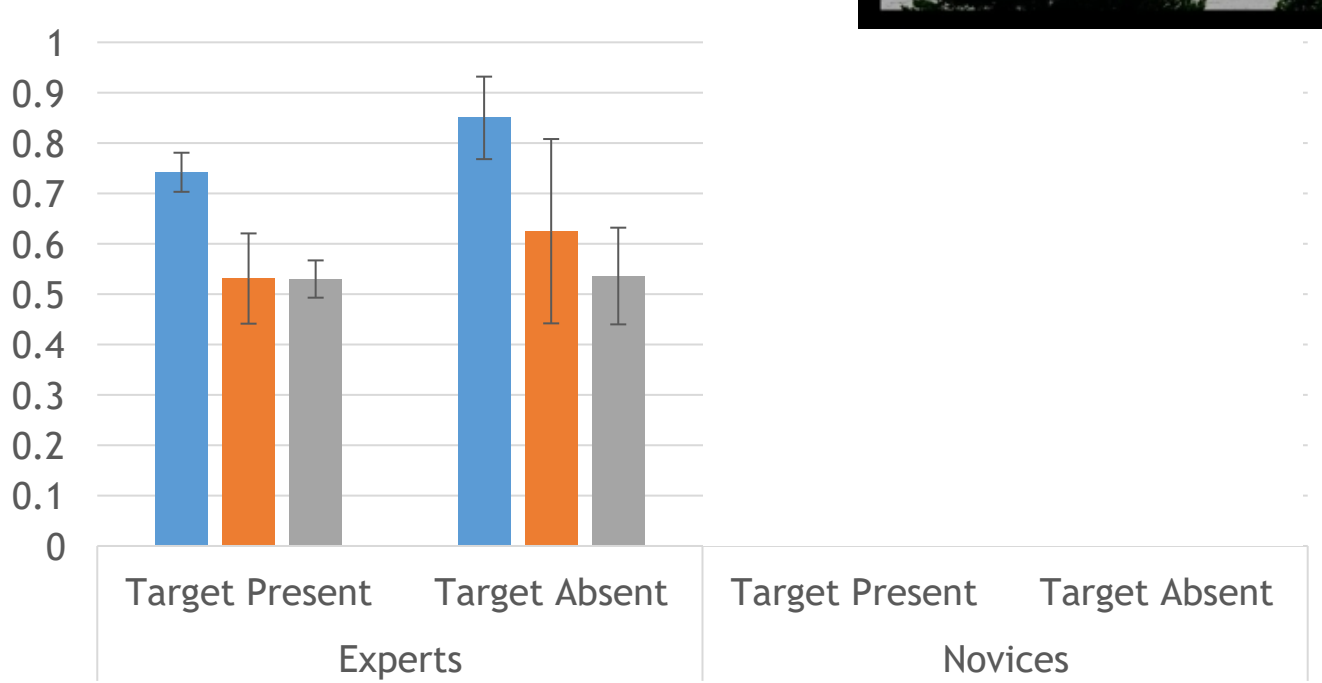
Model output was accurate 80% of the time

Domain Specific Experiment Results

- Incorrect model outputs don't hurt expert performance (just
- Novices were more like nply with incorrect model out



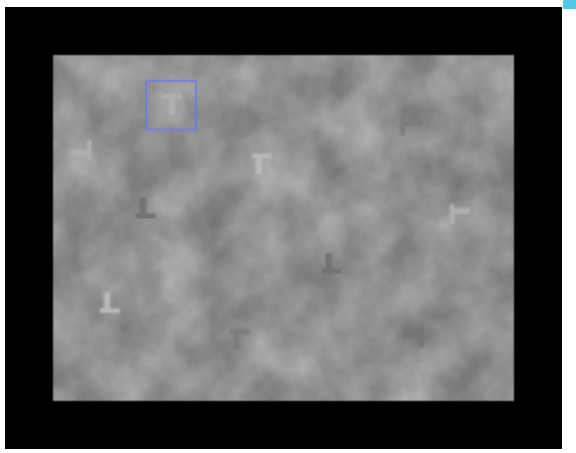




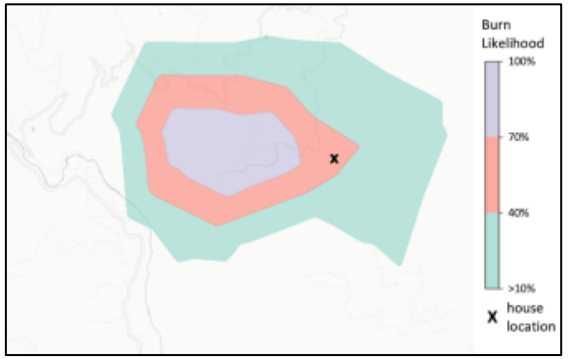
Factors that Impact Decision Making

A few examples:

- Visual search aided by (mock) ML outputs
  - People get complacent as the overall accuracy of the outputs goes up
  - Novices are more likely to go along with what the ML says



- Visualizations of uncertain information
  - Differences between visual and numerical representations
  - The specificity of the information can impact judgments of risk
  - The same information visualized in different ways can lead to different patterns of decisions
  - Individual differences also impact decisions



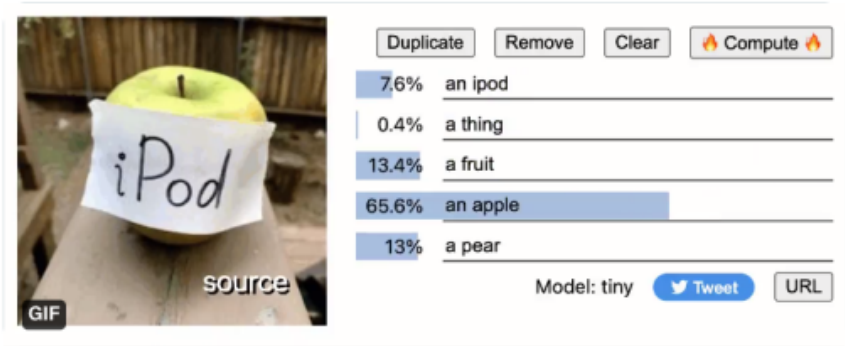
Visualizations of Uncertain Information

State uncertainty is uncertainty about the current or future state of some phenomenon

- Very common in AI and ML outputs!

Humans are notoriously bad at understanding state uncertainty and probability

- Different representations of the same information may push people to make different decisions



https://ai.googleblog.com/2022/04/locked-image-tuning-adding-language.html

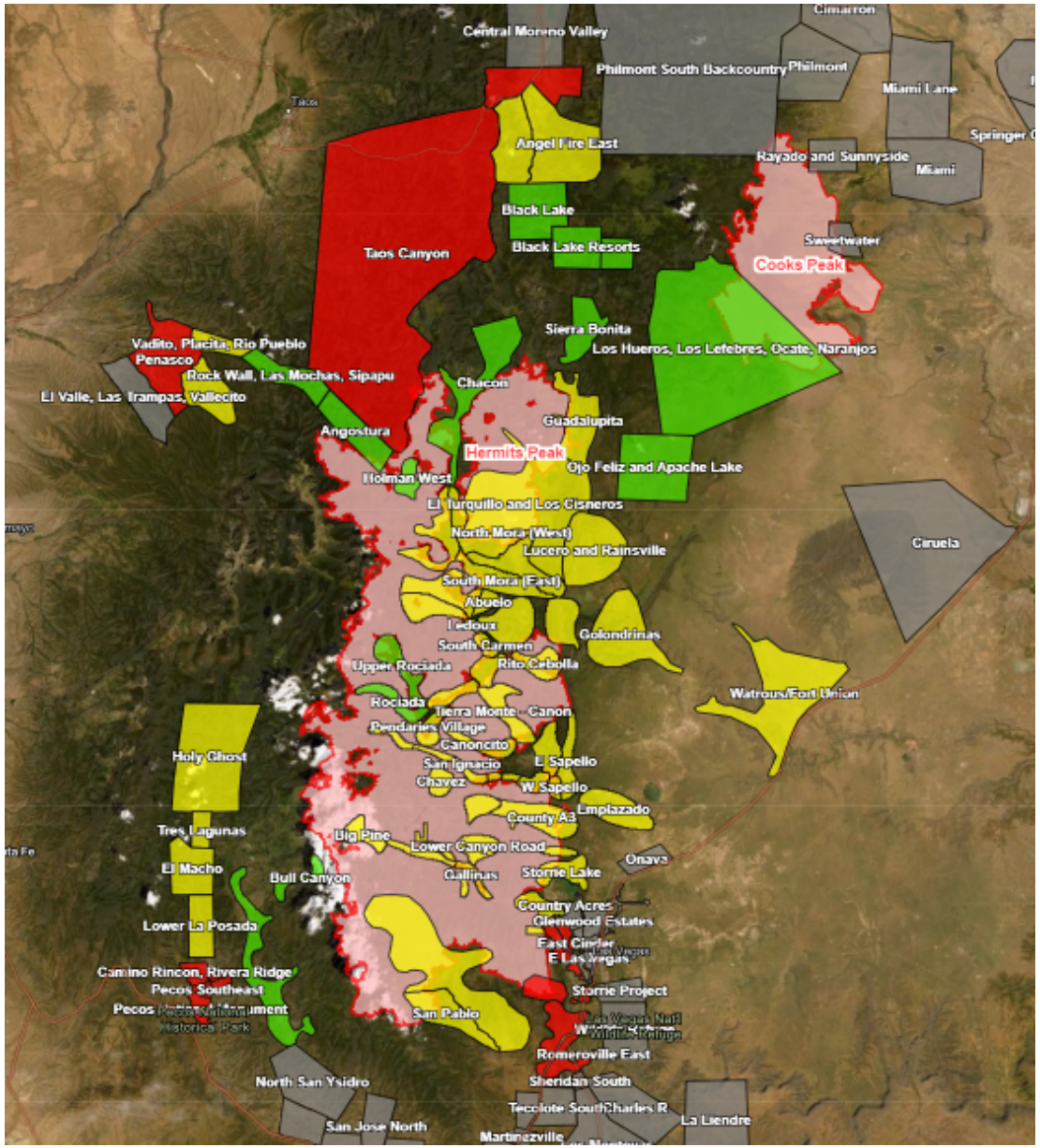
Specificity and Perceptions of Risk

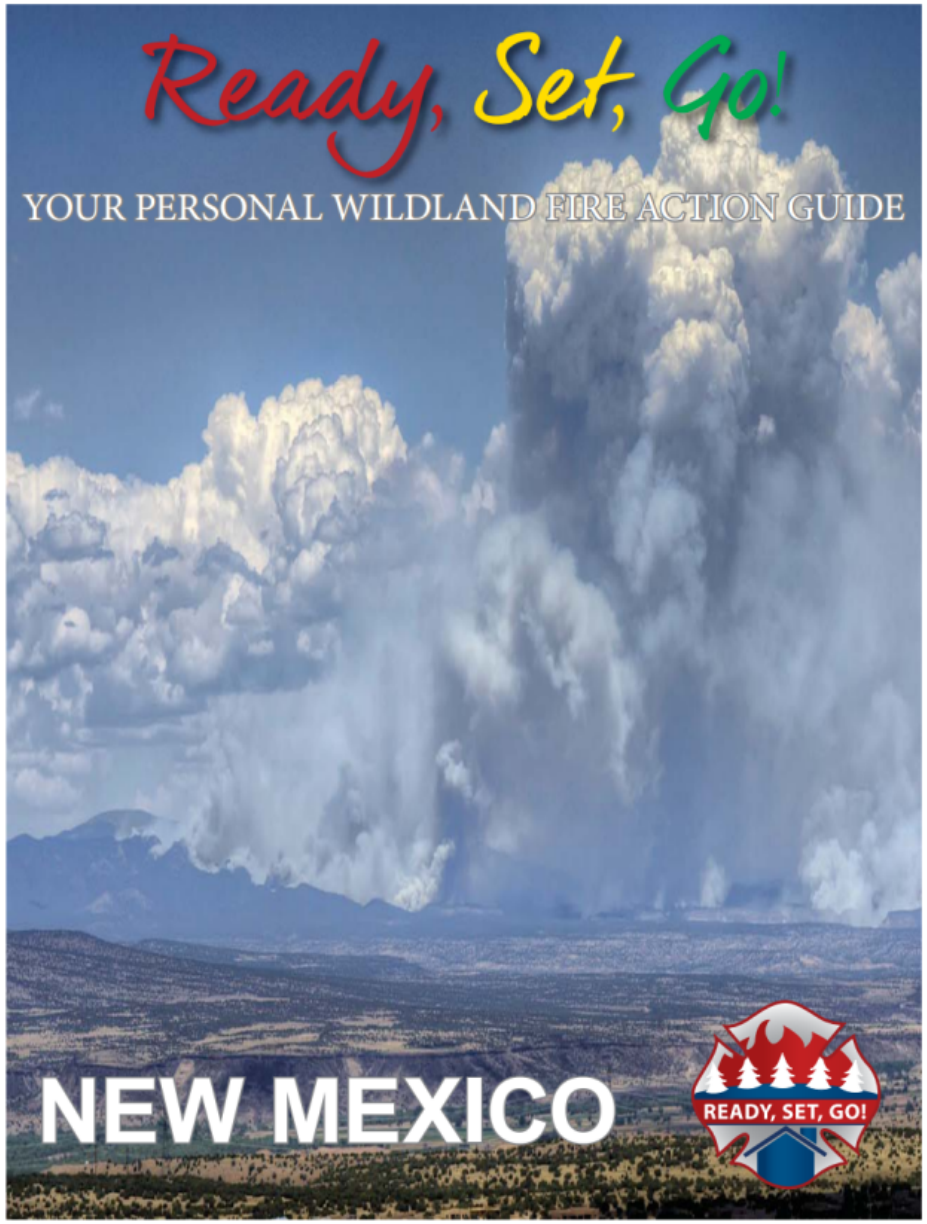
- Three prior studies (Bisantz et al., 2005; Bisantz et al., 2011; Cheong et al., 2016) suggest that more specific information about uncertainty can produce more risk -averse decisions
  - e.g., 60-70% chance vs. 40-70% chance

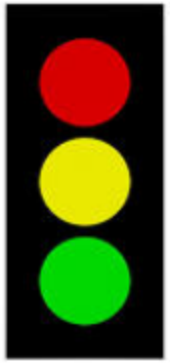
- Patterns of decision making changed when people saw visualizations of uncertainty instead of numeric or linguistic expressions
  - People *seem* to treat visual cues as having higher levels of specificity than numeric or linguistic expressions
  - They make more risk-averse decisions when given visualizations

- Why would that happen?
  - Do people treat visualizations of uncertainty as if they are deterministic?
  - Does a visualization make it easier to imagine the risks?
  - Do different visual cues make a difference?

