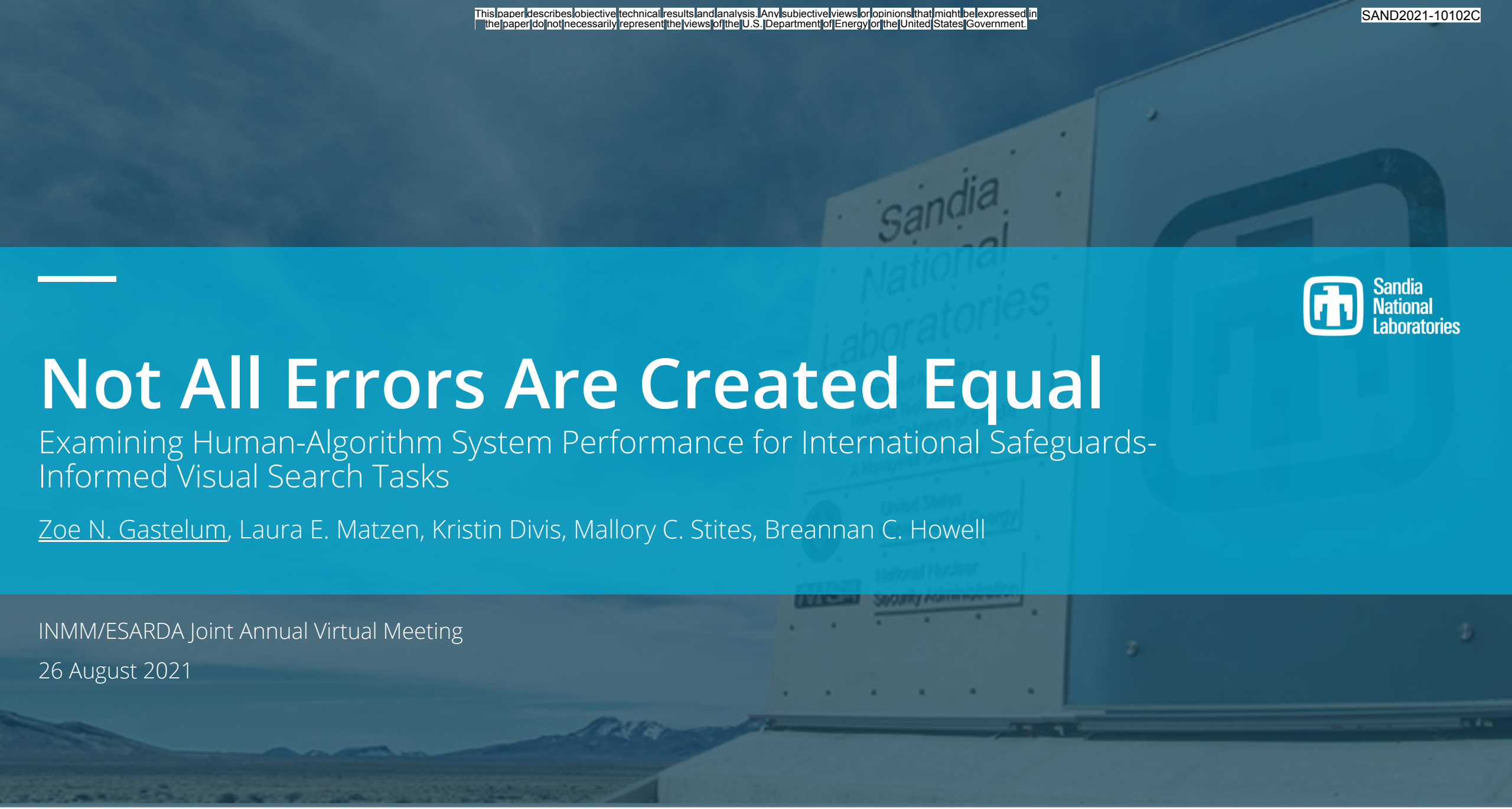
SAND2021-10102C



Not All Errors Are Created Equal

Examining Human-Algorithm System Performance for International Safeguards-Informed Visual Search Tasks

Zoe N. Gastelum, Laura E. Matzen, Kristin Divis, Mallory C. Stites, Breannan C. Howell

INMM/ESARDA Joint Annual Virtual Meeting

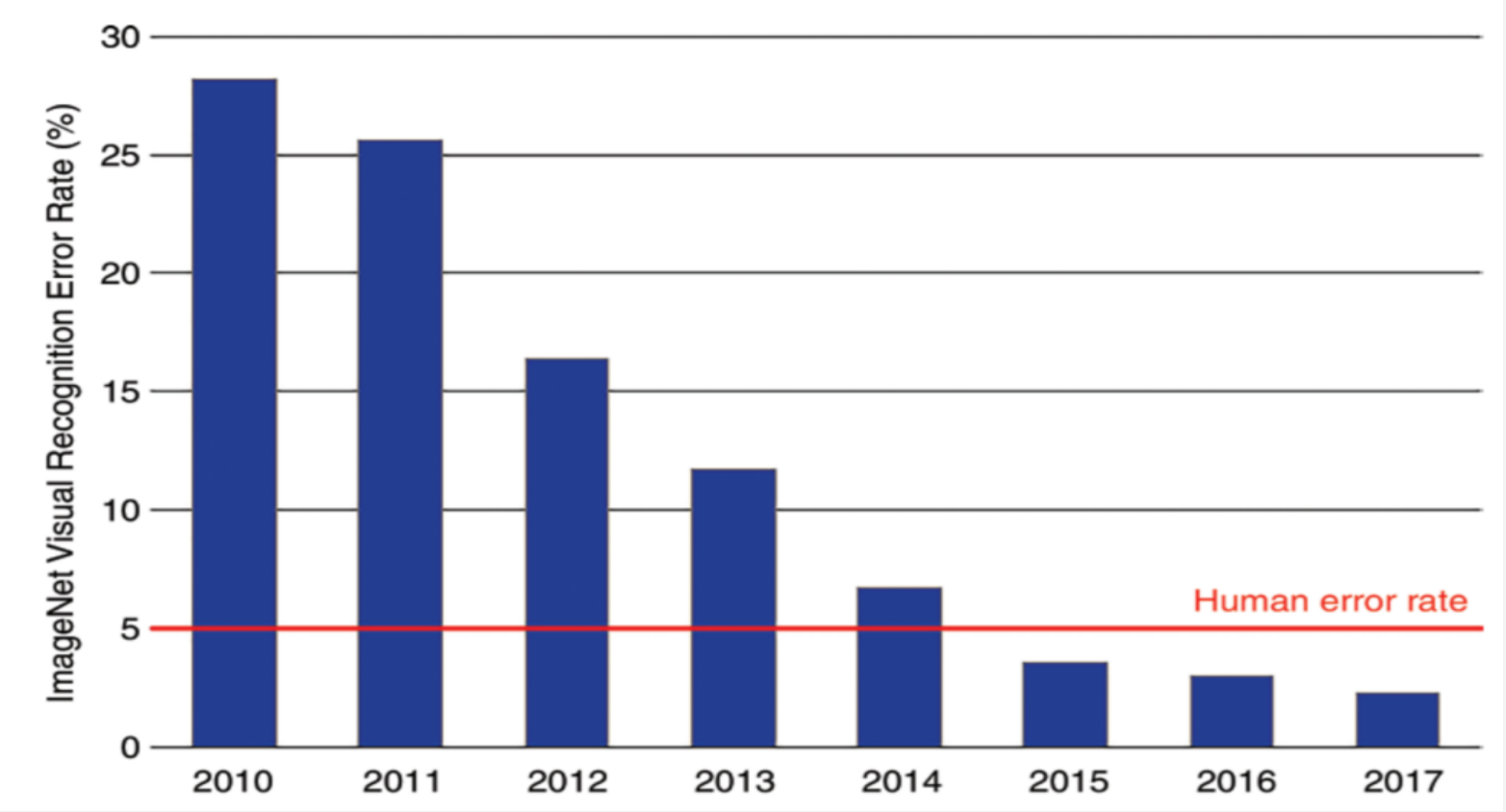
26 August 2021

Main Points

1) All models make errors, but some are useful.
2) Users respond to different types of errors in different ways.
3) Domain makes a difference, and expertise is the culprit.

All models make errors, but some are useful.

Algorithms can outperform humans on some tasks.



(Langlotz, et al. 2018)

Sometimes, the consequence of model errors is low.

Images: Smithsfoodanddrug.com

Sometimes, the consequence could be very high.

