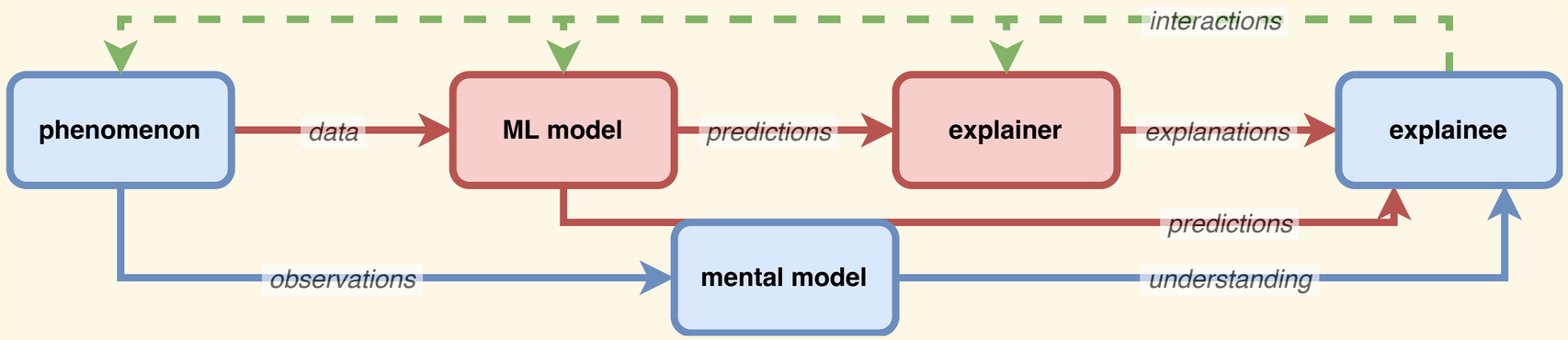
Where Does the Understanding Come From When Explaining Automated Decision-making Systems?

Kacper Sokol

Automated Decision-making

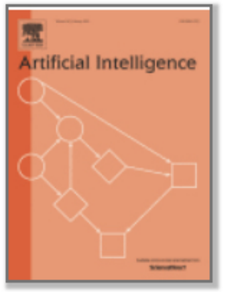

interactions
phenomenon
data
ML model
predictions
explainer
explanations
explainee
predictions
observations
mental model
understanding

Explainability

Human-Centred Explainability

Artificial Intelligence

Volume 267, February 2019, Pages 1-38



Explanation in artificial intelligence: Insights from the social sciences

Tim Miller ✉

Show more

+ Add to Mendeley Share Cite

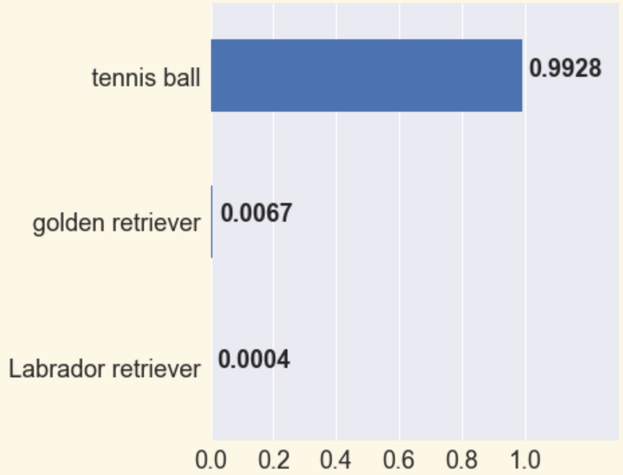
https://doi.org/10.1016/j.artint.2018.07.007

