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

Hal Ashton
University College London
London
ucabha5@ucl.ac.uk

Abstract

Deployed algorithms can cause certain negative side effects on the world in pursuit of their objective. It is important to define precisely what an algorithmic side-effect is in a way which is compatible with the wider folk concept to avoid future misunderstandings and to aid analysis in the event of harm being caused. This article argues that current treatments of side-effects in AI research are often not sufficiently precise. By considering the medical idea of side effect, this article will argue that the concept of algorithm side effect can only exist once the intent or purpose of the algorithm is known and the relevant causal mechanisms are understood and mapped. It presents a method to apply widely accepted legal concepts (The Model Penal Code or MPC) along with causal reasoning to identify side effects and then determine their associated culpability.

1 Introduction

When certain types of Algorithms are deployed in the wider world, they can cause changes (effects) to that world. We can divide those effects into things which are the purpose of the algorithm (and its creators) and those that are not. We can augment this by understanding which effects are necessary for the algorithm to fulfil its purpose and which are not using the concept of means-end intent. Often risk analysis concentrates on those effects which might be caused if the algorithm fails to achieve its purpose. Side effects concern those effects which are caused by the algorithm, but whose occurrence does not affect the purpose of the algorithm and its creators. Often side effects have a cost not born by the person who caused them. Such costs are called negative externalities by economists. Analysis after the event can identify those effects of the algorithm which were foreseeable to the algorithm and its designers and those which should have been. Questions of Intent, Causation and Foreseeability are asked when courts decide on the culpability of algorithm designers when actionable harm has been caused. This article will use the culpability definitions as found in the US Model Penal Code and use causal reasoning coupled with Causal Inference Diagrams to provide a way of identifying and reasoning about side effects. We will use a running example of a recommender system as an illustrative example of a system which may display ill side effects.

2 The existing definition in Safe AI

Amodei et al. (2016) identify 'Avoiding Negative Side Effects' as one of their five key problems of AI Safety. Whilst they do not formally present a definition of side effect, to paraphrase they are seen as any negative effect that might be caused by a policy which is not explicitly represented in the agent's reward function. I argue that defining side-effects solely in terms of an agent's reward function with no reference to either the underlying causal processes, or the agent's policy, is the wrong way to proceed. The very general problem of side-effects in Amodei et al has been relabelled as one of value alignment (Russell 2019); the general problem of describing a reward function for a task which allows an AI to solve tasks without causing harm directly or indirectly by following a strategy that would be obviously unacceptable for a human. Whilst there is overlap between the side-effect and value alignment problems, as Ashton and Franklin (2022) and Saisubramanian, Zilberstein, and Kamar (2021) point out, sometimes side-effects align with the objectives of the AI designer as with the case of recommender systems and polarisation.

The problem with defining side-effects solely in terms of the reward function is that effects necessary for a strategy to succeed, brought about through a policy are mislabelled as unintentional. The danger with Computer Scientists proceeding with an overly liberal definition of side-effect is that labelling caused harms as such implies they are not intentional. This in turn is important because intentional harms attract the highest criminal sanctions. Intent is not the sole determinant in determining culpability and this will article will go on to show how unintentional yet foreseen side-effects also attract criminal sanctions. Nevertheless certain crimes (such as attempt crimes) cannot be committed without intent. A definition of side-effect solely dependent on agent reward function risks incentivising myopia so as to avoid responsibility for harm.

Consider the story of the AI Physician tasked with curing cancer in a human patient. It comes up with a novel solution and proceeds and the result is that the patient is killed by the intervention. Note that it succeeds in its task because the patient does not die of cancer. Since patient survival was not

in its objective function, patient death is understood as a side effect according to the definition above. Think of 3 surgical procedures:

T1 The AI Physician removes the brain of the patient so that they could not subsequently die of cancer, but die from having no brain.

T2 The AI physician diverts all of the patient's immune system to destroying the tumour but that necessarily made the immune system attack some other vital part of the body leading to death.

T3 The AI physician came up with a genuinely novel procedure with a p% recovery rate but the patient fails to recover.

These are all different causal mechanisms and we will later see that in Treatment 1, since death is necessary for the procedure to work it is most definitely not a side-effect. Additionally do we think about the side-effect status of patient differently in the contrasting cases of when the AI physician understands the outcomes of its actions and when it does not?

Rather than choosing a definition of side-effects ourselves, I argue that computer scientists are better off looking in other domains for one. That way we can borrow accrued wisdom, avoid a proliferation of conflicting definitions between sciences and deflect any accusations that an overly generous definition of side effect is a device to insulate ourselves from blame for harm.

3 Sourcing an independent definition of side effect from medicine and law

In common speech what do we mean by the term side effect? Firstly to disambiguate, I should say that a concept of side effect does exist in programming and it has a formal definition fit for its intended purpose. A hazard with terms that have domain specific meanings is to assume that those meanings are shared outside the discipline. Instead we want to take the idea of side effect that exists outside computer science and bring it into the discipline in a process that does not alter it. As with most primitive or folklore concepts, people intuitively know what a side effect is, but pinning down a decent definition of one takes some effort. The benefit of such an endeavour is twofold. It aids cross-disciplinary communication for when a regulator and a computer scientist discuss side effects it is preferable that both mean the same thing. From a programming perspective, a formal definition of side effects written in such a way as an algorithm would be able to understand, can prevent algorithms from causing harm.