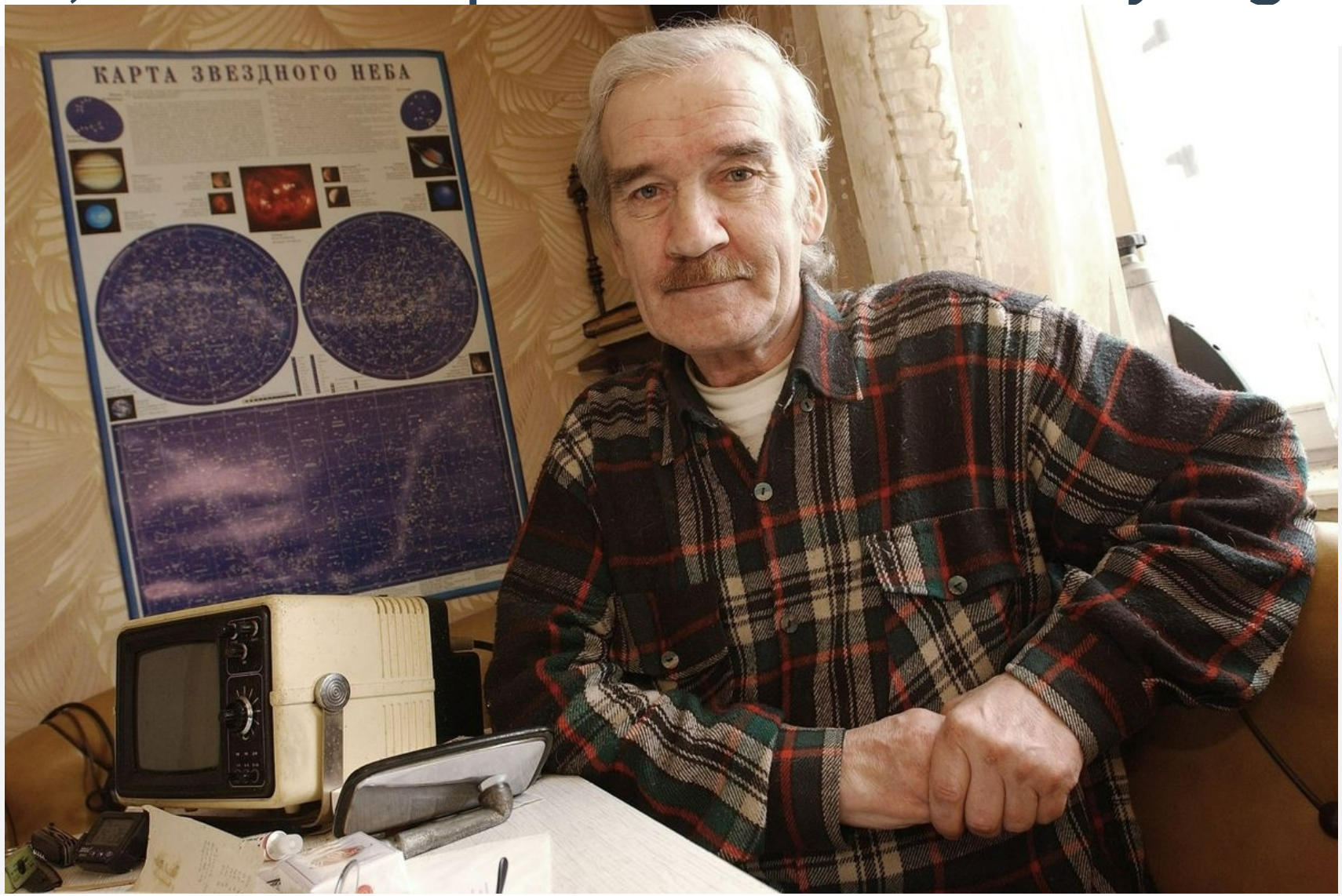
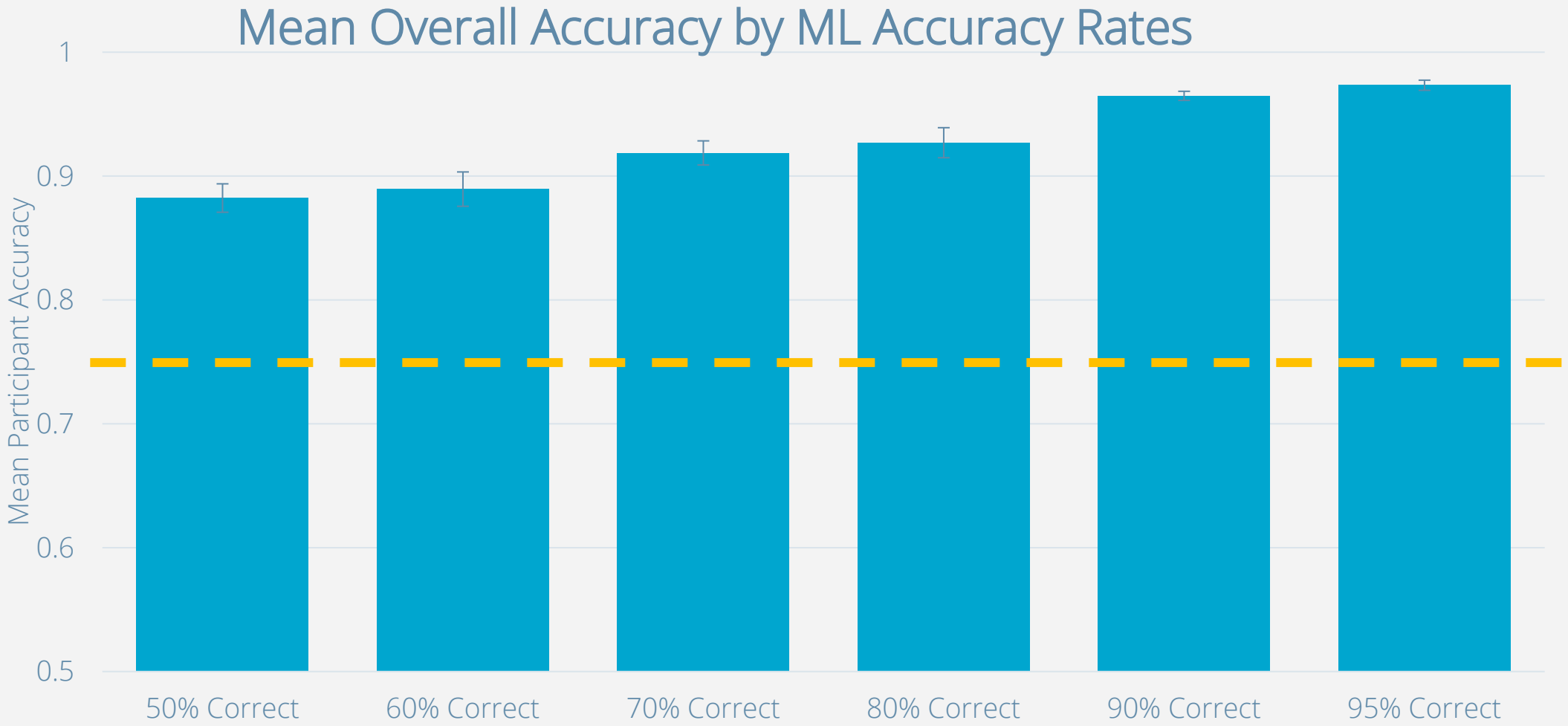
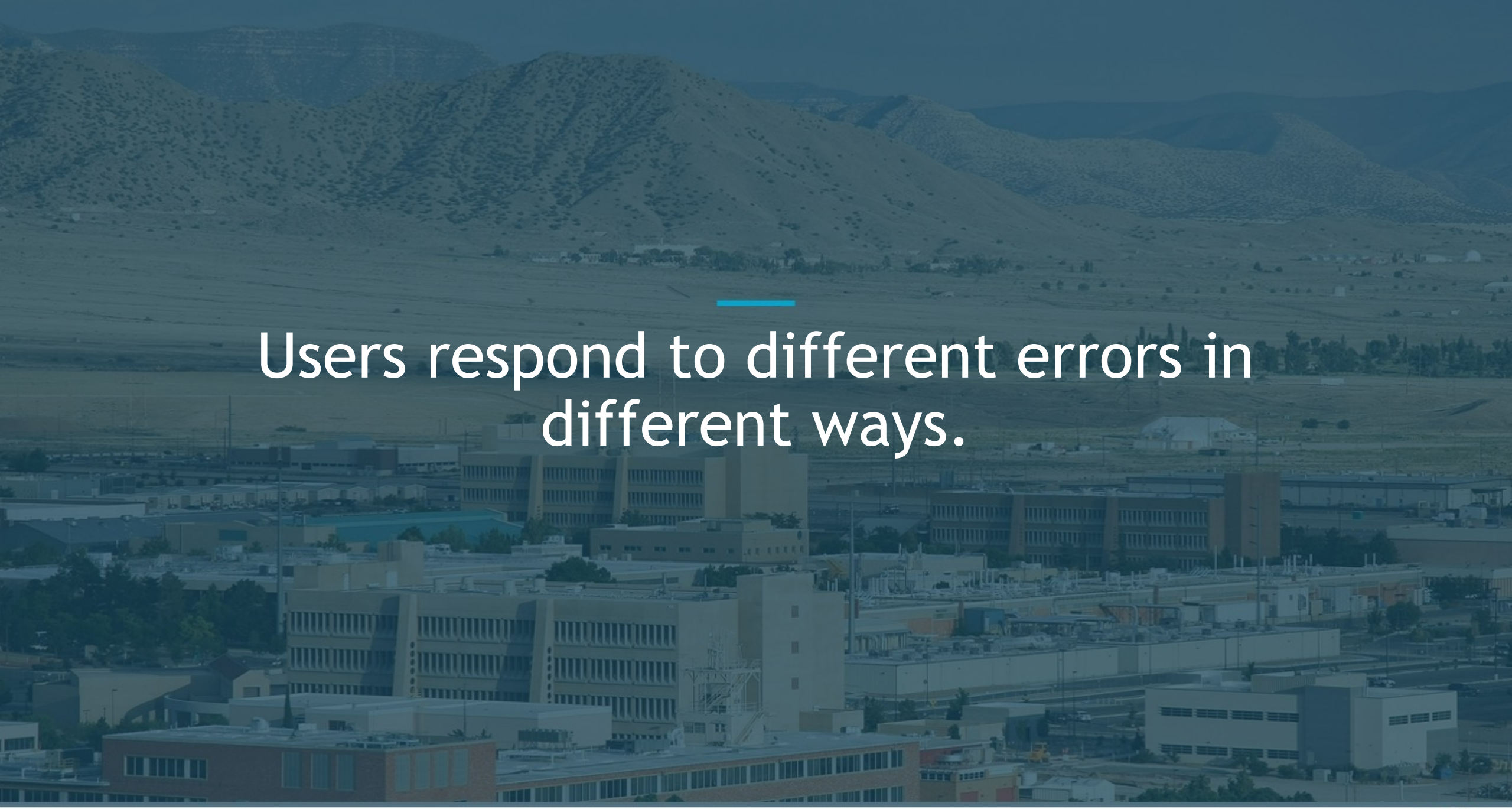


