The promise of AI

Automation + Collaboration
Promising Human-AI Collaborations

Decision-Making

Productivity

Creativity

Science
Automation vs. Collaboration
What is a good collaborator?

<table>
<thead>
<tr>
<th>Human Collaborator</th>
<th>AI Collaborator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capable</td>
<td>Accurate</td>
</tr>
<tr>
<td>Efficient</td>
<td>Fast</td>
</tr>
<tr>
<td>Reliable</td>
<td>Reliable, Robust</td>
</tr>
<tr>
<td>Good communicator</td>
<td>Intelligible, Transparent</td>
</tr>
<tr>
<td>Consistent over time</td>
<td>Backward Compatible</td>
</tr>
<tr>
<td>Diverse skillset</td>
<td>Complementary</td>
</tr>
<tr>
<td>Fun</td>
<td>Usable + Interactive + more</td>
</tr>
</tbody>
</table>
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: 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)

https://github.com/gagb/caja
Beyond Accuracy: Simple Error Boundaries

Performance decreases with the number of conjunctions.

Performances increases as num. of literals increase.

More specific errors.
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

People take over.

More difficult to take over when AI fails.
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.

---

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

V1
Accuracy=80%

Seems trustable on elderly patients.

Update

V2 should not be trusted on elderly patients.

V2
Accuracy=90%

AI wrong
AI correct
\[ \text{BTC}(v1, v2) = \frac{\#(v1=\text{Right} \cap v2=\text{Right})}{\#(v1=\text{Right})} \]

**Goal:** \(v2\) should maintain trust. 
How much trust is preserved?

*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]
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]

$$\text{BEC}(v_1, v_2) = \frac{\#(v_2 = \text{Wrong} \cap v_1 = \text{Wrong})}{\#(v_2 = \text{Wrong})}$$

Goal: $v_2$ should not introduce any new errors. What portion of errors are not new?
Trust Compatibility Score

\[
\frac{(v1=\text{Right} \cap v2=\text{Right})}{(v1=\text{Right})}
\]

V1
Accuracy=80%

V2_not_compatible
Accuracy=90%

Trust Compatibility = 7/8 = 0.88

V2_compatible
Accuracy=90%

Trust Compatibility = 8/8 = 1.0
Updates can break team performance

![Graph showing update styles and their impact on score, stability, learning, update disruption, and accuracy.]

- **Update style**
  - same error boundary
  - compatible error boundary
  - incompatible error boundary
  - no update

- Score:
  - 0.0
  - 0.5
  - 1.0
  - 1.5
  - 2.0
  - 2.5
  - 3.0
  - 3.5
  - 4.0

- Cycles:
  - 0
  - 25
  - 50
  - 75
  - 100
  - 125
  - 150

- **Accuracy**:
  - 80% accurate
  - 85% accurate

- **Phases**:
  - learning
  - stability
  - update disruption
  - stability
Putting models into a system perspective

**Software System**: component-component collaboration

ML Model A  I/O  ML Model B  I/O  ML Model C

**Sociotechnical System**: Human-AI collaboration

ML Model API  I/O  ML Model API
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?
Reformulated loss function
\[ L_c = L + \lambda_c \cdot D(v_1, v_2) \]
Dissonance

New-error dissonance
\[ D(x, y, v_1, v_2) = 1(v_1(x) = y) \cdot L(x, y, v_2) \]

Imitation dissonance
\[ D(x, y, v_1, v_2) = L(x, v_1, v_2) \]

Strict imitation dissonance
\[ D(x, y, v_1, v_2) = 1(v_1(x) = y) \cdot L(x, v_1, v_2) \]
Compatibility can be planned

Exploration graphs

![Graph showing V2 performance (mortality prediction) vs. Compatibility score with different compatibility scores and types: New error (D), Imitation (D'), and Strict-Imitation (D'') at λ_c = 0 and λ_c = 7.](image)
Backward Compatibility Analysis

https://github.com/microsoft/backwardcompatibilityML
with: Xavier Fernandes, Juan Lema, Nicholas King

**New Error**
\[ 1 \cdot (v_1(x) = y) \cdot L(x, y, v_2) \]

**Strict Imitation**
\[ 1 \cdot (v_1(x) = y) \cdot L(x, v_1, v_2) \]

VISUALIZATION TOOL

PyTorch

TensorFlow
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

CIFAR-10

get_instance_image()
get_instance_metadata()
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

ML Model
Readmission Predictor

Patient

Advice
Yes / No

Decision

+ Confidence / Uncertainty

Should the patient be placed in a special outpatient program?

Simplification of error boundary.

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

<table>
<thead>
<tr>
<th>Meta-decision/Decision</th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept</td>
<td>1.0</td>
<td>-1.0</td>
</tr>
<tr>
<td>Solve</td>
<td>0.5</td>
<td>-1.5</td>
</tr>
</tbody>
</table>
Optimizing AI for teamwork

Being accurate where it matters

<table>
<thead>
<tr>
<th>Meta-decision/Decision</th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept</td>
<td>1.0</td>
<td>-1.0</td>
</tr>
<tr>
<td>Solve</td>
<td>0.5</td>
<td>-1.5</td>
</tr>
</tbody>
</table>

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

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

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

$a$ : accuracy of user  
$\beta$ : cost of mistake  
$\lambda$ : cost of handoff
Expected Team Utility

\( a \) : accuracy of user

\( \beta \) : cost of mistake

\( \lambda \) : 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

\[ a : \text{accuracy of user} \]
\[ \beta : \text{cost of mistake} \]
\[ \lambda : \text{cost of handoff} \]

\[(a = 1.0, \beta = 1.0, \lambda = 0.5) \rightarrow \tau = 0.75\]
Expected vs. Empirical Team Utility

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Δ Accuracy</th>
<th>Δ Expected Util.</th>
<th>Δ Emp. Util.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fico</td>
<td>-0.247</td>
<td>0.013</td>
<td>-0.075</td>
</tr>
<tr>
<td>German</td>
<td>-0.015</td>
<td>0</td>
<td>-0.019</td>
</tr>
<tr>
<td>MIMIC</td>
<td>-0.004</td>
<td><strong>0.066</strong></td>
<td><strong>-0.035</strong></td>
</tr>
<tr>
<td>Moons</td>
<td>-0.02</td>
<td>0.079</td>
<td>-0.006</td>
</tr>
<tr>
<td>recidivism</td>
<td>-0.17</td>
<td>0.015</td>
<td>-0.02</td>
</tr>
<tr>
<td>Scenario1</td>
<td>-0.165</td>
<td>0.102</td>
<td>0.061</td>
</tr>
</tbody>
</table>

Expected utility increases
Empirical utility decreases
HAIC and Machine Learning Optimization

Expected vs. Empirical Team Utility

- Non-informative gradients.
- Loss-metric mismatch.
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
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]
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 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?

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

Decision-Making

Productivity

Creativity

Science

Comparative studies: Human vs. Machine representations

Human-interpretable representations

Concept/Discovery summarization