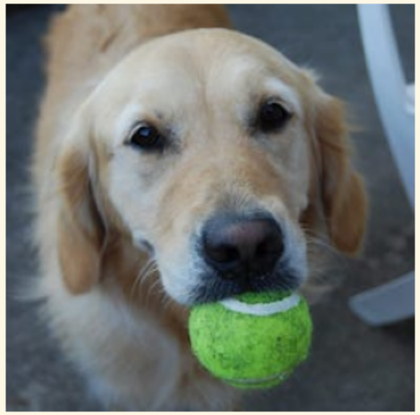
Generating Explanations

Model Prediction

In [7]:
```
classification
```

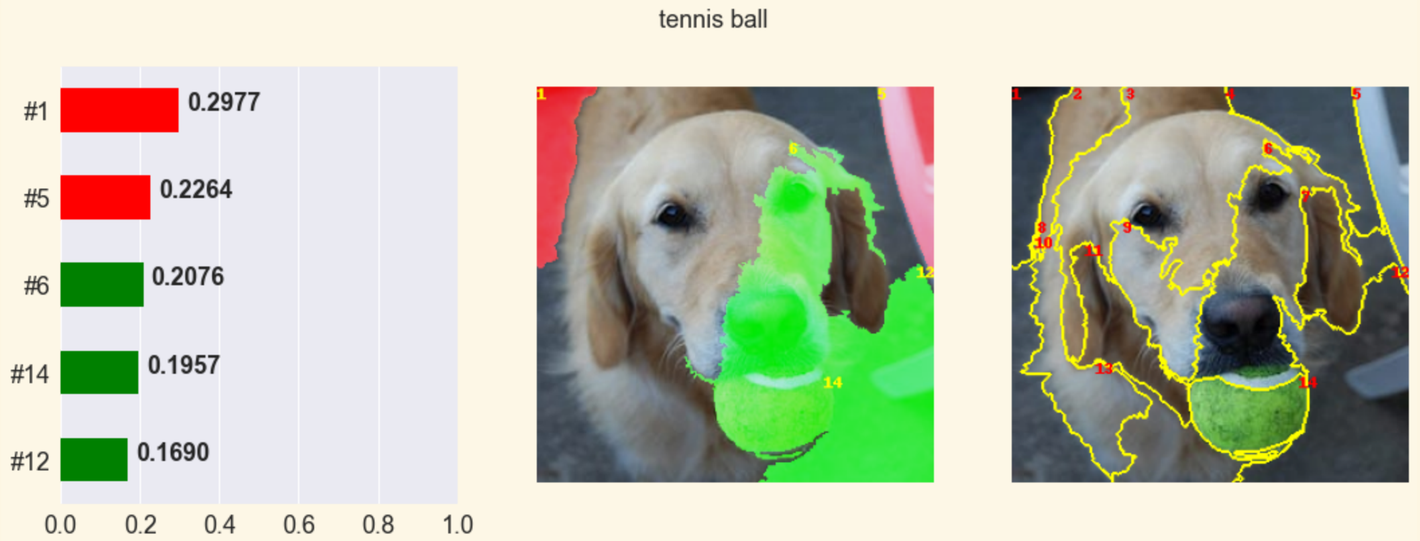
Out[7]:





Prediction Explanation

In [9]:
```
exo.plot_image_explanation(blimey_image, explain_classes[0])
```



Prediction Explanation

In [10]:
```
exo.plot_image_explanation(blimey_image, explain_classes[1])
```

Explainer Demo

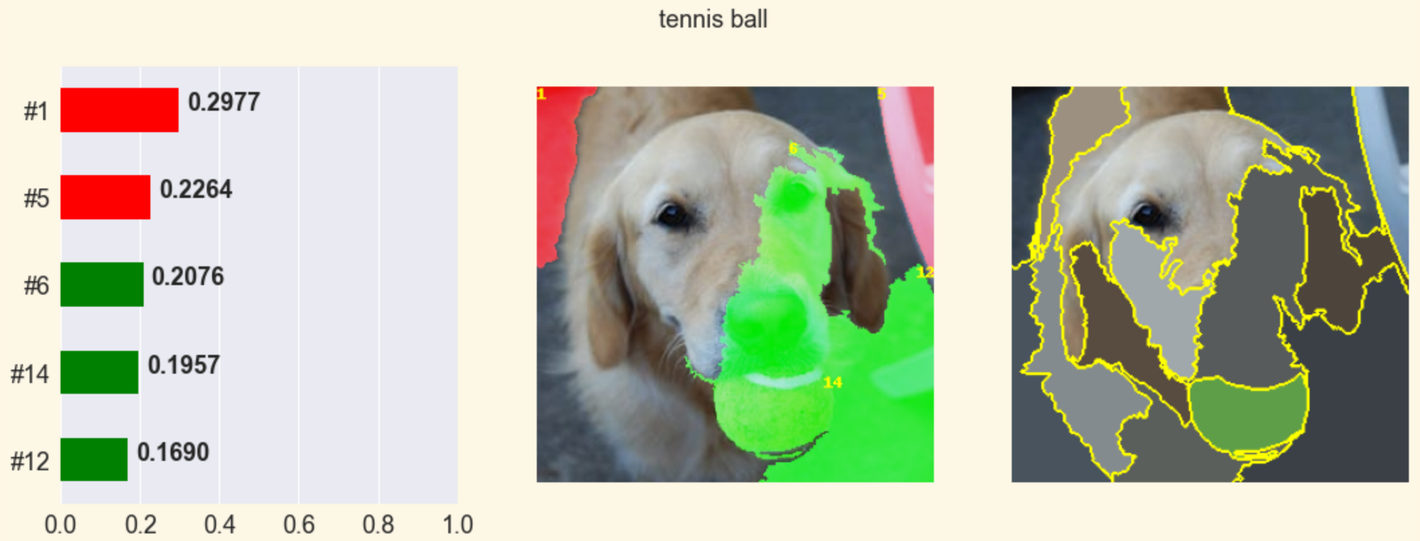
In [15]:
```
surrogate_image_explainer
```

