

Explainable AI is Dead, Long Live Explainable AI!

Hypothesis-driven Decision Support using
Evaluative AI

Charles Vin

Sorbonne Université - 21216136

Quick summary

Arg for a Paradigm Shift in (X)AI for Decision Support

→ Evaluative AI Concept

Goals:

- Human-centered Approach
- Going Beyond Recommendations
- **Mitigating Over-Reliance**
- Support for Hypothesis Evaluation
- Machine-in-the-Loop Paradigm

Over/Under-reliance

Definitions

- **Over-reliance:** Decision makers accept a machine recommendations, even when it is wrong, but would be rejected if coming from a human.
 - *The machine "must be right" because it's a machine*
- **Under-reliance:** Machine outputs are consistently rejected, even when it is correct, but would be accepted if coming from a human.

See "Automation bias" \Rightarrow Problems after deployment:
AI systems ignored OR over-reliance related problems.

Over/Under-reliance

Causes

- Over-reliance: ;
- Under-reliance: Algorithmic aversion.

When adding current XAI tools for more explanation
⇒ Confirmation bias (called *fixation* in the paper).

Over/Under-reliance

Solutions

1. Cognitive forcing
 - Eg. forcing people to give a decision before seeing a recommendation ;
 - Slightly mitigated overreliance, but not enough to lead to a statistically significant differences ;
 - Least preferred method by participant : people not wanting to exert mental energy.
2. Changing the XAI framework 🤔🧐💡💡

What makes a good decisions?

In a simple way:

- Identify options
- Compare options
- Choose an option

In a less simple way: the 10 "cardinal decision issue" outlined by Yates and Potworowski

- Needs, mode, Investment, Options, Possibilities, Judgements, Value, Trade-offs, Acceptability, Implementation

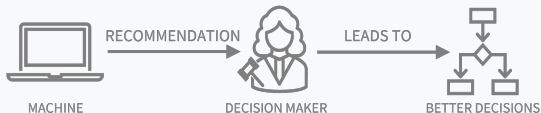
What makes a good decisions support system?

Summed up

- Options: Help to identify options, well as help to narrow down the list of feasible or realistic options
- Possibilities: Help to to identify possible outcomes
- Judgement: Help to judge which outcomes are most likely and what will be the positive and negative impacts
- Trade-offs: Help to make trade-offs on the above criteria for each options
- Understandable: Help to understand how and why the tools works as it does, and when it fails

Does current decision support align with those
criteria?

Giving recommendations with no explanatory information



A model of giving recommendations for decision support.

This assumes that decision makers will carefully consider recommendations. However, empirical evidence suggests this is not the case.

- × Options
- $1/n$ Possibilities & Judgement
- × Trade-offs
- × Understandable

Giving recommendations with explanatory information



A model of explainable AI for decision support. This assumes that the problem of distrust can be mitigated by giving reasons or explanations for decisions. However, empirical evidence suggests people do not pay careful attention to the reasons/explanations.

- × Options
- × Possibilities & Judgement
- × Trade-offs
- × Understandable

Giving recommendations with cognitive forcing



Figure: A model of cognitive forcing.

Withholding recommendations & giving an "explanation"

→ forces people to engage

→ limit over-reliance.

× still a "recommend and defend" approach.

× Options

× Possibilities & Judgement

× Trade-offs

× Understandable

The evaluative AI framework

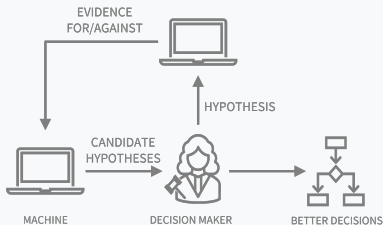
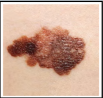


Figure: A model of Evaluative AI.

- Align with decision making processes
- Keep the decision maker in control
- Ask users to rely on evidence instead of recommendations.

The evaluative AI framework

Lesion


Notes
Patient reports itchiness and bleeding.
Lesion has changed colour.

Lesion location

☐ Head

☐ Upper arm

☐ Face

☐ Hand/Lower Arm

☒ Back

☐ Upper Leg

☐ Front Torso

☐ Foot/Lower Leg

Your hypothesis

Melanoma

Melanocytic Nevus

Basal Cell Carcinoma

Actinic Keratosis

Benign Keratosis

Dermatofibroma

Vascular Lesion

Evidence for

Lesion location

Colour

Scarred

Bleeding

Evidence against

Asymmetric shape

Changed colour

Itchiness

- ✓ Options
- ✓ Possibilities & Judgement
- ✓ Trade-offs
- ✓ Understandable

Figure: A simple prototype of a diagnostic interface using evaluative AI.

The evaluative AI framework

Zoom on properties

- ✓ Options
 - Show the most likely options (with or without probabilities) ;
 - Not a single recommendation.
- ✓ Possibilities & Judgement
 - The machine provide feedback on humain judgement only.
- ✓ Trade-offs
 - Offer real trade-offs between *any* set of two options ;
 - Evaluative AI provides evidence both for and against each option, irrelevant of the judged likelihood of that option ;
 - *Option awareness* in the literature.

The evaluative AI framework

Differences with cognitive forcing

- **Control** Permit to explore hypothesis, not a single recommendation ;
- Built on the way we makes decision (identify, compare, choose)

Long live XAI

- Evaluative AI is designed only toward decision making, Evaluative AI \subset XAI ;
- XAI is still needed and more adapted to many situation (eg. making decision at scale) ;
- Recommendation based models : base of any XAI techniques ;
- Many existing XAI tools are already adapted to Evaluative AI
 - Constrastive explanation ;
 - Feature importance (eg. SHAP).

Limitation

- Why would people pay attention to evidence this time?
 - Evaluative AI
 - Better control ;
 - Process built on the way we makes naturally decision
 - people would naturally follow
 - ≠ Contrary to recommendation-driven approches ;
 - × Proof?
- Cognitive load remain a problem
 - Evaluative AI still reduce the quantity of information the decision maker needs (only revelant information are presented)
 - × Still the less prefered solution by decision makers

Bibliography

Template from here 

A plain frame has no headline

<code>\Alegreya</code>	Lorem ipsum dolor sit amet
<code>\AlegreyaMedium</code>	Lorem ipsum dolor sit amet
<code>\AlegreyaExtraBold</code>	Lorem ipsum dolor sit amet
<code>\AlegreyaBlack</code>	Lorem ipsum dolor sit amet
<code>\AlegreyaSansThin</code>	Lorem ipsum dolor sit amet
<code>\AlegreyaSansLight</code>	Lorem ipsum dolor sit amet
<code>\AlegreyaSans</code>	Lorem ipsum dolor sit amet
<code>\AlegreyaSansMedium</code>	Lorem ipsum dolor sit amet
<code>\AlegreyaSansExtraBold</code>	Lorem ipsum dolor sit amet
<code>\AlegreyaSansBlack</code>	Lorem ipsum dolor sit amet

Alert! A *plain* frame does not show the progress bar but still appears in it unless the frame comes after `\End`

A ***STANDOUT*** frame can be used to focus attention

In combination with *plain*,
it makes a nice thank-you slide!



<https://github.com/piazzai/arguelles>
<https://ctan.org/pkg/beamertheme-arguelles>