Financial Computing, Faculty of Engineering Sciences

UCL

Defining and Identifying the Legal Culpability of Side Effects Using Causal Graphs

SafeAI22

Hal Ashton

Motivation - a semantic exercise in alignment

The study of Side-effect avoidance is a recent research topic in Safe AI research.

What constitutes a ‘side-effect’ has been left somewhat vague.

Problem 1:

With no definition it is difficult to guarantee everyone is talking about the same thing.

Solutions to avoid ‘side effects’ may not be transferable.

Problem 2:

By labelling something as a ‘side-effect’, responsibility for that harm is tacitly lowered

Objective: When we talk about side-effects as Computer Scientists: 1) we should agree what they are 2) Society should agree as well.

Side Effects Background Research in Safe AI

Side effect Research is preoccupied with the safety of vases.

Amodei et al (2016) identify - 'Avoiding Negative Side Effects' as one of their top 5 problems in AI Safety - Side effect = any negative effect that might be caused by a policy which is not represented explicitly in the agent's reward function.

Impact Regularisers - (Armstrong and Levinstein, 2017) - Complete task by changing world as little as possible.

Related concept of reversibility - (Eysenbach et al, 2017)

Krakovna et al (2020) generalise to policies which must ensure possible future tasks are still possible.

Saisubramanian et al (2021) present a taxonomy of side-effect mitigation techniques - Severity, Reversibilty, Avoidability, Frequency, Stochasticity, Observability

Hang on, shouldn't we just concentrate on stopping bad stuff?

Harms will occur despite best efforts.

Outside cyber physical systems Harms are not always immediately obvious.

Identifying what are true side-effects can help with understanding why a particular policy of an algorithmic agent is going wrong.

Determining the intentional status of caused harm is a key part to the functioning of the law to which even engineers are beholden.

Unpeeling the Side-effects Onion 1 - Medical Definition

Side-effects are a major concern of regulated medicine, so why not use the same definition:

"Any reaction secondary to the intended therapeutic effect that may occur following administration of a drug or treatment" - APA

Unpeeling the Side-effects Onion 2 - Intent Definition

"Any reaction secondary to the intended therapeutic effect that may occur following administration of a drug or treatment" - APA

For something to be considered a side-effect - It is not intended, it is not the purpose of the therapy.

MPC (Model Penal Code) definition of purpose (aka intent):

"A Person act purposefully with respect to a material element of an offense when…it is his conscious object to engage in conduct of that nature or to cause such a result"

To identify side effects we need to understand what caused effects are intended.

Unpeeling the Side-effects Onion 2 - Means End Consistency

Philosophy and Law both place a consistency requirement on outcomes intermediate to the intended outcome.

If an intermediate outcome is necessary to occur for the intended outcome to occur, then it is also said to be intended (if it is also caused by the agent)

The omelette requirement.

This prevents agents from claiming innocence by saying that they only intended some final result.

This is particularly relevant to policies found through machine learning where objectives are given and policies are learned with minimal guidance as to how the objective should be reached.

Side Effect Identification Procedure

Side-Effects are not intended, therefore they cannot be necessary for intended outcomes to occur.

1. Identify all intended outcomes of an action policy.
2. Identify all intermediate outcomes necessary for intended outcomes.
3. Side effects of a policy correspond to remaining realisations of variables causally dependent on a policy.
4. They may not realise if the intended outcome realises.

Intended outcomes which fail to realise and take some other value are not side-effects - perhaps adverse outcomes?

Culpability for side-effects

Labelling a caused harm as a side effect diminishes the culpability for that harm because side-effects are never intended.

The Law generally defines different levels of culpability for caused harms dependent on the mental state of the actor.

- **Purpose (Intent)** - Outcome of actions is the actor's aim to achieve,

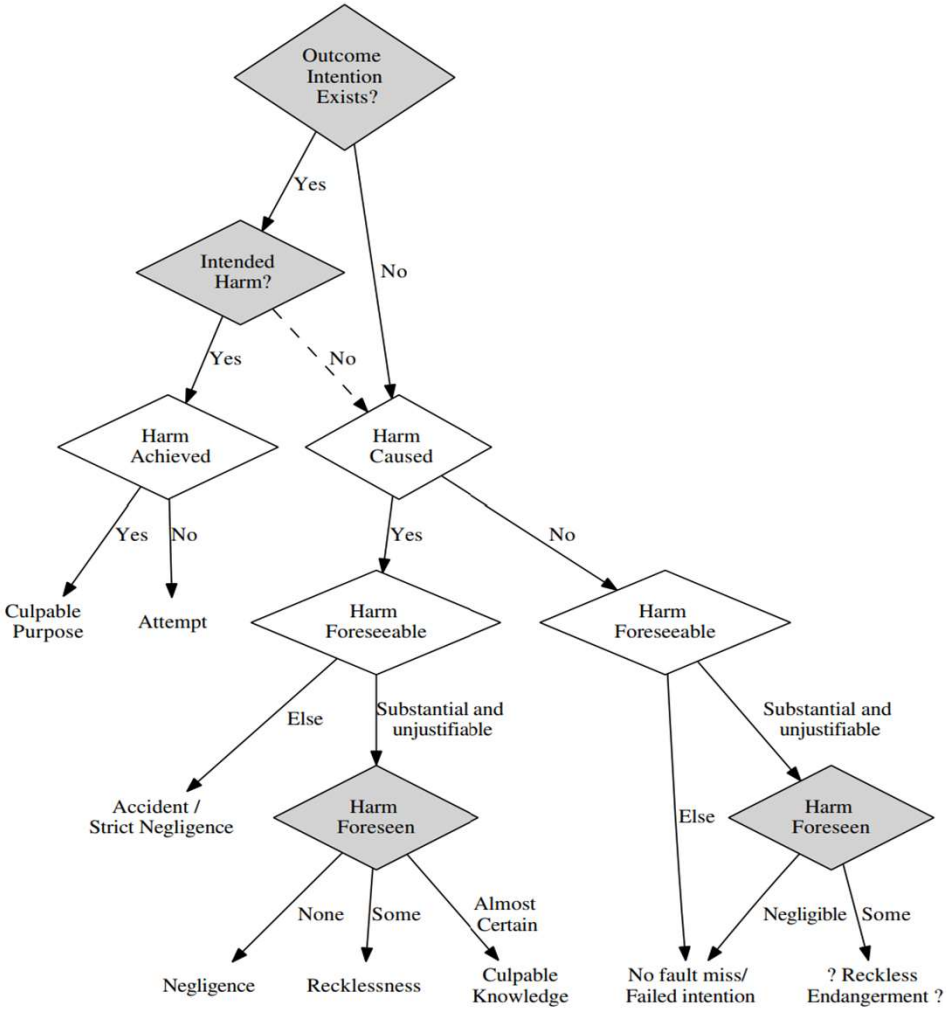
- **Knowledge** - Almost certain outcome of actions is known to the actor at point of commission.

- **Recklessness** - Outcome of actions is foreseeable to the actor, they continue anyway.

- **Negligence** - Outcome of actions is foreseeable to a reasonable person

Factors to determine culpability of side effects

1. Intention
2. Causation
3. Foreseeability of harm (to a reasonable person)
4. Harm foreseen by Actor



Financial Computing, Faculty of Engineering Sciences

Defining and Identifying the Legal Culpability of Side Effects Using Causal Graphs

SafeAI22

Hal Ashton

ucabha5@ucl.ac.uk