

University of Bamberg



Working Paper

# The Effect of Explainable AI-based Decision Support on Human Task Performance: A Meta-Analysis

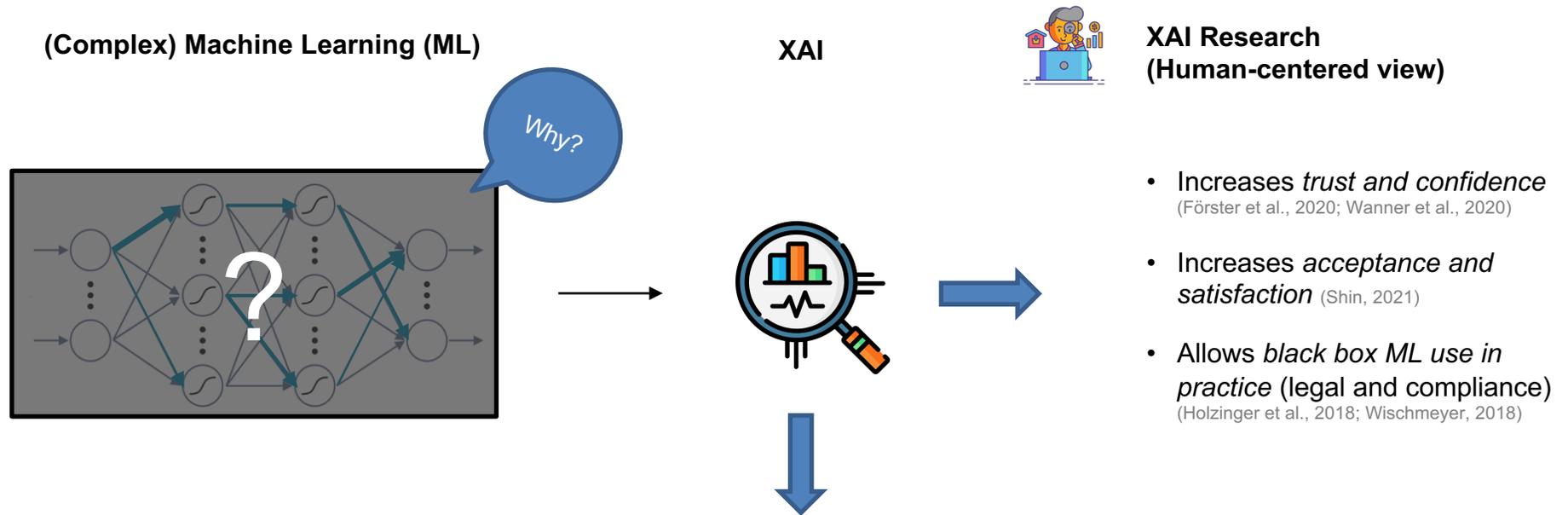
Felix Haag

Chair of Information Systems and Energy Efficient Systems

Research Discussion of the WhiteBox-AI Group – Bamberg | 2024-12-13

---

# Explainable AI (XAI) for Decision Support!



## Another important stream: XAI to improve human decision-making

- Price estimations (Bauer et al., 2023)
- Diabetes detection (van der Waa et al., 2021)
- Industrial process design (Senoner et al., 2021)

# Some studies report improved human task performance, while others report negligible effects

- XAI research is experiencing an increasing number of studies evaluating the **effect of XAI-assisted decision-making on human task performance** (Bansal et al., 2021; Buçinca et al., 2022; Hemmer et al., 2022; van der Waa et al., 2021; Zhang et al., 2020)



