The Utopia of Human-AI Collaboration

Besmira Nushi
Microsoft Research






Besmira Nushi
Microsoft Research

Gagan Bansal
University of Washington

Megha Srivastava
Stanford University

Ece Kamar
Microsoft Research

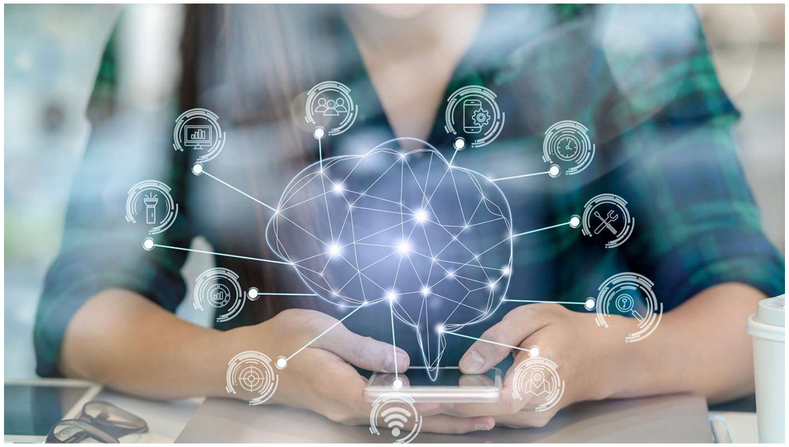
Dan Weld
University of Washington
Allen Institute for AI

Eric Horvitz
Microsoft Research

The promise of AI

Automation

+

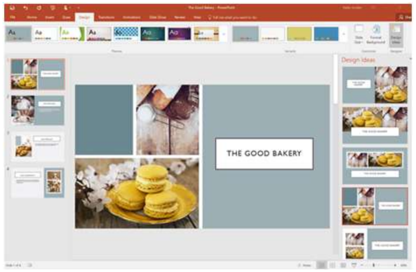


Collaboration

Promising Human-AI Collaborations

Decision-Making

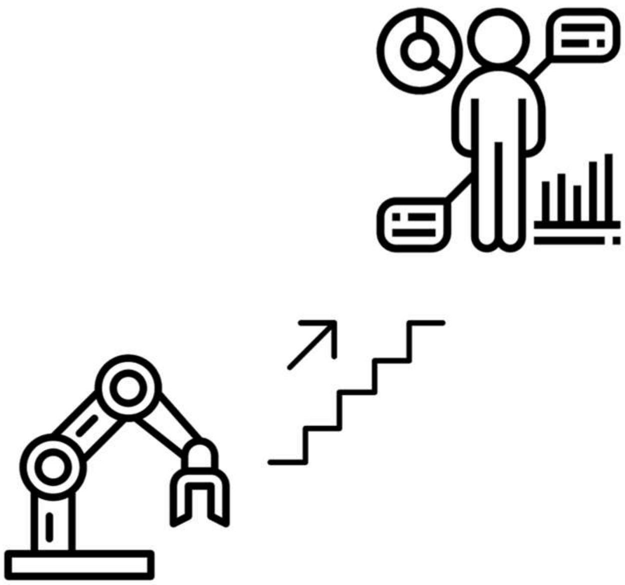
Productivity



Creativity

Science

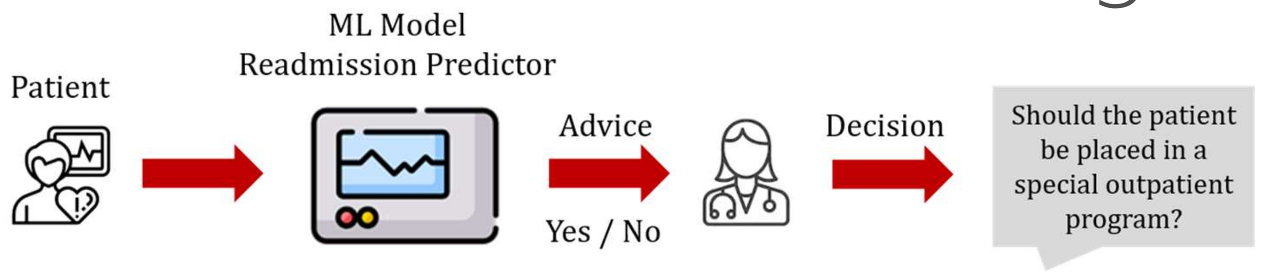
Automation vs. Collaboration

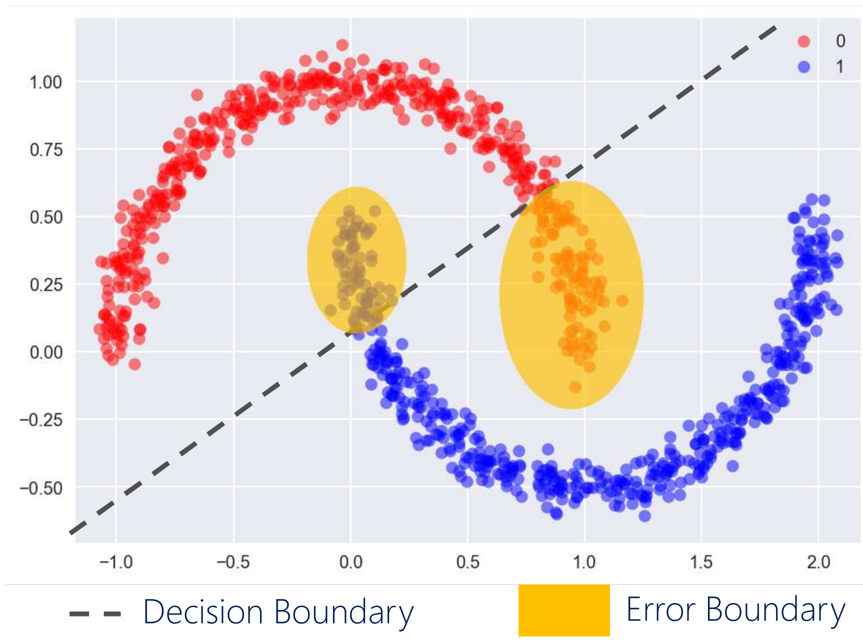


What is a good collaborator?

Human Collaborator	AI Collaborator
Capable	Accurate
Efficient	Fast
Reliable	Reliable, Robust
Good communicator	Intelligible, Transparent
Consistent over time	Backward Compatible
Diverse skillset	Complementary
Fun	Usable + Interactive + more

AI-Assisted Decision-Making



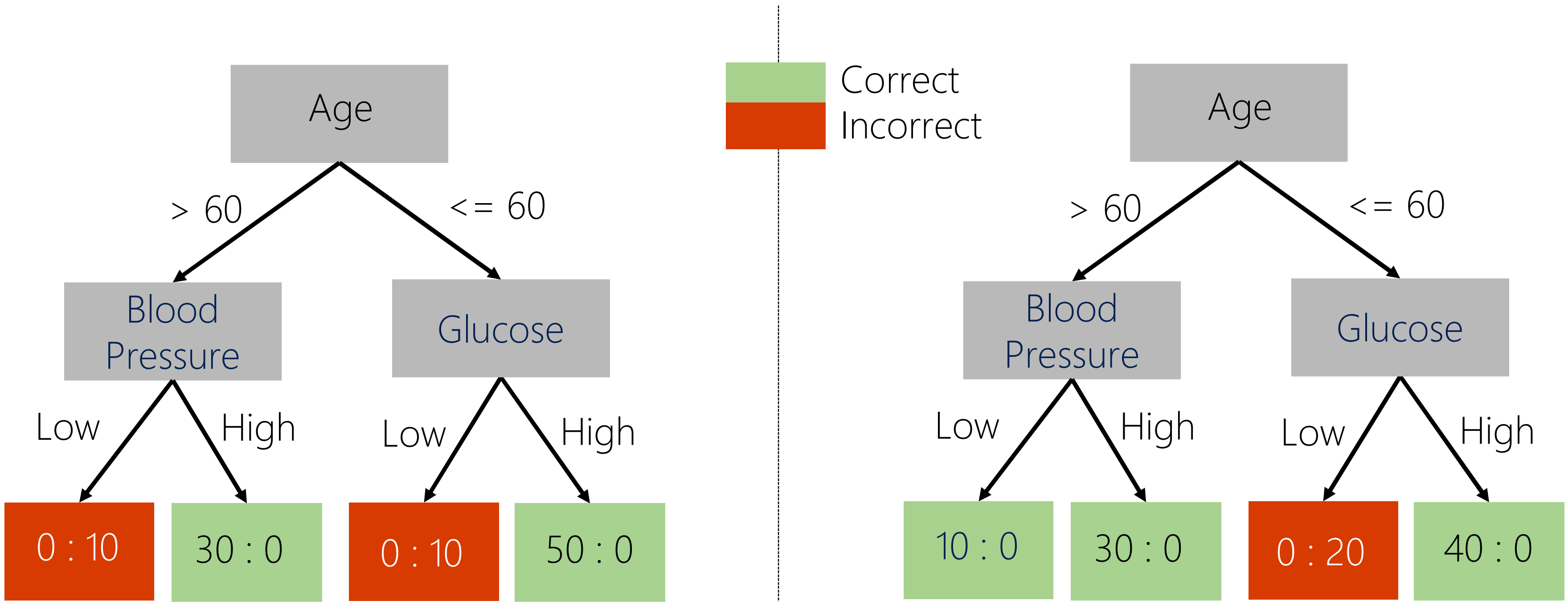


What is a good collaborator?

Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance [Bansal et al., HCOMP 2019]

Beyond Accuracy: Simple Error Boundaries

Accuracy = 80%



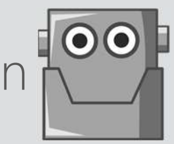
1) High blood pressure
2) Low glucose

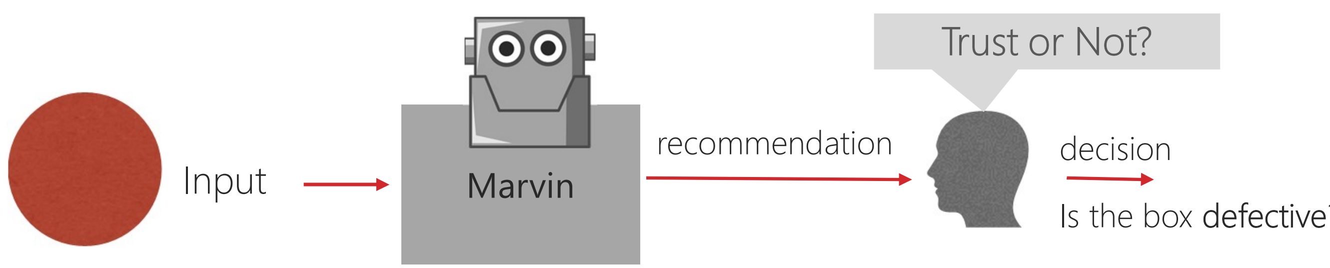
1) Low glucose

Caja

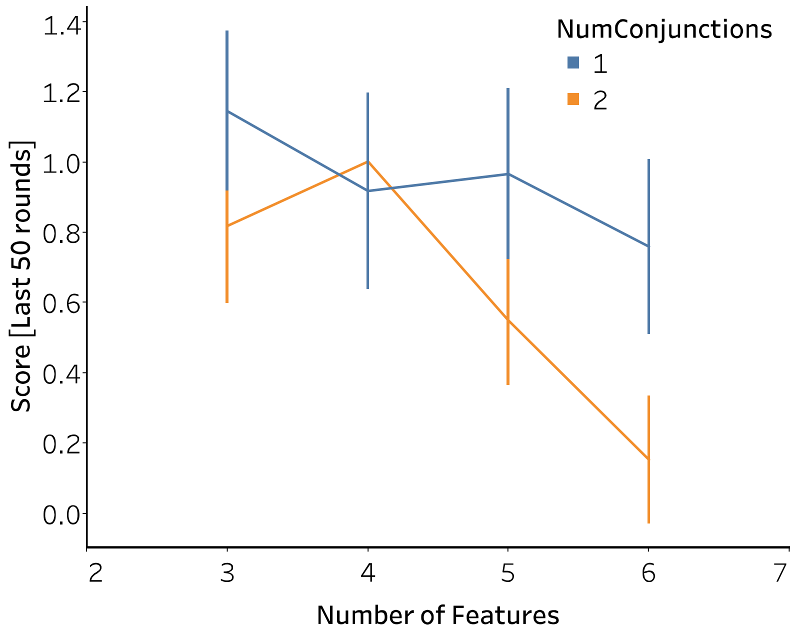
https://github.com/gagb/caja

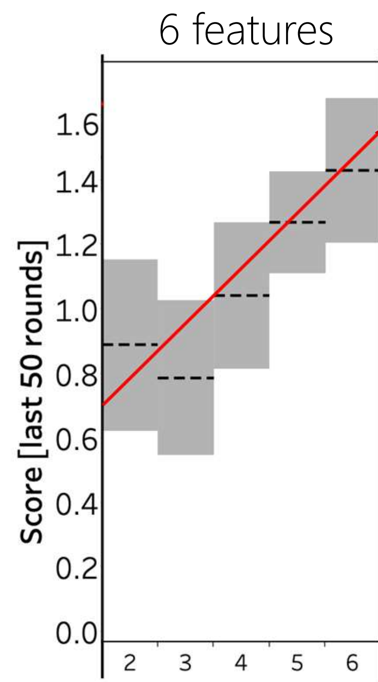
Caja: a platform for user studies

1. Imagine you are a factory worker…
2. On an assembly line, boxes with various features arrive one-by-one…
3. You have a robot assistant named Marvin 
4. Decide which objects are defective
5. Mistakes are costly ($0.04 correct, -$0.16 wrong)



Beyond Accuracy: Simple Error Boundaries





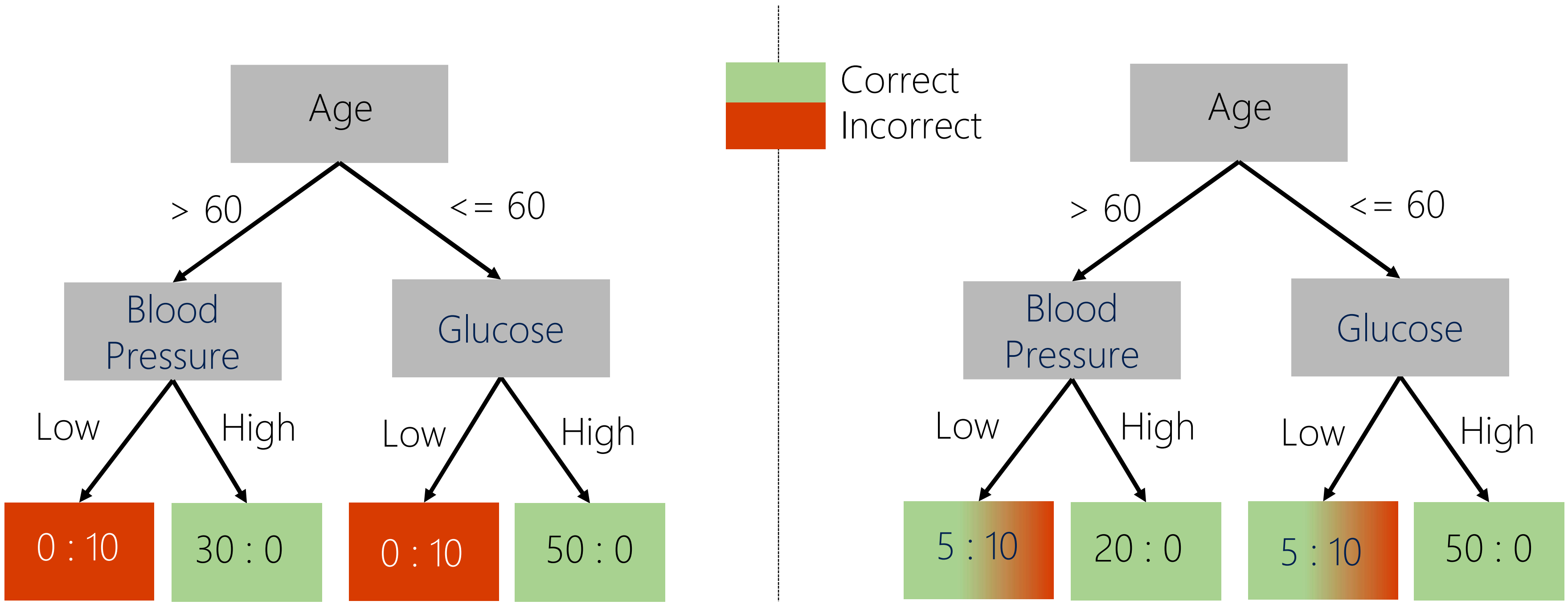
More specific errors.

Performance decreases with the number of conjunctions.

Performances increases as num. of literals increase.

