Appendices

A

In this appendix we present the general version of Definition 3 allowing harm and benefit to be measured along specific causal paths.

The path-specific counterfactual harm measures the harm caused by an action $A = a$ compared to a default action $A = \bar{a}$ when, rather than generating the counterfactual outcome by including all causal paths from $A = \bar{a}$ to outcome variables Y, we consider only the effect along certain paths g. This is somewhat analogous to the path specific causal effect [5], as we are using the g-specific intervention $A = \bar{a}$ on Y in the counterfactual world relative to reference $A = a$ (the factual action).

Definition 9 (Path-specific counterfactual harm & benefit)**.** *Let G be the DAG associated with model $\mathcal{M}$ and g be the edge sub-graph of G containing the paths we include in the harm analysis. The path specific harm caused by action $A = a$ compared to default action $A = \bar{a}$ is given by*

$$h_g(a, x, y; \mathcal{M}) = \int_{y^*} P(Y_{\bar{a},\mathcal{M}_g} = y^* | a, x, y; \mathcal{M}) \max\{0, U(\bar{a}, x, y^*) - U(a, x, y)\} \tag{12}$$

$$= \int_{y^*, e} P(Y_{\bar{a}} = y^* | e; \mathcal{M}_g) P(e | a, x, y; \mathcal{M}) \max\{0, U(\bar{a}, x, y^*) - U(a, x, y)\} \tag{13}$$

Where $Y_{\bar{a},\mathcal{M}_g}$ is the counterfactual outcome Y under intervention $do(A = \bar{a})$ in model $\mathcal{M}_g$ where $\mathcal{M}_g$ is formed from $\mathcal{M}$ by replacing the causal mechanisms for each variable $f^i(pa^i, e) \to f^i_g(pa^i(g)^*, e) = f^i(pa^i(g)^*, pa^i(\bar{g}), e)$, where $Pa^i(\bar{g})$ is the set of parents of $V^{(i)}$ that are not linked to $V^{(i)}$ in g and $pa^i(\bar{g})$ is the factual state of those variables. $E = e$ is the joint state of the exogenous noise variables in $\mathcal{M}$. Likewise, the expected benefit is

$$b_g(a, x, y; \mathcal{M}) = \int_{y^*} P(Y_{\bar{a},\mathcal{M}_g} = y^* | a, x, y; \mathcal{M}) \max\{0, U(a, x, y) - U(\bar{a}, x, y^*)\} \tag{14}$$

Note that if we following the construction of $\mathcal{M}_g$ in [5] we get that $\mathcal{M}_g$ is formed from $\mathcal{M}$ by i) partitioning the parent set for each variable $V^{(i)}$ in $\mathcal{M}$ into $\text{Pa}^i = \{\text{Pa}^i(g), \text{Pa}^i(\bar{g})\}$ where $\text{Pa}^i(g)$ are the parents that are linked to $V^{(i)}$ in g and $\text{Pa}^i(\bar{g})$ is the complimentary set, ii) replacing the mechanisms for each variable with $f^i(\text{pa}^i, e^i) \to f^i_g(\text{pa}^i, e^i) = f^i(\text{pa}^i(g)^*, \text{pa}^i(\bar{g}), e^i)$ where $\text{pa}^i(\bar{g})$ takes the value of $PA^i(\bar{g})_z$ in $\mathcal{M}$ where $A = z$ is the reference action. However, in (12) and (14) we condition on the state of all factual variables and assume no unobserved confounders, and the reference action is the factual action state. Therefore the state of $PA^i(\bar{g})_a$ in $\mathcal{M}$ is equal to the factual state of these variables, giving our simplified construction for $\mathcal{M}_g$.

We give examples of computing the path-specific harm in Appendices B-C.

B

In this appendix we discuss the omission problem and pre-emption problem [13], and the preventing worse problem [15], and show how these can be resolved using our definition of counterfactual harm (Definition 3 and its path-specific variant Definition 9).

Omission Problem: Alice decides not to give Bob a set of golf clubs. Bob would be happy if Alice had given him the golf clubs. Therefore, according to the CCA, Alice's decision not to give Bob the clubs causes Bob harm. However, intuitively Alice has not harmed Bob, but merely failed to benefit him [13].