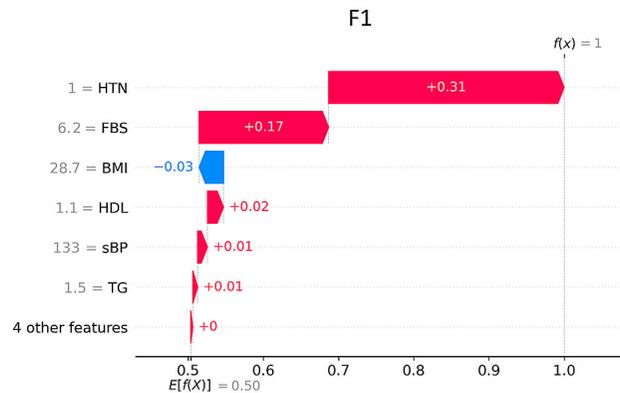
# Popular forms of XAI explanation types

## Method 1: Feature Attributions

Describe causal attributions (*why?*)

*“X had occurred, because of Y”*

### ML prediction: Class 1

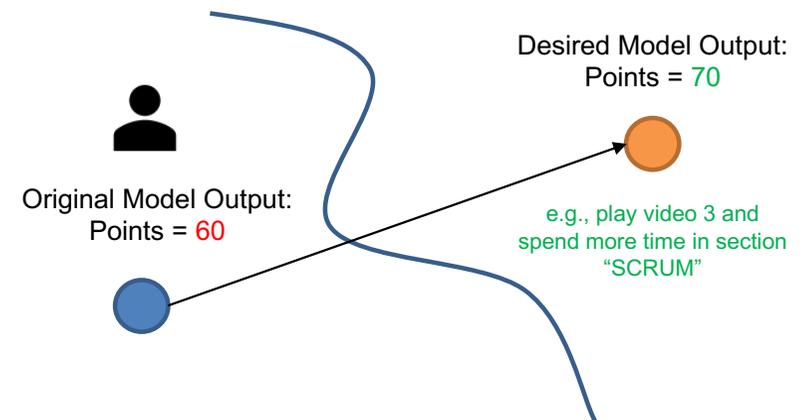


## Method 2: Counterfactuals

Describe causal situations (*what if?*)

*“If X had occurred, Y would have occurred”*

### ML prediction: 60 Pts.



(Mohseni et al., 2021; Herm, 2023)

# Popular forms of XAI explanation types

**Method 3: Example-based**

Describe similar instances (*what-else?*)

*“X had occurred, because Y is very similar to X”*

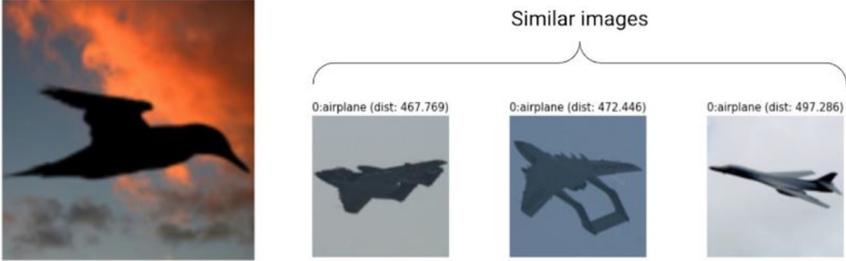


FIGURE 2. USE EXAMPLE-BASED EXPLANATIONS FOR MISCLASSIFICATION ANALYSIS

(Mohseni et al., 2021; Herm, 2023)

# Some studies report improved human task performance, while others report negligible effects

- XAI research is experiencing an increasing number of studies evaluating the **effect of XAI-assisted decision-making on human task performance** (Bansal et al. 2021; Buçinca et al. 2022; Hemmer et al. 2022; van der Waa et al. 2021; Zhang et al. 2020)



- In this work, I'd like to present an **initial synthesis of existing research on user studies examining the effect of XAI-assisted decision-making on human task performance**
  1. The **overall/main effect** of XAI on human **task performance** (already done with 9 articles; see Schemmer et al. 2022)
  2. The **conditions** under which **XAI is helpful** (or not): Subgroup analyses

# Research Question

## Overarching RQ

**RQ 1a:** To what extent does XAI-based decision support affect human performance in classification tasks, considering the current body of empirical studies?