Beyond Accuracy: Non-stochastic Error Boundaries

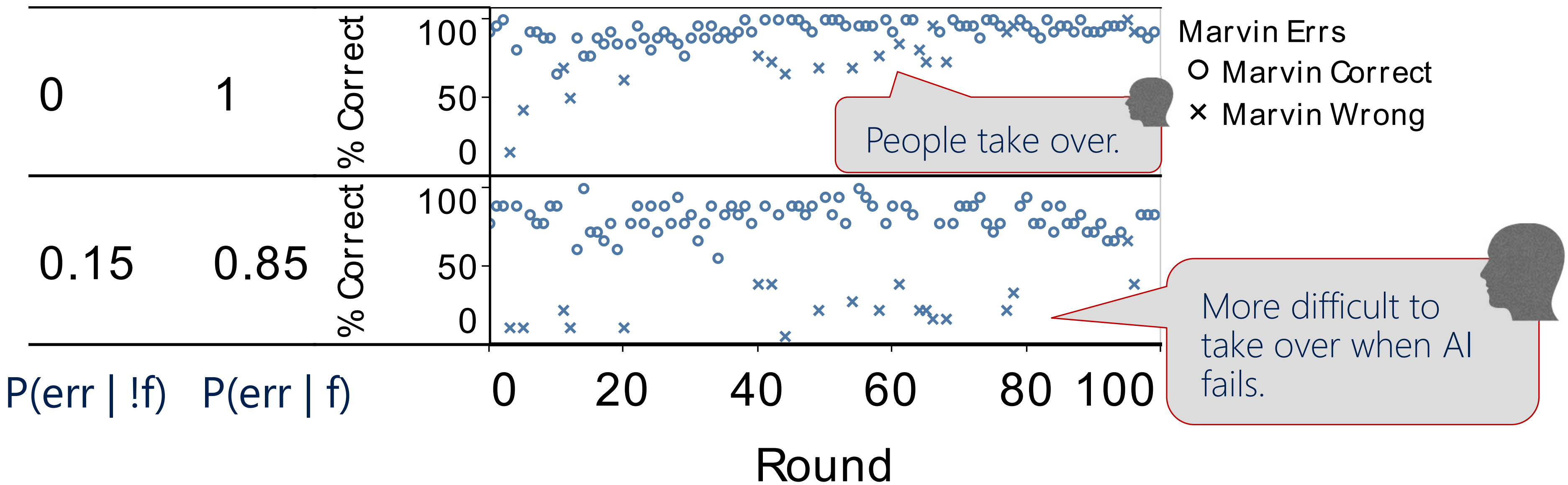
Accuracy = 80%



1) High blood pressure
2) Low glucose

1) High blood pressure (p = 0.67)
2) Low glucose (p = 0.67)

Beyond Accuracy: Non-stochastic Error Boundaries



Updates in Human-AI Collaboration

TRANSPORTATION | CARS | TESLA

83

Tesla can change so much with over-the-air updates that it's messing with some owners' heads

Praise for a recent software fix to the Model 3's braking is met with worry that different update slowed some customers' cars

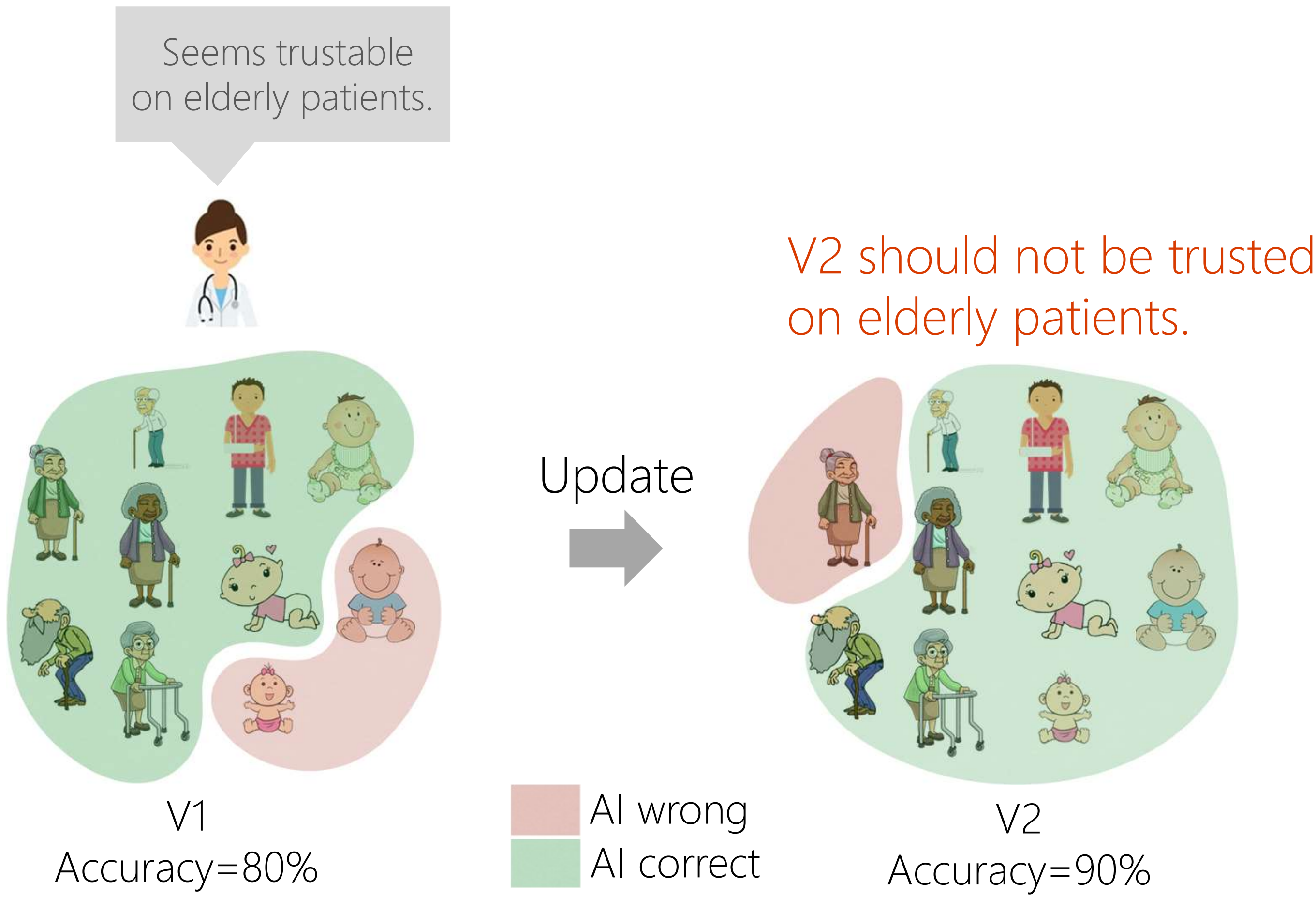
By Sean O'Kane | @sokane1 | Jun 2, 2018, 1:00pm EDT

This week was different, though, because it showed just how far the company can go with those updates. With a swift change in the software, the company showed it can reach as deep as the systems that control the brakes. It creates the feeling that you could get out of your car one night, and by the time you get back in the next morning, the car could do some things — maybe everything — in a totally different way.

OUR PRODUCTS "ARE A BIT MYSTERIOUS, AND DO COOL THINGS, AND SOMETIMES THEY DO SOMETHING CREEPY OR HARMFUL," RINESI SAYS

Rinesi says it's also hard to define "software" in the first place since much of what modern technology does relies on things that live outside the physical object — in this case, the car. "You don't buy a car, or a phone, or soon enough a house or a medical implant or whatever: you buy an interface to, or an aspect of, a huge platform-company-ecosystem-whatever that changes by the minute," he says.

Beyond Accuracy: Backward Compatible Error Boundaries



Trust Compatibility Score

Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff
[Bansal et al., AAAI 2019]

An Empirical Analysis of Backward Compatibility in Machine Learning Systems
[Srivastava et al., KDD 2020]

$$\mathrm{BTC}(v1, v2) = \frac{\#(\mathbf{v1{=}Right} \cap \mathbf{v2{=}Right})}{\#(\mathbf{v1{=}Right})}$$

Goal: v2 should maintain trust.
How much trust is preserved?