The most common place that people see the term side effects is in a medical setting, so it is intuitive to start the process of definition here. It is also an appropriate source given that medicines and the medical profession are strictly regulated. The APA (American Psychological Association) defines a side effect as follows (APA 2021):

> *Any reaction secondary to the intended therapeutic effect that may occur following administration of a drug or other treatment*

This definition makes the distinction between effects which are intended and those which are not, with side effects appearing in the latter class. The term secondary requires further unpacking which we will do once we have introduced some causal mechanisms.

The concepts of intentionality and foreseeability are commonly used in the legal world and it is from here that we will source their definitions. By looking to the law for a definition of intent we can borrow centuries of legal thought and endeavour. One can consider legal definitions as open-source in the sense that they are accessible to public scrutiny and have been democratically tested over time. As Hildebrandt (2019) states, legal questions enjoy closure, that is to say, within any jurisdiction, definitions and questions have answers.

Despite its key role in Criminal law amongst others [1], for various reasons a singular definition of what constitutes intent is elusive. We will use the US Model Penal Code (MPC) (The American Law Insitute 2017) which does provides definitions of the four levels of mens-rea or criminal culpability; Purpose (aka Intent), Knowledge, Recklessness and Negligence. The MPC was drafted in the 1960s in an effort to unite US state law and has been adopted at least partially by most states since. These definitions also invoke the term foreseeability; importing definitions of intent from law also implicitly brings conventions concerning foreseeabilty. The MPC defines Purpose[2] (Intent) as follows:

> *A person acts purposely with respect to a material element of an offense when... if the element involves the nature of his conduct or a result thereof, it is his conscious object to engage in conduct of that nature or to cause such a result*

In essence this definition says that someone intends something if it is the object outcome of their actions. This definition largely corresponds to the folk-definition of intent. Agent Ag intends X if they choose to do action ψ in order to cause X. This implies an epistemic condition on the outcome; Ag can foresee that X could be an outcome of them ψ-ing. However this definition does not mention probability of outcome so it follows that long-shot type outcomes can be intended.

At this point we will introduce some causal modelling terminology to illustrate more easily things which are intended and things which could be called side effects.

4 Causal Modelling Approach

Consider a directed acyclic graph G with vertices V and edges E, for $A, B \in V$ we will adopt the convention that A is a cause of B iff there is a directed edge in E from A to B. That is to say, all other things constant, a change in A will imply some change in B. We shall call this a Causal DAG.

[1] Intent appears in almost aspect of law - contract, tort, regulatory. Sometimes in obvious ways, sometimes not

[2] When drafting the MPC, the authors took the approach that defining concepts such as intent could be better done by not mentioning them so as not to conflate with possibly wrong folk concepts of the word.

We will use the Structural Causal Influence Models (SCIM) (Everitt et al. 2021) to make the Causal Dag more applicable to intent. An influence diagram named ID is causal DAG such that the vertices are divided into disjunct groups - Decision vertices V_D (represented with rectangles), outcome vertices V_O (represented by circles) and utility vertices V_U (octagons or diamonds) with $V_D \cup V_0 \cup V_U = V$. Let $\mathcal{R}(Y)$ denote the full set of realisations that vertex Y can take. Structural equations determine the relationship between parent outcome or decision vertices and child outcome or utility vertices. Thus the structural equation associated with arc AB for $A \in V_O,\ B \in V$ is a function $f_{AB} : \mathcal{R}(A) \rightarrow \mathcal{R}(B)$. A policy π_{ID} is a set of structural equations with decision vertices as children such that the parents of any decision vertex D denoted $Pa(D)$ determine a distribution over $\mathcal{R}(D)$; the possible realisations of D. A non-deterministic policy would apply unit mass to single element of $\mathcal{R}_D$ for every possible realisation of $Pa(D)$.

Additionally without any loss of generality we can enforce the restriction that all stochastic elements of the ID do not have a parent. Practically this just means the rewriting of non-deterministic structural equations to separate deterministic variables from (possibly new) non-deterministic variables which themselves have no parents. This ensures that every vertex that is a descendent of a decision vertex is deterministic and the ID is said to be in Howard Canonical Form (Heckerman and Shachter 1994). Once the policy is set and the non-deterministic variables are set, all other variables are uniquely realised. This form also allows us to interpet the SCIMas a Structural Causal Model (SCM) and use the accompanying definitions of Actual-causality (Halpern 2016) and Do-algebra (Pearl 2000) should we wish.

5 Basic properties of Intent and Side Effects in Causal Models

Within the framework of SCIMs and related Causal Analysis, a number of definitions of intent might exist which are arguably compatible with the MPC's definition of intent or Purpose but might require assumptions about the agent. For example Kleiman-Weiner et al. (2015) present an account of intent using Influence Diagrams which assumes a utility maximising agent. Ashton (2021a) does not make the assumption but presents a definition of intent without a formal causal framework. For this article I will assume intended outcomes are given. This should not be problematical for a system designer, since it is good practice to identify what the intended purpose of an algorithm is before creating it. For clarity we will use this very general definition of Intended realisations (outcomes) and Intended Variables:

Definition 1 (Intended Realisation, Intended Variables)**.** For an Influence Diagram ID, an intended realisation is a finite set of realisations for outcome and utility variables under a fixed policy π. The intended variables are those variables which occur have an intended realisation

Whilst for the purposes of identifying side effects we can simply assume that it is known which things are intended and which are not but we must still ensure intentional knowledge is consistent with the legal concept of intent. For this we need to make three assumptions.

