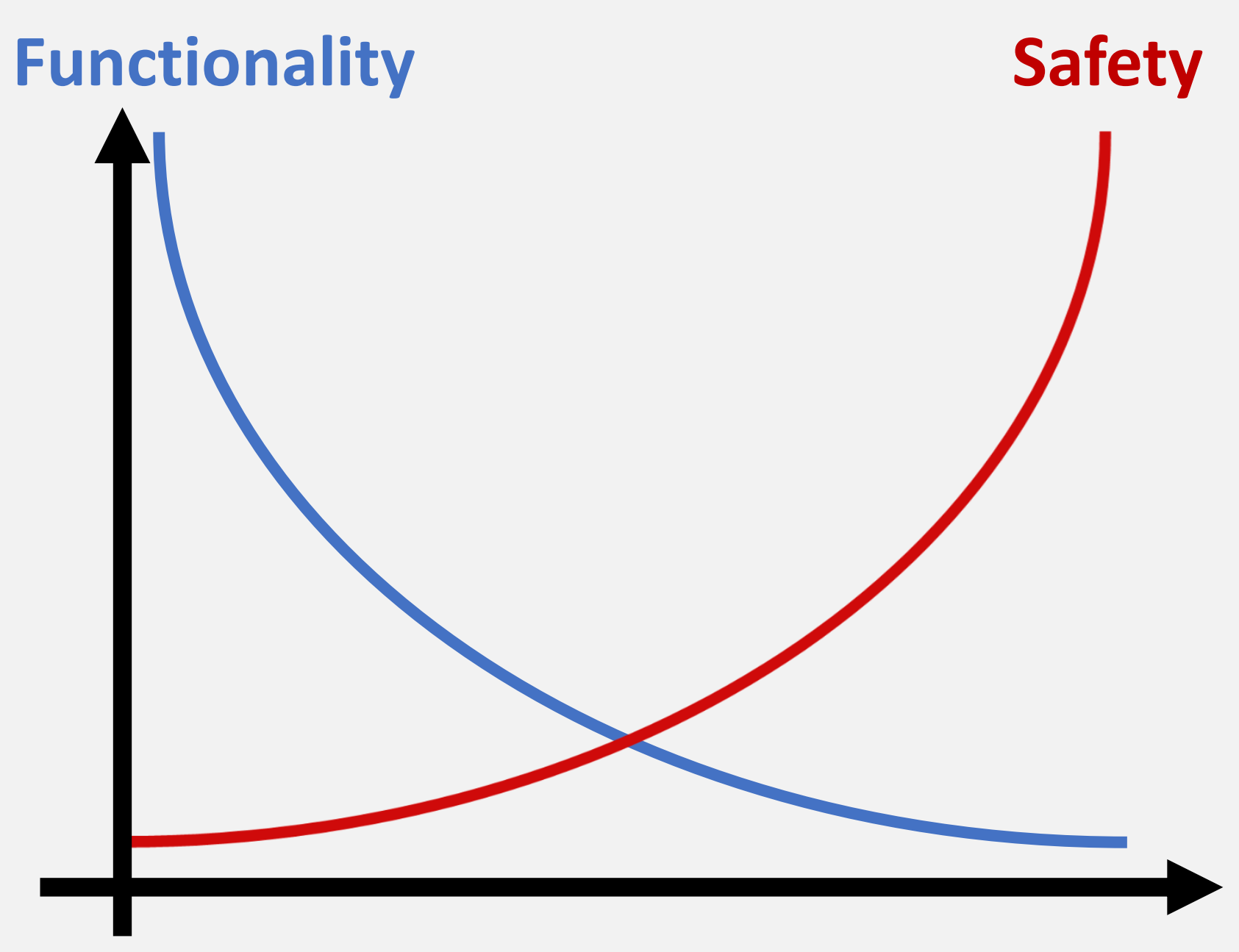
Safety-by-Design for Deep Learning Methods at the Example of Autonomous Driving

Y. Kees*[1], F. Köster[1], S. Hallerbach[1],

[1] German Aerospace Center (DLR), Institute for AI Safety and Security
*yannick.kees@dlr.de