Error Compatibility Score

Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff
[Bansal et al., AAAI 2019]

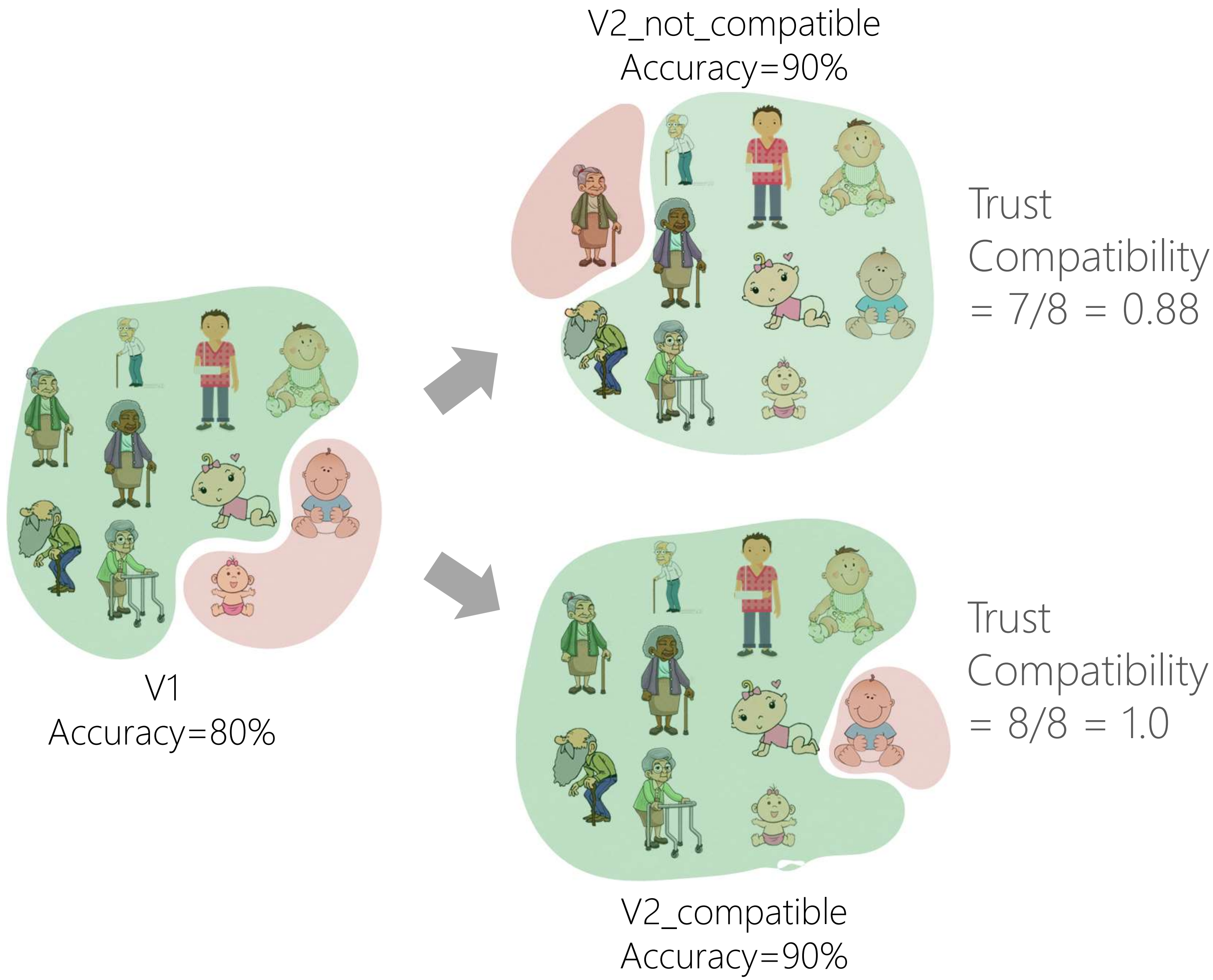
An Empirical Analysis of Backward Compatibility in Machine Learning Systems
[Srivastava et al., KDD 2020]

$$\mathrm{BEC}(v1, v2) = \frac{\#(\mathbf{v2{=}Wrong} \cap \mathbf{v1{=}Wrong})}{\#(\mathbf{v2{=}Wrong})}$$

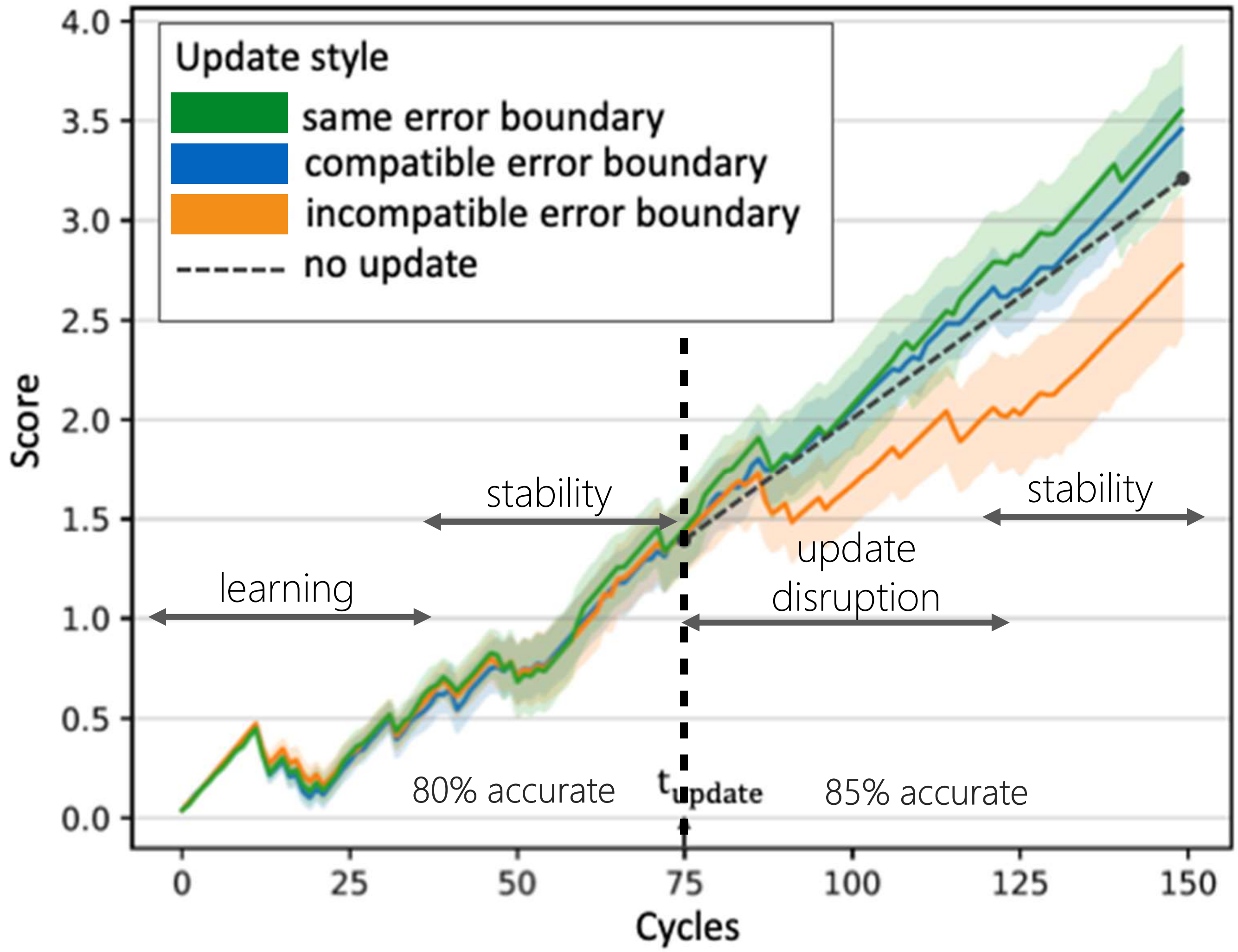
Goal: v2 should not introduce any new errors.
What portion of errors are not new?