Wildfire Evacuation Task

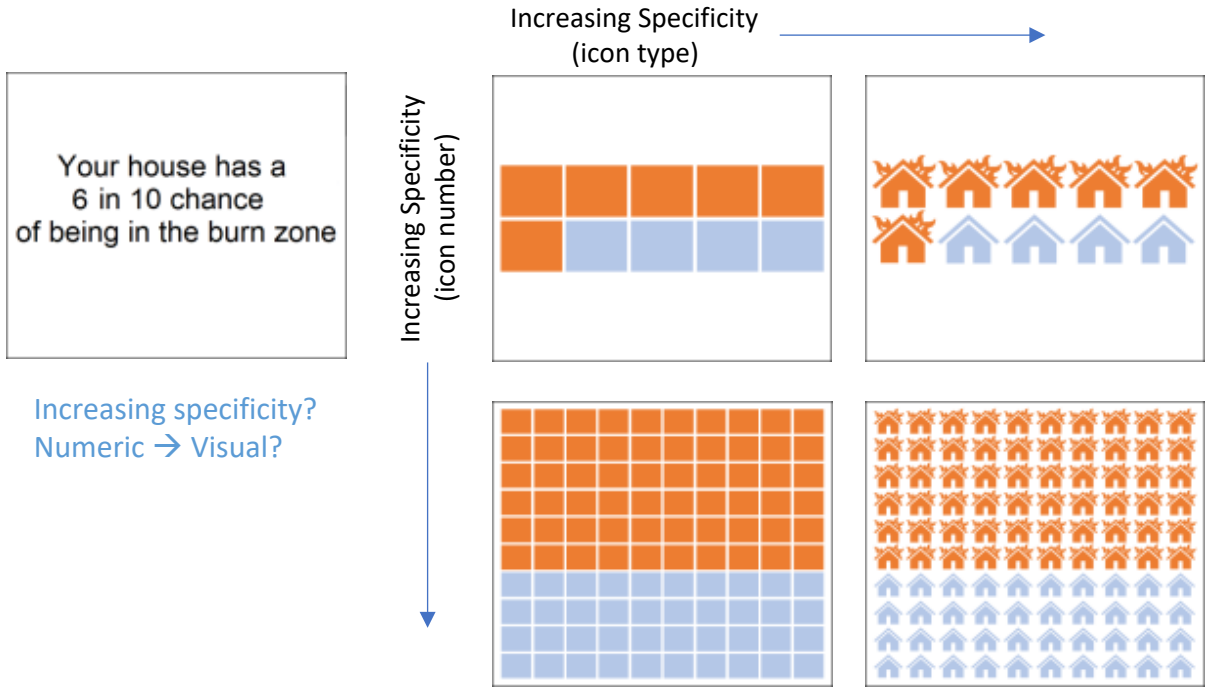


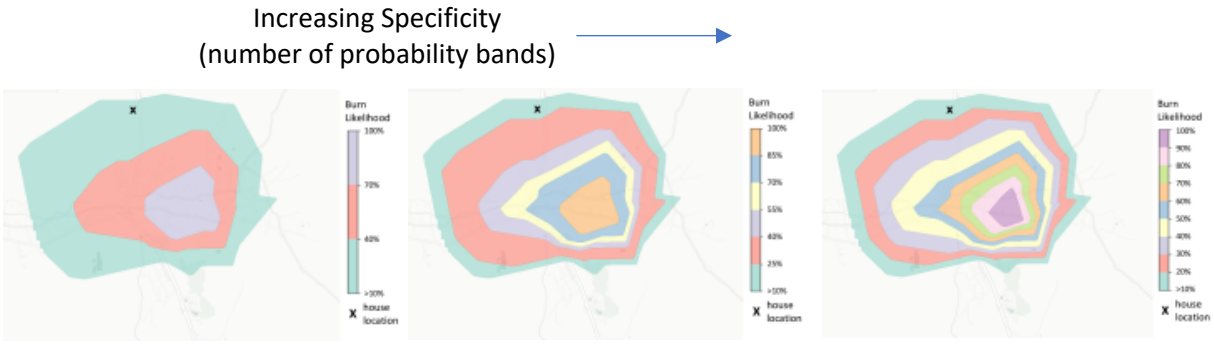




Wildfire Evacuation Task

- Show people the probability that their house will burn down in a wildfire. Ask them if they will stay or evacuate. Evacuation costs money, staying in a house that burns down costs money, participants receive a real money based on their decisions
  - Does increasing specificity of the representation change their decision threshold (does it make them more risk-averse?)
  - Do we get the same effect when we compare visual representations to numeric ones?
    - Does visualizing the info increase the apparent specificity?



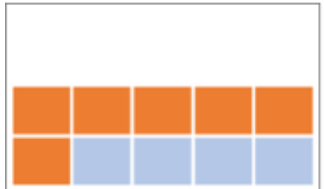
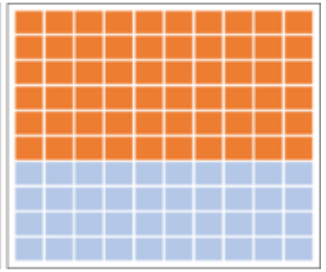
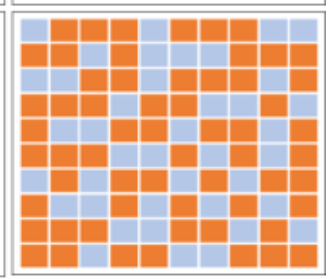
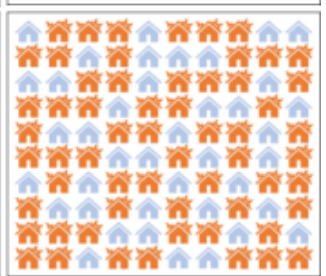


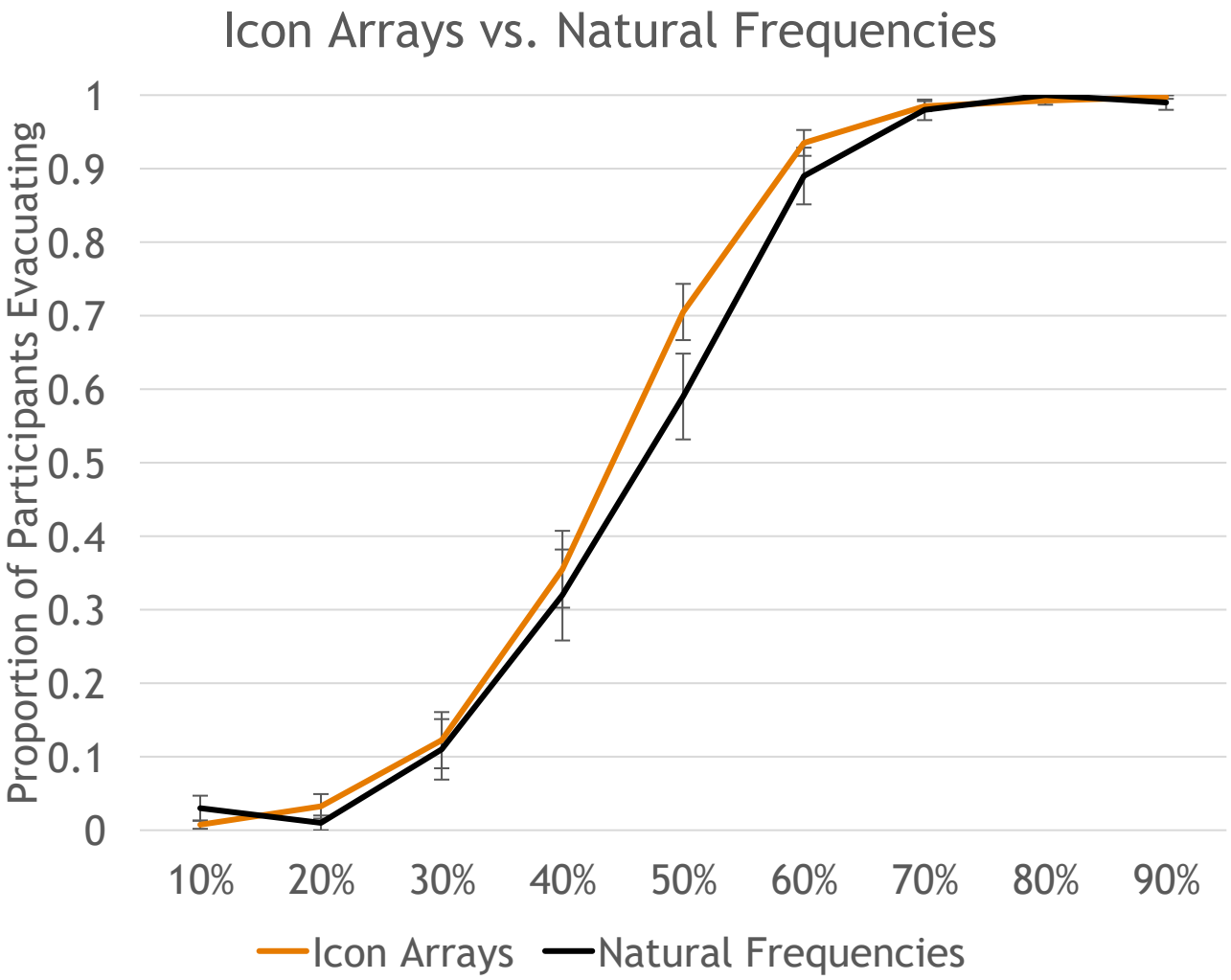
Icon Array Experiment

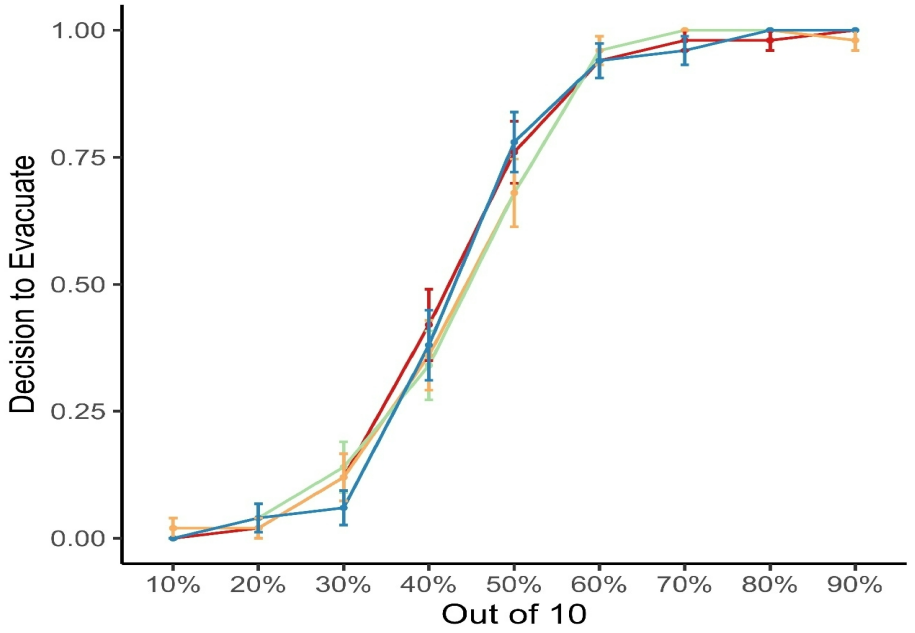
- Widely used to support risk communication in medical decision making
  - Useful for helping people understand probability, especially people with lower numeracy skills

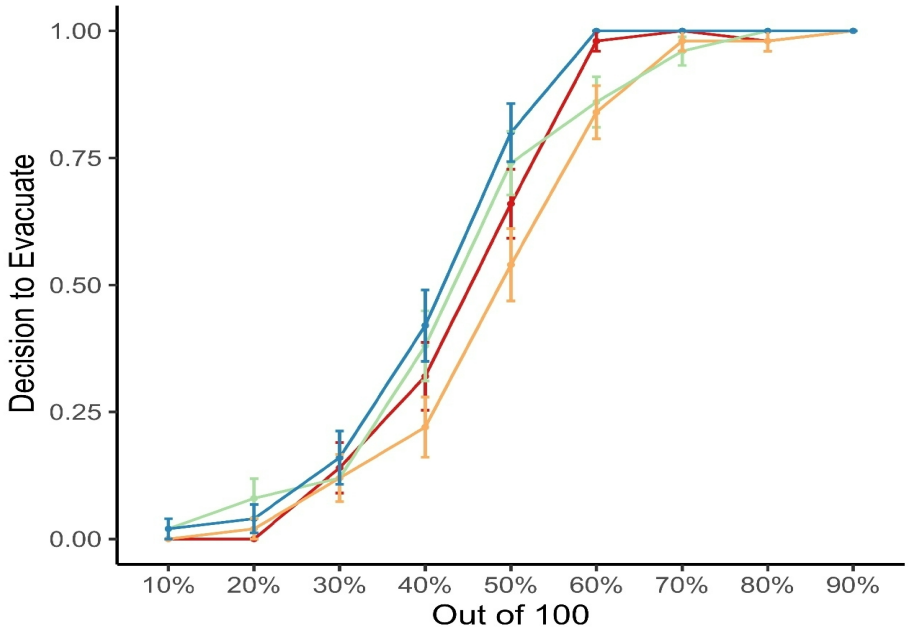
- Increasing the specificity of icon arrays:
  - Increase the number of icons (10 vs 100) to give more precise percentages

- Manipulations that might increase the perceived specificity of icon arrays:
  - Iconicity
  - Randomization of icons

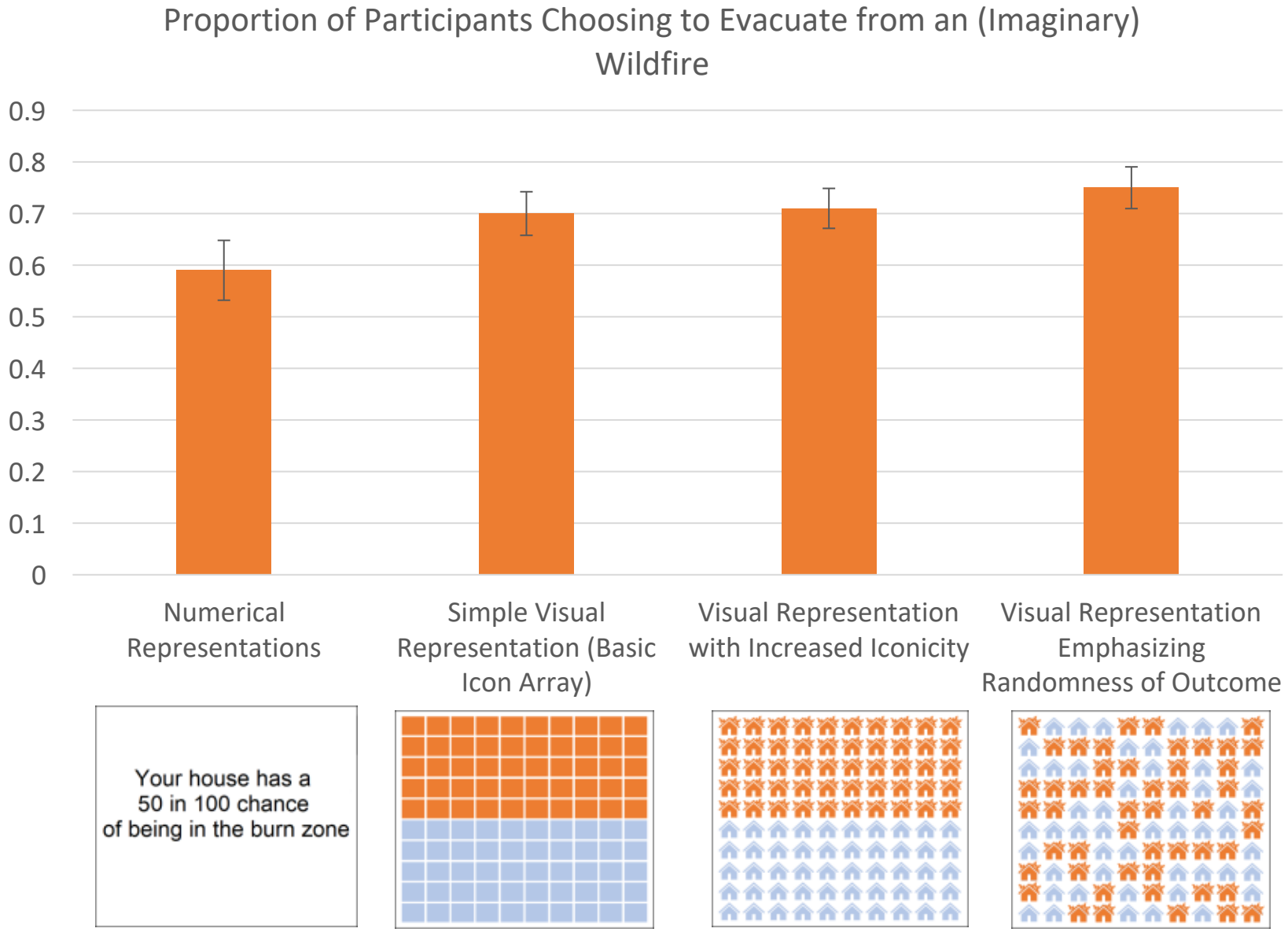
Your house has a 6 in 10 chance of being in the burn zone	Your house has a 60 in 100 chance of being in the burn zone
	
