Guidelines for Human-AI Interaction

Saleema Amershi, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Derek DeBellis, Ruth Kikin-Gil, Shamsi Iqbal, Paul Bennett, Dan Weld, Jina Suh, Kori Inkpen, Jaime Teevan, and Eric Horvitz

https://aka.ms/aiguidelines
Guidelines for Human AI Interaction
Learn more: [https://aka.ms/aiguidelines](https://aka.ms/aiguidelines)

**INITIALLY**

1. Make clear what the system can do.
2. Make clear how well the system can do what it can do.

**DURING INTERACTION**

3. Time services based on context.
4. Show contextually relevant information.
5. Match relevant social norms.
6. Mitigate social biases.

**WHEN WRONG**

7. Support efficient invocation.
8. Support efficient dismissal.
9. Support efficient correction.
10. Scope services when in doubt.
11. Make clear why the system did what it did.

**OVER TIME**

12. Remember recent interactions.
13. Learn from user behavior.
14. Update and adapt cautiously.
15. Encourage granular feedback.
16. Convey the consequences of user actions.
17. Provide global controls.
18. Notify users about changes.
Agenda

Intro to the guidelines
Findings and impact
Engineering and AI implications
Challenges for Intelligible AI
Agenda

- Intro to the guidelines
- Findings and impact
- Engineering and AI implications
- Challenges for Intelligible AI
Creating good AI user experiences is hard

Hey autocorrect, stop correcting my swear words you piece of shut. 😤 It really annoys people who can actually spell. #odetoautocorrect #autocorrectfail #icanspell #doh

Autocorrect makes me say things I didn’t Nintendo.

Culver City Firefighters
@CC_Firefighters

While working a freeway accident this morning, Engine 42 was struck by a #Tesla traveling at 65 mph. The driver reports the vehicle was on autopilot. Amazingly there were no injuries! Please stay alert while driving! #abc7eyewitness #ktla #CulverCity #distracteddriving
AI is fundamentally changing how we interact with computing systems
The Consistency Principle

Consistent interfaces and predictable behaviors saves people time and reduces errors.
AI systems are probabilistic and can change over time

Behaviors may change over time

Behaviors may differ in subtly different contexts
Creating the Guidelines for Human-AI Interaction
ACM CHI 2019, Best Paper Honorable Mention Award

Phase 1. Consolidation
Identified themes across 150+ recommendations

Phase 2. Team Evaluation
Modified heuristic evaluation over 13 common AI products

Phase 3. User Evaluation
Systematic analysis of 20 AI products with 49 UX practitioners

Phase 4. Expert Review
Final review with 11 UX practitioners
The guidelines are not a checklist

Additional guidelines may be needed in some scenarios

You are using them “the right way” if you consider them during development
Guidelines for Human AI Interaction
Learn more: https://aka.ms/aiguidelines

INITIALLY

1. Make clear what the system can do.
2. Make clear how well the system can do what it can do.

DURING INTERACTION

3. Time services based on context.
4. Show contextually relevant information.
5. Match relevant social norms.
6. Mitigate social biases.

WHEN WRONG

7. Support efficient invocation.
8. Support efficient dismissal.
9. Support efficient correction.
10. Scope services when in doubt.
11. Make clear why the system did what it did.

OVER TIME

12. Remember recent interactions.
13. Learn from user behavior.
14. Update and adapt cautiously.
15. Encourage granular feedback.
16. Convey the consequences of user actions.
17. Provide global controls.
18. Notify users about changes.
Examples from common AI-based products

MySearchEngine: AI used for query processing, ranking results, filtering spam...
MyAssistant: AI used for speech processing, task support....
MyEmail: AI used for email sorting, entity detection, response generation...
MySocialNetwork: AI used for filtering feed, recommending ads...
<table>
<thead>
<tr>
<th>INITIALLY</th>
<th>DURING INTERACTION</th>
<th>WHEN WRONG</th>
<th>OVER TIME</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Make clear what the system can do.</td>
<td>2. Make clear how well the system can do what it can do.</td>
<td>3. Time services based on context.</td>
<td>4. Show contextually relevant information.</td>
</tr>
<tr>
<td>9. Support efficient correction.</td>
<td>10. Scope services when in doubt.</td>
<td>11. Make clear why the system did what it did.</td>
<td>12. Remember recent interactions.</td>
</tr>
<tr>
<td>17. Provide global controls.</td>
<td>18. Notify users about changes.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Set the right expectations

Coverage: Many people think “everything is on the web”*

Quality: 33% of people use the term “magic” when explaining how search works*

Can be problematic when people overestimate search capabilities for high-stakes tasks

*Dan Russell. The Joy of Search.
Set the right expectations – What can you do?