Image: Scott Peterson/Getty Images, 2004

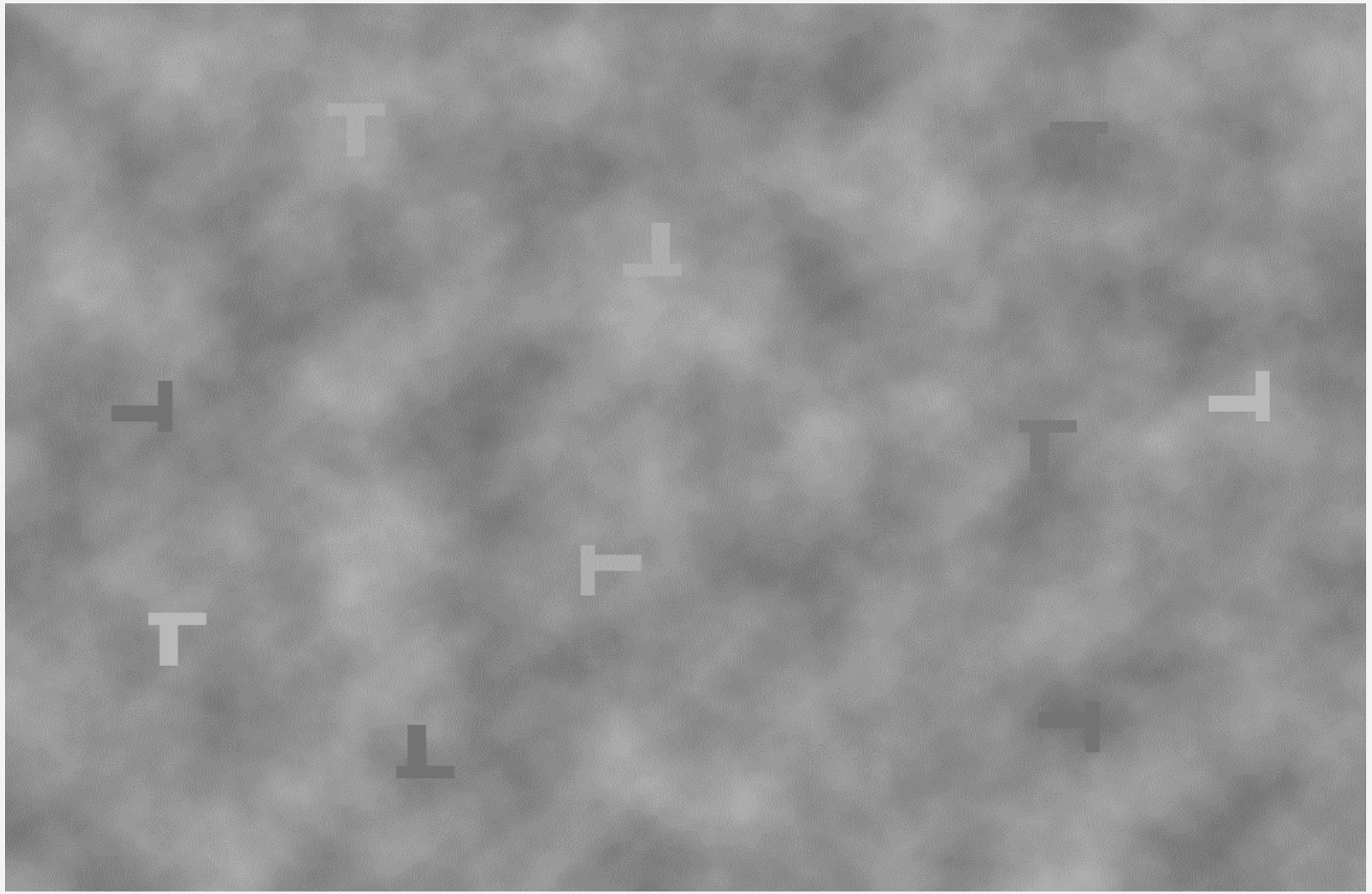
Even imperfect ML can help users identify targets.