	
	
	


Icon Arrays vs. Natural Frequencies
Proportion of Participants Evacuating
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
10%
20%
30%
40%
50%
60%
70%
80%
90%
Icon Arrays
Natural Frequencies

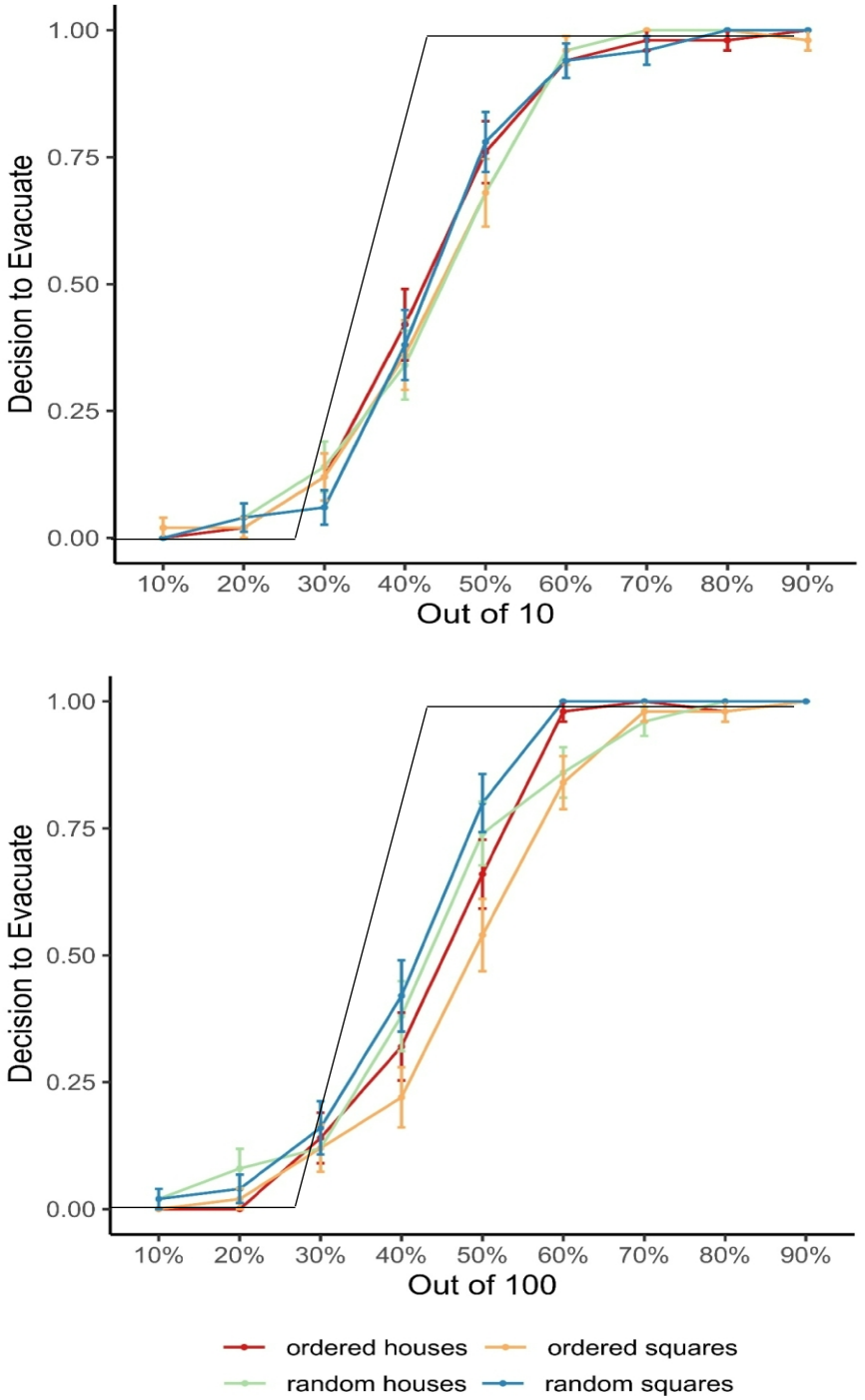

Decision to Evacuate
1.00
0.75
0.50
0.25
0.00
10%
20%
30%
40%
50%
60%
70%
80%
90%
Out of 10


Decision to Evacuate
1.00
0.75
0.50
0.25
0.00
10%
20%
30%
40%
50%
60%
70%
80%
90%
Out of 100
ordered houses
ordered squares
random houses
random squares

Results – 50% probability

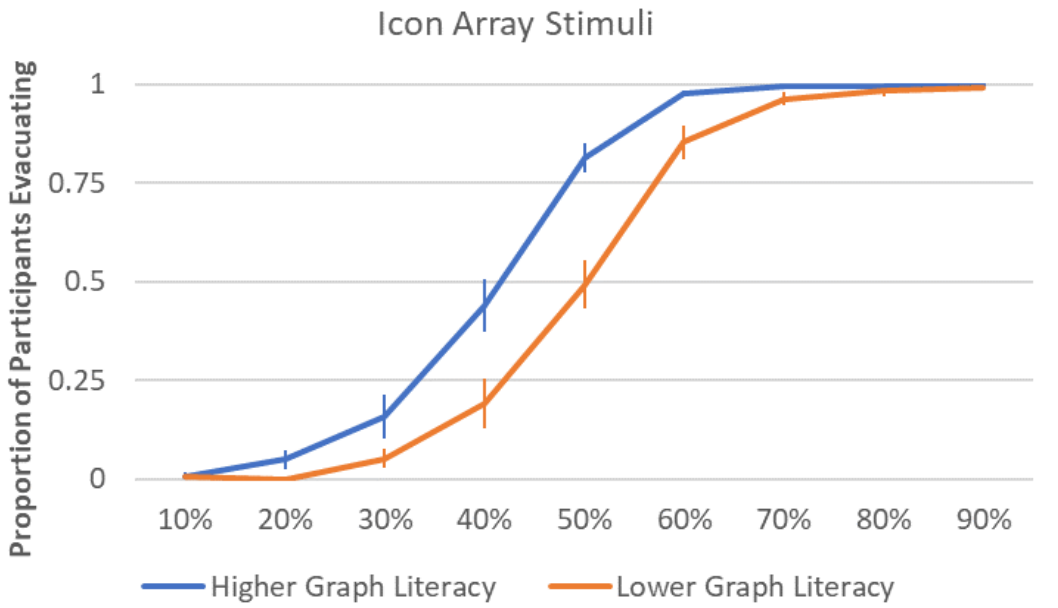


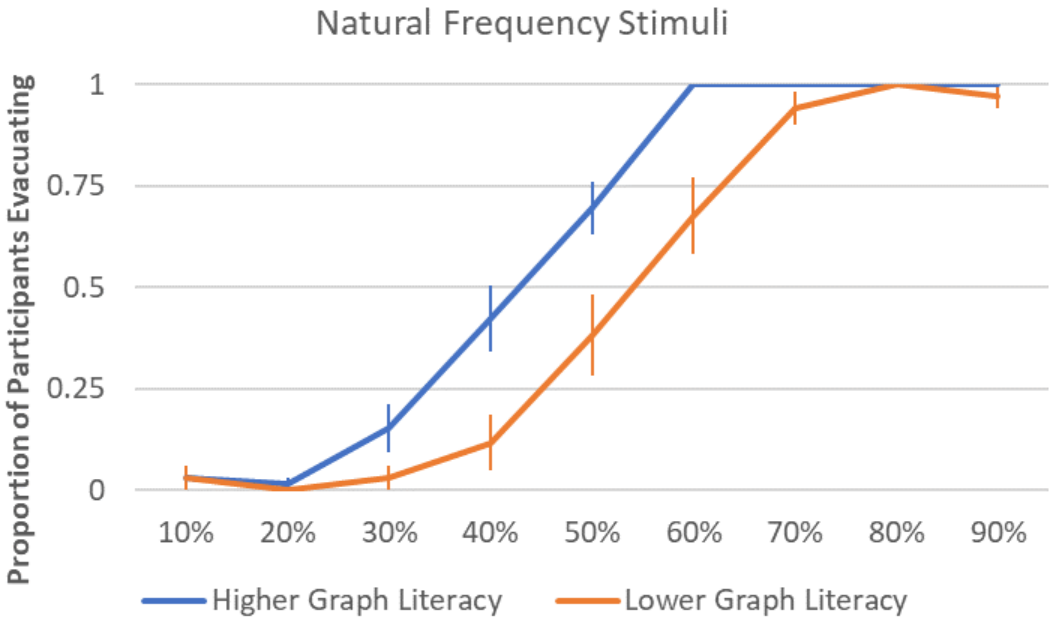
- Results:
  - People were more likely to evacuate when shown icon arrays than when shown natural frequencies
  - People more likely to evacuate when shown houses than when shown squares
  - People more likely to evacuate when shown randomized icons than when shown ordered icons
  - No differences for different numbers of icons



Individual Differences

- Objective Numeracy Scale
- Subjective Numeracy Scale
- Short Graph Literacy Scale
- Risk Propensity Scale
- Need for Cognition



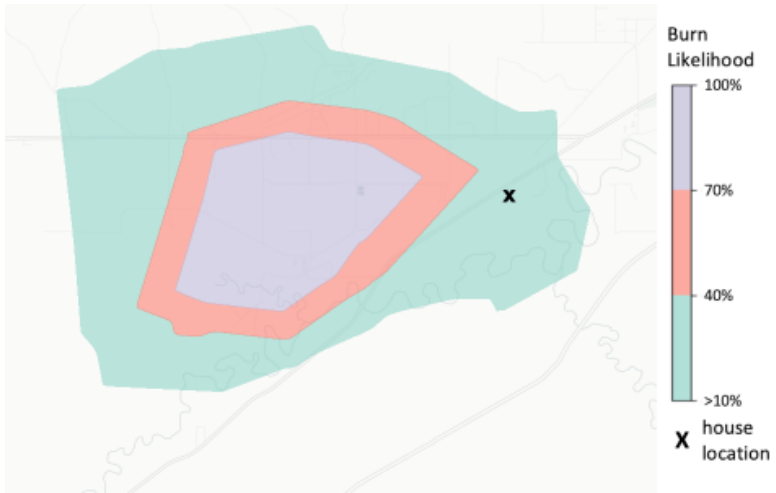


Results

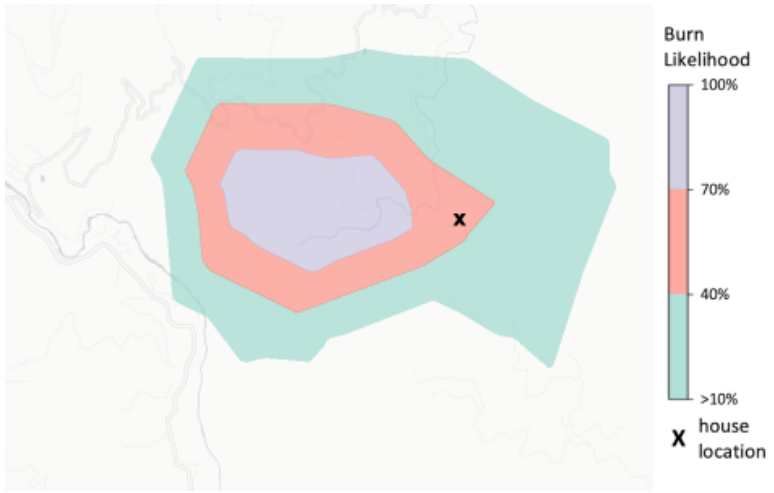
- The way in which state uncertainty is represented impacts decision making
  - More specific representations (i.e. “61 out of 100” instead of “6 out of 10”) tend to produce more risk-averse decisions
  - Visualizing uncertain information produces more risk-averse decisions than numerical representations with the same level of specificity
    - **People seem to interpret visualizations as if they are more specific than other types of representations**
  - The specifics of the visual encoding also matter
    - **Visual design choices can increase the *perceived* specificity of the visualization, which makes people *even more* risk averse**

Spatial Uncertainty – Maps

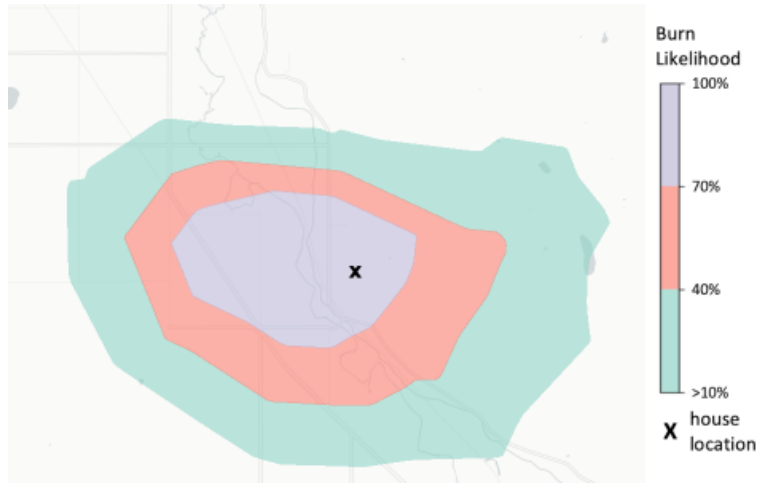
Your house is located in the
10 to 40%
burn likelihood zone.