Trust Compatibility Score

$$\frac{\#(v1=\text{Right} \cap v2=\text{Right})}{\#(v1=\text{Right})}$$

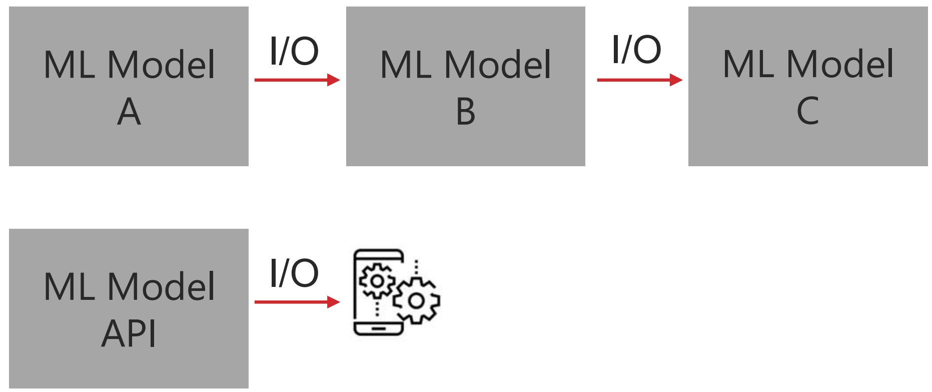


Updates can break team performance



Putting models into a system perspective

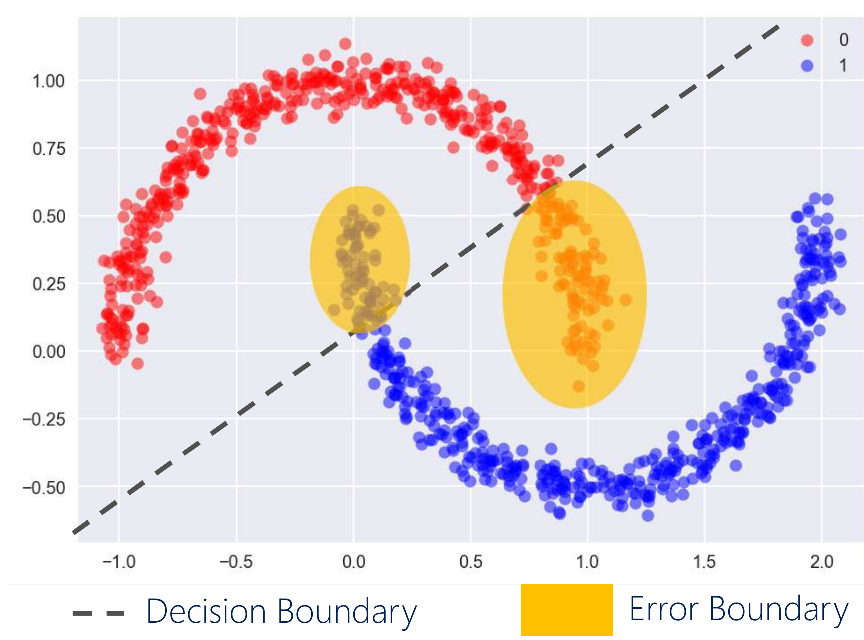
Software System: component-component collaboration



Sociotechnical System: Human-AI collaboration

What is a good collaborator?

Desirable properties beyond accuracy



Simple
Non-stochastic
Backward Compatible

Error Boundaries

Human-Centered ML Optimization
i.e. Good collaborators and where to find them?

Training Compatible Models

Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff
[Bansal et al., AAAI 2019]

Reformulated loss function

$$L_c = L + \lambda_c \cdot \underbrace{\mathcal{D}(v_1, v_2)}_{\text{Dissonance}}$$

New-error dissonance

$$\mathcal{D}(x, y, v_1, v_2) = 1\,(v_1(x) = y) \cdot L(x, y, v_2)$$

Imitation dissonance

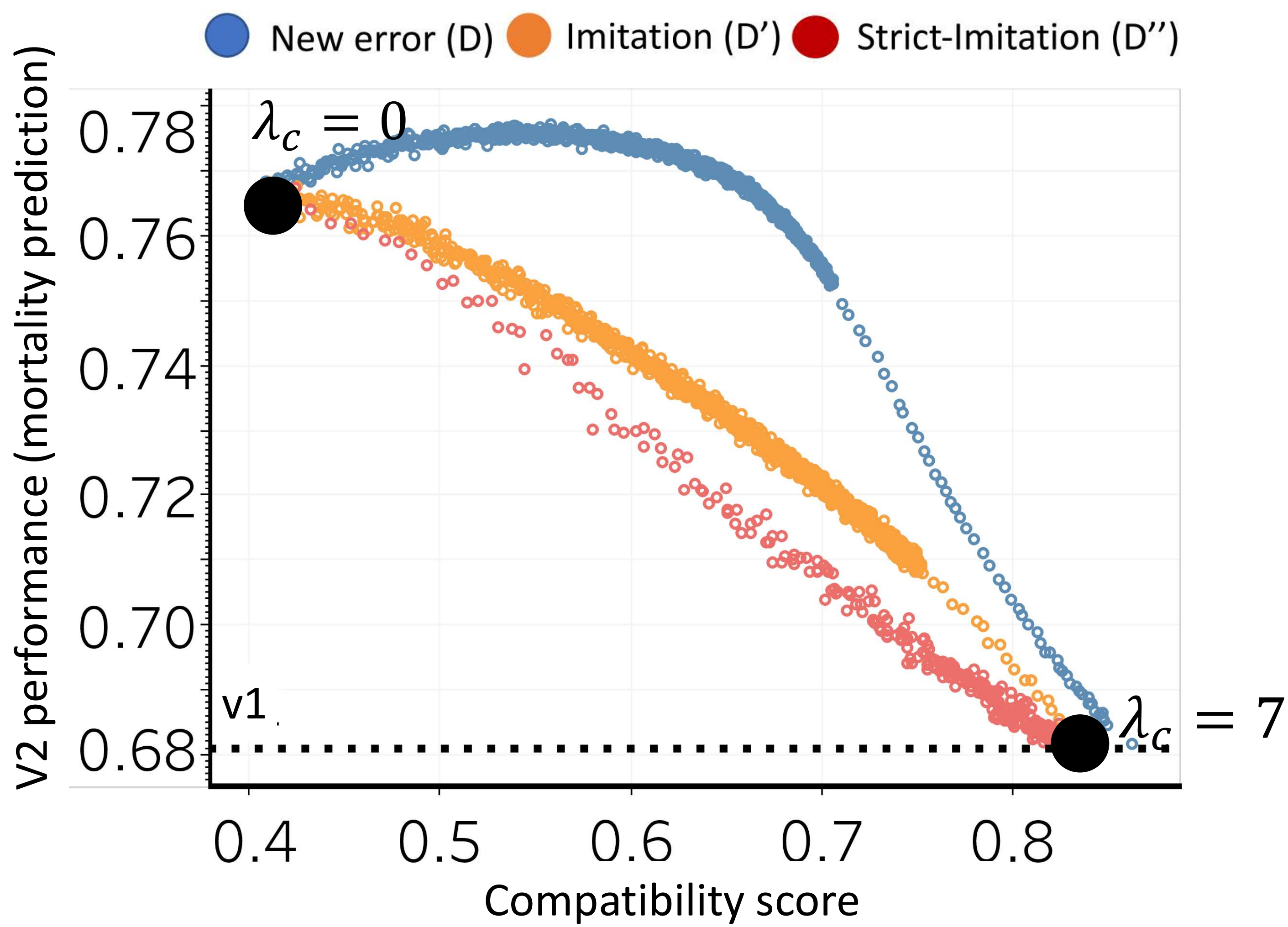
$$\mathcal{D}(x, y, v_1, v_2) = L(x, v_1, v_2)$$