Solution: The omission problem relies on the judgement that Alice does not have a ethical obligation to provide Bob with golf clubs, therefore her choice not to do so does not constitute harm to Bob. In our definition of harm, this judgement is encoded by Alice not giving Bob clubs by default, i.e. the desired harm query is the harm 'compared to the world where Alice does not give Bob clubs'. To compute the harm we construct the model $\mathcal{M}$ comprising of two variables; Alice's action $A \in \{0, 1\}$

where $A = 0$ indicates 'Bob not given clubs' and $A = 1$ 'Bob given clubs', and outcome $Y \in \{0, 1\}$ where $Y = 1$ indicates 'Bob has clubs' and $Y = 0$ indicates 'Bob does not have clubs'. By default, Alice is not expected to give Bob clubs, which is encoded by choosing the default action $A = \bar{a}$ where $\bar{a} = 0$. The causal mechanism for Y is $y = a$, i.e. Bob has clubs iff he is given them. Whatever utility function describes Bob's preferences, the action $A = 0$ causes no harm in this model (Lemma 3 Appendix J) as $P(Y_0 = y^*|A = 0, Y = y) = \delta(y^* - y)$ (factual a and counterfactual $\bar{a}$ are identical) and for non-zero harm we require $y^* \neq y$.

Note there are other reasonable scenarios where Alice's actions would constitute harm. For example, if Alice was a clerk in a golf shop and Bob had pre-paid for a set of golf clubs, we could claim that 'the clerk Alice harmed Bob by not giving him golf clubs'. In this case, we would expect Alice to give Bob the clubs by default (she has a ethical obligation to do so) and the harm query we want (implied by our ethical assumptions about clerks) is where the default action is $\bar{a} = 1$. By choosing not to—$A = 0$—Alice causes harm to Bob. For example, if Bob's utility is $U(y) = y$ (i.e. 1 for clubs, 0 for no clubs), then the harm caused by Alice is $P(Y_{A=1} = 1|A = 0, Y = 0) = 1$. So we can see that the choice of default action is vital for expressing these different normative assumptions.

Preemption Problem: Alice robs Bob of his golf clubs. A moment later, Eve would have robbed bob of his clubs. Therefore, Alice's action does not cause Bob to be worse off as he would have lost his clubs regardless of her actions, and so by the CCA Alice does not harm bob by robbing him. However, intuitively Alice harms Bob by robbing him, regardless of what occurs later [13].

Let $A = \{1, 0\}$ denote Alice {robbing, not robbing} Bob respectively, and similarly $E = \{1, 0\}$ for Eve. $B = \{1, 0\}$ denotes Bob {has clubs, does not have clubs}. Assume Bob's utility is $U(b) = b$. The causal mechanisms are $e = 1 - a$ (Eve always robs Bob if Alice doesn't) and $b = 1 - a \vee e$ (Bob has no clubs if either Alice of Eve robs him). See Figure 3 for the causal model depicting these variables.

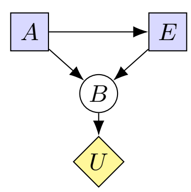

Figure 3: SCM depicting the preemption problem.

Note that while Alice's action is an actual cause of Bob not having clubs, it is also an actual cause of Eve not robbing Bob, which is an event equally as bad as Alice robbing Bob. Intuitively, when we claim that Alice robbing Bob was harmful, we are making a claim about the effects of Alice's actions on Bob independently of their effect on Eve's actions (independent of the effect that her action has mediated through Eves action, preventing Eve from robbing Bob), i.e we are concerned with the direct harm caused by Alice's actions on Bob.

The relevant harm query is the path-specific harm where we compare to the default action where Alice does not rob Bob, $\bar{a} = 0$. We want to determine the harm caused by Alice's action independently of its effect on Eve's action, which we do by blocking the path $\bar{g} = \{A \to E\}$. Applying Definition 9 amounts to replacing the mechanism for E with $f^E(a) \to f^E_g(A = 1) = 0$, i.e. E is evaluated for the factual value of A. We then compute the harm using the counterfactual default action $A = 0$, giving the counterfactual $B(A = 0, E = 0) = 1$, which gives a counterfactual utility of 1 compared to a factual utility of 0. Therefore Alice directly harmed Bob by robbing him.