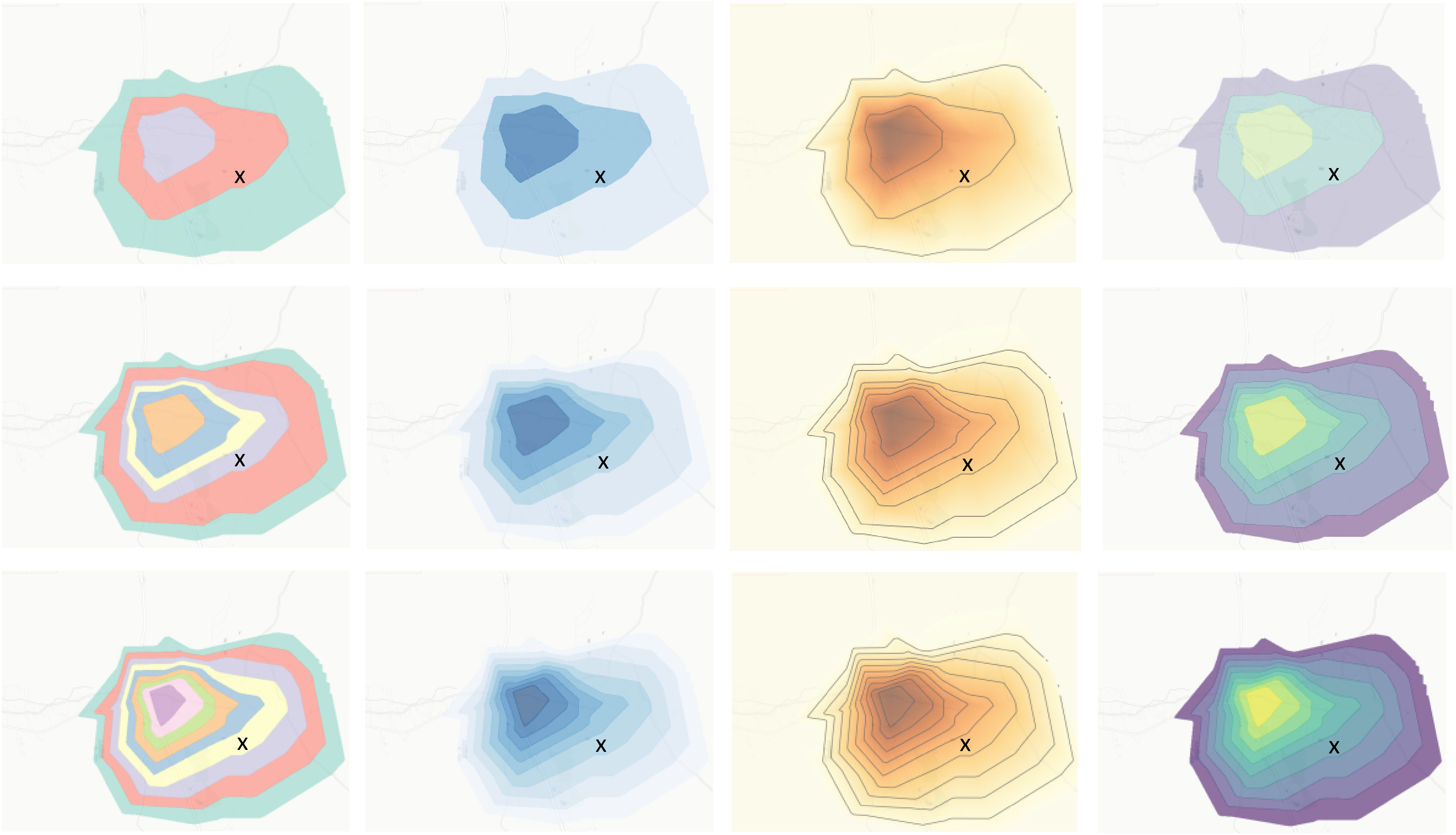


Your house is located in the
40 to 70%
burn likelihood zone.

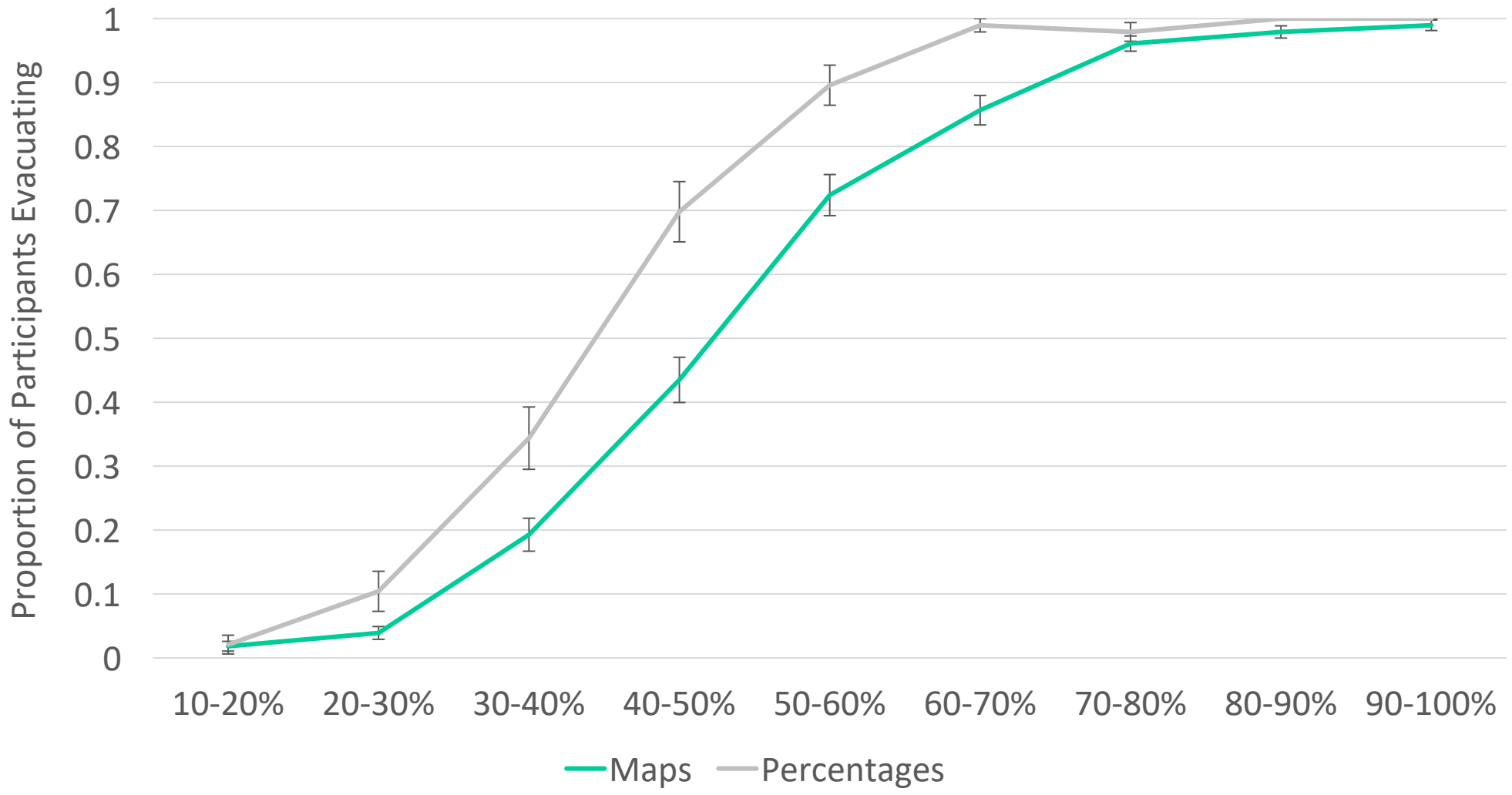


Your house is located in the
70 to 100%
burn likelihood zone.