Strict imitation dissonance

$$\mathcal{D}(x, y, v_1, v_2) = 1\,(v_1(x) = y) \cdot L(x, v_1, v_2)$$

Exploration graphs

Compatibility can be planned



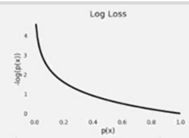
Backward Compatibility Analysis

https://github.com/microsoft/backwardcompatibilityML

with: Xavier Fernandes, Juan Lema, Nicholas King

LOSS FUNCTIONS + METRICS

New Error



$$1\,(v_1(x) = y) \cdot L(x, y, v_2)$$

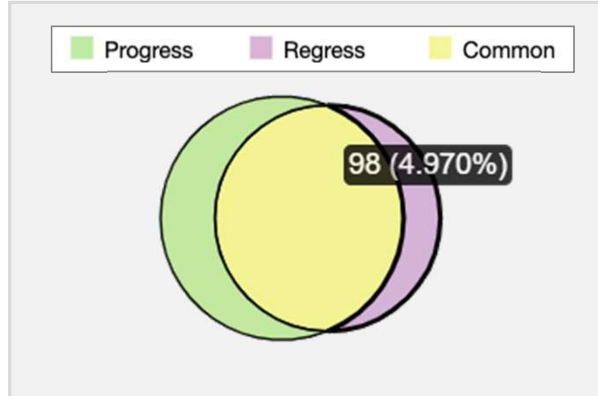
Strict Imitation

$$1\,(v_1(x) = y) \cdot L(x, v_1, v_2)$$

PyTorch

TensorFlow

VISUALIZATION TOOL



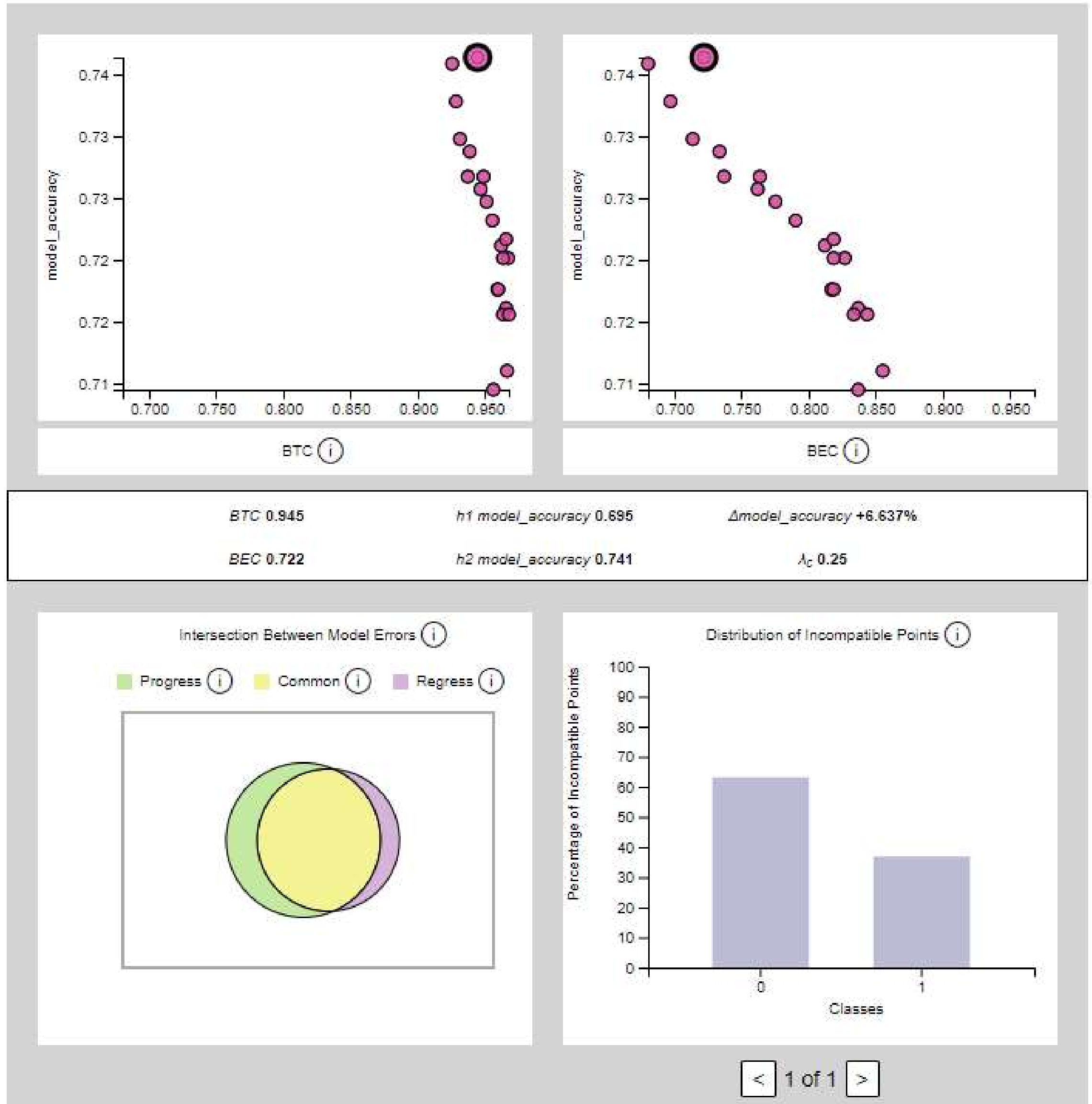
jupyter

Backward Compatibility Analysis

https://github.com/microsoft/backwardcompatibilityML

with: Xavier Fernandes, Juan Lema, Nicholas King

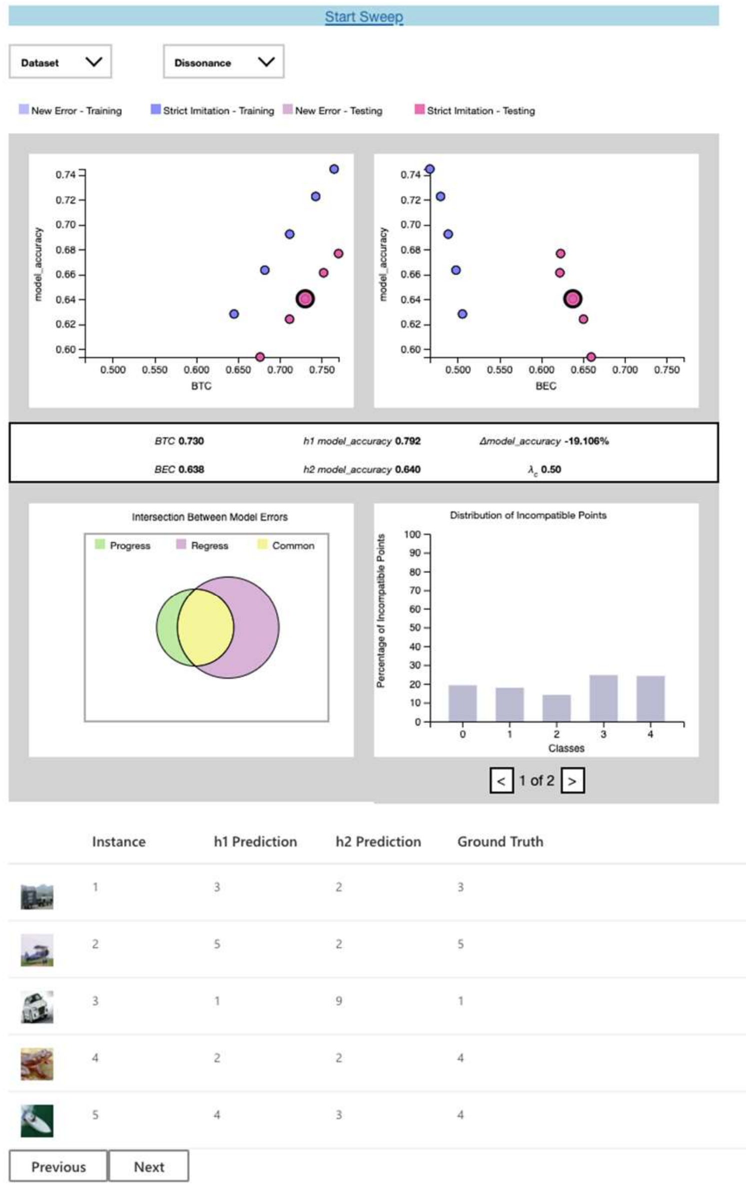
FICO
Credit Risk Prediction



Backward Compatibility Analysis

https://github.com/microsoft/backwardcompatibilityML

with: Xavier Fernandes, Juan Lema, Nicholas King



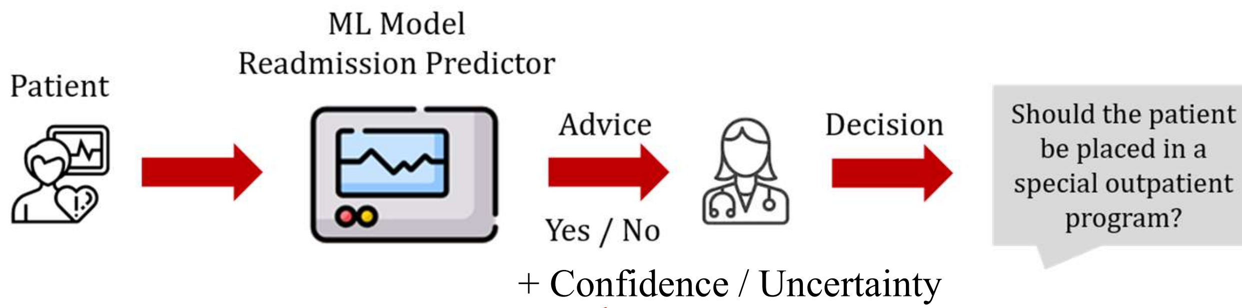

```
get_instance_image()
get_instance_metadata()
```

CIFAR-10

Being accurate where it matters

Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork
[Bansal et. al, AAAI 2021]

Optimizing AI for teamwork



Simplification of error boundary.

Utility Matrix
(Cost of human effort $\lambda = 0.5$, Cost of mistake $\beta = 1$)

Meta-decision/Decision	Correct	Incorrect
Accept	1.0	-1.0
Solve	0.5	-1.5

Being accurate where it matters

Optimizing AI for teamwork

Utility Matrix
(Cost of human effort $\lambda = 0.5$, Cost of mistake $\beta = 1$)