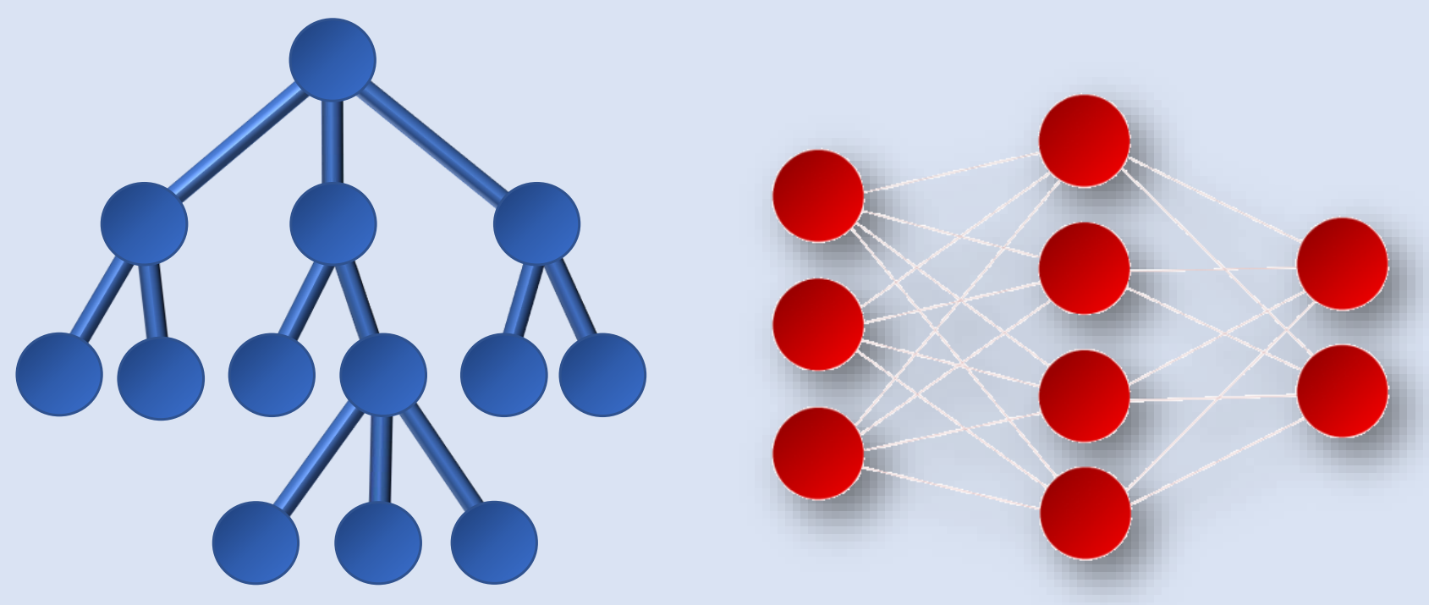
Abstract. The development of autonomous vehicles and autonomous driving functions has increased rapidly in recent years. A driving factor for this is the recent breakthroughs in deep learning. One problem with these methods is that they are **black-box systems**. If we have a system that we cannot understand precisely, this raises the question of how we can use it in safety-critical applications, such as the traffic sector. Often, engineers focus too much on fulfilling the requirements of their stakeholders, so safety-relevant aspects take a back seat. In contrast, we want to consider them early in the **design** and **concept phase**. This approach is known as **safety-by-design** (ISO 26262). Ensuring safety does not mean the absence of risk factors. It means that risks are considered and assessed, and a strategy exists for how best to deal with them (ISO 21448). Recently, much effort has been put into applying these results to AI components (ISO 8800). The field that deals with applying engineering safety arguments to AI models is known as **AI engineering**. Based on these findings, we want to model a **trade-off** between **safety aspects** and the system's **functionality**.