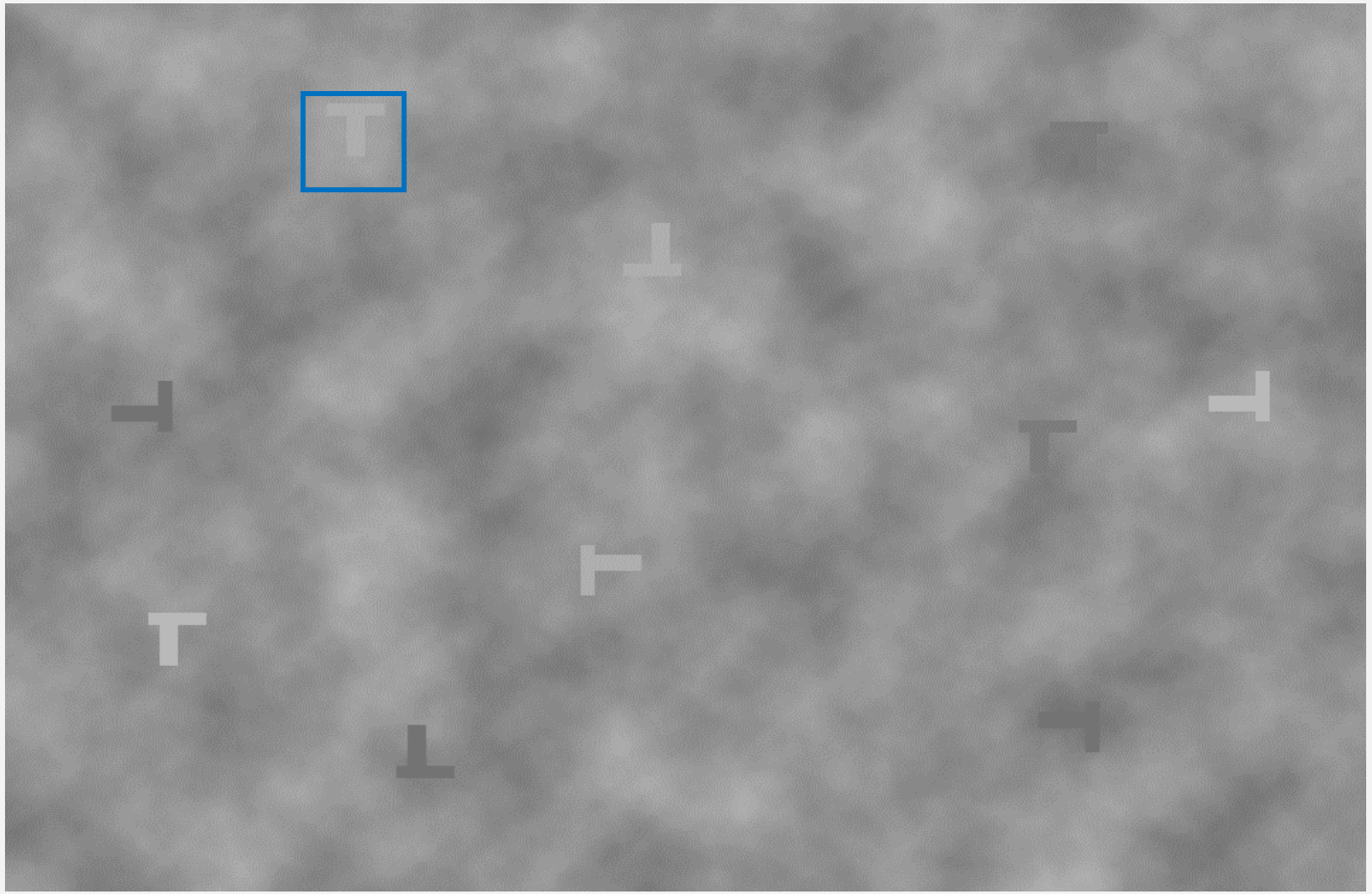


Users respond to different errors in different ways.

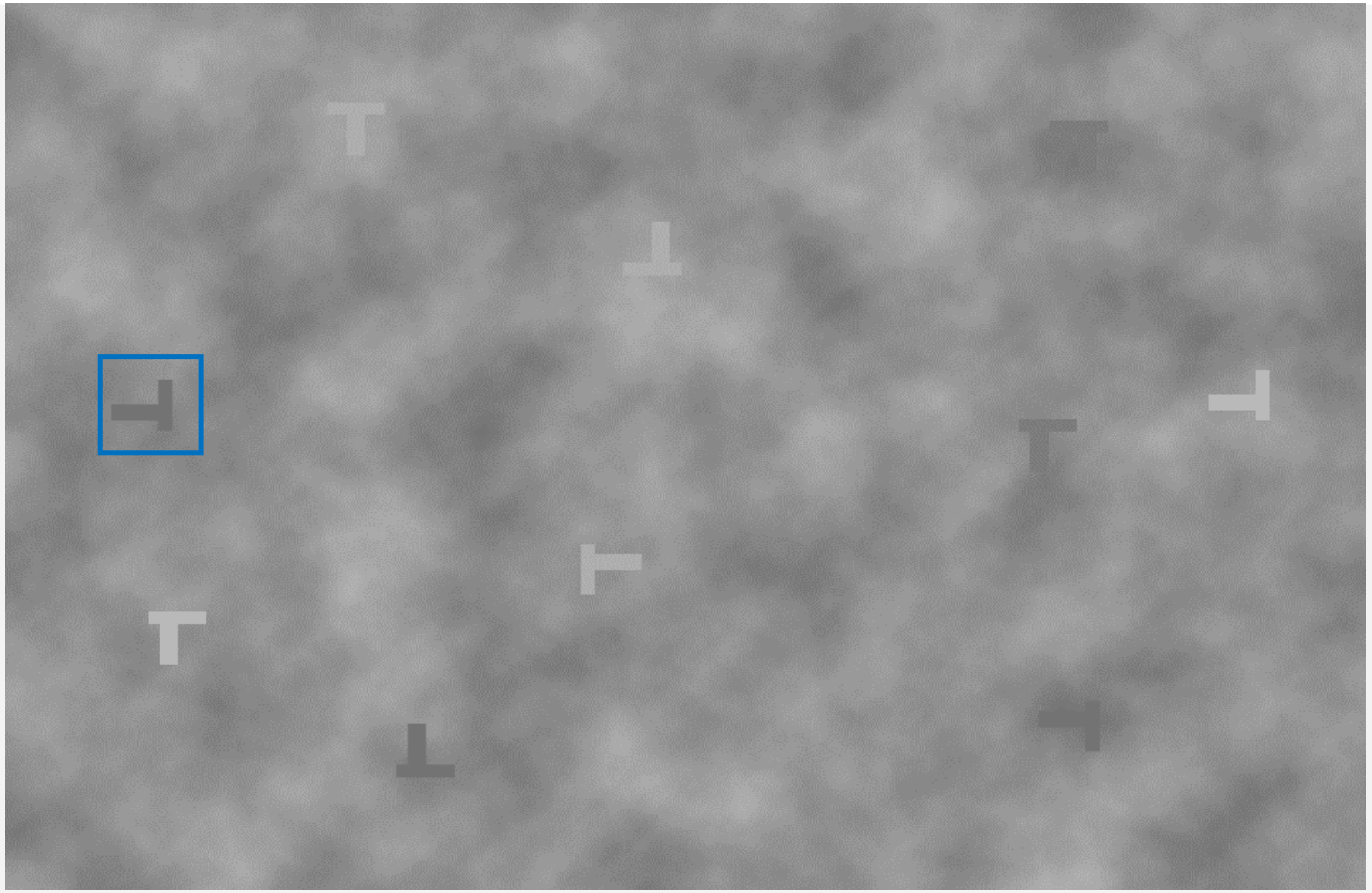
We measured human performance in finding target “T”



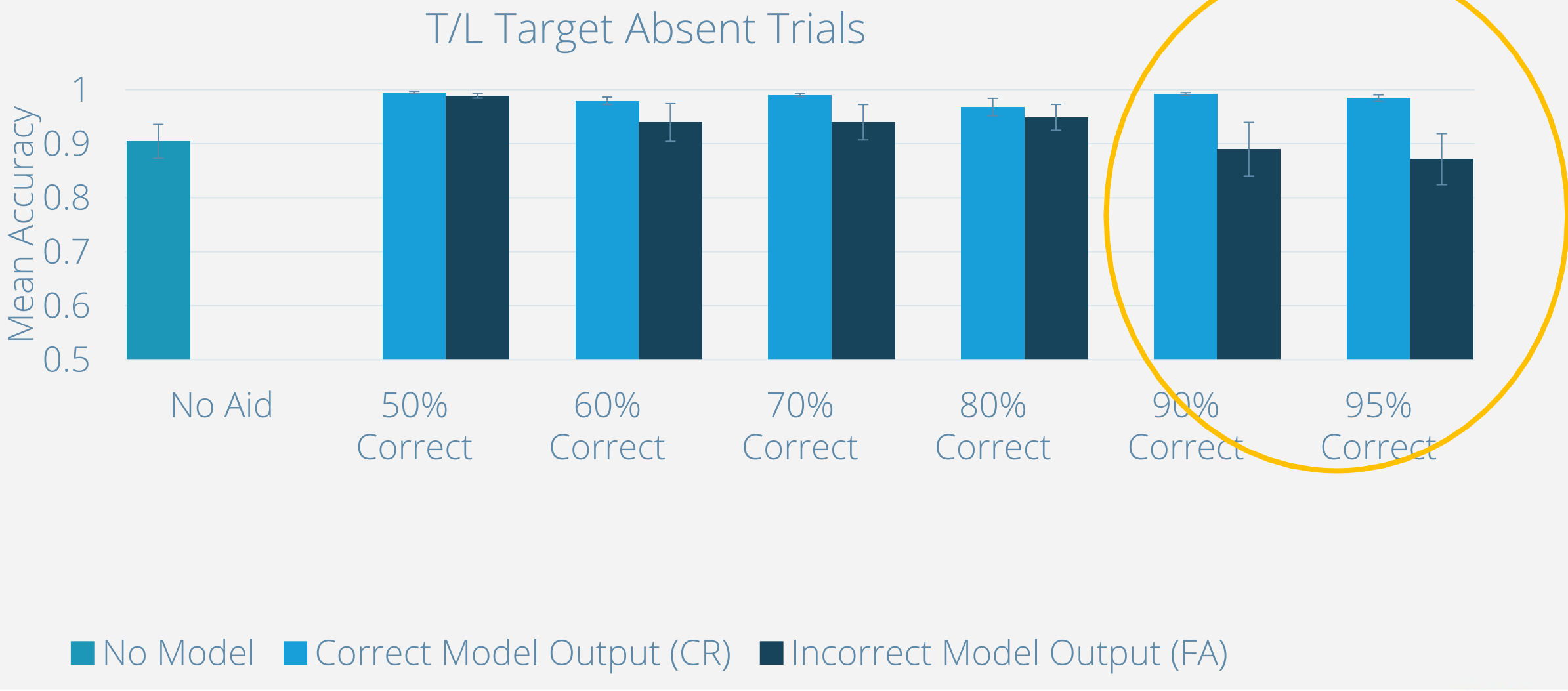
and compared it to performance with “ML” support.