Note we can also choose a different model where we explicitly represent the outcomes of the two agents decisions and the temporal order in which they occur (Figure 4). In this case the relevant harm query is essentially the same; the path specific harm where we determine the harm caused by Alice's action independently of the effect it has on whether or not Eve robs Bob (i.e. $\bar{g} = \{R_A \to R_E\}$).

Preventing worse: We provide two versions of the preventing worse problem [15] which have identical causal models but intuitively different harms attributed to Alice's action.

Case 1: Bob has \$2. The thief Alice is stalking Bob in the marketplace and notices that Eve (a more effective thief) is also stalking Bob. Seeing Eve before Eve notices her, Alice decides to make her

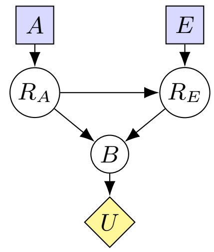

Figure 4: SCM depicting the preemption problem explicitly representing the temporal asymmetry between Alice and Eve's actions effecting Bob.

move first. She steals \$1 from Bob. Eve was going to steal \$2 from Bob, but is incapable of doing so if someone else robs him first (e.g. Bob realizes he's been robbed and call for the police, making further robbery impossible). Seeing that Bob was robbed by Alice she decides not to rob him.

Case 2: Eve has captured Bob and intends to torture him to death. Alice sees this, and is too far away to prevent Eve from doing so. She has a line of sight to Bob (but not Eve) and can shoot him before eve has a chance to torture him to death, resulting in a painless death.

The causal model describing both of these cases is depicted in Figure 5. Let $E = \{1, 0\}$ denote if Eve is present or not, $A \in \{1, 0\}$ be Alice's action (rob, shoot) or not, $AB \in \{1, 0\}$ denote the outcome following Alice's action (Bob is robbed of \$1 / bob is shot, or not) and let $EB \in \{1, 0\}$ denote Eves action on Bob (Bob is robbed of \$2 / Bob is tortured, or not). Let $Y \in \{0, 1, 2\}$ denote Bob's outcome, with 2 being the best (Bob has \$2 in Case 1, Bob survives in Case 2), 1 being the second worst (Bob has \$1 in Case 1, is killed painlessly in Case 2), and 0 the worst (Bob has \$0 in Case 1, died painfully in Case 2). The causal mechanisms are $a = e$ (e.g. Alice shoots/robs if Eve is present), $ab = a$ (Alice's bullet hits with certainty / successfully robs with certainty), $eb = e(1 - ab)$ (Eve tortures Bob if she is present and he is not shot / eve robs Bob if she is present and hasn't been robbed already), and $y = ab + 2(1 - ab)(1 - eb)$ (Case 1: if Bob is shot he dies quickly, else if Eve tortures him he dies slowly, else he lives, Case 2: Bob has \$2 if not robbed, \$1 if robbed by Alice, \$0 if not robbed by Alice and robbed by Eve).

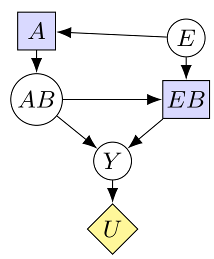

Figure 5: SCM depicting the preventing worse problem.

In Case 1, Alice intuitively harms Bob by robbing him. The argument supporting this is that Alice's robbery caused Bob to lose \$1, regardless of the fact that Alice's action prevented a worse robbery by Eve. However, for Case 2 it is argued in [15] that Alice intuitively didn't harm Bob. While Bob died due to Alice shooting him, this action was intended to prevent a worse outcome from occurring (Bob being tortured to death), which would have happened with certainty had Alice not shot him. However, these two scenarios are described by equivalent causal models—only the variables have been re-labeled. However, the ethical assumptions differ between Case 1 and 2.

