Value alignment and control

- King Midas, paperclips and trolleys
- Measure what matters and manage tradeoff
- Prediction versus optimization and control
- **Maintaining human oversight**

Draft European AI regulations

- *Unacceptable-risk AI systems:*
  - subliminal, manipulative, or exploitative systems that cause harm
  - real-time, remote biometric identification systems used in public spaces for law enforcement
  - all forms of social scoring, such as AI or technology that evaluates an individual’s trustworthiness based on social behavior or predicted personality traits.
- *High-risk AI systems:*
  - Systems evaluating consumer creditworthiness, recruitment, or employee management
  - Systems used in the administration of justice
  - Systems utilising biometric identification in nonpublic spaces
  - Safety-critical systems or systems that put the health of citizens at risk due to failure

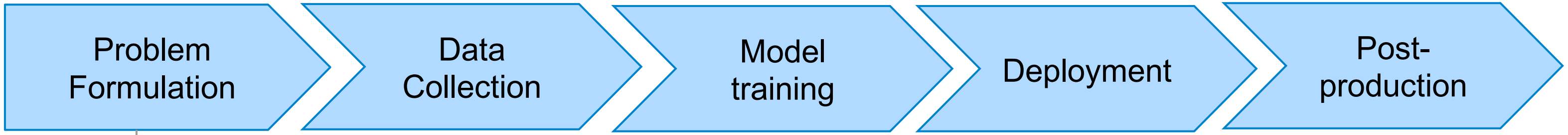
https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/what-the-draft-european-union-ai-regulations-mean-for-business
https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

High-risk AI systems must be under human oversight

- Implementation of data governance and risk-management systems
- Technical documentation, record keeping and logging
- Transparency and explainability
- Accuracy, robustness and cybersecurity (safety)
- **Human oversight**
  - *Throughout the AI system's lifecycle*
  - *Appropriate oversight measures designed before the system goes to production*
  - *Operational constraints that cannot be overridden by the system itself*
  - *Responsiveness to the human operator*
  - *Human supervisor has the necessary competence, training and authority to carry out that role*

https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/what-the-draft-european-union-ai-regulations-mean-for-business
https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

“Human oversight throughout the AI system’s lifecycle”

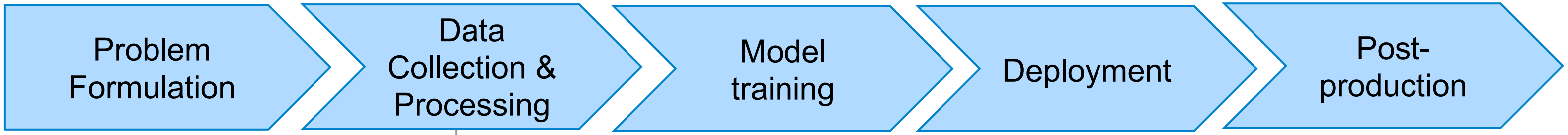


Problem Formulation → Data Collection → Model training → Deployment → Post-production

Value Alignment Tasks

- Target variable selection
- Choose whether the task is prediction, decision, or optimization ?
- Choose loss / reward function *a priori* where possible, otherwise estimate it from human preferences
- Establish additional metrics
- Estimate relative preferences *a priori* where possible (otherwise inspect Pareto front *a posteriori* in training)
- Estimate difficulty / resourcing needed for model development

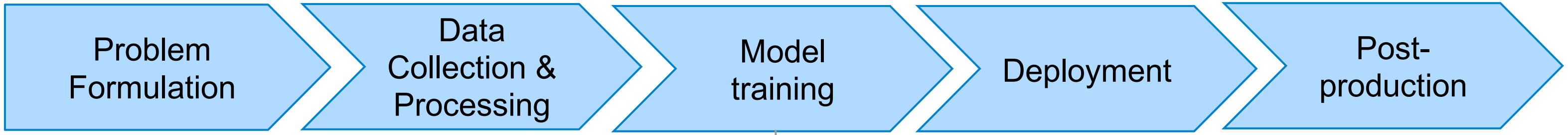
“Human oversight throughout the AI system’s lifecycle”