Overall Error



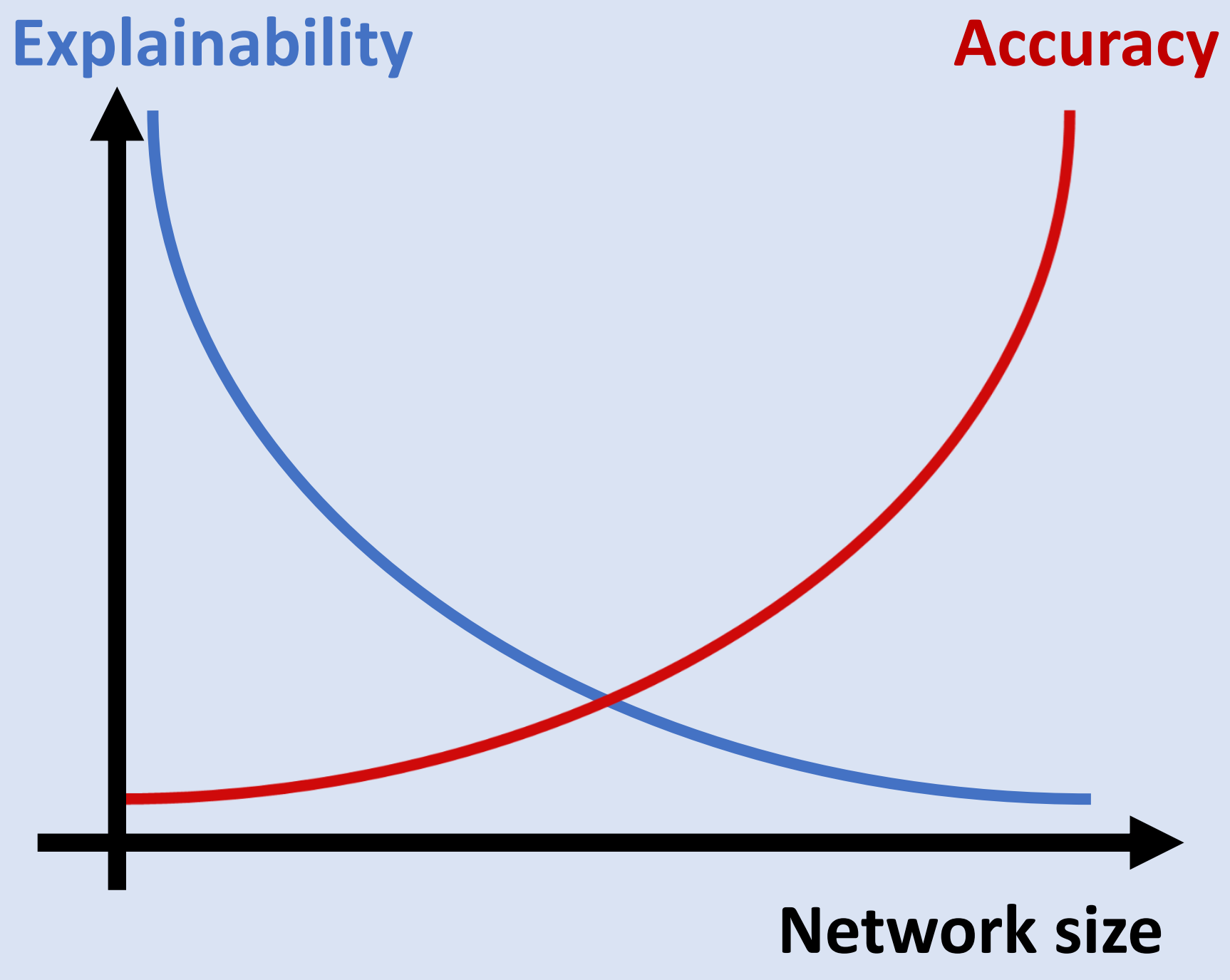
Approximation Error

- ❑ Minimal distance between target function and all possible functions the network can represent
- ❑ Strongly connected to explainability of the system
- ❑ Specific network architectures work better with different sensor data, i.e., CNN for perception modules
- ❑ Explainability decreases with increasing network complexity
- ❑ Need to **compare explainability between systems**