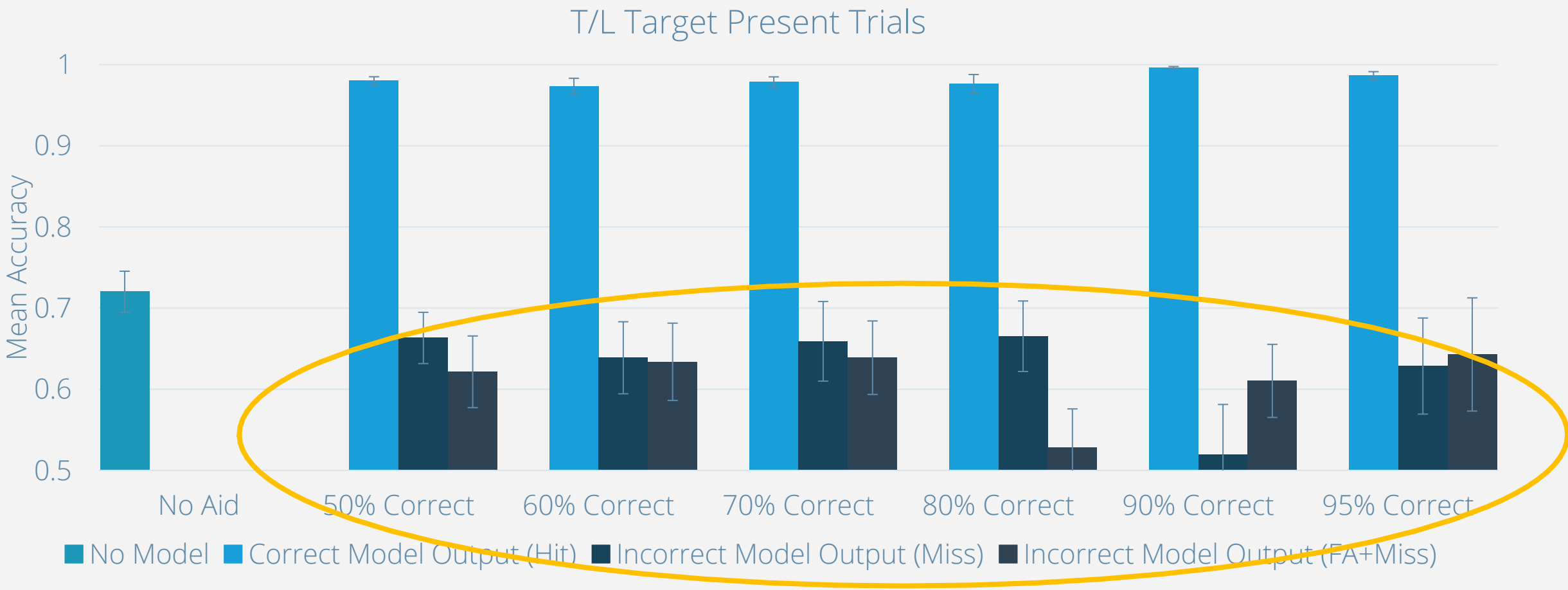
Sometimes the “ML” was wrong. We varied how it was wrong, and how often.



For target absent, false alarms decreased human performance only at the highest model accuracy.