- Provide documentation (use sparingly)
- Show examples
- Introduce features at appropriate times
- Give people controls
### Guidelines for Human AI Interaction

Learn more: [https://aka.ms/aiguidelines](https://aka.ms/aiguidelines)

<table>
<thead>
<tr>
<th>Initially</th>
<th>During Interaction</th>
<th>When Wrong</th>
<th>Over Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Make clear what the system can do.</td>
<td>2. Make clear how well the system can do what it can do.</td>
<td>3. Time services based on context.</td>
<td>4. Show contextually relevant information.</td>
</tr>
<tr>
<td>9. Support efficient correction.</td>
<td>10. Scope services when in doubt.</td>
<td>11. Make clear why the system did what it did.</td>
<td>12. Remember recent interactions.</td>
</tr>
<tr>
<td>17. Provide global controls.</td>
<td>18. Notify users about changes.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Contextual Mismatches

3. Time services based on context.
4. Show contextually relevant information.
5. Match relevant social norms.
6. Mitigate social biases.

Remember to call your mom.
Contextual Mismatches – What can you do?

1. Understand and infer critical contexts
2. Monitor appropriate signals, model critical contexts, take appropriate actions

3. Time services based on context.
4. Show contextually relevant information.
5. Match relevant social norms.
6. Mitigate social biases.
Contextual Mismatches – What can you do?

3 Time services based on context.
4 Show contextually relevant information.
5 Match relevant social norms.
6 Mitigate social biases.

Understand and infer critical contexts
Monitor appropriate signals, model critical contexts, take appropriate actions
Contextual Mismatches – What can you do?

Understand and infer critical contexts
Monitor appropriate signals, model critical contexts, take appropriate actions
Develop and test with diversity in mind

3 Time services based on context.
4 Show contextually relevant information.
5 Match relevant social norms.
6 Mitigate social biases.

“Information is not subject to biases, unless users are biased against fastest routes”

“There’s no way to set an avg walking speed. [The product] assumes users to be healthy”
### Guidelines for Human AI Interaction

#### Initially
1. Make clear what the system can do.
2. Make clear how well the system can do what it can do.

#### During Interaction
3. Time services based on context.
4. Show contextually relevant information.
5. Match relevant social norms.
6. Mitigate social biases.

#### When Wrong
7. Support efficient invocation.
8. Support efficient dismissal.
9. Support efficient correction.
10. Scope services when in doubt.
11. Make clear why the system did what it did.

#### Over Time
12. Remember recent interactions.
13. Learn from user behavior.
14. Update and adapt cautiously.
15. Encourage granular feedback.
16. Convey the consequences of user actions.
17. Provide global controls.
18. Notify users about changes.

Learn more: [https://aka.ms/aiguidelines](https://aka.ms/aiguidelines)
Common errors: false positives, false negatives, partially correct, uncertain...
Model Errors – What can you do?

Common errors: false positives, false negatives, partially correct, uncertain...

Consider the costs of errors and provide appropriate mitigation strategies.

7. Support efficient invocation.
8. Support efficient dismissal.
9. Support efficient correction.
10. Scope services when in doubt.
11. Make clear why the system did what it did.
Model Errors – What can you do?

Common errors: false positives, false negatives, partially correct, uncertain...

Consider the costs of errors and provide appropriate mitigation strategies (or explanations)

- Support efficient invocation.
- Support efficient dismissal.
- Support efficient correction.
- Scope services when in doubt.
- Make clear why the system did what it did.
INITIALLY

1. Make clear what the system can do.
2. Make clear how well the system can do what it can do.

DURING INTERACTION

3. Time services based on context.
4. Show contextually relevant information.
5. Match relevant social norms.
6. Mitigate social biases.

WHEN WRONG

7. Support efficient invocation.
8. Support efficient dismissal.
9. Support efficient correction.
10. Scope services when in doubt.
11. Make clear why the system did what it did.

OVER TIME

12. Remember recent interactions.
13. Learn from user behavior.
14. Update and adapt cautiously.
15. Encourage granular feedback.
16. Convey the consequences of user actions.
17. Provide global controls.
18. Notify users about changes.
Cupcakes are muffins that believed in miracles.