Segmentati... low

Occlusion c... mean | black | white | randomise-patch | green

Explained cl... tennis ball | golden retriever | Labrador retriever

✔ Explain!



Presenting Explanations

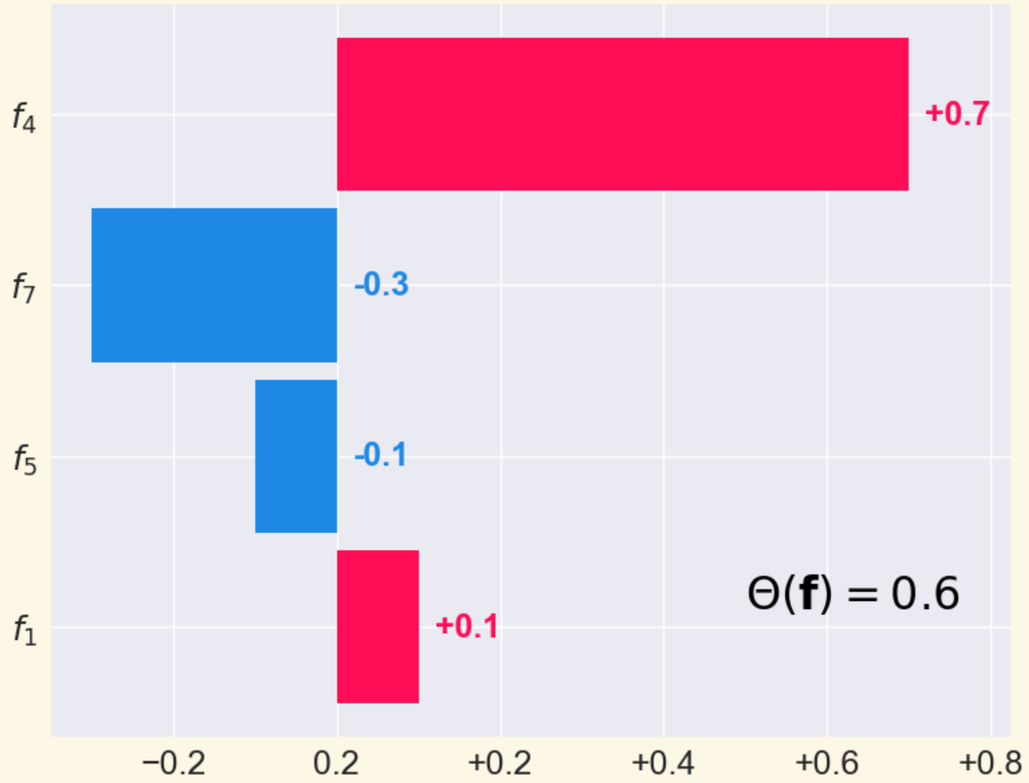
$$\Theta(\mathbf{f}) = 0.2 \quad + \quad 0.25 \times f_1 \quad + \quad 0.7 \times f_4 \quad - \quad 0.2 \times f_5 \quad - \quad 0.9 \times f_7$$