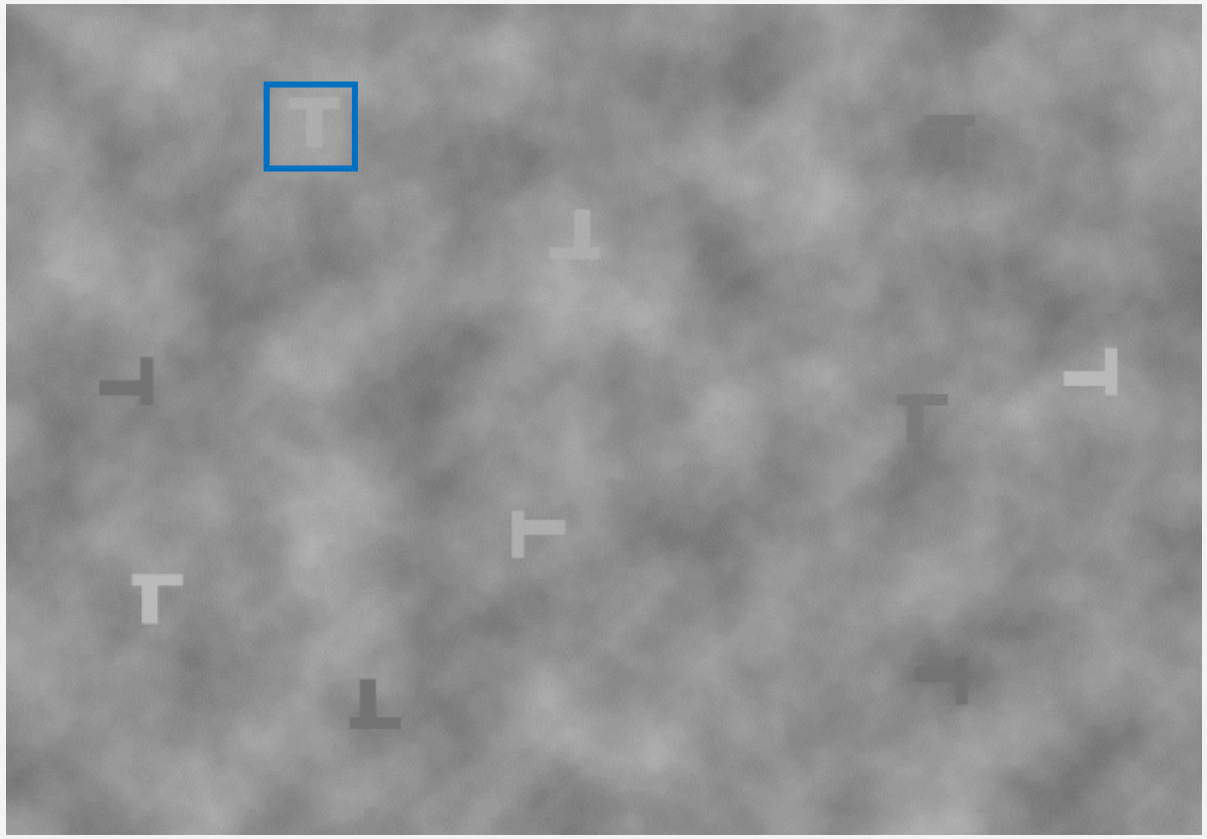


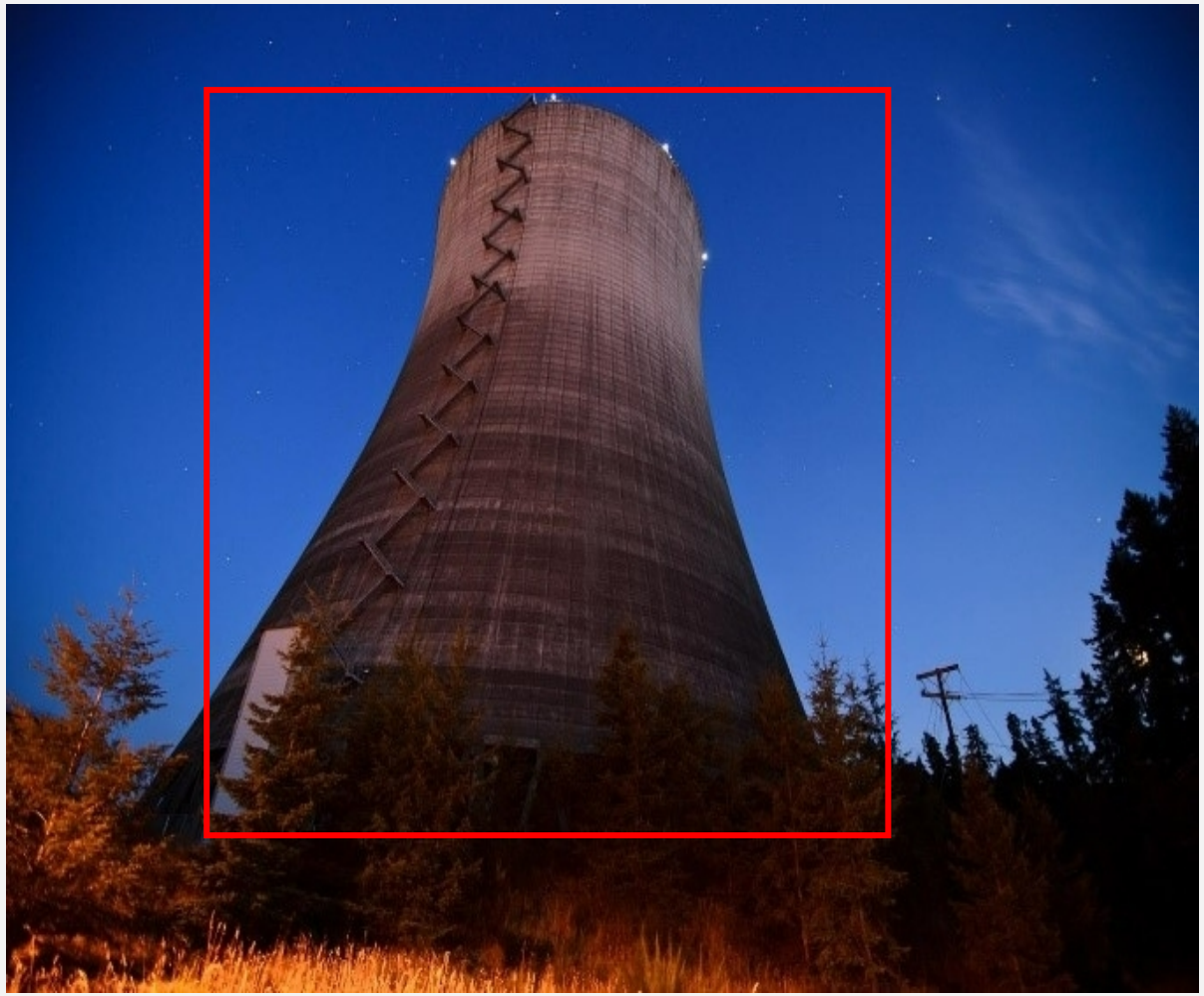
For target present, misses (two types) significantly decreased performance across model accuracies .



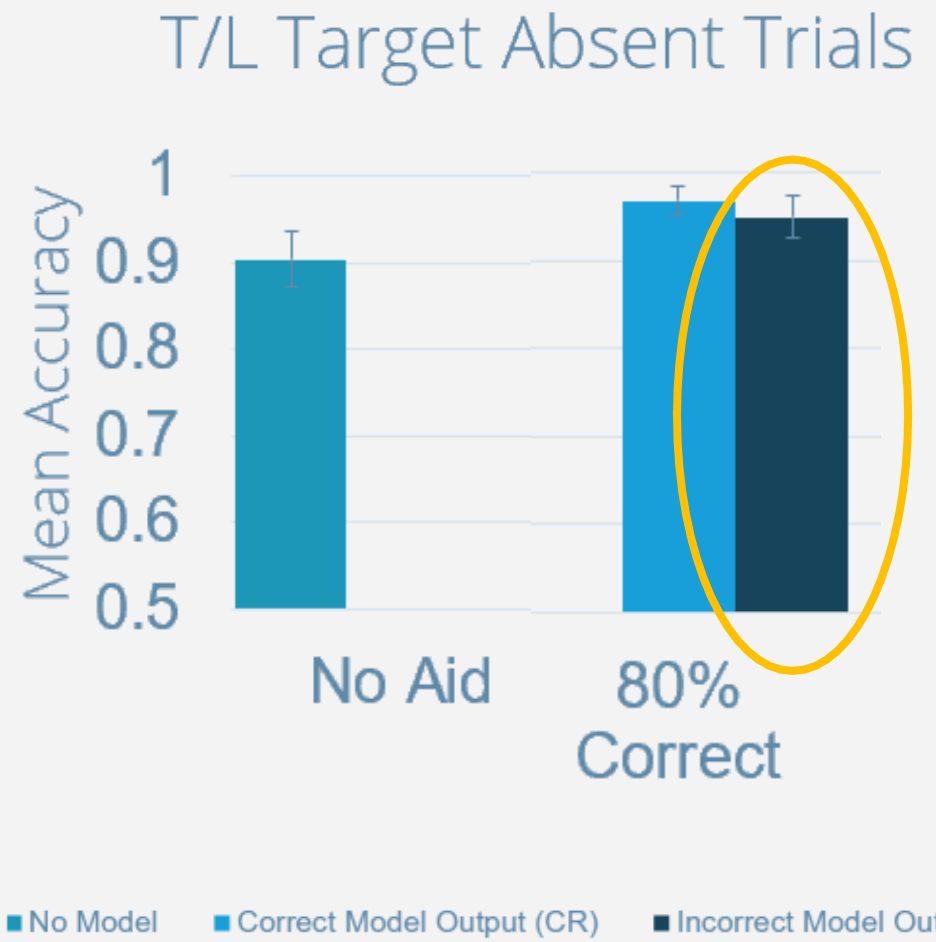
ML impact differs by domain. Expertise is likely culprit.

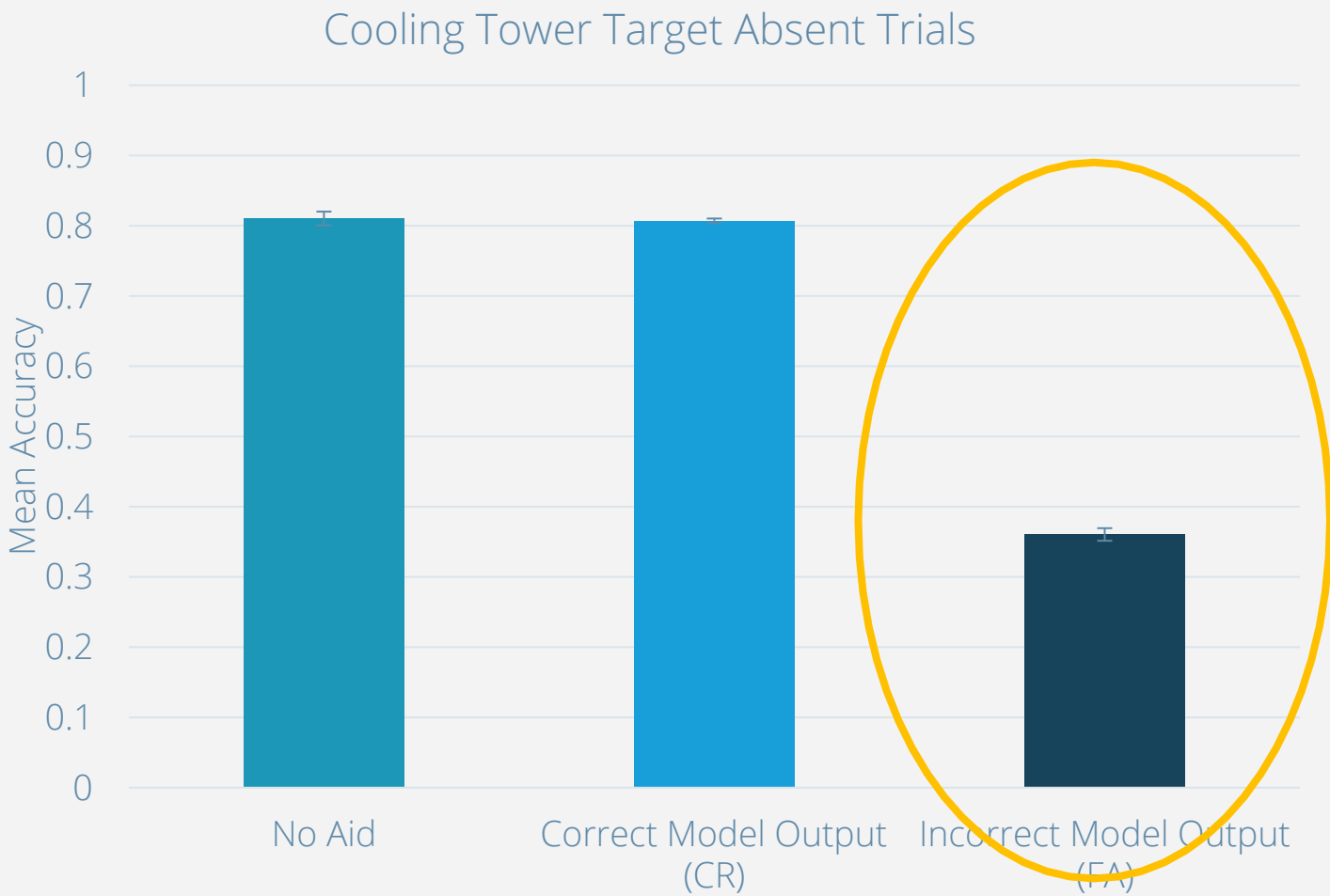
Do wrong model impacts translate to a safeguards use case? Experiments at the 80% model accuracy level.



