AI & Models: The Bridge to Ethics & Philosophy
Models & AI: Relationship Between Humans and AI
Models & AI: Relationship Between Humans and AI

Human Perceived Reality

Physical Reality

Human

Perception & Interaction

Reality
Models & AI: Relationship Between Humans and AI
Models & AI: Relationship Between Humans and AI

- Human Perceived Reality
- AI Perceived Reality
- Physical Reality

Reality

Human

Machine

DR-Model Learning

Digital Reality

Simulation of Reality

Problem Domain
(entirely part of reality)
Models & AI: Relationship Between Humans and AI

- Human Perceived Reality
- AI Perceived Reality
- Physical Reality

Problem Domain
( entirely part of reality)

Digital
Reality

Human

Machine

H-Model?

M-Model

Model Alignment & Explainable AI

Communication about H- vs M-Model !!

Perception & Interaction

Reality

DR-Model

Learning

Simulation of Reality

Communication about H- vs M-Model !!
Models & AI: Relationship Between Humans and AI

Meta Domain

Lay Person

Lay-Person-View: Meta Model of Problem Domain

Philosophy, others

Philosophy-View: Meta Model of Problem Domain

Artificial Intelligence

AI-View: Meta Model of Problem Domain

Reality

Digital Reality

Simulation of Reality

Perception & Interaction

Communication about H- vs M-Model!!

Human

H-Model?

Machine

M-Model

Model Alignment & Explainable AI

DR-Model Learning

Problem Domain
(entirely part of reality)
Models & AI: Relationship Between Humans and AI

- **Philosophy, others**
  - Philosophy-View: Meta Model of Problem Domain

- **Lay Person**
  - Lay-Person-View: Meta Model of Problem Domain

- **Artificial Intelligence**
  - AI-View: Meta Model of Problem Domain

- **Human**
  - H-Model?

- **Machine**
  - M-Model

- **Reality**
  - DR-Model Learning
  - Perception & Interaction
  - Communication about H- vs M-Model

- **Digital Reality**
  - Simulation of Reality

- **Problem Domain** (entirely part of reality)

- **Model Alignment & Explainable AI**

- **Model Alignment**
Models & AI: Relationship Between Humans and AI

**Models & AI:**

**Reality**

Human

Machine

AI Perceived Reality

Human Perceived Reality

Physical Reality

**Ethics**

Problem Domain (entirely part of reality)

DR-Model Learning

Simulation of Reality

Perception & Interaction

Communication about H- vs M-Model !!

Model Alignment & Explainable AI

DR-Model

H-Model?

M-Model
Ethics/Morality & AI (Provocative!)

- Agreement on a Goal
  - **Wellbeing** of the individual and society [e.g. Sam Harris: The Moral Landscape]

- Evaluating the (likely) outcome of our actions in reality (with respect to goal)
  - E.g. by simulation of action with respect to a *model of reality*

- I argue, this is also what we humans/societies do (*at least* in a rational setting)
  - Model might be implicit/emotional (slow & fast thinking [Kahnemann])

- In principal, AI could/should do the same – but ...
  - Our models of the world are mediocre (but so were early weather models ...)
  - AI cannot deal with the complexity (yet ...)
  - The metric for evaluation is not clear (but might be a function of the goal ...)