RESPONSIBLE ARTIFICIAL INTELLIGENCE: DESIGNING AI FOR HUMAN VALUES

This paper explores the impact of AI in the case of the expected effects on the European labor market, and propose the accountability, responsibility and transparency (ART) design principles for the development of AI systems that are sensitive to human values.

Argues that there are no approaches to the responsible design of AI (maybe because it is from 2017)

QUOTE: However, concrete approaches to the responsible design of AI are mostly non-existent

Greater autonomy must come with greater responsibility

The impact AI can have on us is yet to be determined

QUOTE: current studies on the influence of AI on the jobs market vary from a Utopian society in which nobody has to work, to the ending of economic growth in the western world

QUOTE: but there is no clear European policy vision on potential harmful effects on the jobs market

QUOTE: few studies have provided a clear estimate of the amount of jobs that will be created or on the nature of future jobs

Responsibility is a must

QUOTE: responsibility is fundamental to intelligence and no system can be truly intelligent if it cannot understand responsibility

They argue that responsibility is divided into three principles:

1. Responsibility awareness, or, accountability by the creators
2. Mechanisms that enable AI systems to reason about, and act accordingly to, ethics and human values
3. Participation; understand how different people work and live with AI

AI systems usually are classified by their autonomy, interactivity and adaptability, and they propose to add accountability, responsibility and transparency (ART) to them. Accountability refers to the need to explain and justify one’s decisions and actions. Responsibility refers to the role of people themselves, and to the capability of AI systems to answer for one’s decision and identify errors or unexpected results. Transparency, to the need to describe, inspect and reproduce the mechanisms through which AI systems make decisions (what is behind)

Value Alignment or Misalignment – What Will Keep Systems Accountable?

This paper argues that the proposal to achieve value alignment through inverse reinforcement learning (IRL) is ethically inadequate for a computational architecture designed to guide artificial agents.

The technical challenges IRL has are data bias, generalization issues, and the adequacy of reward functions to represent temporary complex norms

Traditional reinforcement learning is vulnerable to deception on the part of the system.

IRL is the task of inferring a reward or utility function by observing the behaviour of other agents in a RL-like system. Sort of finding an ‘idealized perfect agent’ through modelled behaviour

They argue that it is very difficult to determine the ethics behind the training – although this is common for every technique –

QUOTE: In order for a robust assessment to be made of what an agent or agents did, we must also consider what alternative actions could have been chosen and why: counterfactuals matter in how we judge what an agent did.

Technicalities

QUOTE: If an IRL agent learns from unethical behavior, it will learn to behave unethically

QUOTE: IRL is rightfully admired for its ability to generalize to unseen states, this advantage is counterbalanced by the fact that an IRL agent must learn by observing others’ behavior and thus has no control over what behaviors and which portions of the state space it observes.

QUOTE: we do not believe that IRL by itself can solve the problem of training agents to be moral.

Augmenting Reinforcement Learning with Human Feedback

There is a framework called TAMER+RL which enables human feedback to augment traditional RL agents using MDP (Markov’s decision process) rewards