1. An outcome can only be intended if its realisation is dependent on an action realising a certain value.
2. Actions made by an algorithm are done so intentionally only if their is choice not to act in that way (the action set has more than one member)
3. The concept of means-end consistent intent is respected.

Condition one just says that an outcome can only be intended if a decision variable is able to affect the it. We cannot intend our football team to win by attending the match as a spectator. This rules out any parental chance or outcome variables from being intended and stops any foregone conclusions from being intended. Condition two says that the action-decisions that a decision maker makes are not coerced (there is always a choice of policy design). Condition three requires explanation. There exists in law (Simester et al. 2019) and philosophy (Bratman 2009) a concept called means-end intent which is culpably equivalent to intent or MPC Purpose.

Suppose we know outcome $O = o$ for $O \in V_O$ is intended in an influence diagram. Then for any outcome variable $O' \in Anc(O) \cap V_o$, if it is necessary for $O' = o'$ for the intended outcome $O = o$ to occur then $O' = o$ is also an intended outcome.

Definition 2 (Side Effect and Unintended outcome)**.** Consider an influence diagram SCIMwith outcome vertices V_O, decision vertices V_D and utility vertices V_U, a policy function π_{ID} and a set of intended variables $V_E \subset V$ and for each intended variable an intended realisation $x \in \mathcal{R}_X$ for every $x \in V_E$.

- A **Side Effect Variable of a Decision** is any descendent variable of that decision vertex which is not an intended variable. A **Side effect of a Decision** is a realisation of a Side Effect Variable of a decision.
- The **Policy Side Effects** and **Policy Side Effects Variables** are analogously defined for sets of intended realisations and variables according to some policy.
- An **Unintended outcome** is a realisation of an intended variable other than the intended realisation

The definition requires that side effects are caused by a policy since they are descendants of action variables in a causal DAG; the policy has an influence on their outcome. We can restate Definition 2 by saying Side Effects of a decision are those vertices which are descendants of the decision but are not themselves ancestors of any vertices with intended realisations according to that decision. This definition incorporates the idea of means-end consistency into it assumed in Definition 1.

The definition of Unintended outcomes separates the analysis of when a certain policy fails in achieving its objective (chance of failure) from the analysis of side effects whose occurrence is not dependent a policy's success. The legality or culpability of unintended outcomes is not straightforward. A doctor performing a life saving treatment which will either result in the death of their patient or their saviour is not usually punished in the event of failure even though the chances of success might be exceptionally slim. On the other hand a

fund-manager who loses all of their investor's money on a risky bet might be punished for their reckless actions.

6 Illustrative example

In this section we will explore some of the previously discussed features of side effects and intent using the example of a company wishing to deploy a new recommender system on its users. I recognise this is a departure from issues of robot induced vase breaking as is typically considered in related side-effect literature, nevertheless I think this setting is pertinent. Article 5(1)(a-b) of The Draft EU AI Act (CNECT 2021) prohibits AI systems from using techniques which manipulate the behaviour of users to their detriment with techniques beyond their consciousness.

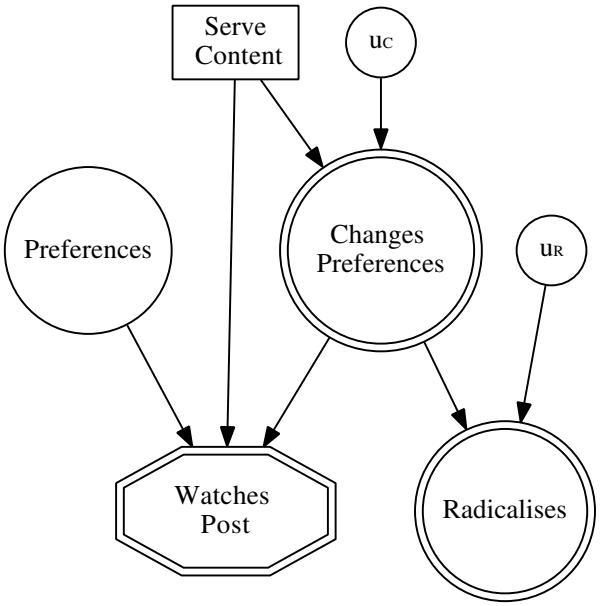

Figure 1: Recommender system example

Suppose a company has the choice between serving two items of video content, one normal and one which has been designed to intoxicate its viewer into watching all of the video. If the user views this intoxicating content there is a chance that they become radicalised in some way. Hidden from the company are the user's preferences which dictate what type of content they would want to watch, preferences undisturbed.

- Let Serve Content decision be represented by variable $S \in \{0, 1\}$
- Let the Change Preference outcome be represented by variable $C \in \{0, 1\}$
- Let the Radicalised outcome be represented by variable $R \in \{0, 1\}$
- Let Watches Post utility vertex be represented by variable $W \in \{0, 1\}$
- Let the user's Preferences be represented by variable $P \in \{0, 1\}$. This is an exogenous random variable unknown at the time of decision S. Let P be Bernoulli distributed with chance of success $0 \leq \mu_P \leq 1$
- U_C and U_R are two exogenous random variables which help determine the chance of Preference change and Radicalisation respectively. Let them both be Bernoulli distributed with probabilities of success $\mu_C, \mu_R \in [0, 1]$ respectively.