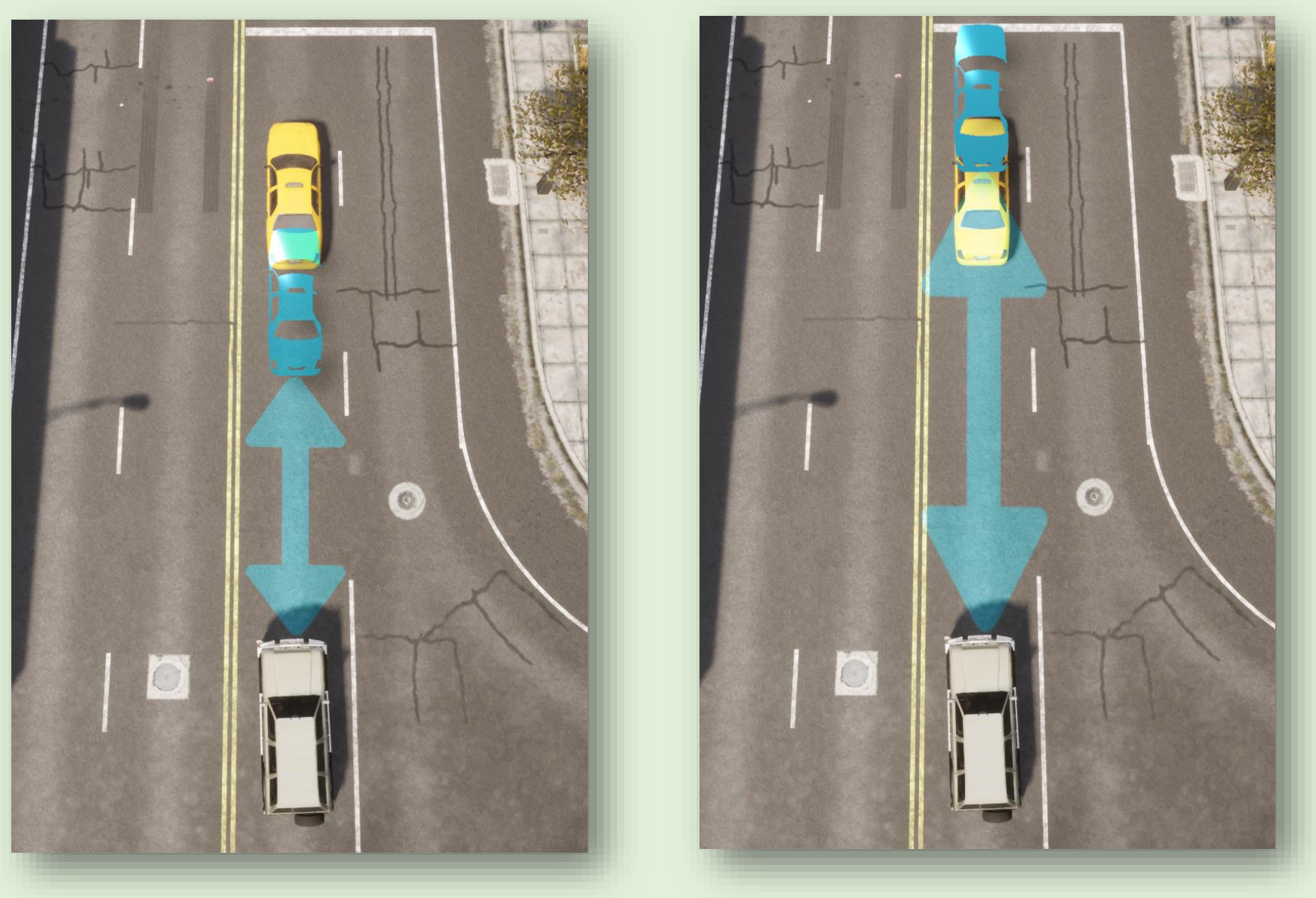
Even though every neural network can be represented as a decision tree, they are considered black boxes because of their sheer size.