Consider changes over time.
Consider changes over time – What can you do?

People and AI models can both change over time

Help people anticipate and guide these changes to suit their needs

12. Remember recent interactions.
13. Learn from user behavior.
14. Update and adapt cautiously.
15. Encourage granular feedback.
16. Convey the consequences of user actions.
17. Provide global controls.
18. Notify users about changes.
Agenda

- Intro to the guidelines
- Findings and impact
- Engineering and AI implications
- Challenges for Intelligible AI
Agenda

- Intro to the guidelines
- Findings and impact
- Engineering and AI implications
- Challenges for Intelligible AI
Findings & Impact

Initial Impact
Opportunity Analysis
Engagements with Practitioners
Opportunity Analysis
Engagements with Practitioners
Guidelines for Human-AI Interaction

18 best practices for human-centered AI design

By Michaela Vierovornu, Akane Amazki, and Peny Collison

Today we’re excited to share our set of Guidelines for Human-AI Interaction. These 18 guidelines can help you design AI systems and features that are more human-centered. Based on more than two decades of thinking and research, they have been validated through a rigorous study published in CHI 2015.

Why do we need guidelines for human-AI interaction?

Guidelines for Human-AI Interaction

By Michaela Vierovornu, Akane Amazki, and Peny Collison

Today we’re excited to share our set of Guidelines for Human-AI Interaction. These 18 guidelines can help you design AI systems and features that are more human-centered. Based on more than two decades of thinking and research, they have been validated through a rigorous study published in CHI 2015.

Why do we need guidelines for human-AI interaction?

Guidelines for Human-AI Interaction

By Michaela Vierovornu, Akane Amazki, and Peny Collison

Today we’re excited to share our set of Guidelines for Human-AI Interaction. These 18 guidelines can help you design AI systems and features that are more human-centered. Based on more than two decades of thinking and research, they have been validated through a rigorous study published in CHI 2015.

Why do we need guidelines for human-AI interaction?
Findings & Impact

Opportunity Analysis

Engagements with Practitioners

Initial Impact
Developing the Guidelines for Human-AI Interaction

Phase 1.
Consolidation
150+ recommendations

Phase 2.
Team Evaluation
13 common AI products

Phase 3.
User Evaluation
49 UX practitioners, 20 AI products

Phase 4.
Expert Review
11 UX practitioners
Developing the Guidelines for Human-AI Interaction

Phase 1. Consolidation
150+ recommendations

Phase 2. Team Evaluation
13 common AI products

Phase 3. User Evaluation
49 UX practitioners,
20 AI products

Phase 4. Expert Review
11 UX practitioners
- Collected of 700+ examples of the guidelines being applied or violated

- 20 different products (both Microsoft and 3rd-party)

- 10 product categories (from fitness trackers to music recommenders)

Phase 3. User Evaluation
49 UX practitioners, 20 AI products

Phase 4. Expert Review
11 UX practitioners
Confirms that the guidelines are applicable to a broad range of products, features and scenarios.
<table>
<thead>
<tr>
<th>Applications</th>
<th>Violations</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>26</td>
</tr>
<tr>
<td>G2</td>
<td>16</td>
</tr>
<tr>
<td>G3</td>
<td>18</td>
</tr>
<tr>
<td>G4</td>
<td>26</td>
</tr>
<tr>
<td>G5</td>
<td>22</td>
</tr>
<tr>
<td>G6</td>
<td>13</td>
</tr>
<tr>
<td>G7</td>
<td>20</td>
</tr>
<tr>
<td>G8</td>
<td>23</td>
</tr>
<tr>
<td>G9</td>
<td>17</td>
</tr>
<tr>
<td>G10</td>
<td>9</td>
</tr>
<tr>
<td>G11</td>
<td>8</td>
</tr>
<tr>
<td>G12</td>
<td>28</td>
</tr>
<tr>
<td>G13</td>
<td>22</td>
</tr>
<tr>
<td>G14</td>
<td>18</td>
</tr>
<tr>
<td>G15</td>
<td>18</td>
</tr>
<tr>
<td>G16</td>
<td>11</td>
</tr>
<tr>
<td>G17</td>
<td>13</td>
</tr>
<tr>
<td>G18</td>
<td>5</td>
</tr>
<tr>
<td>Applications</td>
<td>Violations</td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>G1</td>
<td>26</td>
</tr>
<tr>
<td>G2</td>
<td>16</td>
</tr>
<tr>
<td>G3</td>
<td>18</td>
</tr>
<tr>
<td>G4</td>
<td>26</td>
</tr>
<tr>
<td>G5</td>
<td>22</td>
</tr>
<tr>
<td>G6</td>
<td>13</td>
</tr>
<tr>
<td>G7</td>
<td>20</td>
</tr>
<tr>
<td>G8</td>
<td>23</td>
</tr>
<tr>
<td>G9</td>
<td>17</td>
</tr>
<tr>
<td>G10</td>
<td>9</td>
</tr>
<tr>
<td>G11</td>
<td>8</td>
</tr>
<tr>
<td>G12</td>
<td>28</td>
</tr>
<tr>
<td>G13</td>
<td>22</td>
</tr>
<tr>
<td>G14</td>
<td>18</td>
</tr>
<tr>
<td>G15</td>
<td>18</td>
</tr>
<tr>
<td>G16</td>
<td>11</td>
</tr>
<tr>
<td>G17</td>
<td>13</td>
</tr>
<tr>
<td>G18</td>
<td>5</td>
</tr>
</tbody>
</table>