From this we conclude that to satisfactorily describe these two situations we need two different harm queries. In either case, one of these harm queries is the morally relevant one and the other is not, and to do this we use the path-independent and path-specific harms. Note, this is no different than in causal analysis where in certain problems the casual effect is the desired query and in others the path-specific effect is the desired query [5]. For Case 1 we use the path-specific harm (Definition 9) to determine the harm caused by Alice robbing Bob independently of what effect it had on Eve's action. We block the path $\bar{g} = \{AB \to EB\}$ and use the default action $\bar{a} = 0$. In the counterfactual world, this gives $AB = 0$ and $EB = f^{EB}(A = 1, E = 1) = 0$, and therefore $Y = 2$, and so the direct harm of Alice robbing Bob is 2 - 1 = 1 compared to not robbing him. For Case 2, we note that while Alice shooting Bob is arguably intrinsically harmful (as is captured by the direct harm of 1 caused by $A = 1$ if we calculate the path-specific harm as in Case 1), this is not the morally relevant harm that we are referring to when we say that intuitively Alice did not harm Bob by shooting him. The reason Alice fired the shot was precisely because of its mediating effect on Y through Eve's actions (preventing her from torturing him to death). From this we infer that the morally relevant harm in this case is the path-independent harm. This we calculate using Definition 3 and the default action $\bar{a} = 0$, which in the counterfactual world gives $AB = 0$, $EB = 1$, $Y = 0$ and hence $U = 0$, compared to the factual utility $U = 1$, giving the desired result that Alice did not harm Bob compared to not shooting him. Note that if we favoured the path-independent or path-dependent harm a priori this would either fail to detect harm in Case 1 or incorrectly attribute harm to Alice in Case 2.

We argue from these two examples that there is no single causal formula for harm that is correct in all scenarios—in some the morally relevant measure of harm is path-specific (e.g. the direct harm), in others it is the path-independent harm. This is in contrast to other approaches to define harm with a single causal formula that applies to all scenarios, namely [8], and we discuss this approach and provide counterexamples to it in Appendix D

C

In this Appendix we discuss selecting and interpreting default actions, harmful events, and various edge cases not covered in the main body of our paper such as harmful default actions. Note that while the CCA (Definition 2) states '[the action] had not been performed', this should not be interpreted as 'do nothing', as doing nothing is often a valid action choice and should be included as an element of A. Instead, we argue that statements about harm often implicitly assume some default action, often following from ethical or normative assumptions (although this is not always the case). Indeed, in Appendix D we show in Example 3 that being able to enforce a unique default action is vital in some scenarios to give intuitive results.

Our definition of harm treats the default action as an integral part of the harm query, just as a reference treatment is necessary when defining treatment effects [79]. These default-dependent measures of harm can be converted to default-independent measures if desired, e.g. by taking the max over all default actions, but in all of the examples we explore this is not desirable. We also note that while the examples outlined in the main text assume deterministic default actions, it is trivial to extend our definitions to non-deterministic default actions by replacing $\text{do}(A = \bar{a})$ in Defintion 3 with a soft intervention (e.g. [17]). For examples of how the default action resolves the omission problem, and when path-specific and path-independent harm should be used, see Appendix B.

Default actions: In some cases harm is attributed to an agent by comparing to normative actions or policies, and so the default action is often implied by the situation or determined by normative assumptions (e.g. Example 1 below). For example, in a case of negligence a doctor's actions may be compared to clinical guidelines, or in a randomized control trial the harm caused by a drug is typically determined by comparing to the outcomes that would have occurred if the trial participants had instead been given a placebo. This is not always the case however (Example 2). The relevant harm query can also compare to actions that the agent could never take (Example 3). While some have argued against comparative accounts on the grounds that it is not always clear which comparison is needed [35], this problem arises due to the ambiguity of statements about harm rather than due to a problem with its formal definition (note, we do not consider scenarios where the agent's action alters the user's utility function). Clearly, there is not a single universal comparison or default action that is suitable for all situations (this assumption leads to the omission problem, described in Appendix B), and the ability to explicitly choose the comparison is a feature rather than a fault with the CCA.

Example 1: The claim 'the doctor harmed the patient by not treating them' and 'the bystander with no medical training failed to benefit the patient by not treating them' both tacitly assume different default actions. In the first, the doctor has an ethical obligation to treat the patient (e.g. the Hippocratic oath), and likewise the patient can expect to be treated by the doctor. Hence if they are not treated, harm can occur. In the second, the bystander may have no ethical obligation to help the patient (depending on our ethical assumptions) and so the intuitive choice of default action is to not treat the patient. In both of these examples, the 'correct' default action depends on the situation and in these examples is informed by our assumptions as to the ethical obligations of the agent.