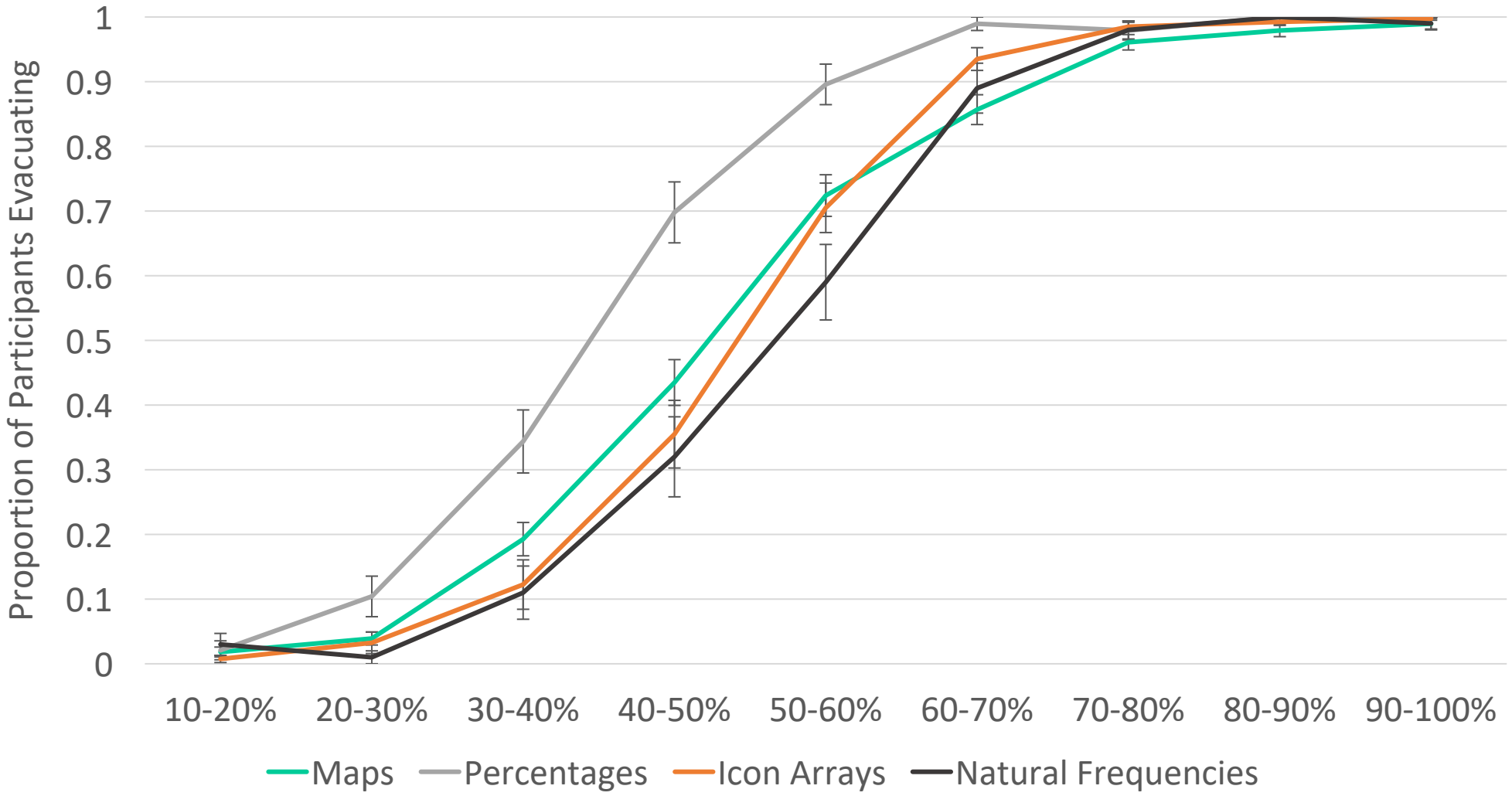




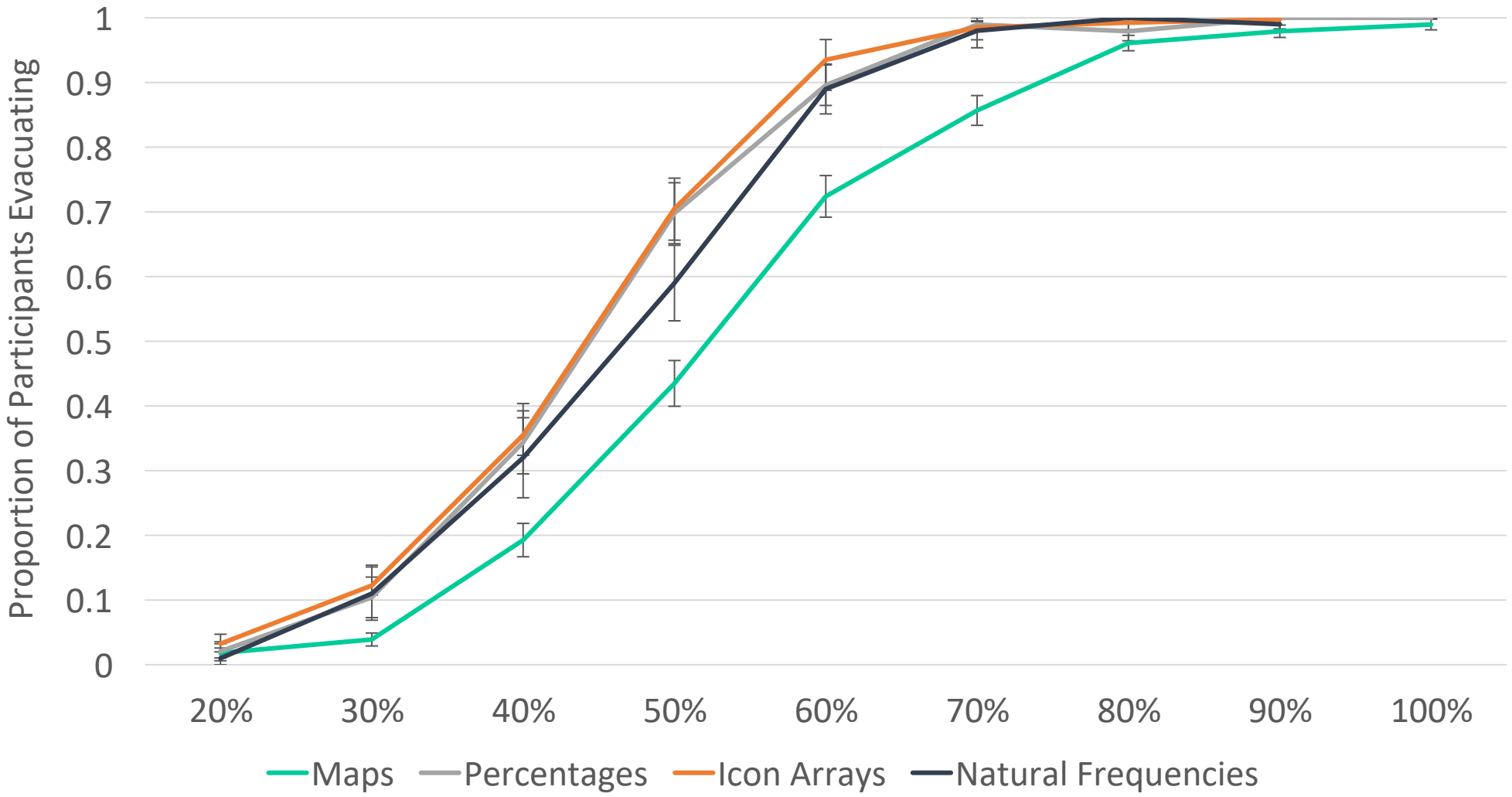
Map Experiment Results

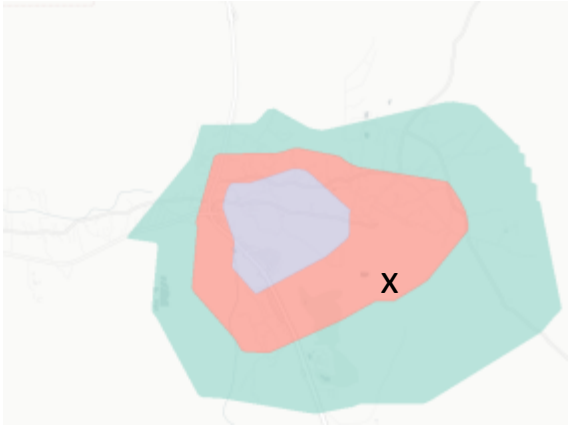


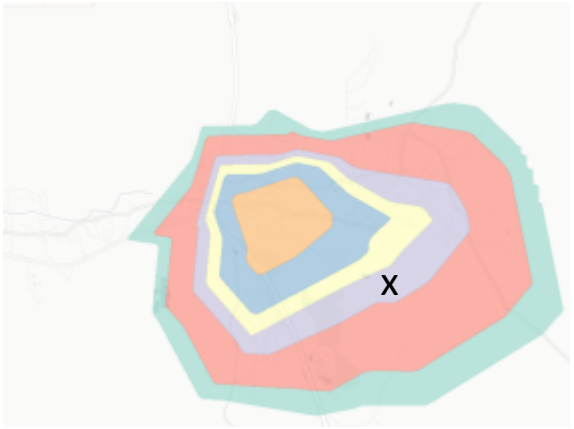
Map Experiment Results

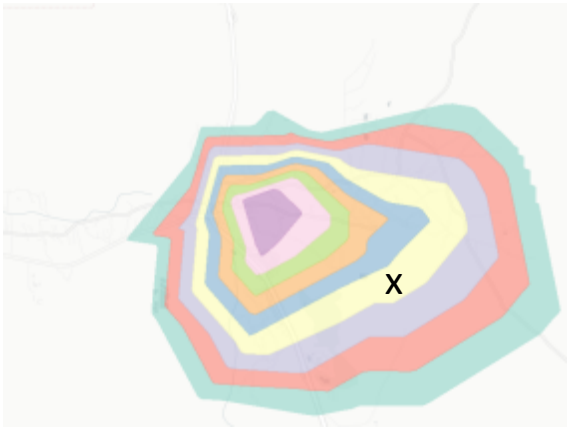


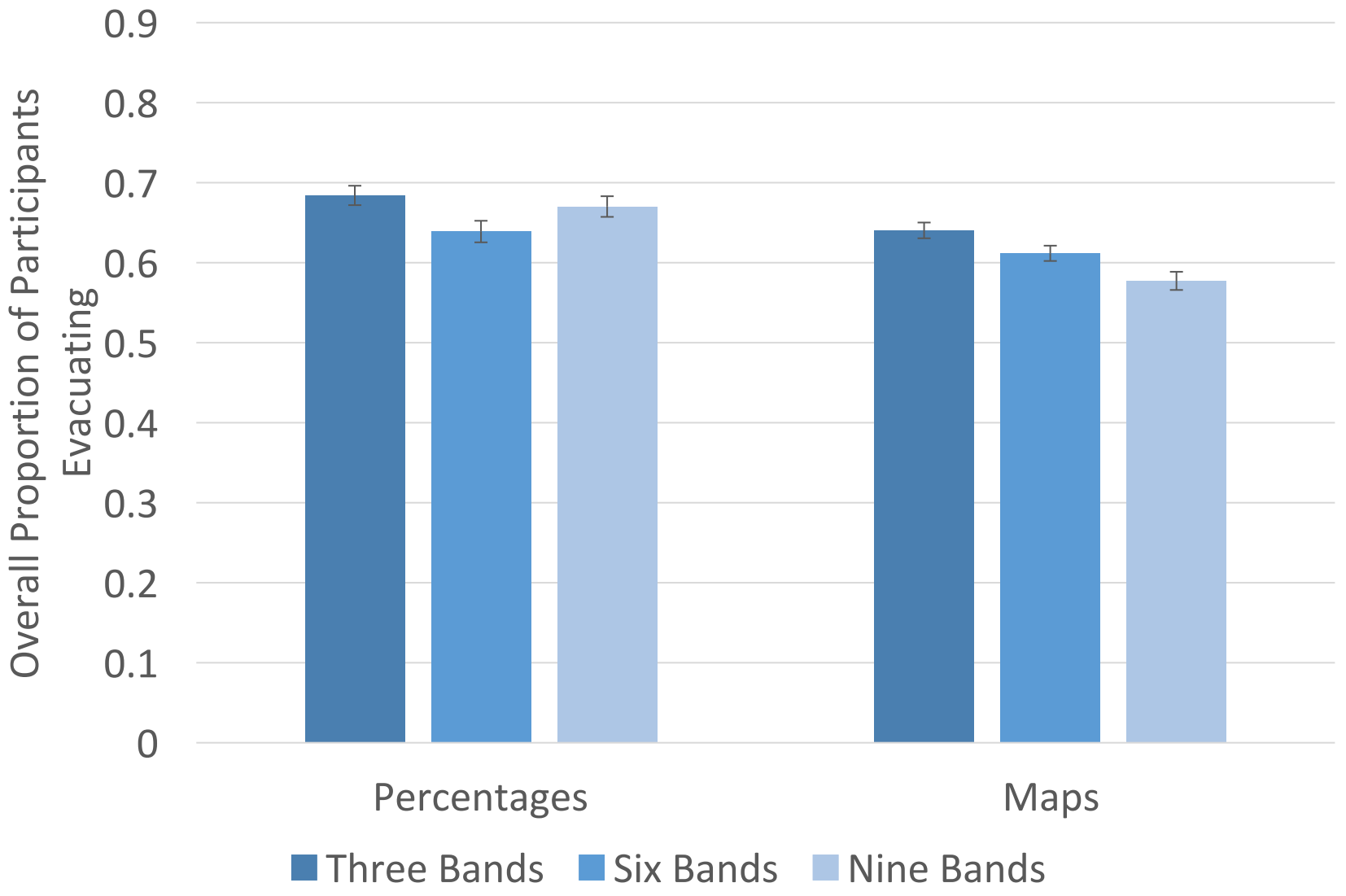
Map Experiment Results




X


X

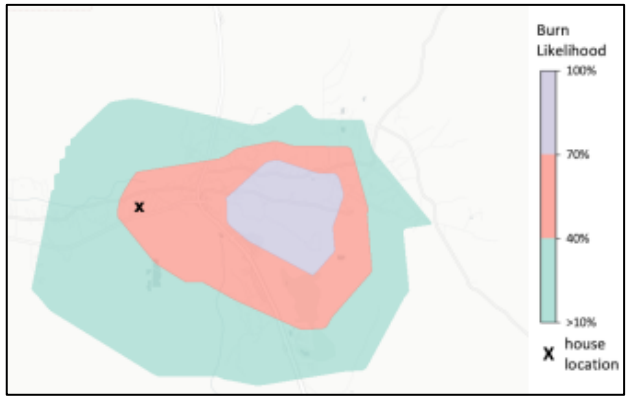

X

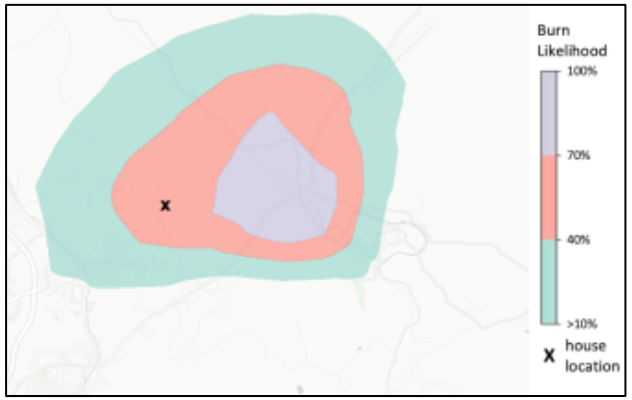

Overall Proportion of Participants Evacuating
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Percentages
Maps
Three Bands
Six Bands
Nine Bands

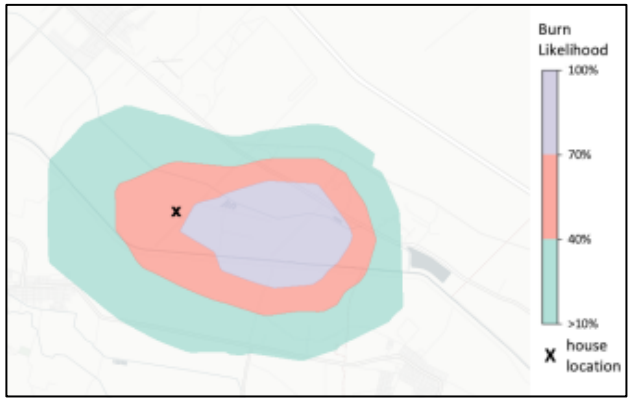
Your house is located in the

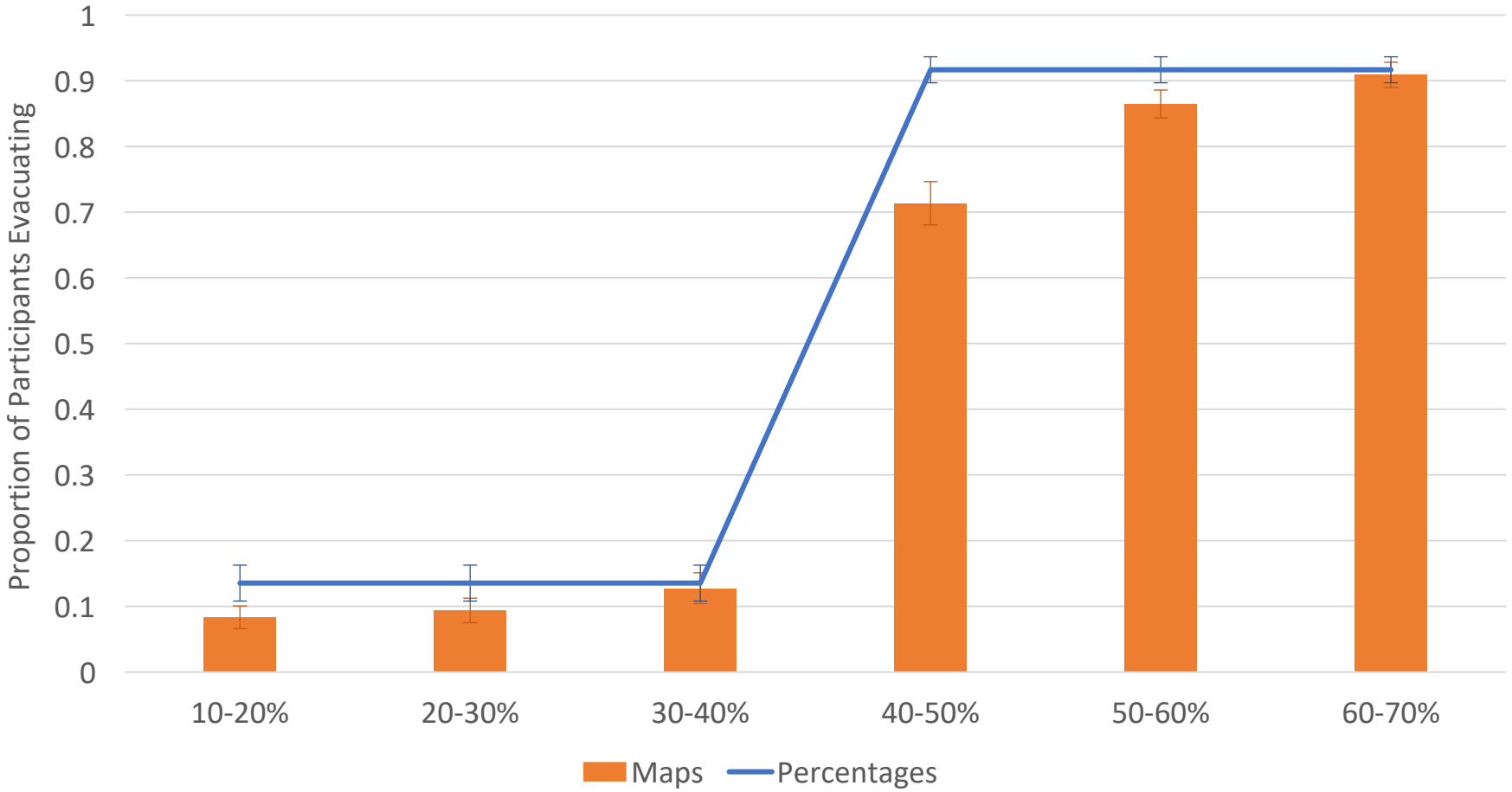
40 to 70%

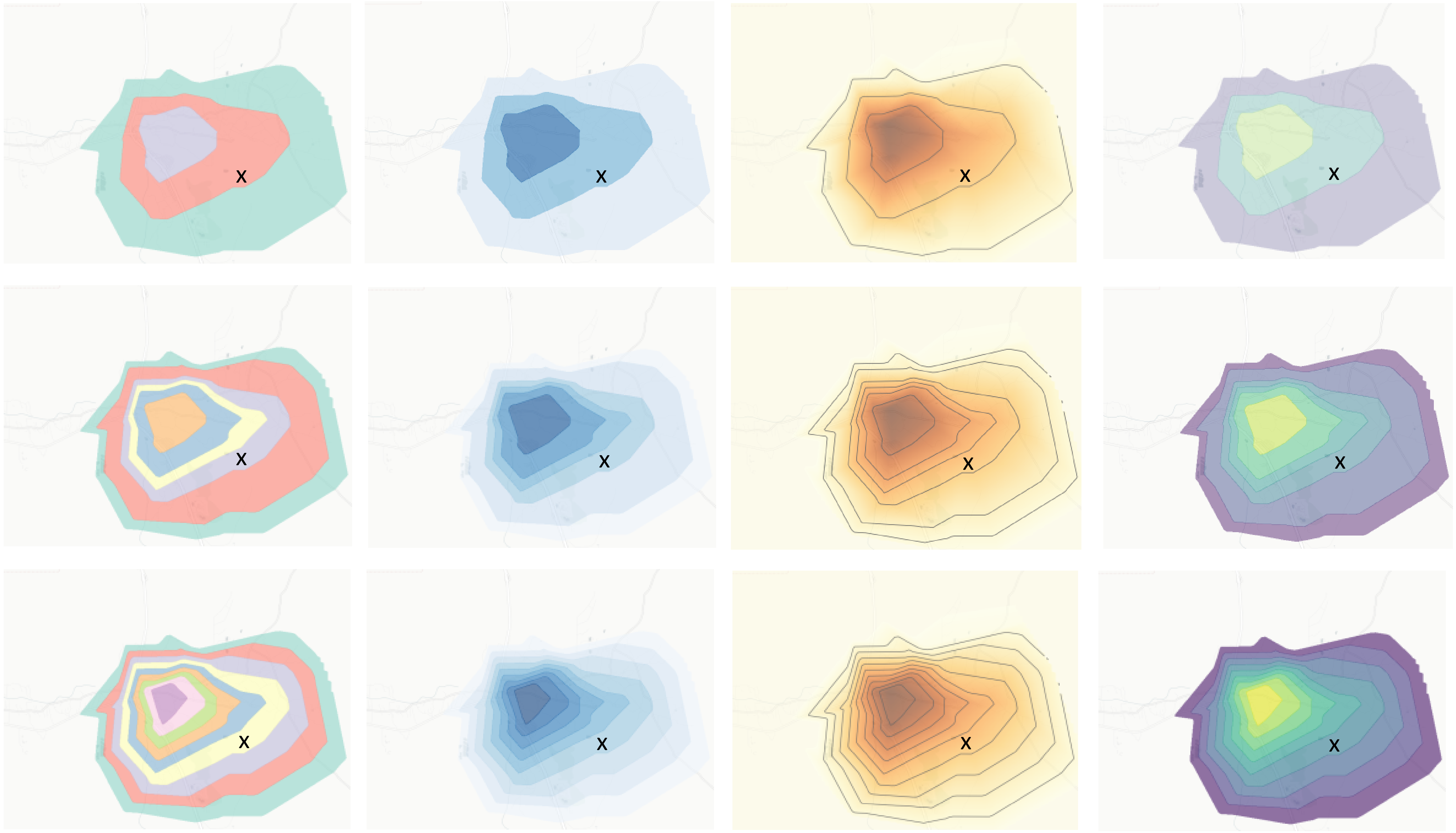
burn likelihood zone.