Remember recent interactions.
Remember recent interactions.

Make clear what the system can do.
1. Make clear what the system can do.

2. Remember recent interactions.

4. Show contextually relevant information.
Provide global controls.
17
Provide global controls.

11
Make clear why the system did what it did.
Provide global controls.

17

Make clear why the system did what it did.

11

Make clear how well the system can do what it can do.

2
Consolidate into a Library (Work in Progress)

Types of content: examples, patterns, research, code

Tagged by guideline and scenario with faceted search and filtering

Comments and ratings to support learning

Grow with examples and case studies submitted by practitioners
Findings & Impact

Initial Impact

Opportunity Analysis

Engagements with Practitioners
Q & A Break
Agenda

Intro to the guidelines
Findings and impact
Engineering and AI implications
Challenges for Intelligible AI
Agenda

- Intro to the guidelines
- Findings and impact
- Engineering and AI implications
- Challenges for Intelligible AI
How can I implement the HAI Guidelines?
Interaction Design for AI requires ML & Eng Support

3 Time services based on context. Hard to implement if the logging infrastructure is oblivious to context.

10 Scope services when in doubt. Does the ML algorithm know or state that it is “in doubt”?

11 Make clear why the system did what it did. Is the ML algorithm explainable?
Setting expectations right – Performance reports

1. Make clear what the system can do.

2. Make clear how well the system can do what it can do.

AI-powered scans can identify people at risk of a fatal heart attack almost a DECADE in advance 'by looking at the entire iceberg and not just the tip'

- The AI predicted heart risk with 90% accuracy, according to data
- Current medical scans are only able to see 'the tip of the iceberg'
- It could benefit around 350,000 in Britain, cardiologists believe
- Government funding will fast track the tech into the NHS in two years
## Setting expectations right – Performance reports

1. Make clear what the system can do.

2. Make clear how well the system can do what it can do.

### Performance Table

<table>
<thead>
<tr>
<th>#</th>
<th>Team Name</th>
<th>Notebook</th>
<th>Team Members</th>
<th>Score</th>
<th>Entries</th>
<th>Last</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PFDet</td>
<td></td>
<td></td>
<td>0.62882</td>
<td>49</td>
<td>1y</td>
</tr>
<tr>
<td>2</td>
<td>Avengers</td>
<td></td>
<td></td>
<td>0.62161</td>
<td>48</td>
<td>1y</td>
</tr>
<tr>
<td>3</td>
<td>kivajok</td>
<td></td>
<td></td>
<td>0.61707</td>
<td>102</td>
<td>1y</td>
</tr>
<tr>
<td>4</td>
<td>XJTU</td>
<td></td>
<td></td>
<td>0.61559</td>
<td>22</td>
<td>1y</td>
</tr>
<tr>
<td>5</td>
<td>ikciting</td>
<td></td>
<td></td>
<td>0.59472</td>
<td>39</td>
<td>1y</td>
</tr>
<tr>
<td>6</td>
<td>Sogou_MM</td>
<td></td>
<td></td>
<td>0.57936</td>
<td>105</td>
<td>1y</td>
</tr>
<tr>
<td>7</td>
<td>QLearning</td>
<td></td>
<td></td>
<td>0.56688</td>
<td>20</td>
<td>1y</td>
</tr>
<tr>
<td>8</td>
<td>[RingUkraine] CloudResearch</td>
<td></td>
<td></td>
<td>0.53742</td>
<td>50</td>
<td>1y</td>
</tr>
<tr>
<td>9</td>
<td>Res101+SoftNMS</td>
<td></td>
<td></td>
<td>0.53413</td>
<td>29</td>
<td>1y</td>
</tr>
<tr>
<td>10</td>
<td>Kyle L.</td>
<td></td>
<td></td>
<td>0.51464</td>
<td>53</td>
<td>1y</td>
</tr>
</tbody>
</table>
## Setting expectations right – Gender Shades study