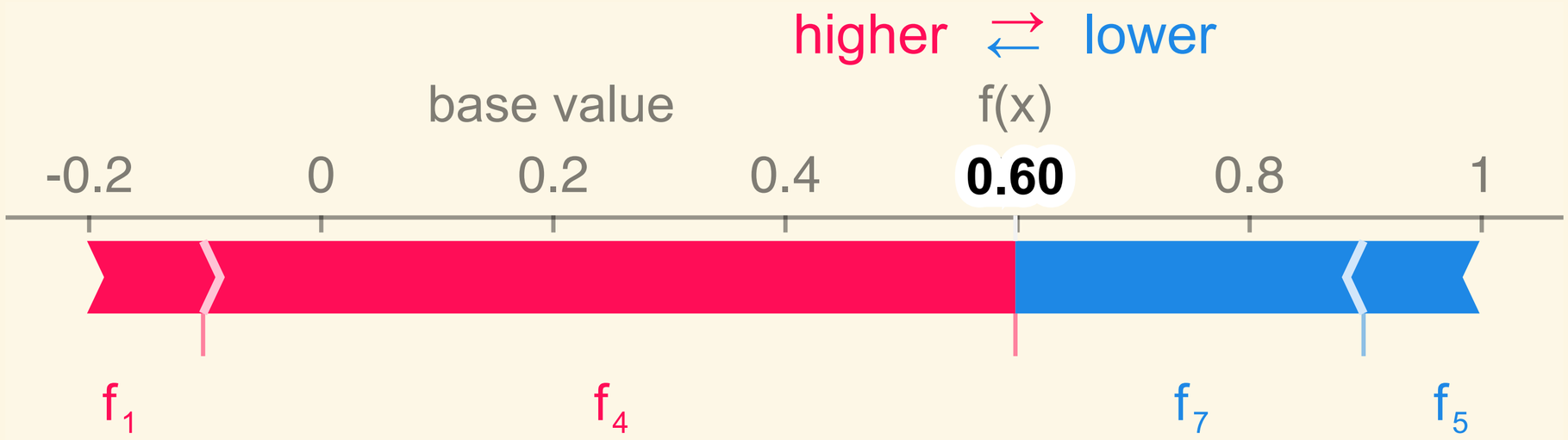
$$\mathbf{f} = (0.4, \dots, 1, \frac{1}{2}, \dots \frac{1}{3})$$

$$\Theta(\mathbf{f}) = 0.2 \quad \underbrace{+0.1}_{f_1} \quad \underbrace{+0.7}_{f_4} \quad \underbrace{-0.1}_{f_5} \quad \underbrace{-0.3}_{f_7} \quad = \quad 0.6$$

In [18]:
```
bar_explanation
```

Out[18]:

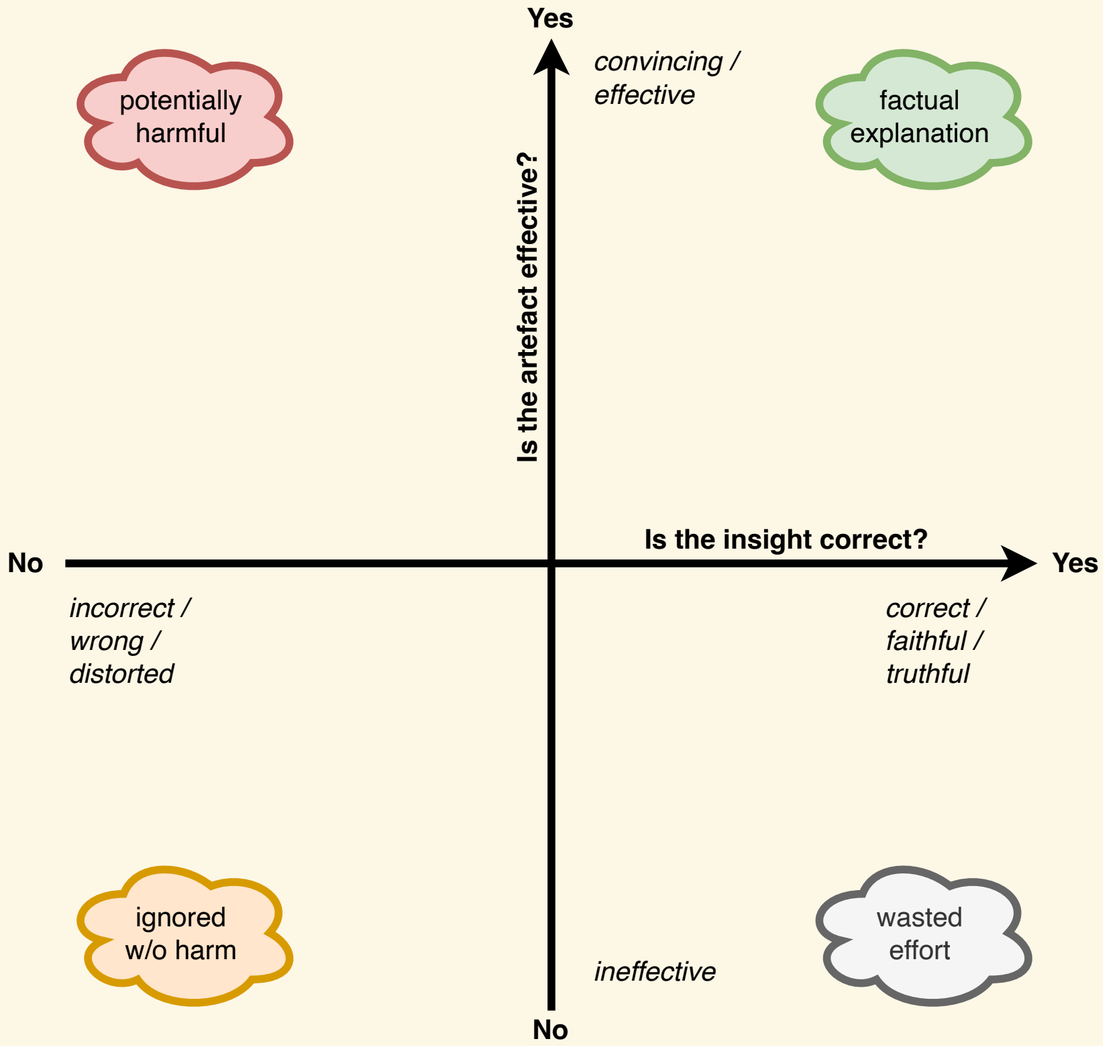



higher ⇄ lower
base value
f(x)
-0.2
0
0.2
0.4
0.60
0.8
1
f1
f4
f7
f5

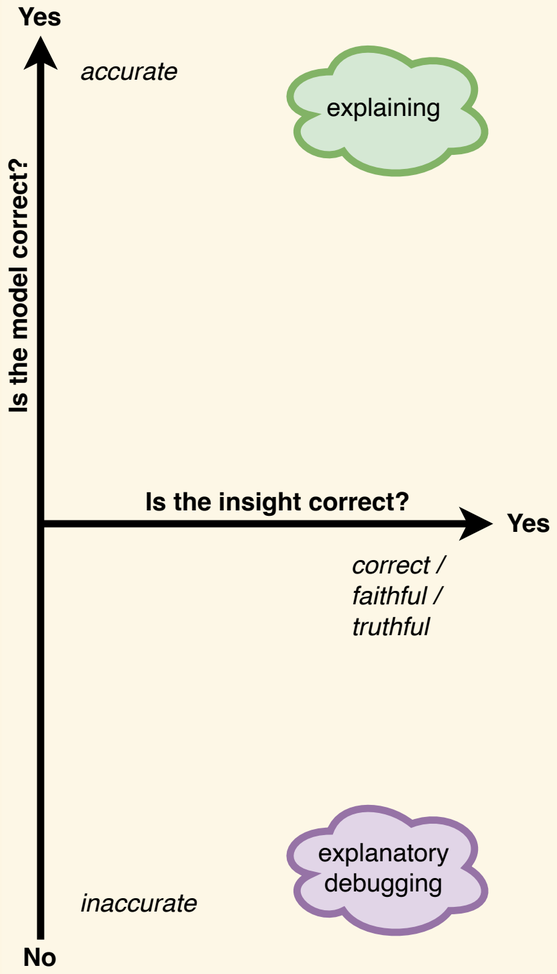
Mapping Explainability

Naïve View

Model and Explanation

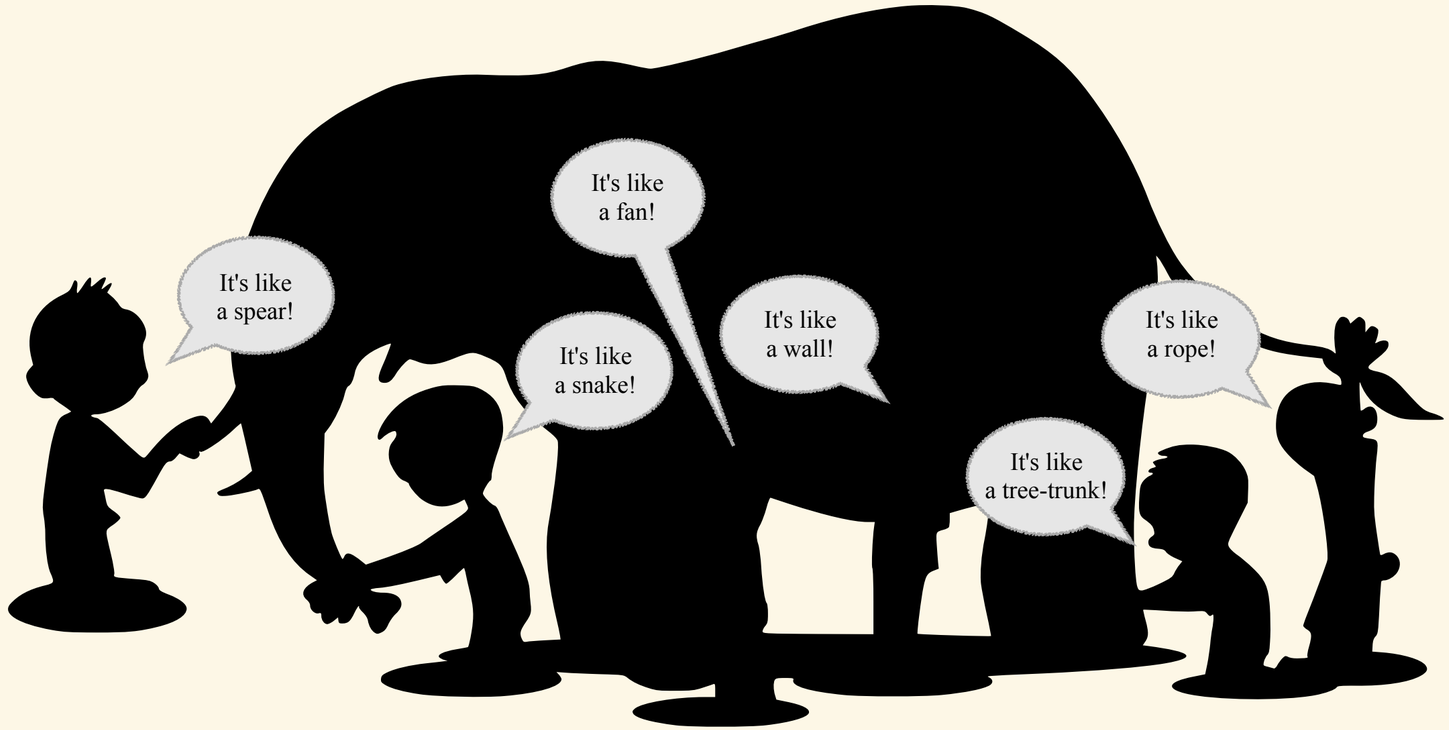


Phenomenon and Explanation

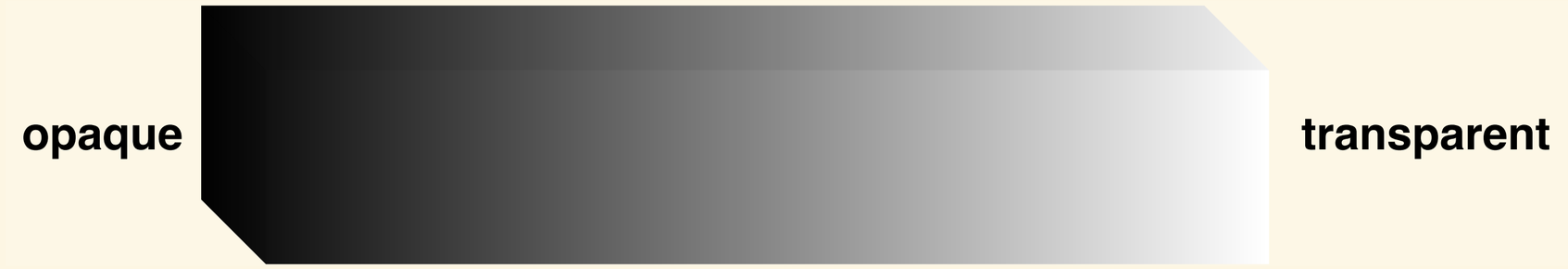


Conceptualising Explainability

The Blind Men and the Elephant.



Explainability is not a binary property; it is a continuous spectrum.



*These are **diagnostic tools** that only become **explainers** when their caveats, properties and outputs are well understood.*

$$\texttt{Explainability} = \underbrace{\texttt{Reasoning}\,(\texttt{Transparency} \mid \texttt{Background Knowledge})}_{\textit{understanding}}$$

- *Transparency* – **insight** (of arbitrary complexity) into operation of a system.
- *Background Knowledge* – implicit or explicit **exogenous information**.
- *Reasoning* – **algorithmic** or **mental processing** of information.

Explainability → **explainee** walking away with **understanding**.

Kacper.Sokol@rmit.edu.au