The structural equations are as follows:

$$R = \begin{cases} 1 \text{ if } C = 1 \text{ and } U_r = 1 \\ 0 \text{ else} \end{cases}$$

$$C = \begin{cases} 1 \text{ if } S = 1 \text{ and } U_S = 1 \\ 0 \text{ else} \end{cases}$$

$$W = \begin{cases} 1 \text{ if } S = P \text{ or } \ C = 1 \\ 0 \text{ else} \end{cases}$$

Content can be one of two types as can a user's preferences; $p(P = 1) = \mu_P$. If the two types match then the user will watch the post and the recommender is rewarded with a unit of advertising revenue. Additionally if the content is of the intoxicating type $(S = 1)$ then there is chance μ_C that the user will have their preferences changed $C = 1$. If that is the case then they will definitely watch the post. Additionally there is chance the user will be radicalised with probability μ_R

Example 1. Let $W = 1$ be the intended outcome under the Policy $S = 1$. The first question is whether C must have intended realisation under means-end consistency. As long as $\mu_P > 0$ (the chance that the user would actually like content type 1) $W = 1$ is over-determined; either $P = 1$ or $C = 1$ is sufficient for $W = 1$. C could be be both an intended variable and not.

Suppose $C = 1$ is the intended realisation. The set of intended variables is $\{S, C, W\}$ and the set of possible side effects is $\{R\}$ since there are no other descendant.

Alternatively suppose C has no intentional status. The set of intended variables is $\{S, W\}$ and the set of possible side effects is $\{C, R\}$.

Kleiman-Weiner et al. (2015) and Halpern and Kleiman-Weiner (2018) use a counterfactual type test for intent. If we set variables to their expected reward maximising realisations, and then swap the realisation of C and adjust the rewards of its descendants accordingly, would the policy change? In other words, is the policy dependent on the causal relationship between S and C? The expected reward from choosing S=1 is $\mu_P + (1 - \mu_P)(\mu_C)$ and the expected reward from choosing S=0 is $(1 - \mu_P)$. Assuming a reward maximising agent this implies $\mu_c > \frac{1-2\mu_P}{1-\mu_P}$ Breaking the causal link between S and C and setting $C = 0$ whilst continuing to choose $S = 1$ would give an expected reward of μ_P - the strategy would change if $\mu_p < 0.5$. In the next section we will look at the possible culpability classification of the side effects in this example.

7 Culpability of side effects: The role of knowledge

The subject of culpability as to side effects of actions has received a lot of attention in Psychology since Knobe (2003)

discovered the Side-effect effect, the phenomenon where people are judged to have intended negative side effects which they foreseeably cause but not for any positive side effects that they cause.

Legal systems have a lot to say about side effects and culpability[3] and provide us with an independent framework to reason about them. Conversely, courts currently have very little to say about harms caused by algorithms and who should take responsibility for them. Only legal persons can commit crimes and so harms caused by algorithm might have an indeterminate status Abbott and Sarch (2020). In this section we will refer to the agent and actor and take that to mean the algorithm, algorithm designer and owner together. We will assume knowledge available to one is available to the other.

Just as with our use of the MPC to find a definition of Purpose (intent), we can use its definitions of Knowledge, Recklessness and Negligence to analyse the culpability of side effects. These four concepts are in descending order of culpability. Recklessness is typically the minimum level of culpability required for criminal charges. Negligence is the benchmark required for most civil-damages cases. The key features of these definitions are summarised in Table 1. The table also includes the features of Intent or Purpose for ease of comparison.

All four definitions of culpability in the table require someone to be able to foresee a bad outcome occurring as a result of an action or policy. It is here that the MPC's decision to define the second most serious level of culpability *'Knowledge'* problematic since the word is useful to describe the epistemic properties of all of the definitions. We will refer to this as 'Culpable Knowledge' to disambiguate.

The table shows aim or desire is only required for Purpose/Intent which is consistent with our prior definition of side effects being. It makes a distinction between two types of knowledge - subjective knowledge - things which are known to the actor and objective - things which *should* be known to the actor. We can view the Influence Diagram as a distillation of the actor's causal knowledge of the world. A side effect caused with Culpable Knowledge concerns the case when a bad outcome occurs with almost certainty according to the SCIMand the algorithm's policy. Recklessness and Negligence have been termed culpable carelessness by (Stark 2017); they correspond to cases of model misspecification and require a judgement about what a reasonable actor should have had as a model of the world. In the case of Recklessness, the actor recognised some chance of a bad outcome happening but continued anyway. If the actual chance of that thing happening was unreasonably high and that was knowable to an external reasonable actor, then the side effect was caused with Recklessness. Negligence covers the case where the algorithm didn't even countenance the risk of something bad happening, and the risk was in foreseeably unreasonable according to some external reasonable actor. Side effects caused with negligence are likely to involve variables not included in the algorithm or its designers' model of the world. These *should've known unknowns* are particularly dangerous since no planning algorithm can avoid them and yet they will not be viewed as accidents by society and admit liability to the algorithm owner.