1. Make clear what the system can do.

2. Make clear how well the system can do what it can do.

[Buolamwini, J. & Gebru, T. 2018]

<table>
<thead>
<tr>
<th></th>
<th>TYPE I</th>
<th>TYPE II</th>
<th>TYPE III</th>
<th>TYPE IV</th>
<th>TYPE V</th>
<th>TYPE VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM</td>
<td>5.1%</td>
<td>7.4%</td>
<td>8.2%</td>
<td>8.3%</td>
<td>33.3%</td>
<td>46.8%</td>
</tr>
<tr>
<td>FACE++</td>
<td>11.9%</td>
<td>9.7%</td>
<td>8.2%</td>
<td>13.9%</td>
<td>32.4%</td>
<td>46.5%</td>
</tr>
</tbody>
</table>

90% accuracy
Setting expectations right – Error Terrain Analysis

1. Make clear what the system can do.
2. Make clear how well the system can do what it can do.

Benchmark data

Internal component features

Content features
- $\text{count}_{\text{OBJECTS}}$
- cat
- people

Descriptive features

Failure explanation models with Pandora
[Nushi et. al. HCOMP 2018]
No makeup (23.7%)

Women (11.5%)

All (5.5%)
Short hair (30.8%)

No makeup (23.7%)

Women (11.5%)
Setting expectations right – Error Analysis

1. Make clear what the system can do.

2. Make clear how well the system can do what it can do.

Error Terrain Analysis \ Pandora
[Nushi et. al. HCOMP 2018]

Errudite
[Wu et. al. ACL 2019]

Manifold
[Zhang et. al. IEEE TVCG 2018]
Setting expectations right: other implications

1. Make clear what the system can do.
2. Make clear how well the system can do what it can do.

- Use multiple and realistic benchmarks
- Estimate the cost and risk of mistakes
- Calibrate and explain uncertainty
Setting expectations right – Uncertainty Calibration

Post-hoc calibration:

Platt scaling, Isotonic regression
[Platt et al., 1999; Zadrozny & Elkan, 2001]

In-built model uncertainty

Bayesian DNNs, Ensemble methods
[Gal & Ghahramani, 2016; Osband et al., 2016]

Setting expectations right – Uncertainty explanation

Explain "Probability of Precipitation"

Forecasts issued by the National Weather Service routinely include a “PoP” (probability of precipitation) statement, which is often expressed as the “chance of rain” or “chance of precipitation”.

EXAMPLE

ZONE FORECASTS FOR NORTH AND CENTRAL GEORGIA NATIONAL WEATHER SERVICE PEACHTREE CITY GA 119 PM EDT THU MAY 8 2008


THIS AFTERNOON, MOSTLY CLOUDY WITH A 40 PERCENT CHANCE OF SHOWERS AND THUNDERSTORMS. WINDY. HIGHS IN THE LOWER 60S. NEAR STEADY TEMPERATURES IN THE LOWER 60S. SOUTH WINDS 15 TO 25 MPH. TONIGHT, MOSTLY CLOUDY WITH A CHANCE OF SHOWERS AND THUNDERSTORMS. LOWS IN THE MID 60S. SOUTHWEST WINDS 5 TO 15 MPH. CHANCE OF RAIN 40 PERCENT.

What does this “40 percent” mean? ...will it rain 40 percent of the time? ...will it rain over 40 percent of the area?

The “Probability of Precipitation” (PoP) simply describes the probability that the forecast grid/point in question will receive at least 0.01” of rain. So, in this example, there is a 40 percent probability for at least 0.01” of rain at the specific forecast point of interest!

https://forecast.weather.gov/
Setting expectations right – Uncertainty explanation

Probably a yellow school bus **driving** down a street

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness

E' stata la migliore delle volte, è stata la peggiore delle volte, era l'età della saggezza, era l'età della follia
Context, Invocation, Dismissal

Time services based on context.

Support efficient invocation.

Support efficient dismissal.