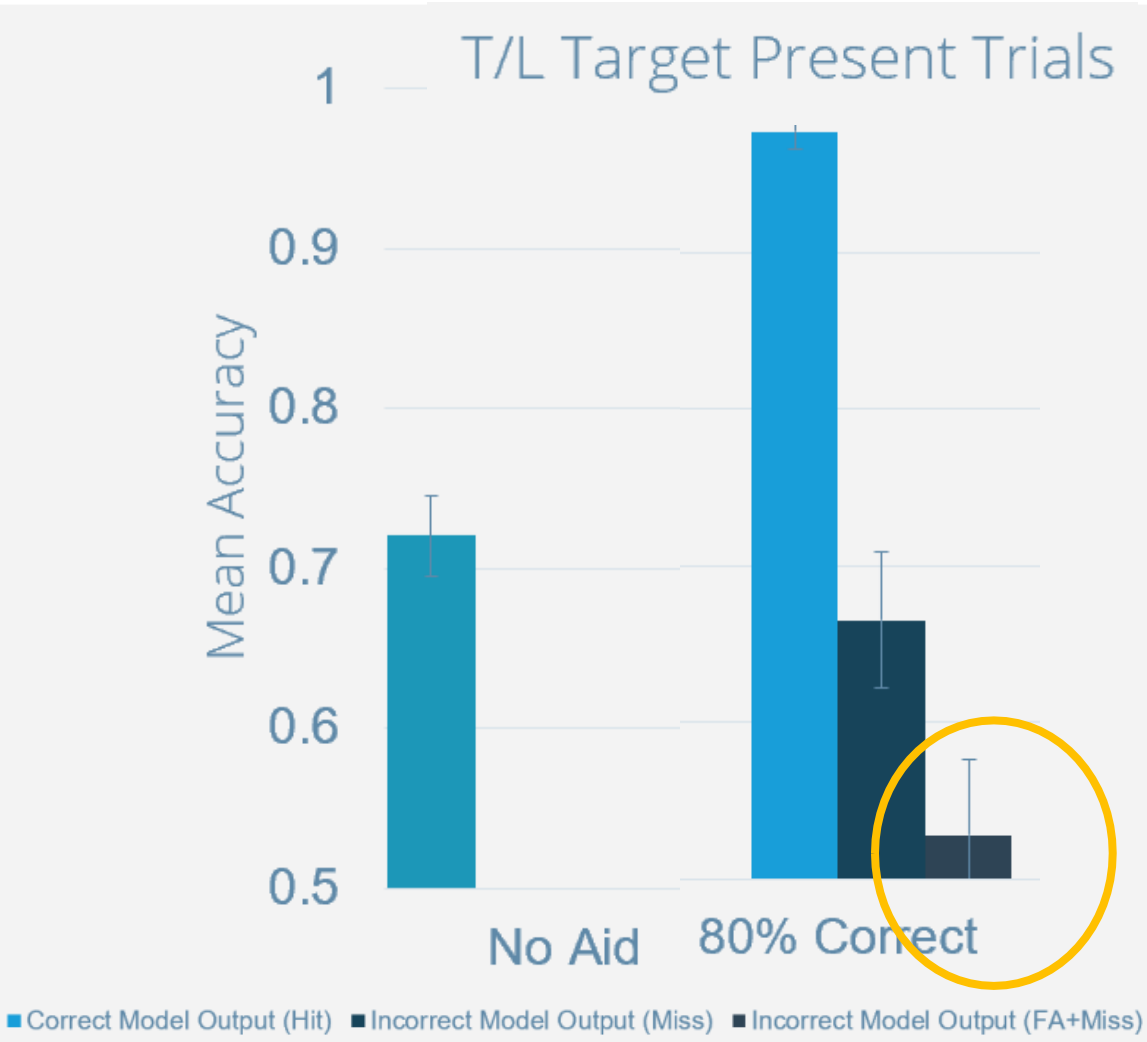


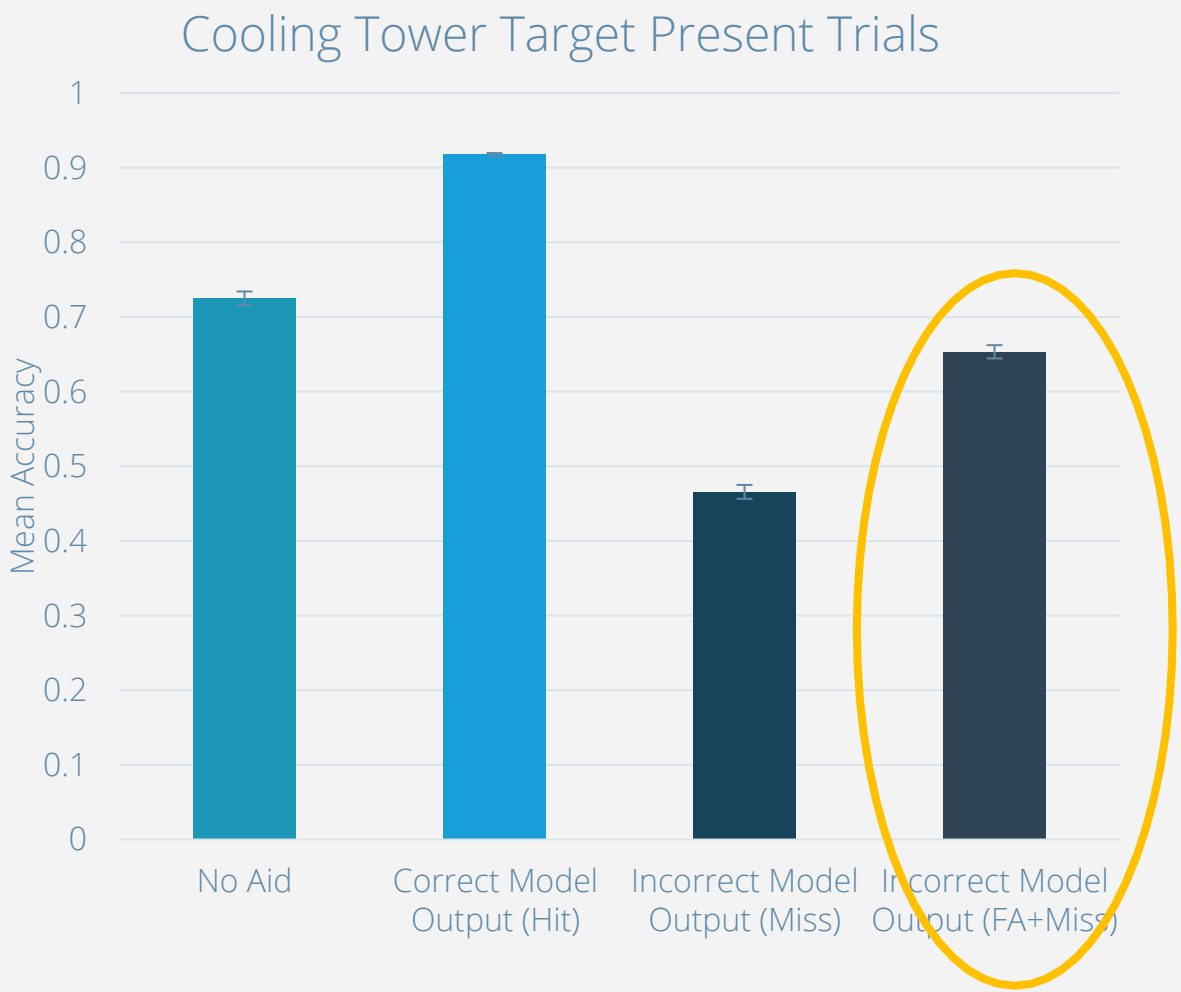
For “target absent”, false alarms had a new negative performance impact.