Value Alignment Tasks

- Labelling data manually (e.g., by experts) or by specifying programmatic rules (weak supervision)
- Consider data selection mechanisms
- Consider historic/systemic bias in data
- Consider bias in proxy target definition
- Quality assurance of engineering pipeline
- Quality assurance of missing data imputation, outlier detection or other processes that modify data in place

“Human oversight throughout the AI system’s lifecycle”



Problem Formulation

Data Collection & Processing

Model training

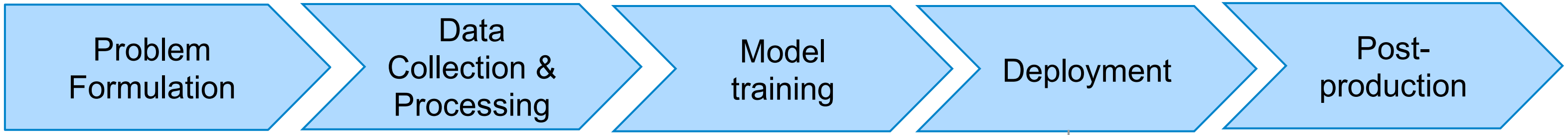
Deployment

Post-production

Value Alignment Tasks

- Recommend distance functions in unsupervised learning
- Assess model explainability vs accuracy
- Ensuring best practices are being kept
- *A posteriori* analysis of Pareto optimal front and model selection
- Obtaining good estimates of uncertainty

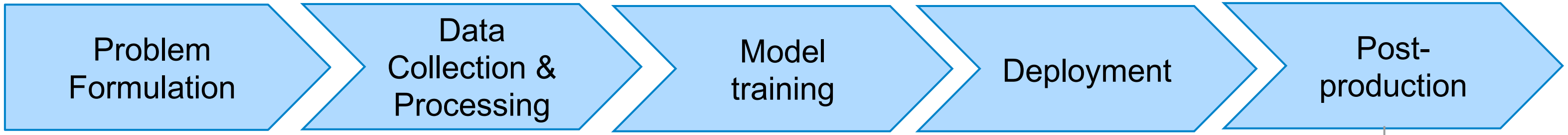
“Human oversight throughout the AI system’s lifecycle”



- Validate via challenger/champion setup
- Decide on active learning strategy / criteria for passing control back to human
- Design suitable “override” systems depending on the use case
- Ensure end user has the ability to give feedback or file complaints
- Decide on retraining strategy

Value Alignment Tasks

“Human oversight throughout the AI system’s lifecycle”