Automated triggering

Explicit invocation

AI System
Context inference

Model compression
[Ba and Caruana 2014; Hinton 2015]

Adaptive networks for inference
[Bolukbasi et. al 2017]
Context, Invocation, Dismissal

3. Time services based on context.

7. Support efficient invocation.

8. Support efficient dismissal.

Automated triggering

Explicit invocation

AI System
Tuning automated triggering

Cost of explicit invocation
user time, accessibility

Cost of wrong invocation
cognitive load, dismissal time

Cost of wrong AI prediction
risk mitigation

Precision

Recall
Incorporating user feedback over time

13 Learn from user behavior.

15 Encourage granular feedback.

14 Update and adapt cautiously.

Content dependent

Context dependent

User dependent

Too slow

Content dependent

Too fast

Static system
Lack of trust/engagement

Forgetting content
Lack of trust/engagement
Feature engineering

13 Learn from user behavior.

15 Encourage granular feedback.

14 Update and adapt cautiously.

- Items
  - Content features
  - User features
  - Context features
Dealing with sparse data

13 Learn from user behavior.

15 Encourage granular feedback.

14 Update and adapt cautiously.

Items

- Content features
- User features
- Context features
Global control support: feedback generalization

- Encourage granular feedback.

- Provide global controls.

Sci-fi

Drama
Global control support: feedback generalization

15 Encourage granular feedback.

17 Provide global controls.

Disney

Hollywood
Global control support: feedback generalization

15
Encourage granular feedback.

17
Provide global controls.

Multiple clusterings

Content features

Sci-fi Docs
Spielberg
After 2000
Q & A

Is there any other functionality you know of or you wish you had in ML & Eng that could simplify Human-AI Interaction?

How much do interaction considerations impact ML & Engineering decisions?

What else do you (or your colleagues) do to support better Human-AI interaction?
Agenda

Intro to the guidelines
Findings and impact
Engineering and AI implications
Challenges for Intelligible AI
Agenda

Intro to the guidelines
Findings and impact
Engineering and AI implications

Challenges for Intelligible AI
Machine Learning Everywhere

11
Make clear why the system did what it did.

12
Remember recent interactions.

13
Learn from user behavior.

Intelligible, Transparent, Explainable AI
Terminology

Caveat: My take – No consensus here

- Predictable ~ (Human) Simulate-able
  - Predict exactly what it will do
- Intelligible ~ Transparent
  - Answer counterfactual
    predict how a \textit{change} to model’s inputs will \textit{change} its output
- Explainable ~ Interpretable
  - Construct rationalization for why (maybe) it did what it did
- Inscrutable \supseteq Blackbox
  - Inscrutable: too complex to understand
  - Blackbox: know \textit{nothing} about it

\textit{Caveat: My take – No consensus here}
Reasons for Wanting Intelligibility

1. The AI May be Optimizing the Wrong Thing
2. Missing a Crucial Feature
3. Distributional Drift
4. Facilitating User Control in Mixed Human/AI Teams
5. User Acceptance
6. Learning for Human Insight
7. Legal Requirements

[Weld & Bansal CACM 2019]
AI Deployments

Autonomous AI

Input ➔ Validation ➔ Decision

Repeat Testing

+ Explanation

Deploy

Human-AI Team

Input ➔ Decision

Input ➔ Decision

+ Explanation

Intelligibility Useful in Both Cases

In this paper, we argue that a classifier that is AI-advised and interacts, or computed in expectation from the system itself, may not work: [Zhang et al., 2020; Lai and Tan, 2018; Feng and Boyd-Graber, 2019].
Reasons for Wanting Intelligibility

1. The AI May be Optimizing the Wrong Thing
2. Missing a Crucial Feature
3. Distributional Drift
4. Facilitating User Control in Mixed Human/AI Teams
5. User Acceptance
6. Learning for Human Insight
7. Legal Requirements

[Weld & Bansal CACM 2019]
Reasons for Wanting Intelligibility

1. The AI May be Optimizing the Wrong Thing
2. Missing a Crucial Feature
3. Distributional Drift
4. **Facilitating User Control in Mixed Human/AI Teams**
5. User Acceptance
6. Learning for Human Insight
7. Legal Requirements

[Weld & Bansal CACM 2019]
The Growing Era of Human-AI Teams
Artificial Intelligence Often Isn’t

But Humans Err as Well
The Space of Errors

Human Errors

Preventable Errors

Joint Errors

AI Specific Errors

AI Errors
The Dream Team

Intelligible AI → Better Teamwork
A Simple Human-AI Team

ML Model Readmission Prediction Classifier

When can I trust it? How can I adjust it?