➢ **Goal**: Trade-off explainability with accuracy



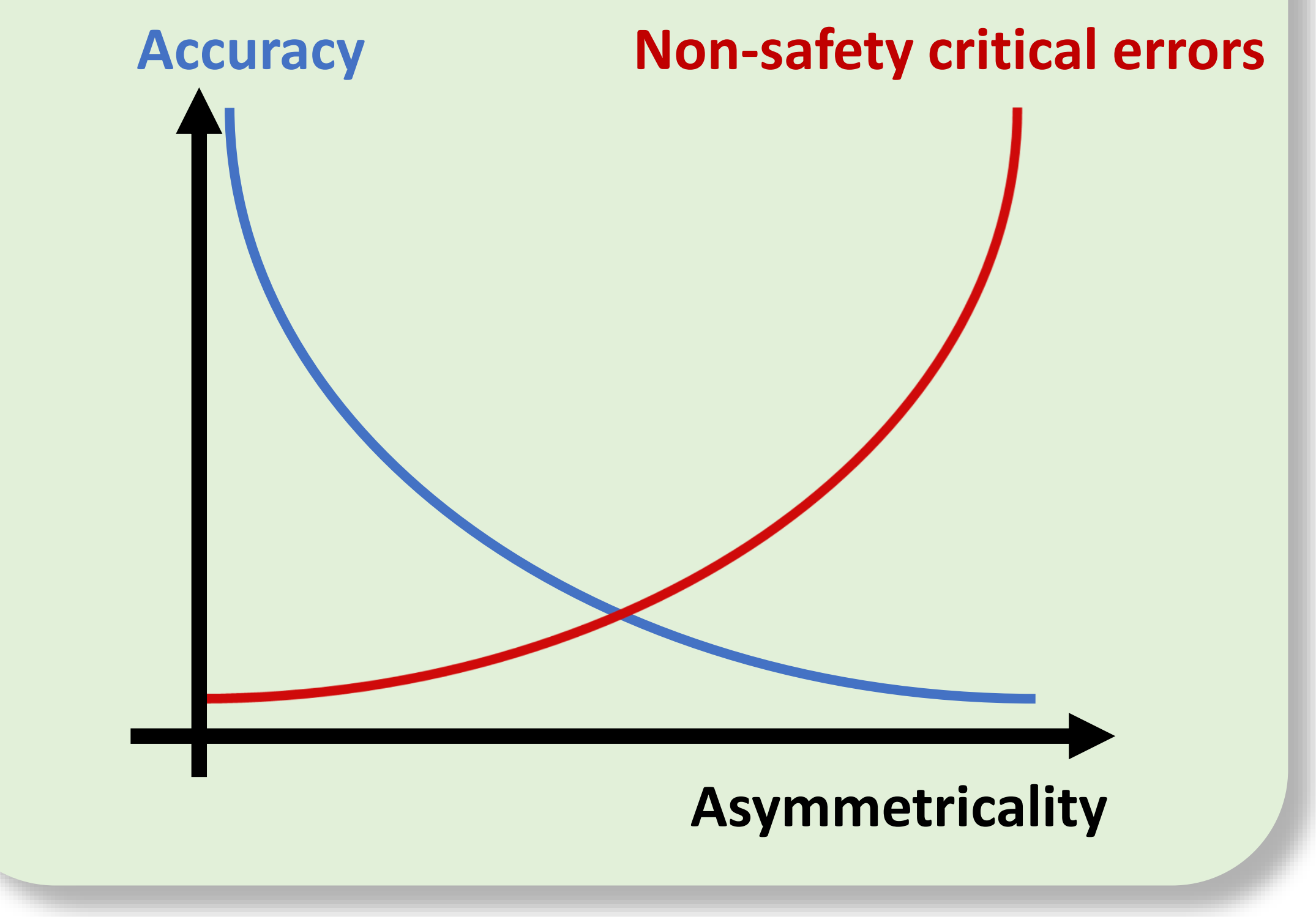
Optimization Error

- ❑ Error arises depending on optimization algorithm
- ❑ Different errors lead to different **safety considerations**
- ❑ Hence, different errors need to be **penalized differently** during training



Example: Estimating a vehicle too far away rather than too close has completely different effects on the vehicle's safety. Hence, this should be represented within the AI training.