## Sub analysis 1

**RQ 1b:** To what extent does the risk of bias in studies affect the outcome of XAI-based decision support on human performance in classification tasks?

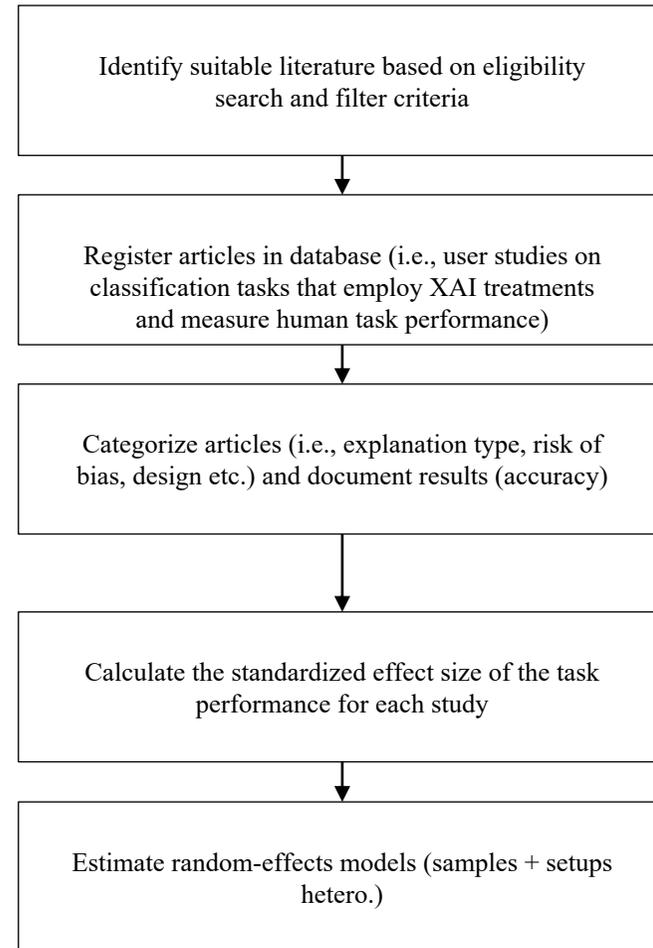
## Sub analysis 2

**RQ 2:** To what extent do (i) feature attribution, (ii) counterfactual, or (iii) example-based explanations affect human performance in classification tasks, considering the current body of empirical studies?

# Methodology Overview: Meta-Analysis

**Part 1** Data collection & categorization

**Part 2** Statistical analysis



# Methodology: Literature Search String

OR		OR		OR
"explainable artificial intelligence"	<b>AND</b>	"instance based"	<b>AND</b>	"task performance"
"xai"		"example based"		"decision performance"
"explainable AI"		"counterfactual"		"human accuracy"
"interpretable machine learning"		"hypothetical"		"human performance"
"interpretable ml"		"causal"		"user study"
"explainable machine learning"		"anchor"		"empirical study"
"explainable ml"		"contrastive"		"field experiment"
("machine learning" OR "artificial intelligence" OR "AI" AND (interpret* OR explain* OR "explanation"))		"feature attribution"		"online experiment"
		"feature importance"		"human experiment"
		"LIME"		"human evaluation"
	"SHAP"	"user evaluation"		
		((behavior* OR behaviour*) AND "experiment"))		

# Methodology: Risk of Bias Assessment of Studies

- The “Risk of Bias 2” tool (RoB 2) is divided into **five risks of bias domains**, where each domain contains a **series of signaling questions** relevant to study bias (Sterne et al., 2019)
- The judgement for the risk posed by a domain is calculated from an algorithm. Judgements on the **risk of bias can be “Low”, “Some concerns”, and “High”**
- Each rating was **double-checked** by a non-affiliated colleague; deviations in the assessment were discussed and resolved