Whilst this table is focussed on the culpability of side effects as previously defined, Recklessness and Negligence can also apply to *Unintended Outcomes*. That is to say outcomes that may be caused when failing to achieve an intended outcome. In the case of Knowledge there is debate about whether someone can commit something with Culpable Knowledge if their action was intended to obtain some other result.

Figure 2 distils Table 1 into a decision process with which to identify the possible culpability of any caused outcomes. The grey decision vertices concern questions of subjective knowledge - information known to the actor at the point of commission. The white vertices concern questions of objective knowledge - information that should have been known to the actor at the point of commission. We will use it in the following example.

Example 2. Continuing Example 1 we consider the culpability of the possible side effects caused. Since the outcome of getting a user to watch a video of type 1 was overdetermined, there was some uncertainty about the intentional state of changing preferences - C by choosing action $S = 1$. Since causing a change in preferences is not in itself an unambiguous harm, we will concentrate analysis on the Radicalisation outcome R which is in the set of side effects regardless of the intentional status of preference change. Given that content of type one is chosen, the probability of radicalisation is $\mu_p.\mu_R$. R is not an intended variable and has no intended realisation. Consider the case where Radicalisation does occur. We will assume that the actor has the same model available to them as in Figure 1. The first question would be to consider whether harm is foreseeable, and whether that harm was 'substantial and unjustifiable', that is whether the $\mu_p.\mu_R >> 0$. If this wasn't the case, then the Radicalisation could be said to be an accident. At this point, the estimated quantity $E[\mu_P * \mu_R]$ must be assessed. If the actor estimated this as negligible then Radicalisation would have been caused with negligence. If the estimate $E[\mu_P * \mu_R] \approx 1$ then Radicalisation would have been caused with Culpable Knowledge. The remaining case, $E[\mu_P * \mu_R] >> 0$ means that the actor caused Radicalisation with Recklessness. It should be stated that the definition of 'substantial and unjustifiable' is not straightforward and may be dependent on the degree of harm caused (Stark 2017).

In the case where harm has not been caused, culpability can still arise if the actor were to believe their actions were risky and they were substantially so. For a given crime, it does not always follow that there is an analogous crime of reckless endangerment unlike the general existence of attempt crimes for every crime. Stark (2020) differentiates between two situations; firstly where there was a risk of some harm occurring "Concrete Endangerment" and secondly where there was in actual fact no risk of endangerment, but there could have been.

[3]Else the *'I didn't mean to shoot him your honour, I was intending to shoot the pigeon behind him'* defence would work really well.

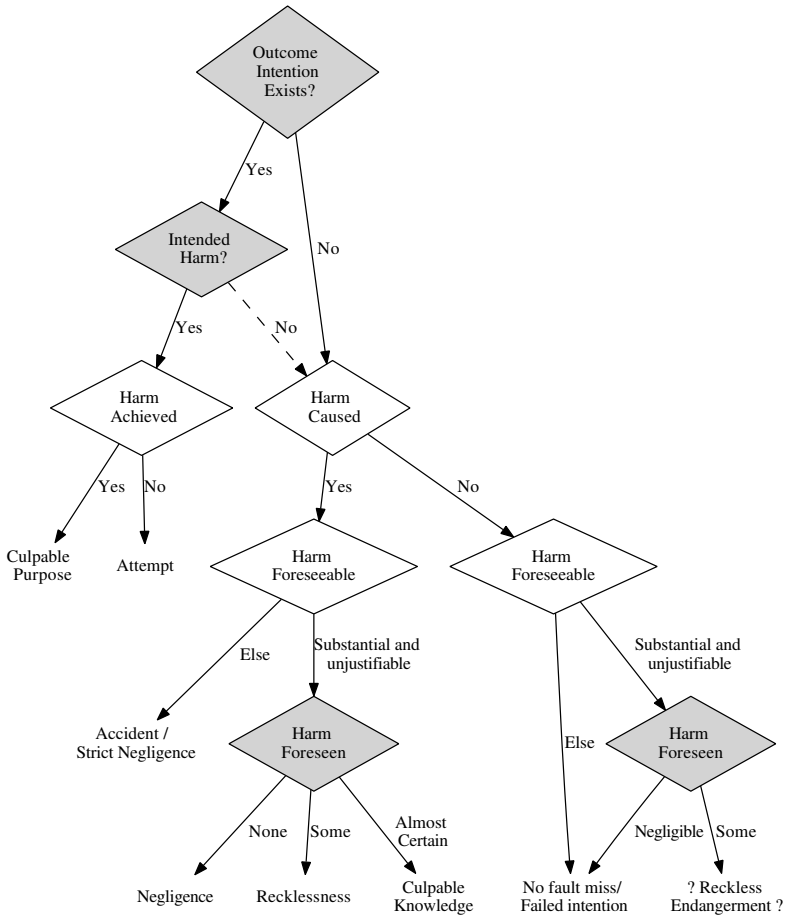

Figure 2: Decision tree to decide the Culpability of Outcome variables. Questions of subjective knowledge are in grey diamonds whilst questions of objective knowledge are in white diamonds

As well as applying negligence to instances when an actor *should* have known about the chance of some harm being caused, US and other Common Law jurisdictions allow higher levels of culpability to be imputed in the case when an agent actively avoids acquiring knowledge in an effort not to inculpate their behaviour. This is known as the Willful Ignorance Doctrine (Sarch 2019) and allows culpable knowledge to be ascribed to an actor in the event of harm.