➢ **Goal**: Trade-off different types of errors



Generalization Error

- ❑ Distance between performance on training data and test data
- ❑ **Normalization** can provide an **abstraction layer** for incoming sensor data of vehicles
- ❑ Normalization is accompanied by a **loss of information**

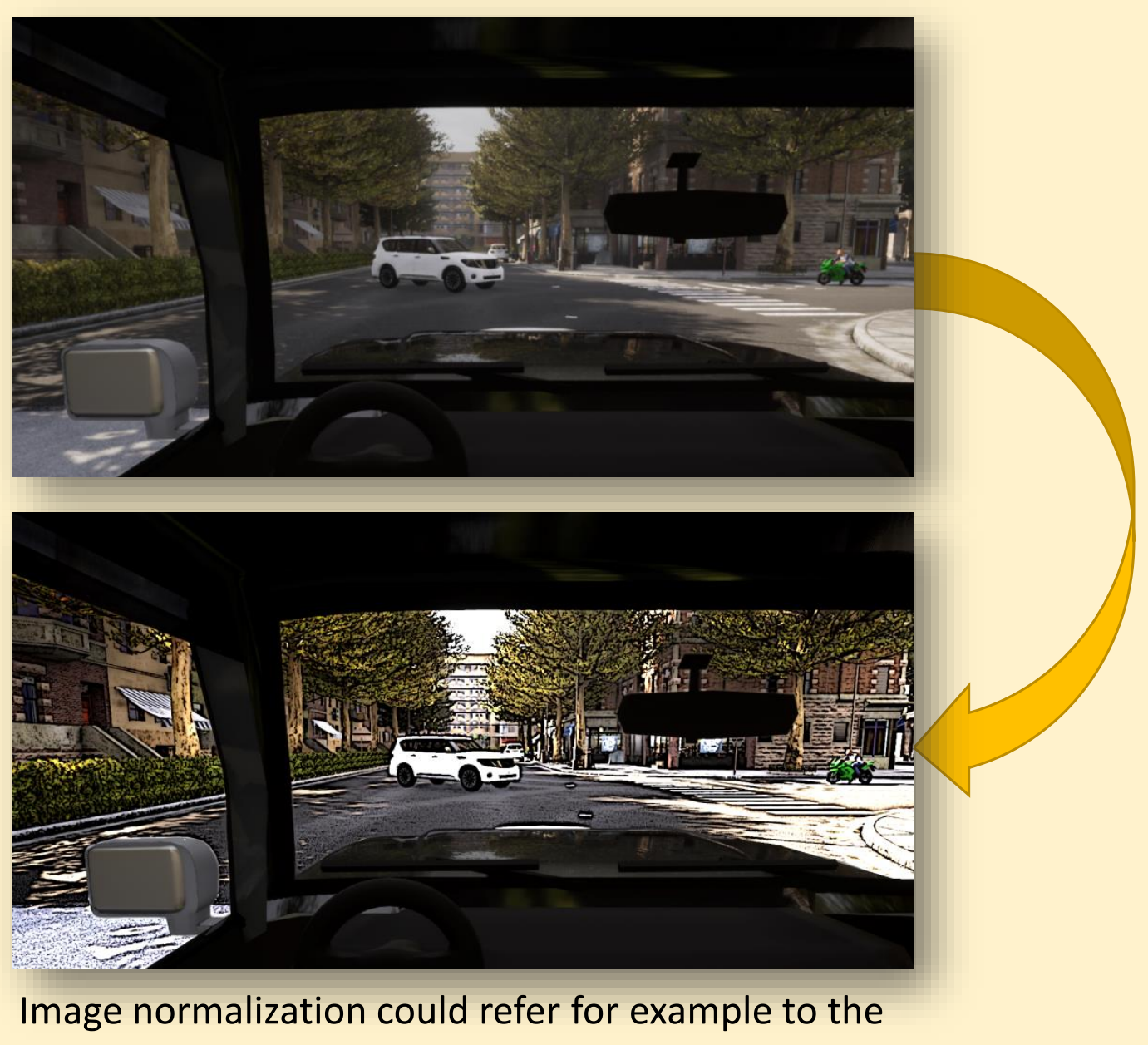
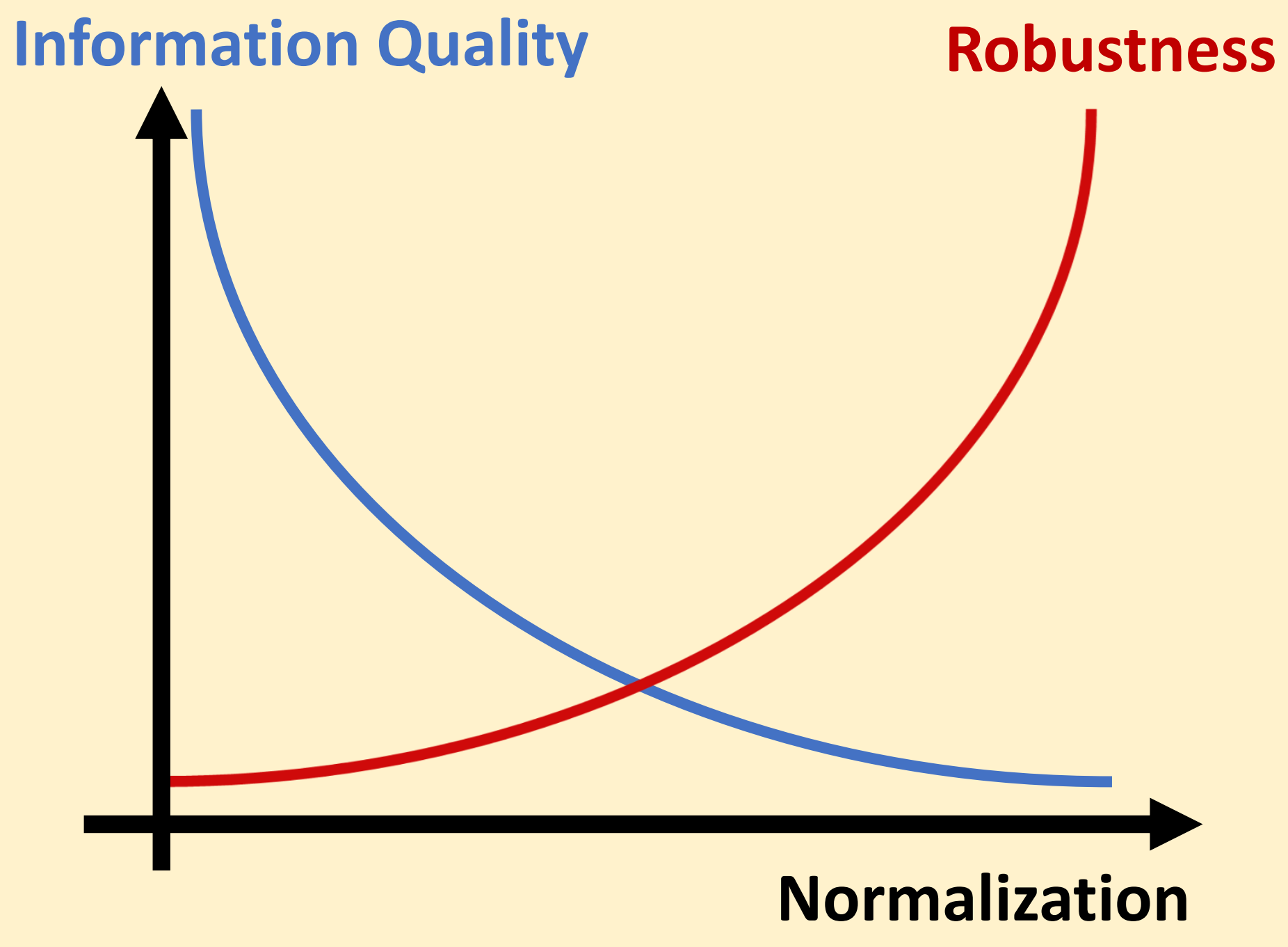


Image normalization could refer for example to the image style, color scale or viewing angle.

➢ **Goal**: Trade-off between information quality of the data and network robustness


Information Quality
Robustness
Normalization