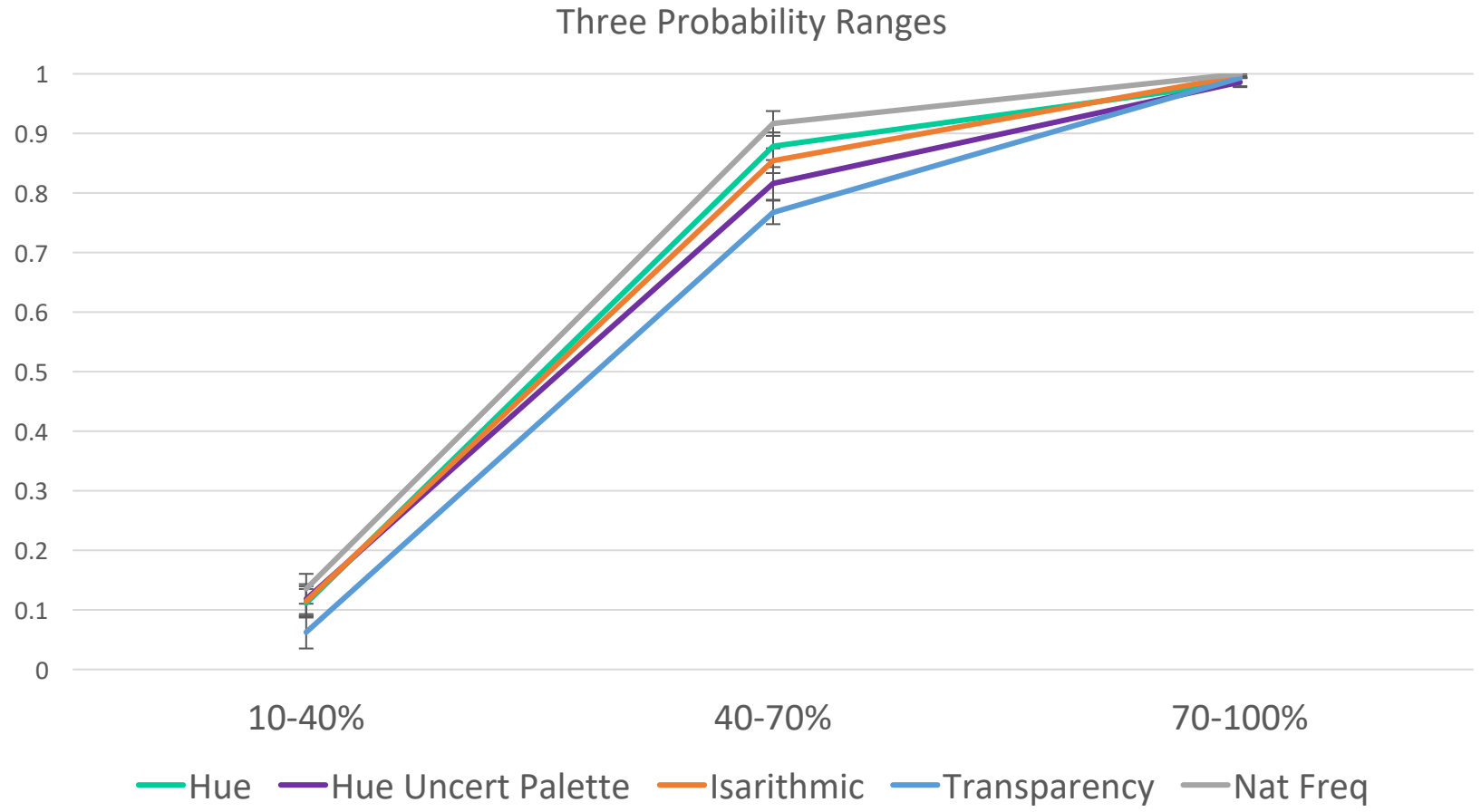


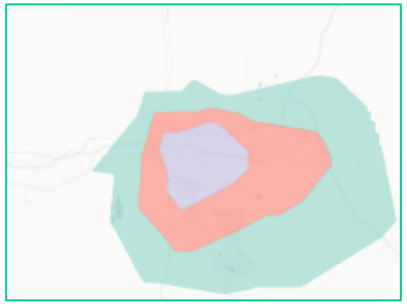


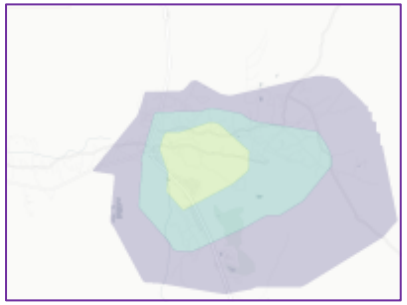


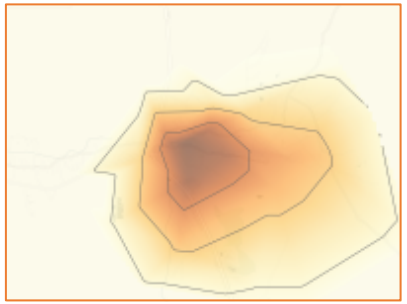

x
x
x
x
x
x
x
x
x
x
x
x

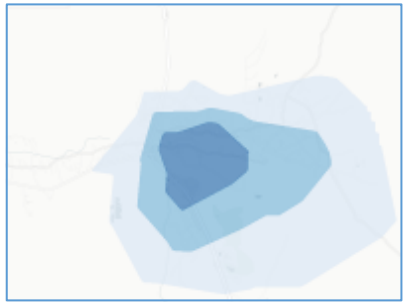
Map Experiment Results



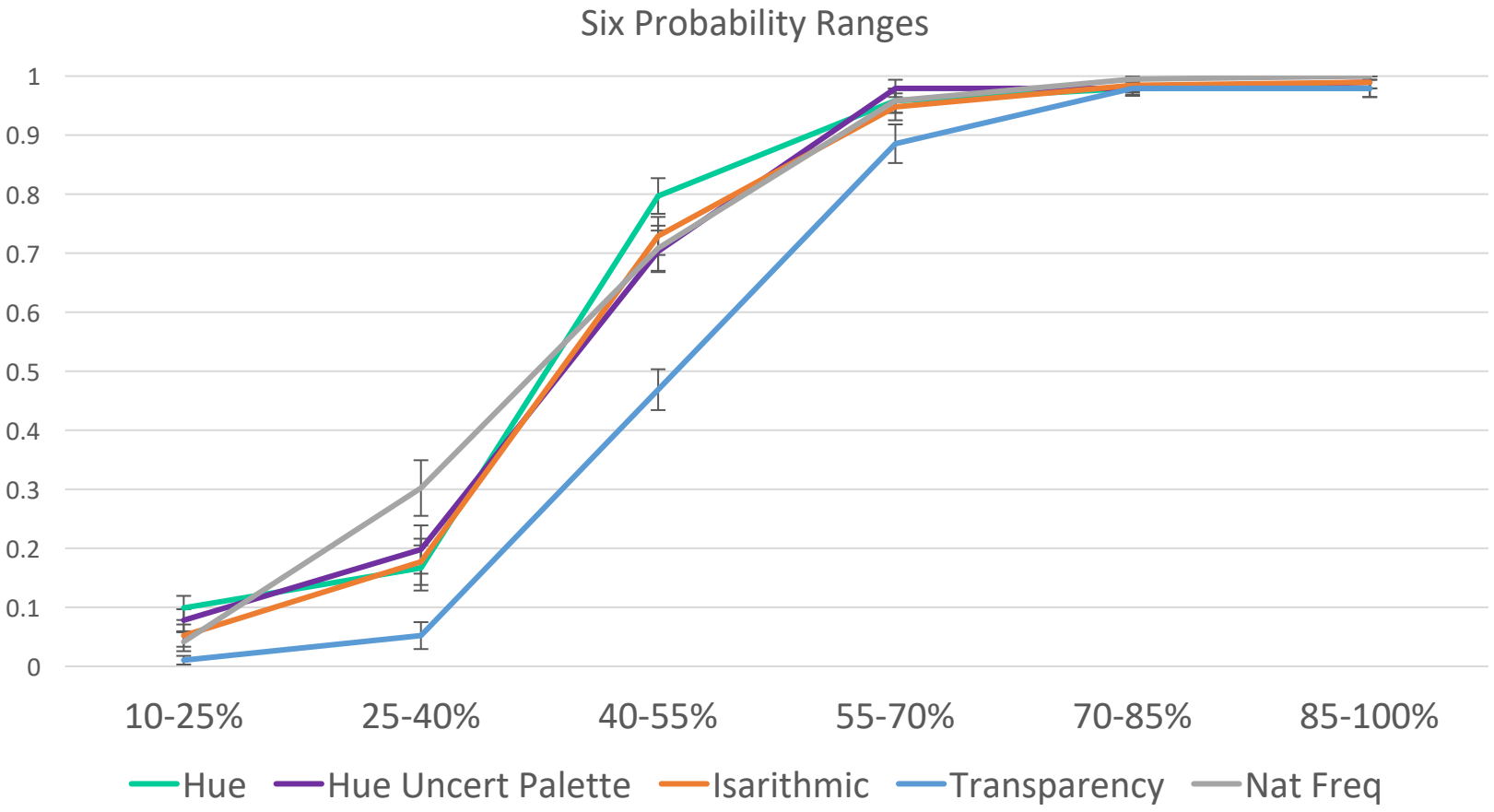


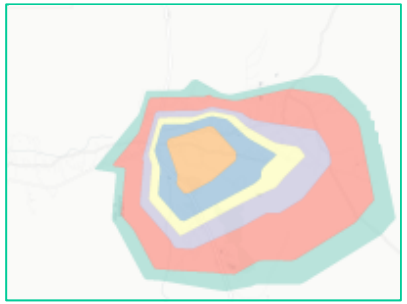


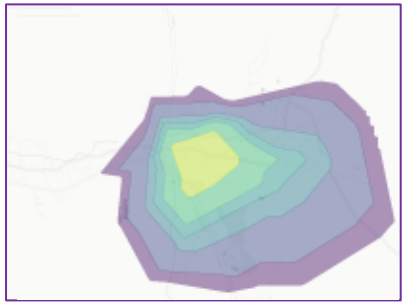


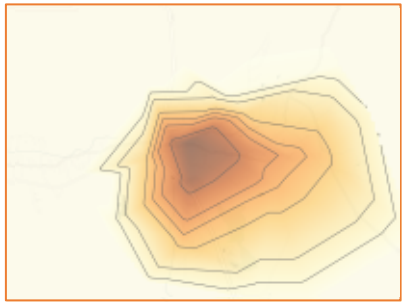


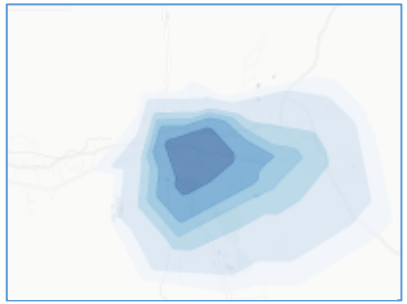
Map Experiment Results



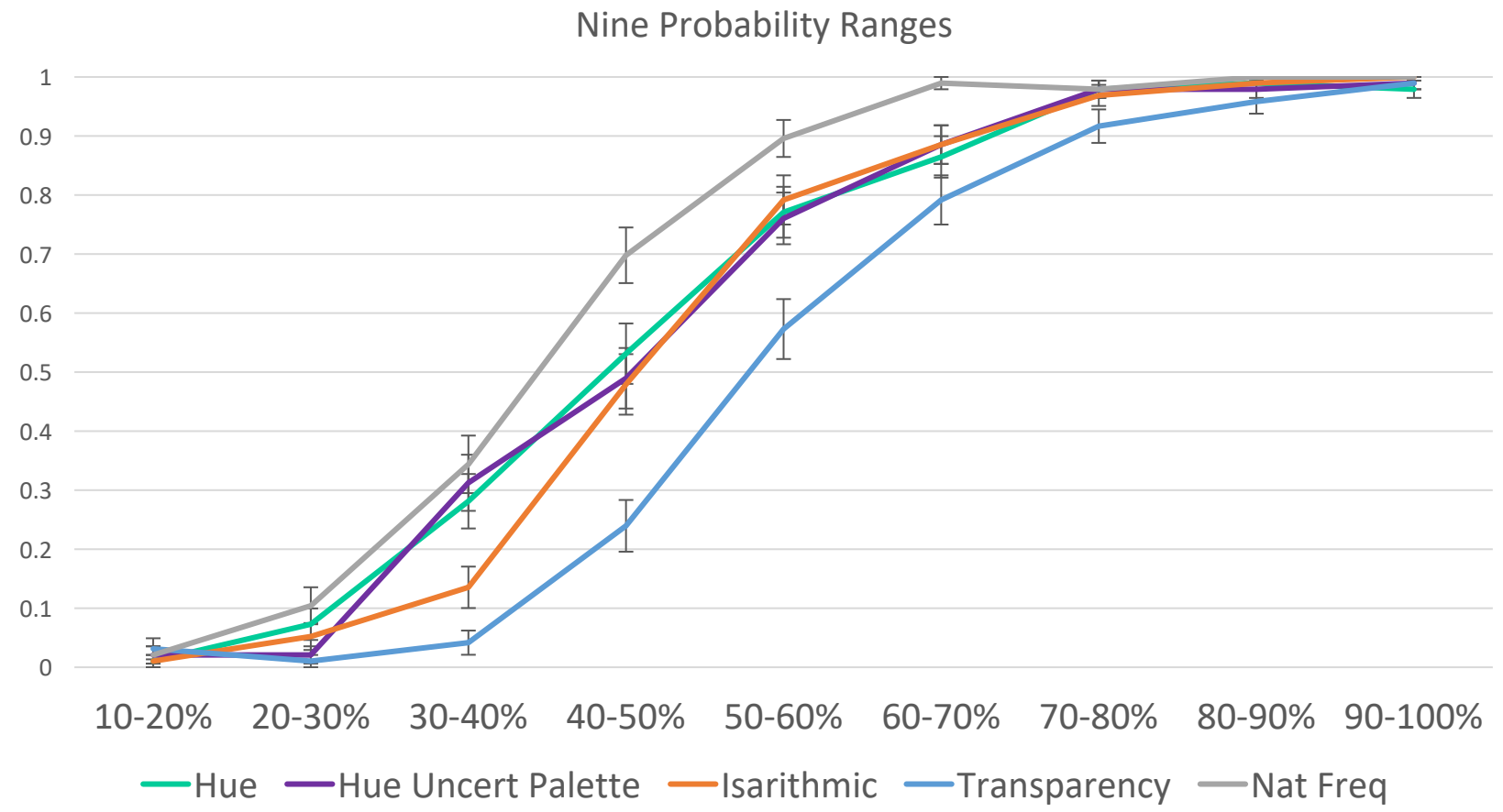


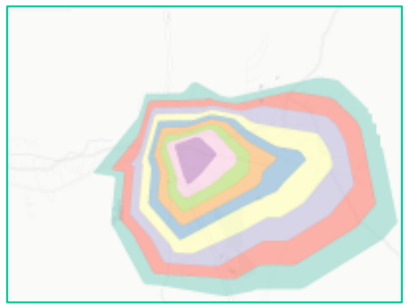


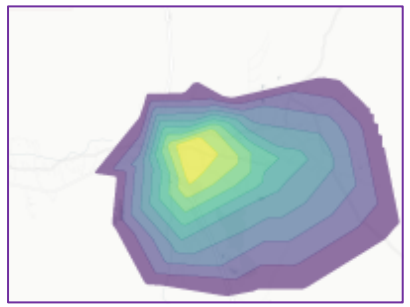


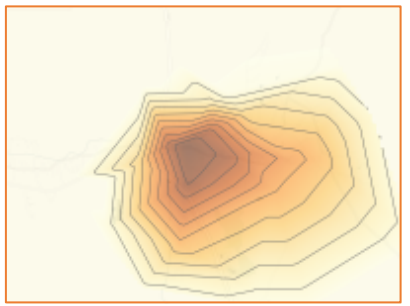


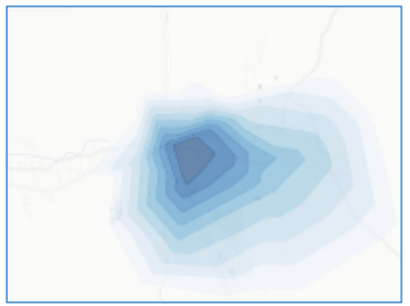
Map Experiment Results





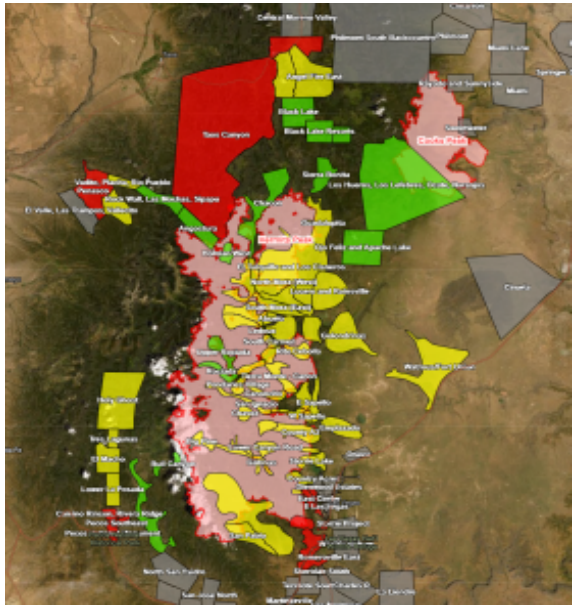


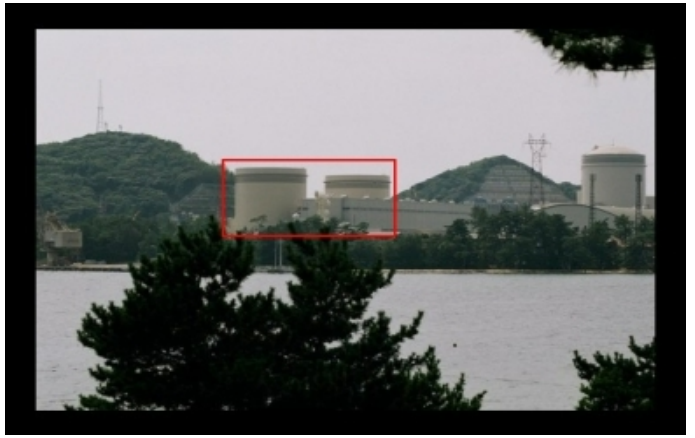


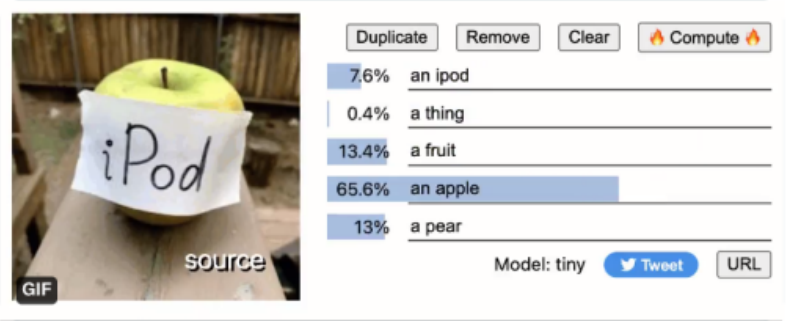


Summary

- It’s complicated. Even before we look at interactivity!
- We need to be aware of the various factors that impact the effectiveness of different representations of ML outputs.
- Different representations may not be equivalent from the perspective of human comprehension and decision making
- What’s going to happen when people interact with the visualizations?



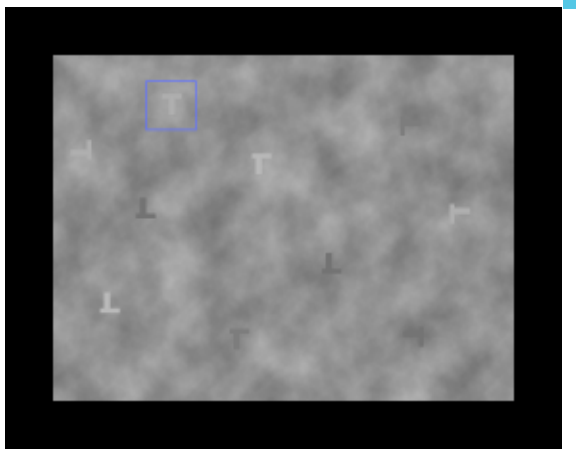




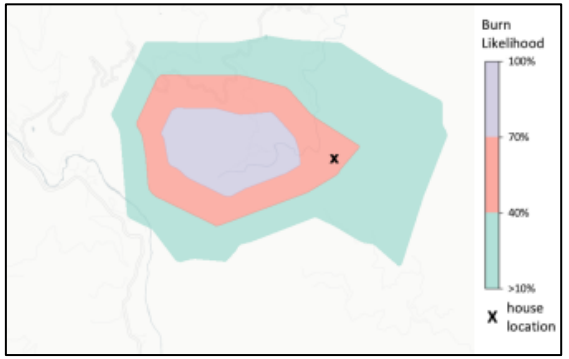
Factors that Impact Decision Making

A few examples:

- Visual search aided by (mock) ML outputs
  - People get complacent as the overall accuracy of the outputs goes up
  - Novices are more likely to go along with what the ML says



- Visualizations of uncertain information
  - Differences between visual and numerical representations