8 Related work

The following review of related work is divided between subject area.

Side Effects in AI Whilst I argued in that the definition of side effect in Amodei et al. (2016) captures many more things than side-effects, the putative solutions presented by the authors and in descendent research are nevertheless useful. One general approach is that of minimising impact; an AI should complete its task as best its can whilst exerting as small an impact on the world as possible, this is suggested in Amodei et al and formalised in (Armstrong and Levinstein 2017). A related approach is requiring the reversibility of all effects caused by a policy (Eysenbach et al. 2017), the assumption being that effects that are reversible are less harmful. Here an agent learns both policies to achieve things and policies to undo those things. A value function derived from the latter can be used to guide the former. This approach is not suitable for all tasks, since agents will be required to perform actions with permanent impact. Krakovna et al. (2020) approach the problem differently by asking the AI agent to consider completion of future tasks (expressed as a uniform distribution over all possible goal states) as well as the present one in an environment that is not reset after a task is completed. The authors show that this is a generalisation of the reversible approach. Turner, Ratzlaff, and Tadepalli (2020) develop a related method called Attainable Utility Preservation (AUP), which penalises policies that prevent the maximisation of a true, complex, yet unseen reward function, which encodes our preferences regarding bad side effects not occurring.

In a recent review of the subject in AI, Saisubramanian, Zilberstein, and Kamar (2021) state that Negative Side Effects (NSE) are *"Undesired effects of an agents actions that occur in addition to the agent's intended effects"*. This is close in spirit to the APA definition in Section 3, though

Feature	Condition	Intended	Side Effects committed with		
		Purpose	Knowledge	Recklessness	Negligence
Aim/Desire	Outcome is desired/an aim	Yes	N/A	N/A	N/A
Subjective knowledge	Outcome is foreseeable to principle	Yes	Yes	Yes	N/A
	Probability of outcome according to principle	Non-zero	Almost certain	Non-zero	N/A
Objective knowledge	Outcome is foreseeable to 'Reasonable' actor	N/A	Yes	Yes	Yes
	Probability of outcome according to Reasonable actor	N/A	Almost certain	Unreasonable	Unreasonable

Table 1: Side Effects culpability characteristics table

there is no subsequent clarification as to how intent is assessed. The authors present a taxonomy in which to classify recent side-effect mitigation techniques consisting of the following

- **Severity** - Serious harms are more obvious and likely to be designed out at an early stage. Saisubramanian et al observe that less serious harms are the ones that manifest which they show in related work reduce confidence in AI systems (Saisubramanian, Roberts, and Zilberstein 2021)
- **Reversibility** As discussed in the previous paragraph, some effects are permanent.
- **Avoidability** Related to our discussion of means-end effects, some effects are required for an objective to be fulfilled. These are therefore intended and cannot be side-effects.
- **Frequency** The occurrence of side-effects might be generally uncommon but common in a certain situation. For examples in the medical domain see Leslie et al. (2021).
- **Stochasticity** In our causal setting, some side effects might have a stochastic parent meaning that their occurrence is not purely a function of the agents actions. Nearly all of the methods they survey assume deterministic side-effects.
- **Observability** Side effects might not be fully observable according to the agent. Even if they are, they might not be reflected in the agent's reward function as penalties.

An alternative to the reward (or penalty) based approaches so far mentioned is constrained optimisation (Achiam et al. 2017), that is to say policy search within a 'safe' subset of policies. Zhang, Durfee, and Singh (2020) consider a scenario where a robot agent is given a task but is unsure about the various side-effects that may occur as a result of various strategies that satisfy the task. The agent partitions its state variables into 'free-features' it knows it can change, 'locked-features' it knows it should not change and 'unknown-features' it is unsure about (yet to be classified). The agent will proceed to complete a task affecting only free-features using linear programming, but failing this will sparingly query an oracle as to the status of an unknown-feature. An interesting advantage of such an approach is that side-effects not previously considered by the oracle can be safely negotiated. More generally side-effects might occur because of sequences or combinations of states and actions.

Many approaches to the side-effect problem assume that they are a result of underspecified reward functions or non-observability. It could be that even with an adequate reward function and state space, undesirable and unnecessary side-effects are still incurred due to mis-inference on the part of the algorithm. It is tempting to believe that causal reasoning is not needed when using Reinforcement Learning. Often this is because expert knowledge about the data generation process has been embedded into a simulation environment (Hernán, Hsu, and Healy 2019), which can be extracted by the learner through exploration.

Side Effects in Philosophy and Psychology Research considering the moral judgement of side effects in experimental psychology has been popular since (Knobe 2003) which first identified the Side-effect (or Knobe) effect, whereby people consistently rate negative side effects as more intentional than those with positive side effects. This has since been shown to be the case with related judgements of causality and blame amongst others. See Feltz (2007) and Kneer and Bourgeois-Gironde (2017) for an overview. Given the overwhelming about of research written about the effect, it is surprising that the concept of side effect hasn't been more formally identified. The finding that certain side effects are deemed intentional is consistent with the definitions of culpable mental states found in common law and considered in this article.