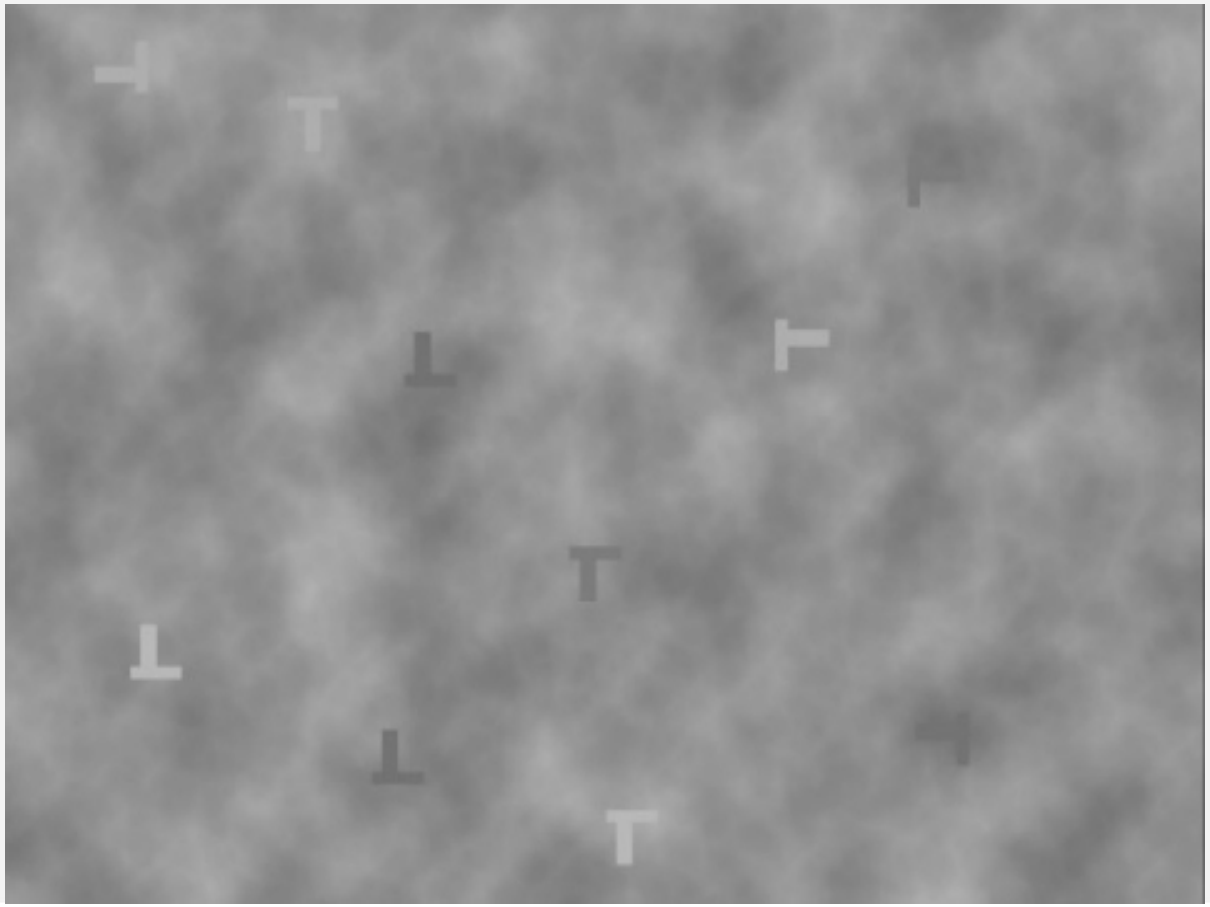
For “target present”, FA+Miss impact appears mitigated.





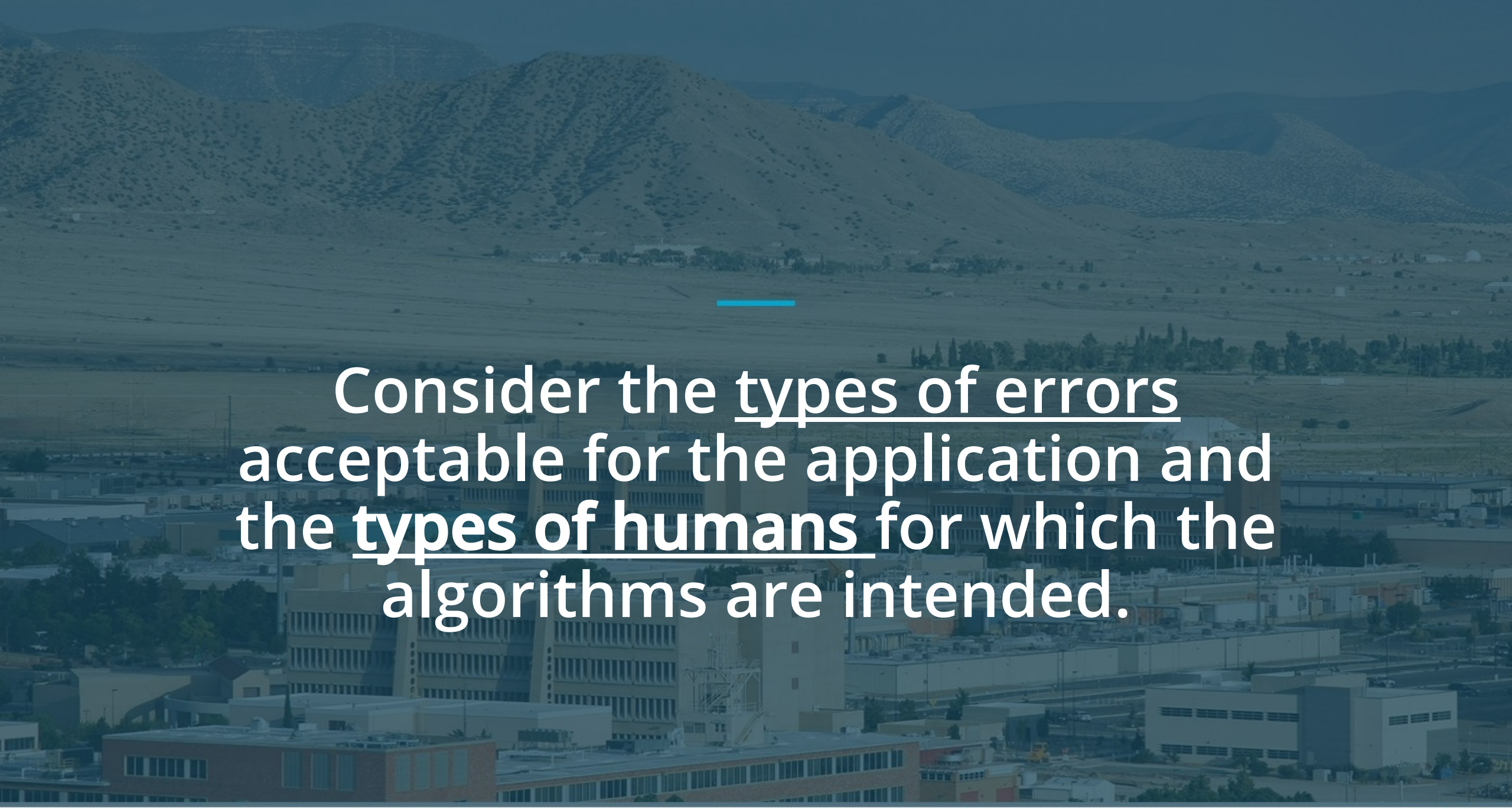
Why the difference?

Ts and Ls are easy to identify.



Cooling towers are more difficult.

Experience plays a role.


Consider the types of errors acceptable for the application and the types of humans for which the algorithms are intended.