  - The specificity of the information can impact judgments of risk
  - The same information visualized in different ways can lead to different patterns of decisions
  - Individual differences also impact decisions



Backup Slides

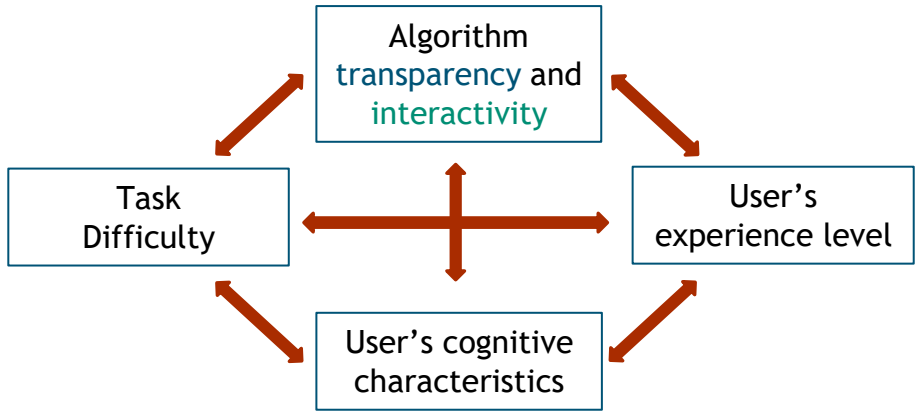
T4T: Tailoring for Appropriate Trust

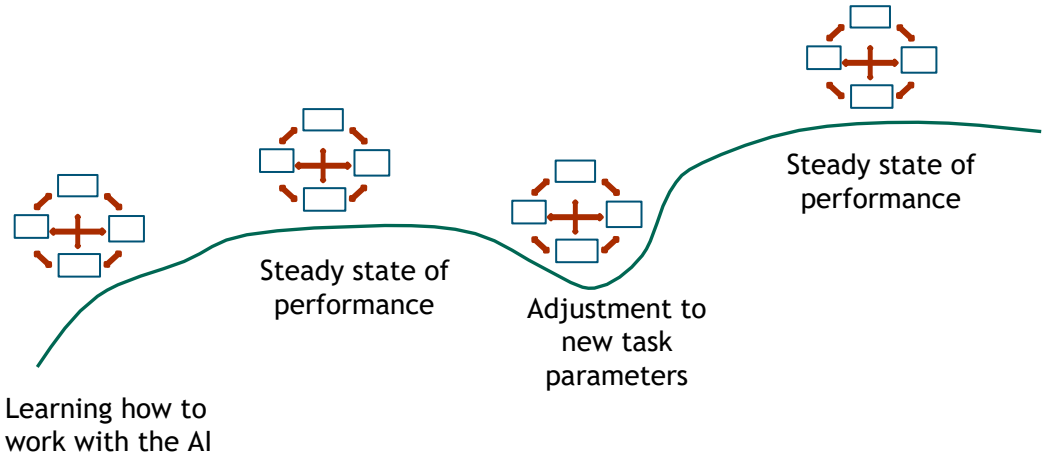
Project Summary

Our goal is to advance the scientific understanding of how people develop appropriate levels of trust in artificial intelligence (AI).

Key Questions:

- Can we develop descriptive and predictive models of how characteristics of the algorithm, the task, and the user interact across different phases of human-AI interaction?
- Can we use those models to tailor an algorithm to optimize task performance and trust for:
  - Specific groups of users (i.e., domain experts)? Specific individuals?
  - Specific phases of an analytical process?
  - Specific individuals during specific phases of an analytical process?
    - And if we can do it, is it worth the effort?





Visualization Cognition

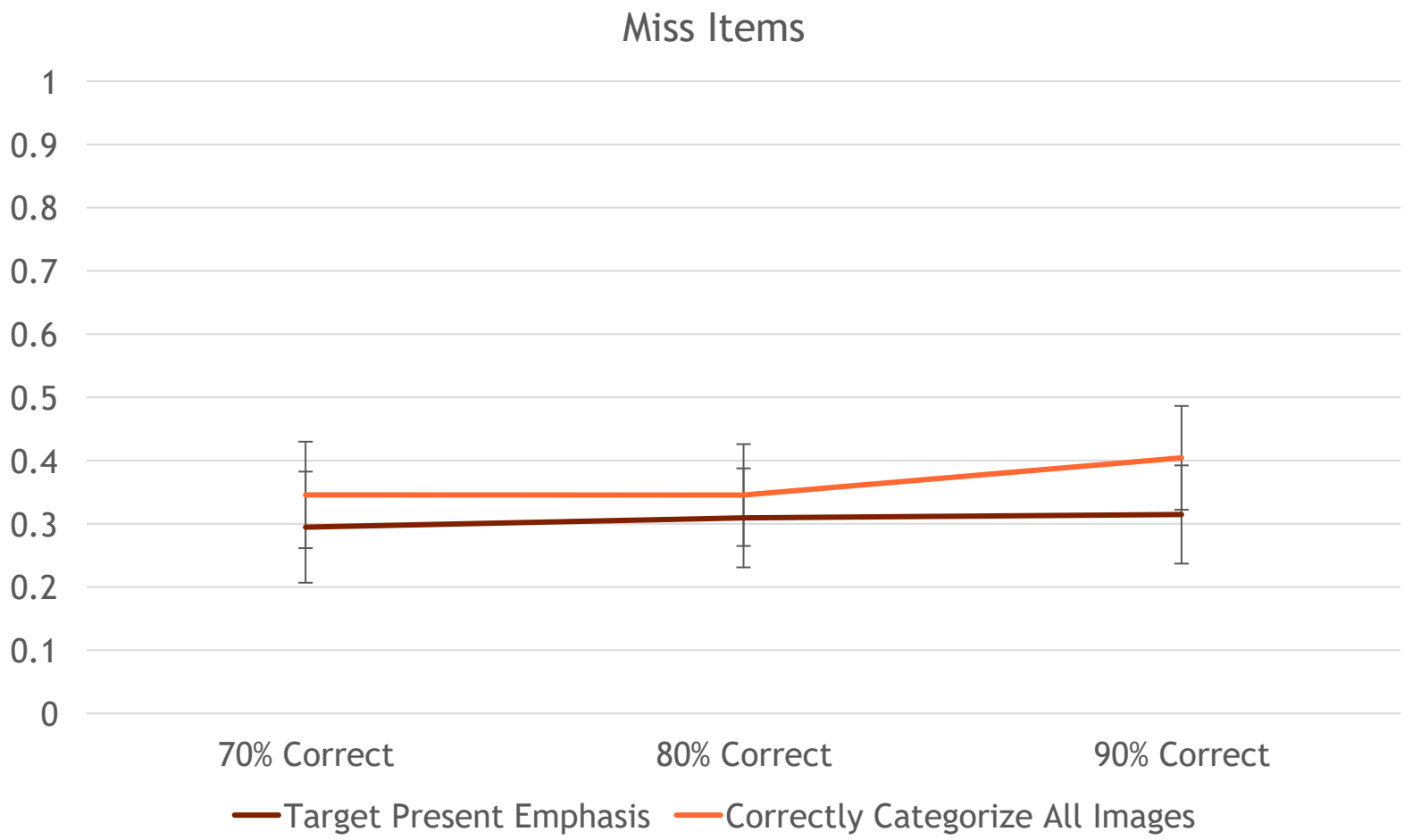
- Task, prior experience, familiarity, low-level visual features, and visual-spatial biases all interact with one another when viewers interpret visualizations
  - Yet carefully controlled experiments can identify systematic patterns in how viewers interpret visual cues
  - Identifying these patterns can help us to understand and mitigate perceptual and cognitive baises

Error Importance Experiment Results

Did emphasizing certain types of errors make a difference?

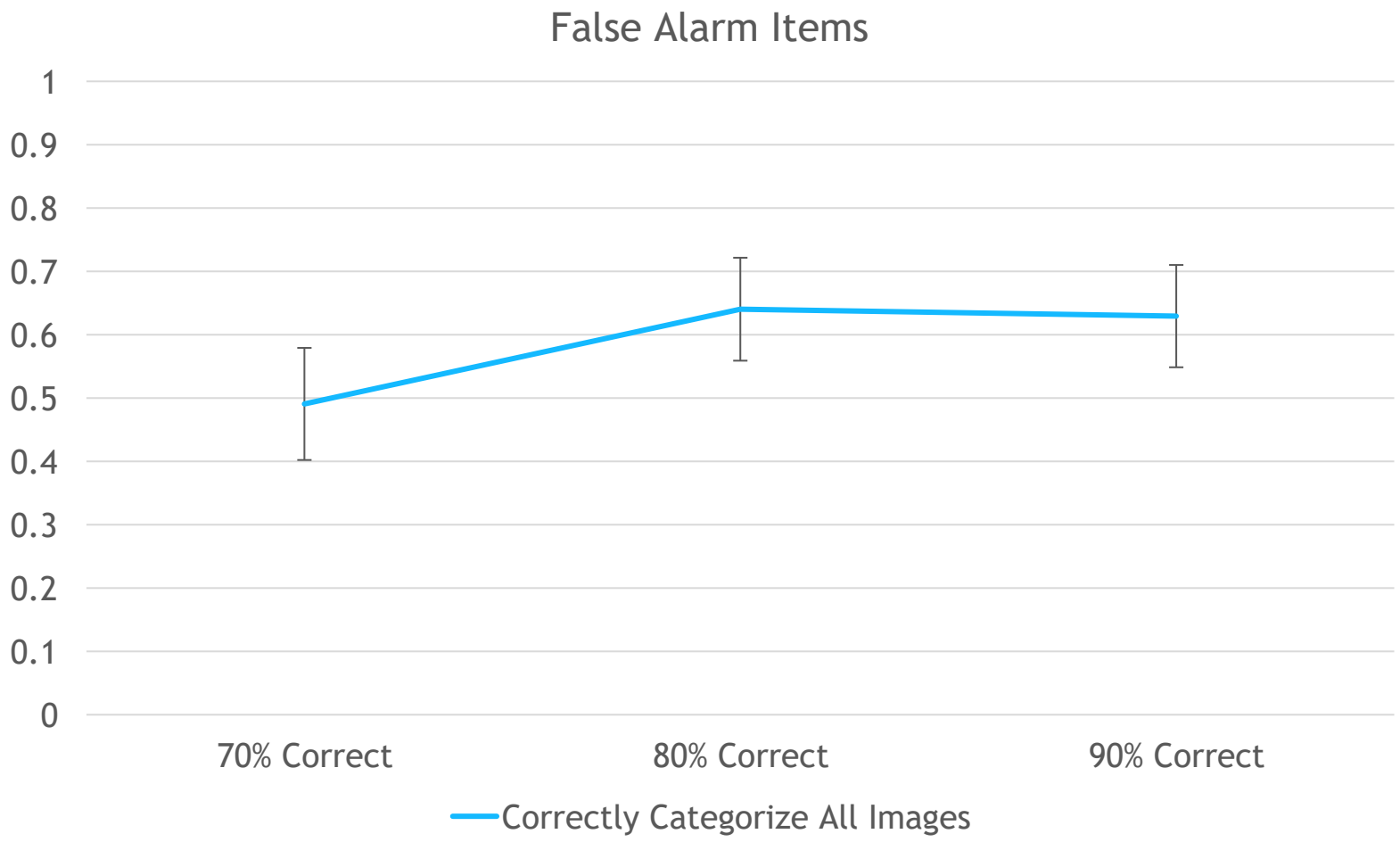
Error Importance Experiment Results

Did emphasizing certain types of errors make a difference?