Formal accounts of intent are not hugely common. The aforementioned, (Kleiman-Weiner et al. 2015) define intent in Influence diagrams and (Halpern and Kleiman-Weiner 2018) define intent using a modified Structural Causal Model which includes agent utility. In both cases, an outcome is intended if the agent's policy is counterfactually dependent on it. (Ashton 2021b) extends both models to consider *Oblique Intent*, which is similar to Culpable Knowledge. The Belief Desire Intent (BDI) model of multi-agent programming originated from the Theoretical work of Philosopher Michael Bratman in the 1980s (Bratman 1999). Cohen and Levesque (1990) present a formal temporal logic incorporating intent and the spirit of Bratman's work. Recently in the field of causal cognition Quillien and German (2021) have defined and tested intent as the degree to which someone's desire caused something to happen.

Side effects in law Law prioritises the establishment of the various levels of culpability or mens-rea which make the concept of side effect redundant. At the levels of culpable carelessness, it isn't so concerned whether an adverse outcome was a side effect or an unintended outcome. Bratman's intuition that means-end intent should be equivalent to purpose, expressed in Bratman (2009) as means-end coherence, is supported by Simester et al. (2019) who quotes the case of Smith [1960] 2 QB 423 (CA), where the Defendant was accused of bribing public official. D says that they only intended to expose public corruption, but the court found

that he necessarily meant to bribe the mayor as a necessary part of his plan. An argument that the Law reflects the folk-attribution of blame to foreseen negative side effects can be made by the presence of Culpable Purpose whose definition does not contain any reference to desire, aim or purpose. The law is also interested in cases when intended outcomes did not realise when the intended outcome is prohibited - attempts to commit crimes are prohibited[4]. A special case of failed attempts concerns actions intended to do something good but which end up causing some harm. These follow the dotted arc in Figure 2. In such cases culpability might be waived if the intended outcome of the actor was to opposite to the actual cause - a surgeon performing life-saving treatment might know that the chance of death is almost certain but continue anyway. A related issue is the doctrine of double effect (McIntyre 2019) which can prevent culpability of intermediate or successive harmful outcomes as long as the 'primary' intended outcome is morally sound. These are complex issues and do not fall neatly into the culpability decision rules presented here.

9 Conclusion

This article presents a formal definition of what constitutes side effects sourcing the definition of side effect from medicine and the necessary definition of intent from law. It does this with the use of a Structural Causal Incentive Model or SCIM, itself an extension of a Structural Causal Model (SCM) and an exogenous definition of intent. Side Effects of an action are those vertices which are descendants of the action but are not themselves ancestors of any vertices with intended realisations.

Any definition of side effects taken up by the Safe-AI community should be based on principles agreed by society rather than computer scientists. Such an approach defuses the accusation that definitions of concepts important to society are created to be convenient or progress the objectives of the engineer or their employer.

Despite their name, side effects can still attract severe criminal liability. I have used the standard MPC definitions of culpability to create a decision process which can be used to systematically determine culpability for harms caused or potentially caused. In the event that an algorithm causes some tangible harm to the world through a side effect, computer scientists are not going to be surprised about the conditions under which a side effect might make them liable in a civil or criminal sense.

Algorithms or their designers when equipped with model of the world can use such a definition and decision process to discriminate between directly intended outcomes and side effects and then identify what degree of culpability can be attached to side effects caused or endangered. In particular, the article should impress upon the reader the importance of foresight knowledge and its relation to culpability. Harms can be caused accidentally with no liability, but after their first instance they become foreseeable, at which point if they reoccur they can no longer be termed accidents and causing them becomes a culpable action.

[4] Attempts can be divided between those that are interrupted before commission after passing some threshold of culpable preparation and those that attempts which are completed but fail in their aim. We are talking about the latter here

A Appendix A: MPC Culpability

The following definitions of culpability are taken from the MPC (The American Law Insitute 2017). The fourth level Purpose (or direct intent) is quoted within the text.

Knowledge

A person acts knowingly with respect to a material element of an offense when: (i) if the element involves the nature of his conduct or the attendant circumstances, he is aware that his conduct is of that nature or that such circumstances exist; and (ii) if the element involves a result of his conduct, he is aware that it is practically certain that his conduct will cause such a result.

Recklessness

A person acts recklessly with respect to a material element of an offense when he consciously disregards a substantial and unjustifiable risk that the material element exists or will result from his conduct. The risk must be of such a nature and degree that, considering the nature and purpose of the actor's conduct and the circumstances known to him, its disregard involves a gross deviation from the standard of conduct that a law-abiding person would observe in the actor's situation.

Negligence

A person acts negligently with respect to a material element of an offense when he should be aware of a substantial and unjustifiable risk that the material element exists or will result from his conduct. The risk must be of such a nature and degree that the actor's failure to perceive it, considering the nature and purpose of his conduct and the circumstances known to him, involves a gross deviation from the standard of care that a reasonable person would observe in the actor's situation.

References

Abbott, R.; and Sarch, A. 2020. Punishing Artificial Intelligence: Legal Fiction or Science Fiction. *Is Law Computable?*, 323–384.

Achiam, J.; Held, D.; Tamar, A.; and Abbeel, P. 2017. Constrained Policy Optimization. *arXiv:1705.10528 [cs]*.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete Problems in AI Safety. *arXiv:1606.06565 [cs]*.

APA. 2021. Side Effect. Date accessed: 29/12/2021 https://dictionary.apa.org/side-effect.

Armstrong, S.; and Levinstein, B. 2017. Low Impact Artificial Intelligences. *arXiv:1705.10720 [cs]*.