Example 2: Consider a drug that a doctor is expected to provide to a patient which rarely causes severe side effects. For a given patient, those side effects occur, and clearly the drug has harmed the patient. Perhaps the most obvious harm measure to capture this would be the total harm caused by the treatment compared to the default action where the doctor did not treat the patient at all, or provided them with a different treatment. Each of these is a different but valid harm query, and the correct one will depend on the situation. For example if we are measuring the harm due to the doctors negligence, we should compare to the normative default action alone (and should find zero harm due to negligence as the doctor followed the correct protocol), whereas if we are trying to establish harm caused by the drug to this patient due to the side effects it caused, we should use the default 'no treatment'.

Example 3: How can we deal with cases where every action available to the agent is harmful? In this case harm is still measured compared to some default action, even if the action is idealized and not actually available to the agent. For example, if a doctor is forced at gun point to choose between administering two poisons that will harm the patient, we can still measure this harm compared to the counterfactual action where the doctor does not treat the patient, even if this action is not available to the doctor.

Harmful events: Finally, we note that while we focus on harmful actions due to our focus on training ethical artificial agents, our results extend trivially to harmful events as actions are formally equivalent to events in the causal models we consider, and instead of default actions we can use default events.

D Comment on Beckers et. al.

In this appendix we discuss an alternative proposal for qualitatively defining harm [8], which was developed following the presentation of our preliminary results. We describe this definition (which we refer to as BCH) and three examples where BCH leads to counter-intuitive results (intuitively harmful actions being identified as not harmful or vice versa). First we present a simplified version of the BCH definition of harm where we restrict our attention to attributing harm to single actions.

Definition 10. *$A = a$ rather than $A = a'$ causes $Y = y$ rather than $Y = y'$ in the model M for exogenous noise state $E = e$ iff;*

1. *$A(e) = a$ and $Y(e) = y$*
2. *The exists a set of environment variables W with factual state $W(e) = w$ such that $Y_{A=a',W=w}(e) = y'$*
3. *$A = a$ is minimal; There is no strict subset of the set of variables $\tilde{A} \subset A$ such that for $\tilde{A} = \tilde{a}$ we can satisfy conditions 1. and 2.*

In the following we will focus on scenarios where we can consider single action variables alone and so we can ignore condition 3 in Definition 10.

Definition 11 (BCH harm)**.** *$A = a$ harms the user in model $\mathcal{M}$ and exogenous noise state $E = e$, if there exists an outcome $Y = y$ an action $A = a'$ such that,*

H1 *$U(y) < d$ where d is the default utility*

H2 *$\exists\, Y = y'$ s.t. $A = a$ rather than $A = a'$ causes $Y = y$ rather than $Y = y'$ and $U(y') > U(y)$.*

H3 *$U(y) \leq U(y'')$ for the unique y'' such that $Y_{a'}(e) = y''$*

If we restrict to deterministic models (i.e. $P(E = e)$ deterministic), attempt to only determine if harm is non-zero rather than quantify how much harm is caused (i.e. map all non-zero harm values to 1), and assume that the users utility function is independent of the agents action A and the context X given the outcome Y, then it is possible to directly compare our harm measure to that proposed in BCH. First, we present three problematic cases where the BCH gives counter-intuitive results. We then attempt to diagnose why our approaches give different answers in these cases.

Example 1: two thieves. (repeat of Case 1 in Appendix B). Bob has \$2. The thief Alice is stalking Bob in the marketplace and notices that Eve (a more effective thief) is also stalking Bob. Seeing Eve before Eve notices her, Alice decides to make her move first. She steals \$1 from Bob. Eve was going to steal \$2 from Bob, but is incapable of doing so if someone else robs him first (e.g. Bob realizes he's been robbed and call for the police, making further robbery impossible). Seeing that Bob was robbed by Alice she decides not to rob him. The causal model for this scenario is described below,

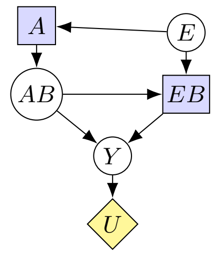

Figure 6: SCM depicting the two robbers problem. $E \in \{1, 0\}$ denote Eve {present, not present}, $A = \{1, 0\}$ denotes Alice decides to {rob, not rob} Bob, $AB \in \{1, 0\}$ denotes Bob is{robbed, not robbed} by Alice, $EB = \{1, 0\}$ Eve attempts to {rob, not rob} Bob. Y denotes how much money Bob has finally. Causal mechanisms $a = e$ (Alice robs if Eve is present), $ab = a$ (Alice always succeeds in robbing Bob), $eb = e(1 - ab)$ (Eve robs Bob if she is present and he hasn't been robbed already), and $y = ab + 2(1 - ab)(1 - eb)$ (if Bob is not robbed at all he has \$2, if Alice Robs him he has \$1, and if Eve robs him and Alice does not he has \$0).