Meta-decision/Decision	Correct	Incorrect
Accept	1.0	-1.0
Solve	0.5	-1.5

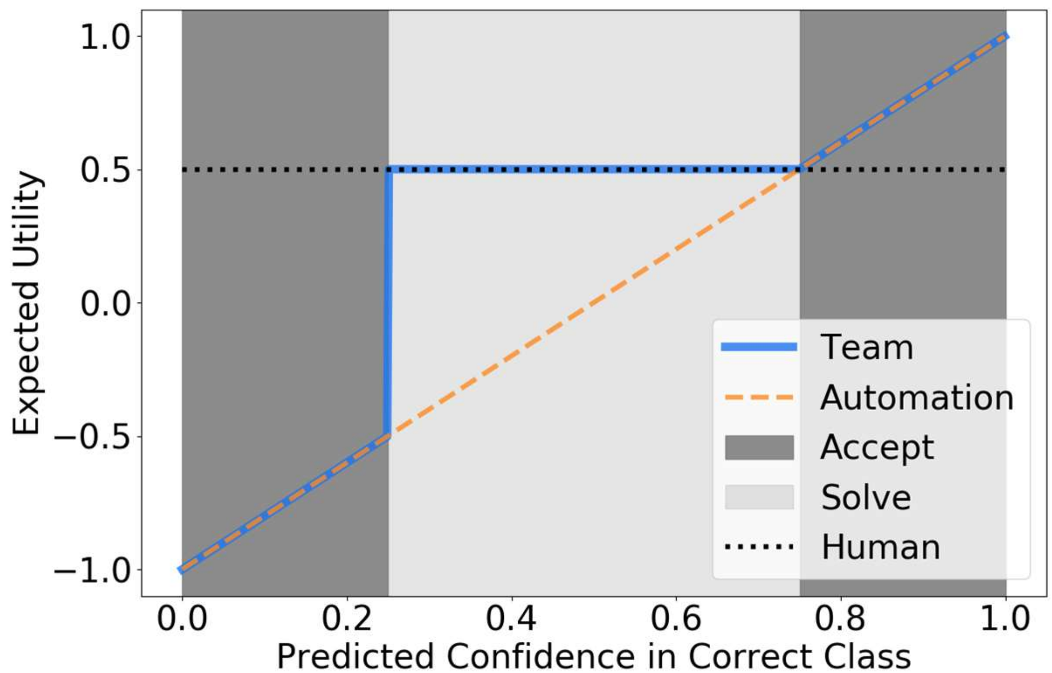
$$P(\mathbf{Accept}) = \begin{cases} 1, \text{if conf} \geq \tau \\ 0, \quad\quad\quad else \end{cases}$$

$$\tau = a - \frac{\lambda}{1 + \beta}$$

a : accuracy of user
β : cost of mistake
λ : cost of handoff

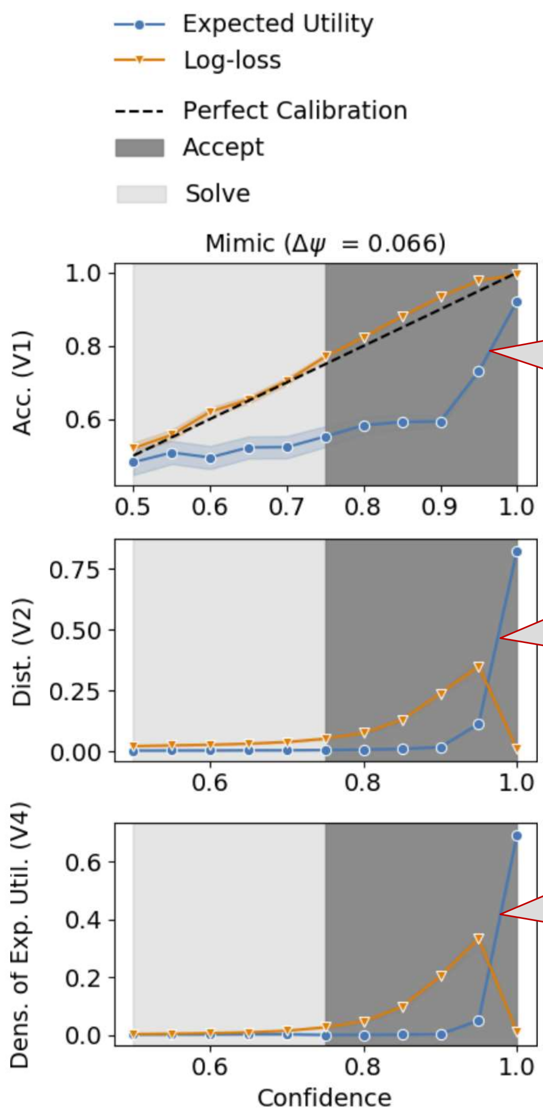
Expected Team Utility

a : accuracy of user
β : cost of mistake
λ : cost of handoff



$$(a = 1.0, \beta = 1.0, \lambda = 0.5) \rightarrow \tau = 0.75$$

Expected Team Utility



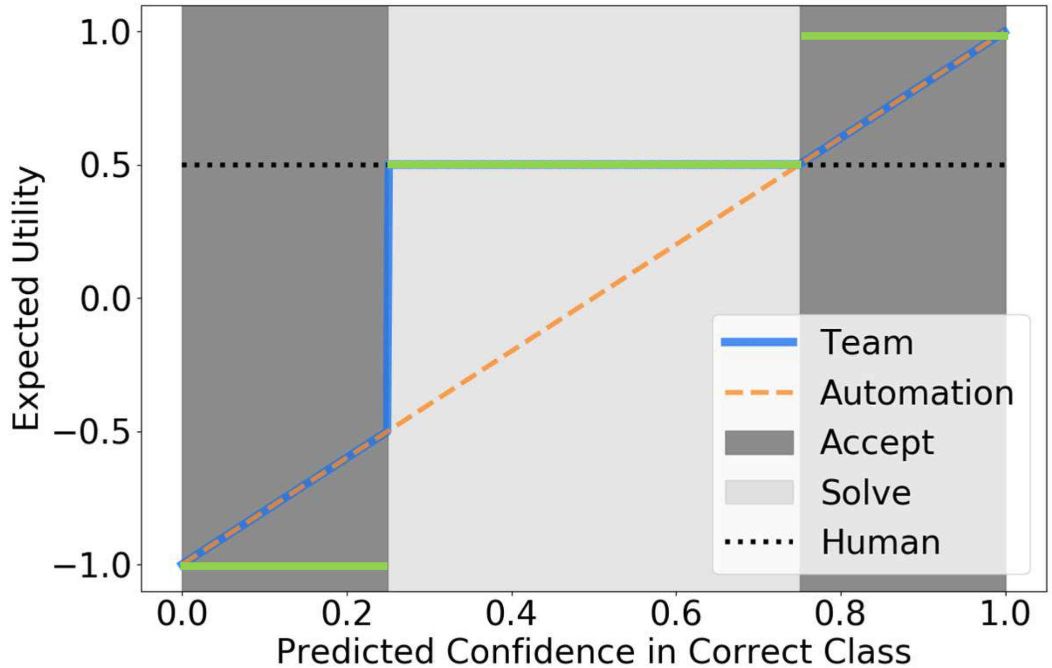
More accurate when the human is likely to trust the model.

Higher density in the higher accuracy, confidence regions.

Higher overall empirical utility.

Expected vs. Empirical Team Utility

a : accuracy of user
β : cost of mistake
λ : cost of handoff



$$(a = 1.0, \beta = 1.0, \lambda = 0.5) \rightarrow \tau = 0.75$$

Expected vs. Empirical Team Utility

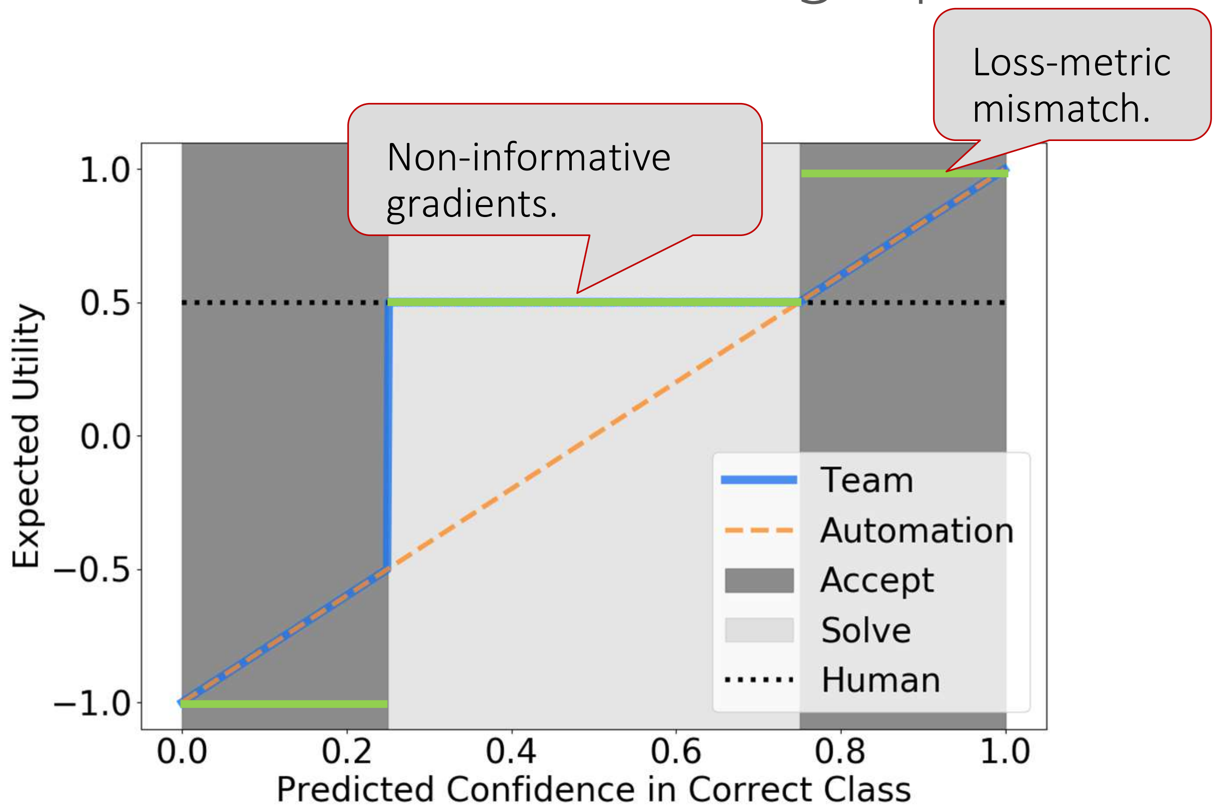
	Expected Utility Loss		
Dataset	**Δ Accuracy**	**Δ Expected Util.**	**Δ Emp. Util.**
Fico	-0.247	**0.013**	-0.075
German	-0.015	0	-0.019
MIMIC	-0.004	**0.066**	-0.035
Moons	-0.02	**0.079**	-0.006
recidivism	-0.17	**0.015**	-0.02
Scenario1	-0.165	**0.102**	**0.061**