## RoB 2 Domains (D1-D5)

- D1:** Bias arising from the randomization process
- D2:** Bias due to deviations from intended interventions
- D3:** Bias due to missing outcome data
- D4:** Bias in measurement of the outcome
- D5:** Bias in selection of the reported result

Study	Risk of bias domains					Overall
	D1	D2	D3	D4	D5	
Steinberg J 2020	+	+	+	+	+	+
Ktuchi M 2018	+	+	+	+	-	-
Romanov A 2016	+	+	+	-	-	-
Ktuchi M 2016	-	X	+	+	-	X
Pokushalov E 2014	+	+	+	+	-	-
Pokushalov E 2012	+	X	-	+	-	-

Domains:  
D1: Bias arising from the randomization process  
D2: Bias due to deviations from intended intervention.  
D3: Bias due to missing outcome data.  
D4: Bias in measurement of the outcome.  
D5: Bias in selection of the reported result.

Judgement  
 High  
 Some concerns  
 Low

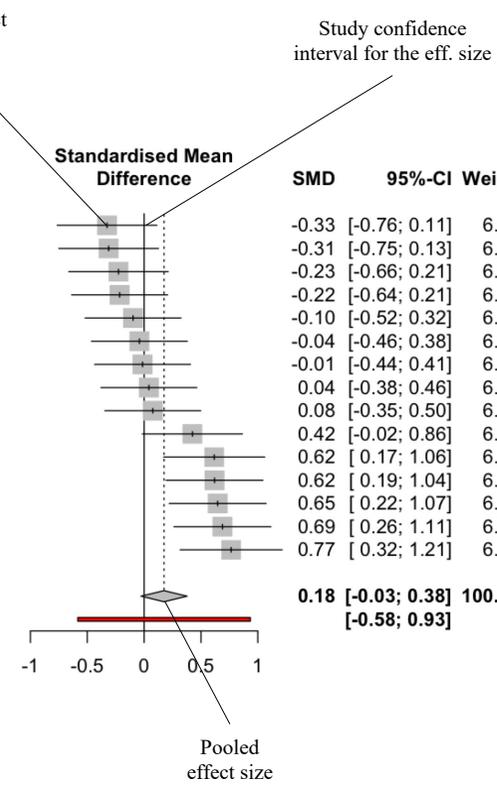
(Templier & Paré, 2018)

# Methodology: Statistical Analysis - Random Effect Models

- The experiments conducted do probably not share a common true effect size (i.e., a fixed effect) because the experimental setups, tasks, and respective samples considerably vary between studies
- Instead of fixed-effects regression models, we employ random-effects regression models to estimate the mean of effects (called “pooled effect”)
- Controlled for multiple control group inclusion (sample N control group / times included in the model)