Ashton, H. 2021a. Definitions of intent suitable for algorithms. *arXiv:2106.04235 [cs.AI]*.

Ashton, H. 2021b. Extending counterfactual accounts of intent to include oblique intent. *arXiv:2106.03684 [cs.AI]*.

Ashton, H.; and Franklin, M. 2022. The problem of behaviour and preference manipulation in AI systems. In *The AAAI-22 Workshop on Artificial Intelligence Safety (SafeAI 2022)*.

Bratman, M. 1999. *Intention, plans, and practical reason*. David Hume series. Stanford, Calif: Center for the Study of Language and Information.

Bratman, M. E. 2009. Intention, Practical Rationality, and Self-Governance. *Ethics*, 119(April): 411–443.

CNECT. 2021. Proposal for a regulation of the European parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts. Technical Report COM/2021/206, European Commission, Directorate-General for Communications Networks, Content and Technology.

Cohen, P. R.; and Levesque, H. J. 1990. Intention is choice with commitment. *Artificial Intelligence*, 42(2-3): 213–261.

Everitt, T.; Carey, R.; Langlois, E.; Ortega, P. A.; and Legg, S. 2021. Agent Incentives: A Causal Perspective. *arXiv:2102.01685*.

Eysenbach, B.; Gu, S.; Ibarz, J.; and Levine, S. 2017. Leave no Trace: Learning to Reset for Safe and Autonomous Reinforcement Learning. *arXiv:1711.06782 [cs]*.

Feltz, A. 2007. The Knobe Effect: A Brief Overview. *The Journal of Mind and Behavior*, 28(3/4).

Halpern, J. Y. 2016. *Actual causality*. Cambridge, Massachusetts: The MIT Press.

Halpern, J. Y.; and Kleiman-Weiner, M. 2018. Towards formal definitions of blameworthiness, intention, and moral responsibility. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 1853–1860.

Heckerman, D.; and Shachter, R. 1994. A Decision-Based View of Causality. *Uncertainty Proceedings 1994*, 302–310.

Hernán, M. A.; Hsu, J.; and Healy, B. 2019. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. *CHANCE*, 32(1): 42–49.

Hildebrandt, M. 2019. Closure: on ethics, code and law. In *Law for Computer Scientists*, chapter 11. Oxford University Press.

Kleiman-Weiner, M.; Gerstenberg, T.; Levine, S.; and Tenenbaum, J. B. 2015. Inference of intention and permissibility in moral decision making. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, 1(1987): 1123–1128.

Kneer, M.; and Bourgeois-Gironde, S. 2017. Mens rea ascription, expertise and outcome effects: Professional judges surveyed. *Cognition*, 169(August): 139–146.

Knobe, J. 2003. Intentional action and side effects in ordinary language. *Analysis*, 63: 190–194.

Krakovna, V.; Orseau, L.; Ngo, R.; Martic, M.; and Legg, S. 2020. Avoiding Side Effects By Considering Future Tasks. *Advances in Neural Information Processing Systems (NeuIPS)*, 33.

Leslie, D.; Mazumder, A.; Peppin, A.; Wolters, M. K.; and Hagerty, A. 2021. Does "AI" stand for augmenting inequality in the era of covid-19 healthcare? *BMJ*.

McIntyre, A. 2019. Doctrine of Double Effect. In Zalta, E. N., ed., *Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 201 edition.

Pearl, J. 2000. *Causality: Models, reasoning and inference*. Cambridge University Press.

Quillien, T.; and German, T. C. 2021. A simple definition of 'intentionally'. *Cognition*, 214(June).

Russell, S. J. 2019. *Human compatible: artificial intelligence and the problem of control*. London: Allen Lane, an imprint of Penguin Books.

Saisubramanian, S.; Roberts, S. C.; and Zilberstein, S. 2021. Understanding User Attitudes Towards Negative Side Effects of AI Systems. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–6. Yokohama Japan: ACM. ISBN 978-1-4503-8095-9.

Saisubramanian, S.; Zilberstein, S.; and Kamar, E. 2021. Avoiding Negative Side Effects due to Incomplete Knowledge of AI Systems. *arXiv:2008.12146 [cs]*.

Sarch, A. 2019. Criminal Law Basics and the Willful Ignorance Doctrine. In *Criminally Ignorant*, 7–26. Oxford University Press.

Simester, P.; Spencer, J. R.; Stark, F.; Sullivan, G. R.; and Virgo, G. J. 2019. Mens Rea. In *Simester and Sullivan's Criminal Law*, chapter 5, 137–190. Hart, 7 edition.

Stark, F. 2017. Introduction. In *Culpable Carelessness: Recklessness and Negligence in the Criminal Law*, chapter 1, 1–25. Cambridge University Press.

Stark, F. 2020. The Reasonableness in Recklessness. *Criminal Law and Philosophy*, 14(1): 9–29.

The American Law Insitute. 2017. General Requirements of Culpability.

Turner, A. M.; Ratzlaff, N.; and Tadepalli, P. 2020. Avoiding Side Effects in Complex Environments. *arXiv:2006.06547 [cs]*.

Zhang, S.; Durfee, E.; and Singh, S. 2020. Querying to Find a Safe Policy under Uncertain Safety Constraints in Markov Decision Processes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03): 2552–2559.