Human Decision Maker

Input Patient

Age

Blood Pressure

Recommendation Yes / No

Decision

Should the patient be placed in a special outpatient program?

[Bansal et al. HCOMP-19]
Inherently Intelligible ML – Example 1

Small decision tree over semantically meaningful primitives

When can I trust it?
How can I adjust it?

Intelligibility threatened if tree grows big
Inherently Intelligible ML – Example 2

Linear model over meaningful primitives

Other models often perform much better
Inherently Intelligible ML – Example 3
GA$^2$M model over semantically meaningful primitives

\[ y = \beta_0 + \sum_j f_j(x_j) \]

1 (of 56) components of learned GA$^2$M: risk of pneumonia death

Part of Fig 1 from R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. “Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission.” In KDD 2015.
Inherently Intelligible ML – Example 3

GA$^2$M model over semantically meaningful primitives

$$y = \beta_0 + \sum_j f_j(x_j) + \sum_{i \neq j} f_{ij}(x_i, x_j)$$

pairwise terms

Part of Fig 1 from R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. “Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission.” In KDD 2015.
When can I trust it? How can I adjust it?

3 (of 56) components of learned GA²M: risk of pneumonia death

Part of Fig 1 from R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." In KDD 2015.
Sometimes you just *need* an inscrutable model

E.g., Medical image analysis

- Deep cascade of CNNs
- Variational networks
- Transfer learning
- GANs

**Kidney MRI**


**Input: Pixels**

Features are not semantically meaningful
Roadmap for Intelligibility

Intelligible?

Yes → Use Directly

No → Map to Simpler Model
  • Explanations
  • Controls

Interact with Simpler Model
Reasons for Inscrutability

- Too Complex
- Features not Semantically Meaningful
Explaining Inscrutable Models

- Too Complex
  - Simplify by currying → instance-specific explanations
  - Simplify by approximating

- Features not Semantically Meaningful
  - Map to new vocabulary

- Usually have to do all of these!
LIME - Local Approximations

To explain prediction for point \( p \)...

1. Sample points around \( p \)
2. Use complex model to predict labels for each sample
3. Weigh samples according to distance from \( p \)
4. Learn new simple model on weighted samples (possibly using different features)
5. Use simple model as explanation

To create **features** for explanatory classifier, compute `superpixels` using off-the-shelf image segmenter. Hope that feature/values are semantically meaningful.

To **sample** points around p, set some superpixels to grey. 

**Explanation** is set of superpixels with high coefficients...

“It’s just looking for...”
Central Dilemma

Understandable

Over-Simplification

Accurate

Inscrutable

Any model simplification is a Lie
What Makes a Good Explanation?

Need Desiderata
Psychology Experiments → Ranking

If you can’t include all details, humans prefer

- Details distinguishing fact & foil
- Necessary causes >> sufficient ones
- Intentional actions >> actions taken w/o deliberation
- Proximal causes >> distant ones
- Abnormal causes >> common ones
- Fewer conjuncts (regardless of probability)
- Explanations consistent with listener’s prior beliefs

Presenting an explanation made people believe P was true
If explanation ~ previous, effect was strengthened
Trust

- Everybody talks about *increasing trust*...

- The psychology literature shows explanations increase trust
  [Miller AIJ-18]

  ... Even when the explainer is *wrong*...

- We *shouldn’t* seek or measure trust...

- We should seek to show the human *when not to trust*
Do Explanations Help *Team* Performance?
Yes!

- Medical Diagnosis
  [Lundberg et al. *Nature biomedical engineering*. 2018]
- Annotation
  [Schmidt & Biessmann. *AAAI Workshop*. 2019]
- Deception Detection
  [Lai & Tan FAT*. 2019]

Except...

In these papers, \text{Accuracy(Humans)} \ll \text{Accuracy(AI)}

So... the rational decision is to \textbf{omit} the humans (not explain)
Are Explanations Helpful??