Expected utility increases

Empirical utility decreases

Expected vs. Empirical Team Utility

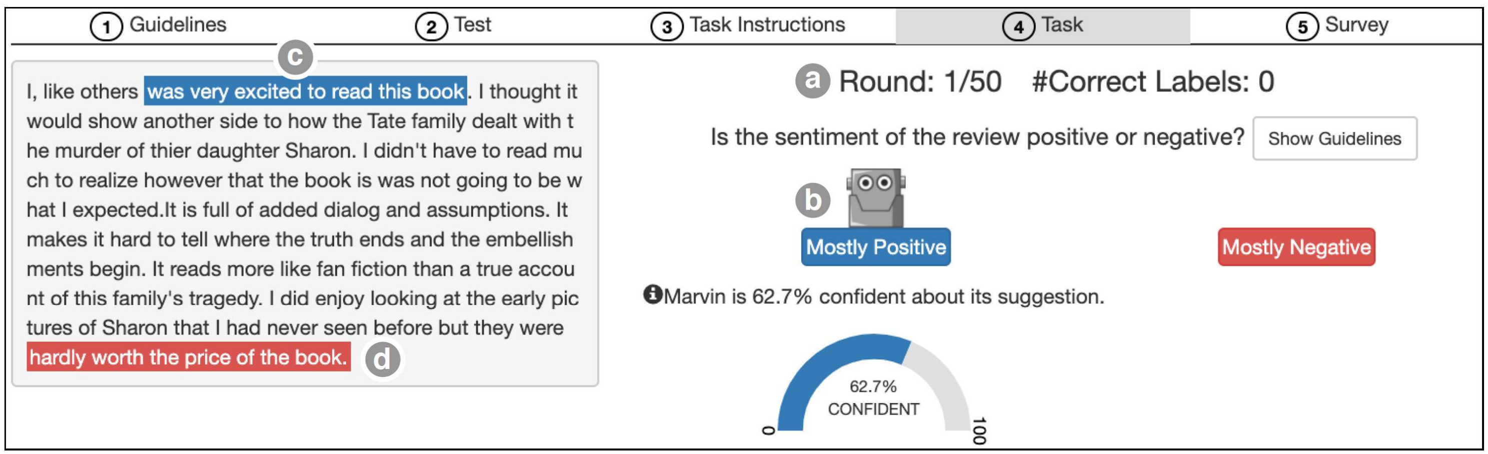
HAIC and Machine Learning Optimization



Explanations for HAIC

Does the Whole Exceed its Parts? The Effect of AI Explanations on **Complementary** Team Performance.
[Bansal and Wu et al., CHI 2021]

NLP Tasks: Sentiment Analysis and SAT Questions



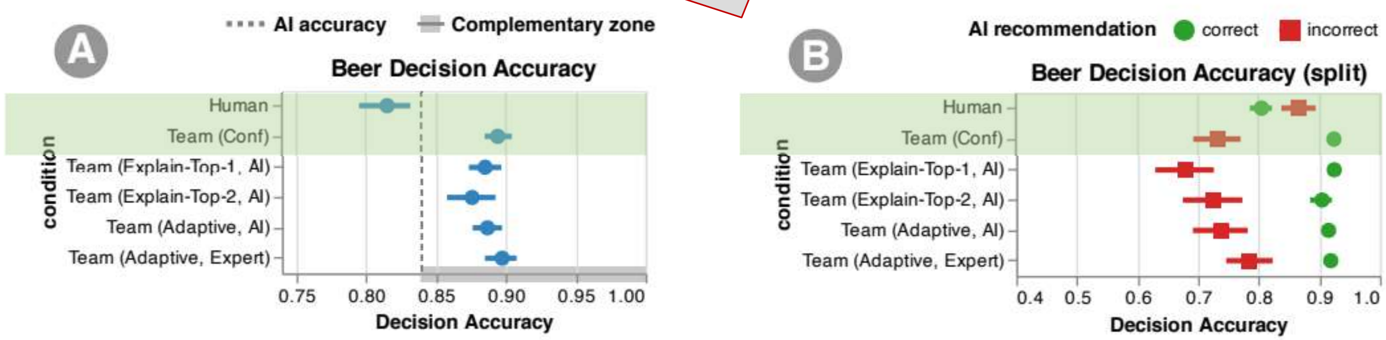
Human alone
AI (conf) + Human
AI (conf + explanations top1) + Human
AI (conf + explanations top2) + Human
AI (conf + explanations adaptive) + Human

Explanations for HAIC

Does the Whole Exceed its Parts? The Effect of AI Explanations on **Complementary** Team Performance.
[Bansal and Wu et al., CHI 2021]

Explainability for Complementary Human-AI teams

Confidence helps for taking over at the right moment.

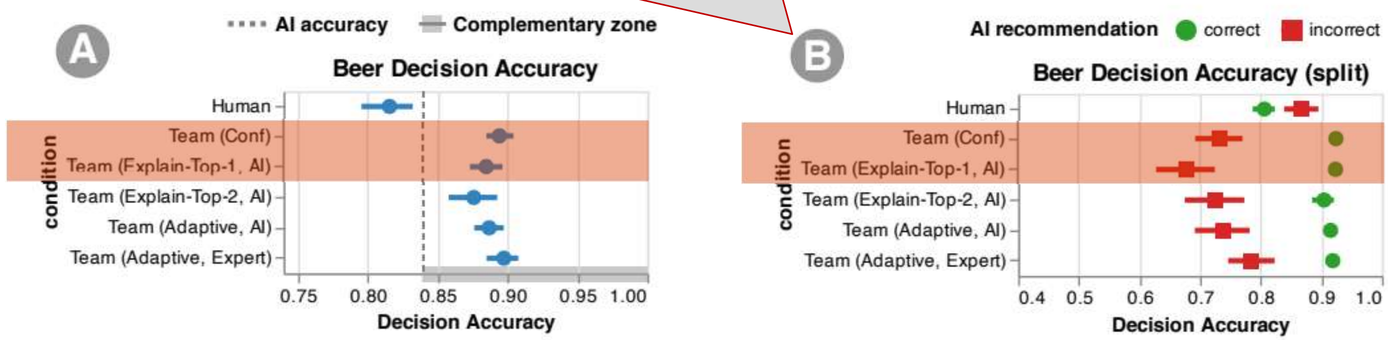


Explanations for HAIC

Does the Whole Exceed its Parts? The Effect of AI Explanations on **Complementary** Team Performance.
[Bansal and Wu et al., CHI 2021]

Explainability for Complementary Human-AI teams

Difficult to improve over confidence via explanations.
People trust AI even when it is wrong.



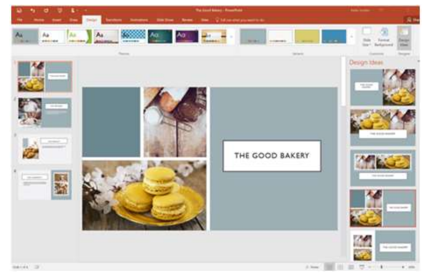
Explainability for handing over control and supporting complementarity. i.e. Building justified trust.

How do we run large-scale experimental studies on real high-stake domains together with decision-making professionals?

Promising Human-AI Collaborations

Decision-Making

Productivity



Creativity

Science

Comparative studies: Human vs. Machine representations

Human-interpretable representations

Concept/Discovery summarization