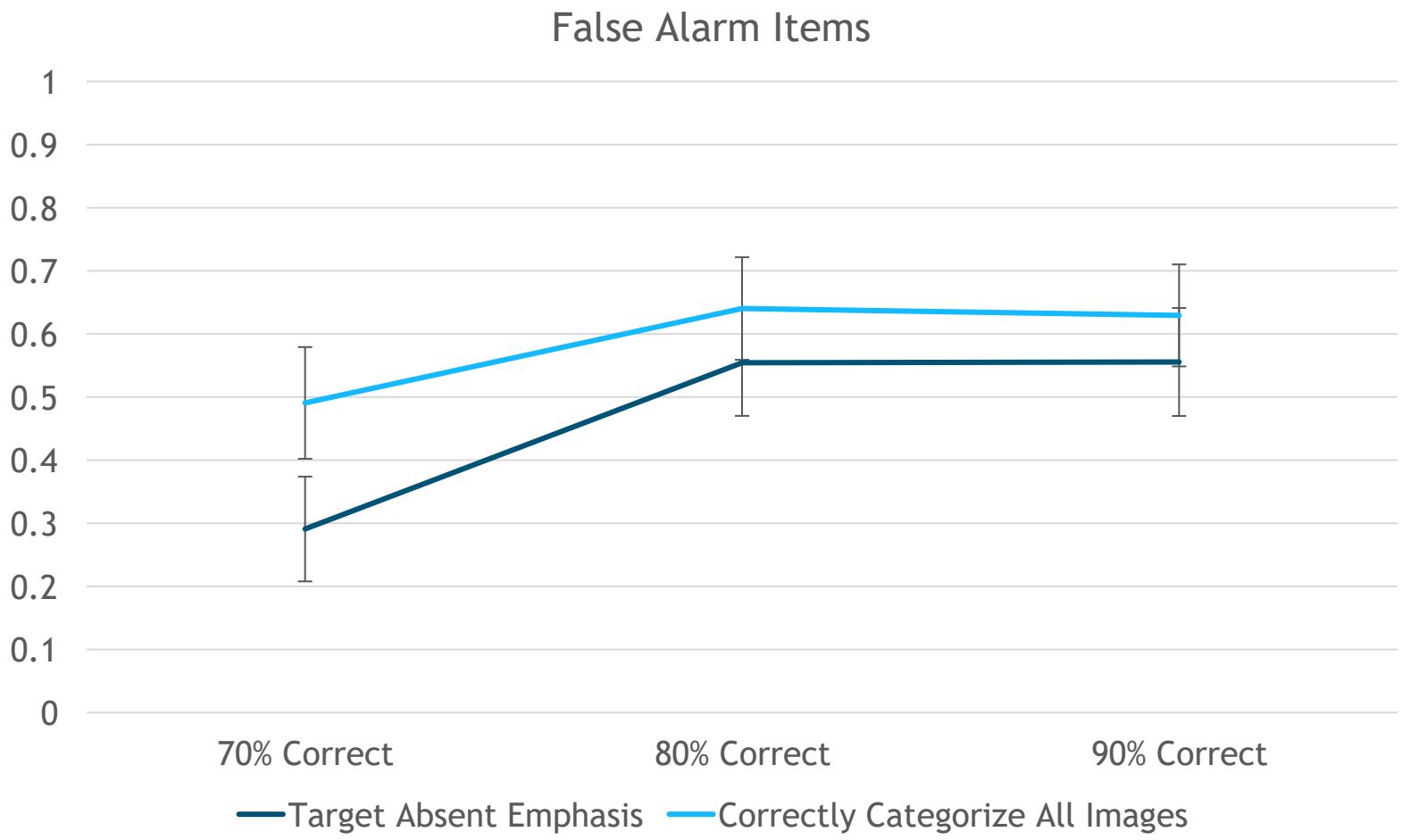
Error Importance Experiment Results

Did emphasizing certain types of errors make a difference?



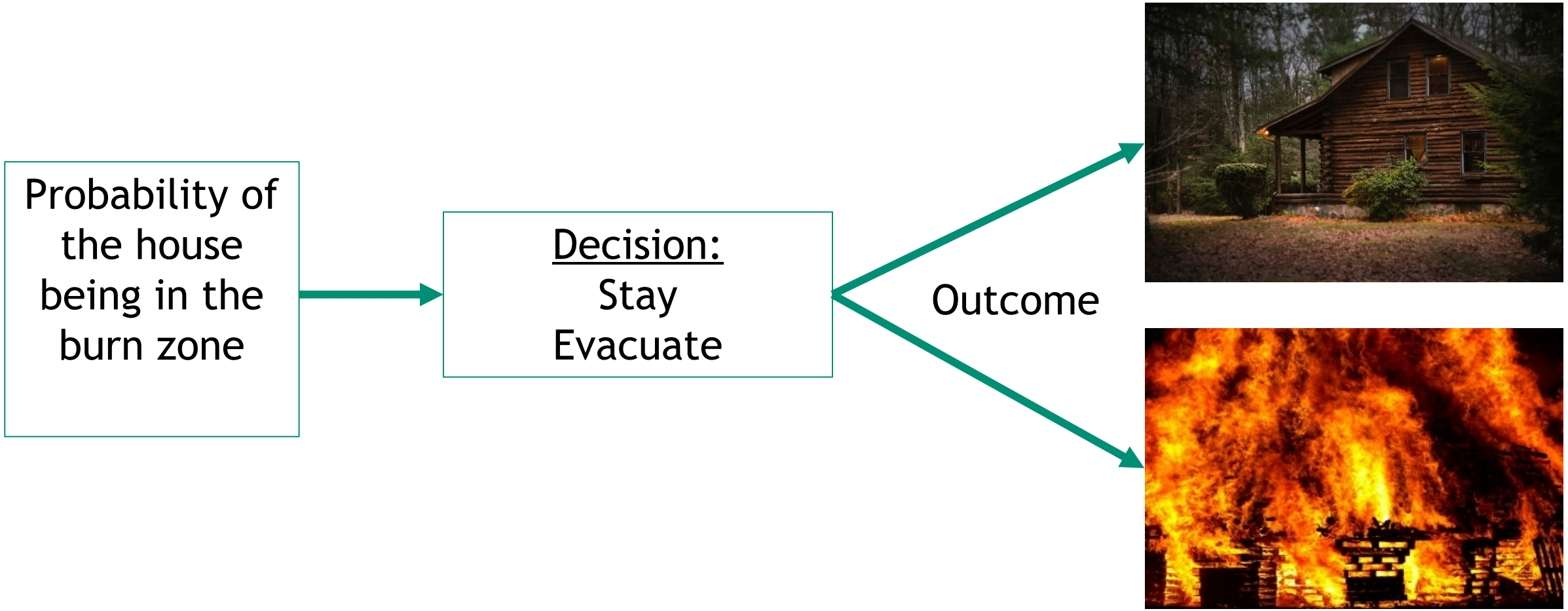
Error Importance Experiment Results

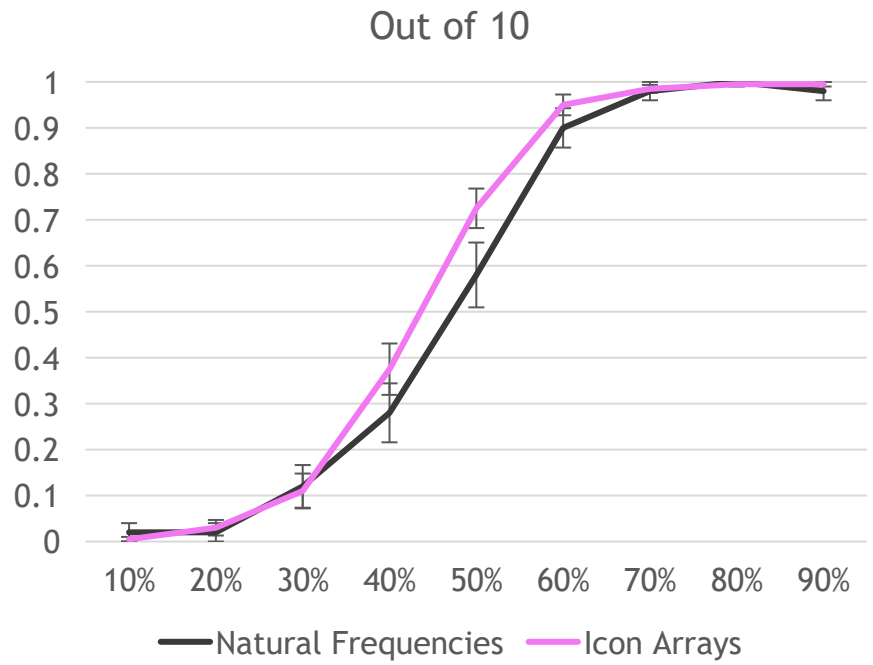
Did emphasizing certain types of errors make a difference?

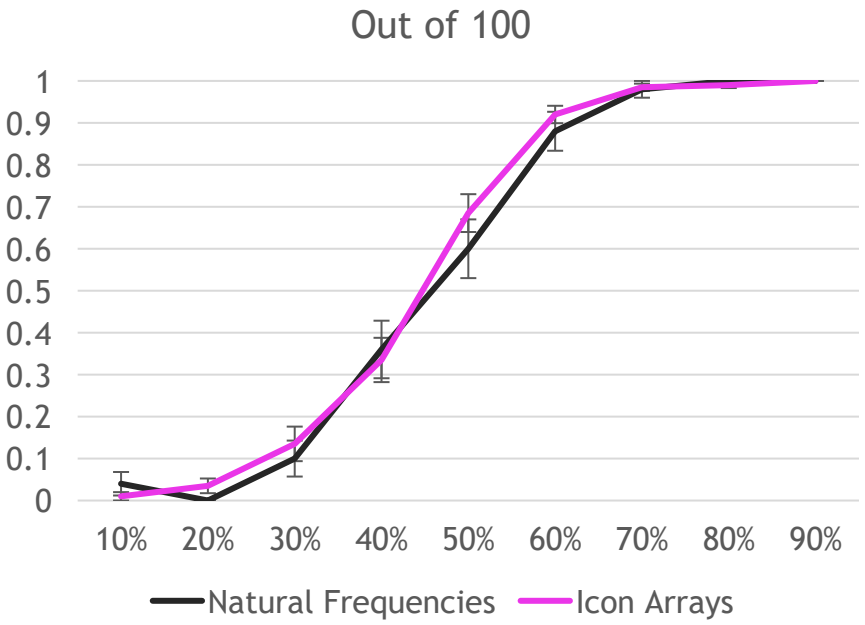


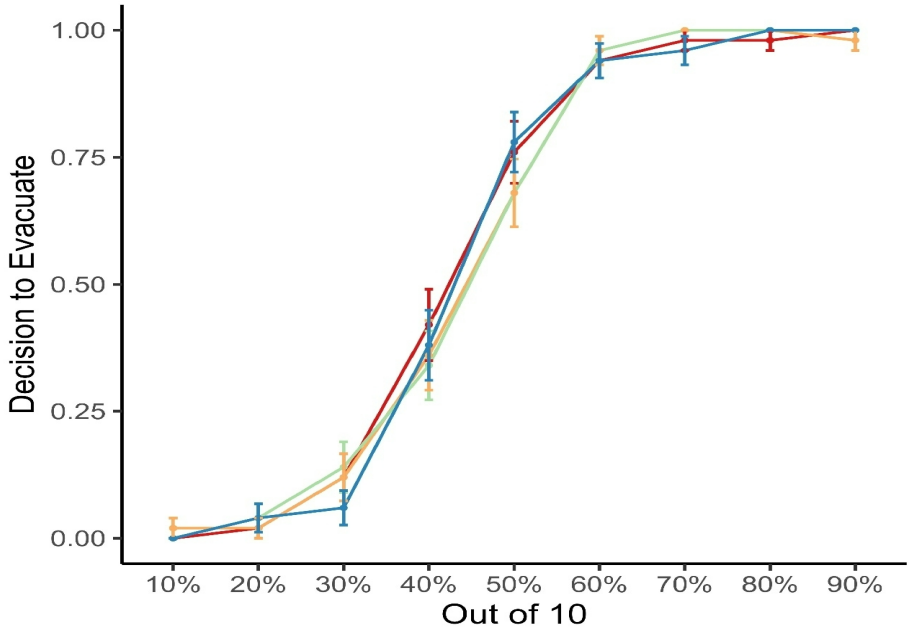
Wildfire Evacuation Task

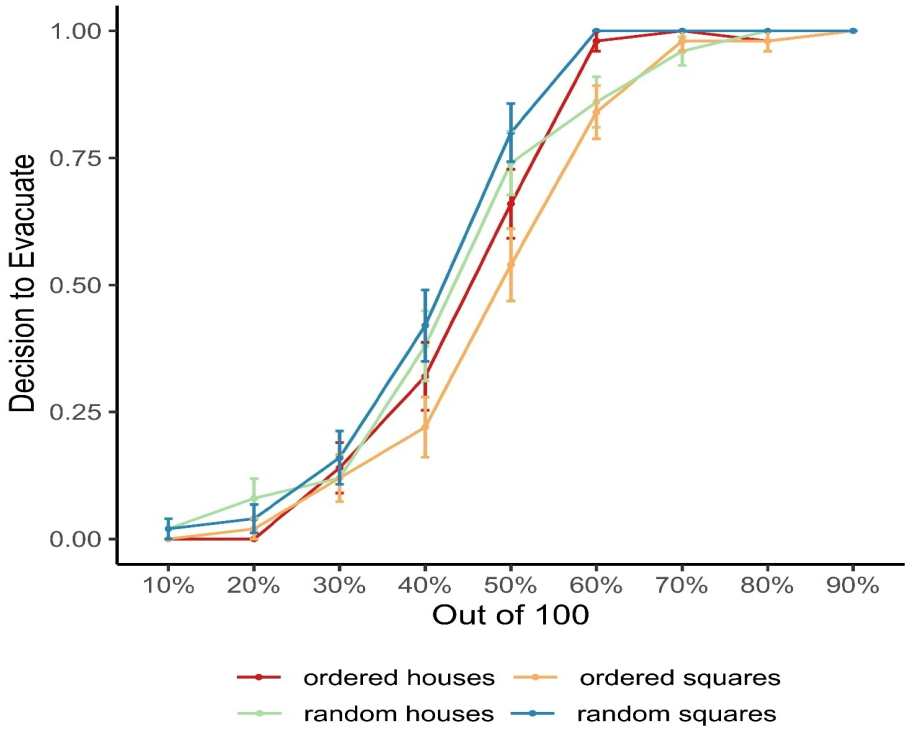
- Participants are asked to pretend that they live in this cabin in the woods, but there is a wildfire in the area. On each trial, they see the probability that their house will be in the burn zone. They must decide whether to stay or evacuate.




Out of 10
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
10%
20%
30%
40%
50%
60%
70%
80%
90%
Natural Frequencies
Icon Arrays


Out of 100
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
10%
20%
30%
40%
50%
60%
70%
80%
90%
Natural Frequencies
Icon Arrays


Decision to Evacuate
1.00
0.75
0.50
0.25
0.00
10%
20%
30%
40%
50%
60%
70%
80%
90%
Out of 10


Decision to Evacuate
1.00
0.75
0.50
0.25
0.00
10%
20%
30%
40%
50%
60%
70%
80%
90%
Out of 100
ordered houses
ordered squares
random houses
random squares

What if we actually make the 100 icon arrays more specific?