We studied a simple human-Al team where

\[
\text{Accuracy(Human)} = \text{Accuracy(Al)} = 0.8
\]

0) Solo Human (No Al)
1) AI Recommends
2) AI also gives its confidence
3) AI also explains (LIME-like)
4) AI gives \textit{human} explanation

[Zhou, Bansal et al. In Prep]
*Not Necessarily...*
Explanations are Convincing

![Diagram showing accuracy comparison between different teams and human performance. The x-axis represents accuracy ranging from 0.4 to 1.0. The green dots represent AI Correct, while the red dots represent AI Incorrect.](image_url)

[Human, Team (Recommendation, R), Team (R+Confidence), Team (R+AI Expl.), Team (R+Convincing Expl.)]

[Zhou, Bansal et al. In Prep]
Explanations are Convincing

Not Necessarily...

Human
Team (Recommendation, R)
Team (R+Confidence)

Accuracy

0.4 0.6 0.8 1.0

AI Correct
AI Incorrect

[Zhou, Bansal et al. In Prep]
Not Necessarily...

Explanations are Convincing

[Zhou, Bansal et al. In Prep]
Not Necessarily...
Explanations are Convincing

Better Explanations are More Convincing
Coming Soon...

• Adaptive Explanations...

[Zhou, Bansal et al. In Prep]
That Other Question...

How can I adjust it?
Tuning

Interpretable Model

Opaque Model

Explanatory Model

Map to Explanatory Model
(e.g., LIME, SHAP)

User

Limeade

User

Explain
Tune

Tune
Explain

[Lee et al. Submitted]
Adaptive Research-Paper Recommendations

Published 2019 in ArXiv

CHIP: Channel-wise Disentangled Interpretation of Deep Convolutional Neural Networks
Xinrui Cui, Dan Wang, Zhen Jane Wang

With the widespread applications of deep convolutional neural networks (DCNNs), it becomes increasingly important for DCNNs not only to make accurate predictions but also to explain how they make...

CONTINUE READING

Published 2019 in ICLR

Beta: s2-sanity.apps.allenai.org

- Deep neural paper embeddings
- Explain with linear bigrams

[Cohen et al. Submitted]
If all one cared about was the explanatory model, one could change this parameters... but not even the features are shared with the neural model!

[Lee et al. Submitted]
Instead... We generate new training instances by varying the feedback feature, weight by distance to $x'$... [Lee et al. Submitted]
**Instead...** We generate new training instances by varying the feedback feature, weight by distance to *Retrain.*

[Lee et al. Submitted]
Evaluation

Good News:

<table>
<thead>
<tr>
<th>Which system...</th>
<th>Baseline</th>
<th>Ours</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>...trust more?</td>
<td>4</td>
<td>17</td>
<td>0.043</td>
</tr>
<tr>
<td>...more control?</td>
<td>0</td>
<td>21</td>
<td>≈0</td>
</tr>
<tr>
<td>...more transparent?</td>
<td>3</td>
<td>18</td>
<td>0.012</td>
</tr>
<tr>
<td>...more intuitive?</td>
<td>12</td>
<td>9</td>
<td>0.664</td>
</tr>
<tr>
<td>...not missing relevant papers?</td>
<td>3</td>
<td>18</td>
<td>0.012</td>
</tr>
</tbody>
</table>

Less Good News:

No significant improvement on feed quality (team performance) as measured by clickthru

[Lee et al. Submitted]
Summary

ML Model
Readmission Prediction
Classifier

When can I trust it?
How can I adjust it?

Human
Decision
Maker

Should the patient be placed in a special outpatient program?

[Bansal et al. HCOMP-19]
Summary

Human-AI Team

Input

+ Explanation

Decision

Current utility helps to make final recommendations, which are computed via human-centered algorithms. If human-made decisions are accurate, then the utility is computed as log-utility. Let’s take a step back and inspect the role of AI advising this human made decision. We would like to show that collaboration may not work: [Zhang et al., 2020; Lai and Tan, 2018; Feng and Boyd-Graber, 2019].
Summary

Intelligible?

Yes → Use Directly

No → Map to Simpler Model
  • Explanations
  • Controls

Interact with Simpler Model
Guidelines for Human AI Interaction

Learn more: https://aka.ms/aiguidelines

Thanks! Questions?
Resources


Learn the guidelines
Introduction to guidelines for human-AI interaction
Interactive cards with examples of the guidelines in practice

Use the guidelines in your work
Printable cards (PDF)
Printable poster (PDF)

Find out more
Guidelines for human-AI interaction design, Microsoft Research Blog
AI guidelines in the creative process: How we’re putting the human-AI guidelines into practice at Microsoft, Microsoft Design on Medium
How to build effective human-AI interaction: Considerations for machine learning and software engineering, Microsoft Research Blog