Problem Formulation

Data Collection & Processing

Model training

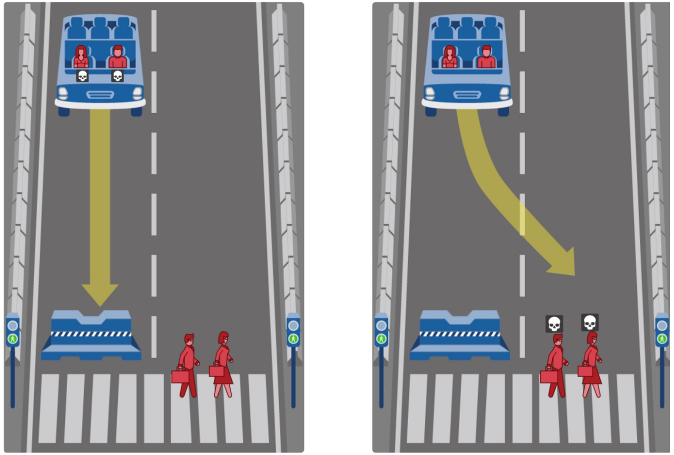
Deployment

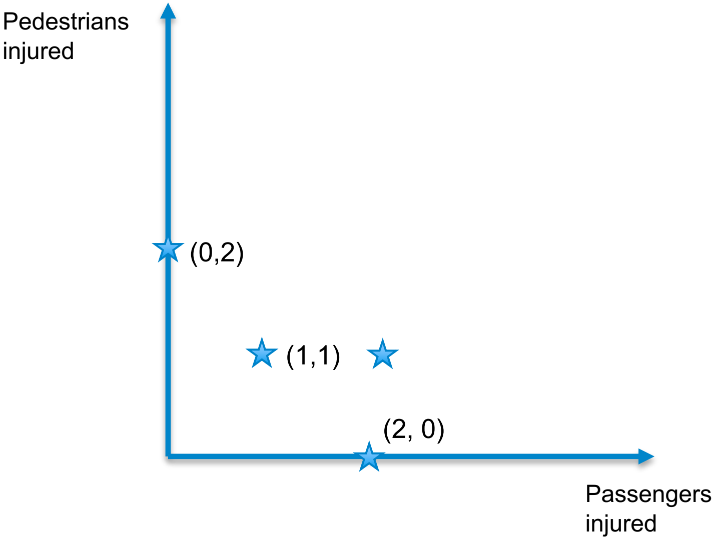
Post-production

Value Alignment Tasks

- Monitor performance along multiple fronts and repeat *a posteriori* Pareto analysis
- Keep track of model lifecycle and ensure maintainer/owner is involved in all usage
- Regular quality assurance reviews and manual error analysis

A priori determination of preferences is sometimes hard





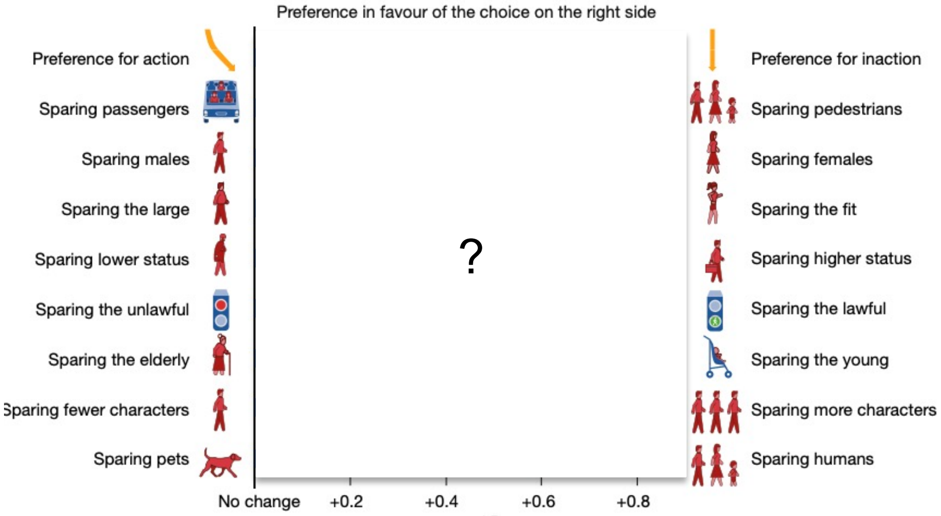
https://www.moralmachine.net/

The Car That Knew Too Much. Can a Machine Be Moral? By Jean-Francois Bonnefon

Imperial College London

Real-life trolley problems

Try to place weights on the relative importance of the following:

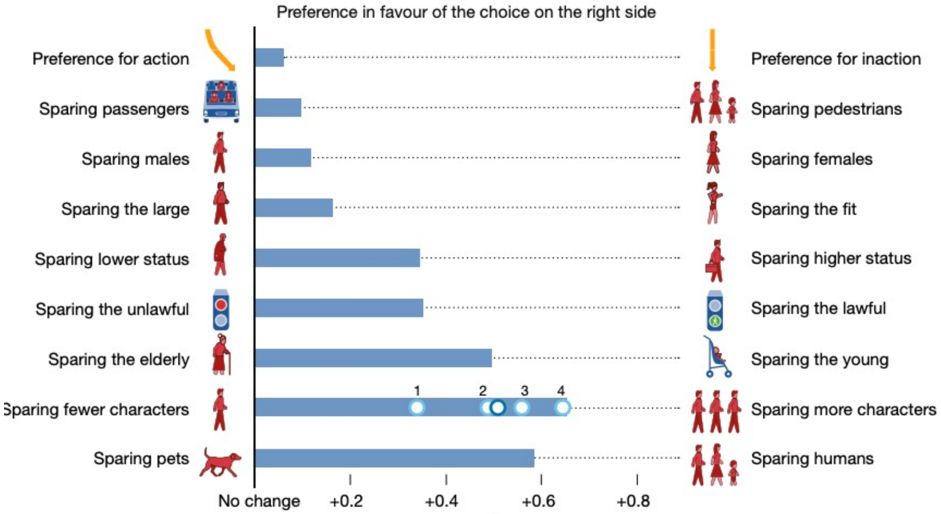


https://www.nature.com/articles/s41586-018-0637-6

Imperial College London

Real-life trolley problems

Try to place weights on the relative importance of the following:



These differed by country (e.g., Eastern countries show greater preference for an equitable attitude between young and old)

https://www.nature.com/articles/s41586-018-0637-6

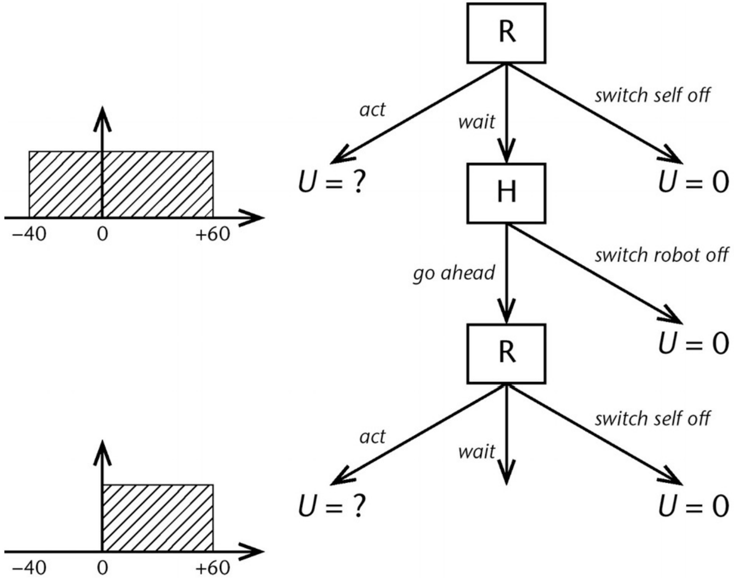
Will a robot allow you to switch it off?

Survival is an instrumental goal for a robot – it needs to survive, to make you tea.

Recent work by Stuart Russell and others is attempting to formalize such scenario

One example is the “switching off game”. A robot is modelled as having the option to switch itself off (give up), perform an action with an expected utility, or ask for permission.

Uncertainty about the reward of its actions (epistemic humility) is a key to safety



https://www.nature.com/articles/s41586-018-0637-6
Russell, Stuart. Human Compatible (p. 197). Penguin Books Ltd

Summary

- Human oversight is likely to be a regulatory requirement in all high-risk AI contexts.
- Value alignment is an appropriate framing for making important choices that are hard to reverse, like the choice of target variable, loss function, and relative preference between that and other metrics.
- Some of these choices can be revisited *a posteriori* in structured ways, like Pareto front inspection.
- Post-deployment, human oversight requires either the AI system handing back control, or manual override by humans. In both cases, *epistemic humility* and faithful reporting of its own uncertainty is a key to achieving safe but efficient human oversight. With increased autonomy, this becomes harder.
- AI safety research is attempting to formalize these questions – but it is only the beginning.

https://www.nature.com/articles/s41586-018-0637-6