## Simple example

Study	Experimental			Control			Standardised Mean Difference	SMD	95%-CI	Weight
	Total	Mean	SD	Total	Mean	SD				
Carton et al.-2020-Features attribution explanation-Keyword Explanation	80	0.52	0.0870	27	0.54	0.0480		-0.33	[-0.76; 0.11]	6.6%
Carton et al.-2020-Features attribution explanation-Full Explanation	80	0.52	0.0680	27	0.54	0.0480		-0.31	[-0.75; 0.13]	6.6%
Carton et al.-2020-Features attribution explanation-Partial Explanation	80	0.53	0.0870	27	0.54	0.0480		-0.23	[-0.66; 0.21]	6.6%
Liu et al.-2021-Features attribution explanation-COMPAS – Interactive/Static	216	56.90	14.9970	24	60.00	4.7470		-0.22	[-0.64; 0.21]	6.7%
Liu et al.-2021-Features attribution explanation-COMPAS – Static/Static	216	58.60	14.9970	24	60.00	4.7470		-0.10	[-0.52; 0.32]	6.7%
Liu et al.-2021-Features attribution explanation-COMPAS – Interactive/Interactive	216	59.40	14.9970	24	60.00	4.7470		-0.04	[-0.46; 0.38]	6.7%
Liu et al.-2021-Features attribution explanation-ICPSR – Interactive/Interactive	216	60.70	14.9970	24	60.90	4.7470		-0.01	[-0.44; 0.41]	6.7%
Liu et al.-2021-Features attribution explanation-ICPSR – Static/Static	216	61.50	14.9970	24	60.90	4.7470		0.04	[-0.38; 0.46]	6.7%
Liu et al.-2021-Features attribution explanation-ICPSR – Interactive/Static	216	62.00	14.9970	24	60.90	4.7470		0.08	[-0.35; 0.50]	6.7%
Lai et al.-2019-Features attribution explanation-Example-based explanation	80	54.40	7.6360	27	51.10	7.9570		0.42	[-0.02; 0.86]	6.6%
Lai et al.-2019-Features attribution explanation-Highlight	80	55.90	7.6520	27	51.10	7.9570		0.62	[0.17; 1.06]	6.5%
Liu et al.-2021-Features attribution explanation-BIOS – Interactive/Interactive	216	72.40	14.9970	24	63.50	4.7470		0.62	[0.19; 1.04]	6.7%
Liu et al.-2021-Features attribution explanation-BIOS – Interactive/Static	216	72.80	14.9970	24	63.50	4.7470		0.65	[0.22; 1.07]	6.7%
Liu et al.-2021-Features attribution explanation-BIOS – Static/Static	216	73.40	14.9970	24	63.50	4.7470		0.69	[0.26; 1.11]	6.7%
Lai et al.-2019-Features attribution explanation-Heatmap	80	57.60	8.5840	27	51.10	7.9570		0.77	[0.32; 1.21]	6.5%
<b>Random effects model</b>	<b>2424</b>			<b>378</b>				<b>0.18</b>	<b>[-0.03; 0.38]</b>	<b>100.0%</b>
<b>Prediction interval</b>									<b>[-0.58; 0.93]</b>	



Random effects model  
 Prediction interval  
 Heterogeneity:  $I^2 = 70\%$ ,  $p < 0.01$

Moderate/high heterogeneity

Reject H0: The true effect size is identical in all studies.

# Results RQ1a: Main Effect – No support vs. XAI

**On request: Please send an e-mail to [felix.haag@uni-bamberg.de](mailto:felix.haag@uni-bamberg.de)**

# Results RQ1a: Main Effect – AI vs. XAI

**On request: Please send an e-mail to [felix.haag@uni-bamberg.de](mailto:felix.haag@uni-bamberg.de)**

# Results RQ1b: Subgroup Analyses – Risk of Bias

**On request: Please send an e-mail to [felix.haag@uni-bamberg.de](mailto:felix.haag@uni-bamberg.de)**

# Results RQ1b: Subgroup Analyses – Risk of Bias

**On request: Please send an e-mail to [felix.haag@uni-bamberg.de](mailto:felix.haag@uni-bamberg.de)**

# Results RQ2: Explanation Type Categorization

**On request: Please send an e-mail to [felix.haag@uni-bamberg.de](mailto:felix.haag@uni-bamberg.de)**

# Results RQ2: Differences in Explanation Types

**On request: Please send an e-mail to [felix.haag@uni-bamberg.de](mailto:felix.haag@uni-bamberg.de)**

# Main Contributions & Next Steps

## Main Contributions

- ➔ **Practical:** Assessment of the value of XAI for decision-making and human task performance (implications for the design of XAI-based DSS)
- ➔ **Theoretical:** Exploration of the conditions under which XAI is effective for decision support (in this paper: which type of explanation is effective)

## Next Steps

- ➔ Further sub-analyses (e.g., task complexity, task domain, (X)AI literacy etc.)
- ➔ Publish the dataset; refine and submit the paper



Thank you very much.