Intuitively Alice harmed Bob by robbing him, but by Definition 11 she did not. The only available counterfactual action for Alice is $\bar{a} = 0$. This counterfactual action (with no contingencies) leads to the counterfactual outcome $Y_{A=0}(e) = 0$, i.e. if Alice doesn't rob Bob then Eve will, resulting in a lower utility $U(Y = 1) > U(Y = 0)$. Therefore H3 is not satisfied and Alice did not harm Bob by robbing him. We discuss this problem further in Appendix B and argue that the morally relevant harm query in this scenario is the direct (path-specific) harm of Alice robbing Bob compared to not robbing him ($\bar{a} = 0$), independent of the benefit caused by preventing Eve from robbing him (blocking $\bar{g} = \{AB \to EB\}$). Applying Definition 9 it is simple to check the path-specific harm described is 1.

Example 2: robber & Samaritan An intuitive property of harm is that the harms caused by one agent's actions should not by default be cancelled out by the another agents beneficial actions—e.g. stabbing someone is harmful, regardless of whether or not a doctor will treat the wound in response. This ability to disentangle agent A's harm from agent B's benefit is vital for determining harm in complex scenarios involving multiple actions or events. For example: Bob has \$1 and Alice steals it. Seeing this, Eve feels bad for Bob and later gifts Bob a dollar, restoring him to his initial funds. Intuitively we would say Alice harmed Bob and Eve benefited him, or at least it would be counter-intuitive to say that Alice robbing Bob was not harmful because at a later time Bob's finances were restored by a second agent (Eve). The causal model describing this situation is depicted below,

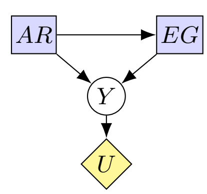

Figure 7: SCM depicting the preventing worse problem. Alice {robs, doesn't rob} Bob (AR) is denoted $AR = \{1, 0\}$. Eve gives (EG) Bob money if she sees he has been robbed ($eg = ar$). Bobs money is his initial money, minus any theft and adding any gifts $y = 1 - ar + eg$. Let $U(y) = y$.

The intuitive default utility for this scenario is $d = 1$ (Bob expects to have \$1), but as the factual outcome is $Y = 1$ then H1 cannot be satisfied. To satisfy H1 we would need to choose a default utility $d > 1$ which amounts to Bob having by default more money than he would have regardless of Alice robbing him (e.g. assuming Bob can expect to become richer following a robbery). We therefore either recover a counter-intuitive answer (Alice did not harm Bob by robbing him), or have to use a counter-intuitive default utility that is hard to justify beyond choosing whichever value gives the desired answer.

Our approach is to measure the direct harm caused by $A = 1$ (Alice robs) compared to $\bar{a} = 0$ (Alice doesn't rob), blocking the path $\bar{g} = \{AS \to EG\}$. This disentangles that harm caused by Alice robbing Bob from the benefit due to this action causing Eve to help Bob. It is simple to check that this results in a harm of 1. As described in Appendix B, this is the intutive choice of harm query as we are interested in the harm caused directly to Bob by Alice robbing him, independent of the indirect effect of causing Eve to benefit him. In other (causally equivalent) scenarios described in Appendix B, the intuitive harm query we desire if the total (path-independent) harm, and as with default actions this has to be implied from the context.

Example 3: omission problem. In this example we present an extension of the omission problem that is violated by the BCH definition of harm. The Phoenicians are a moderately wealthy people, collectively owning \$2. The Romans can decide to gift them an extra \$2 or do nothing, and they have no moral obligation to give them anything. Unbeknownst to the Romans, the Carthaginians decide that if the Romans don't give the Phoenicians anything they will attack them, stealing all of their money. But if the Romans do gift the Phoenicians \$2, the Phoenicians will become too powerful and the Carthaginians wont attack. The Phoenician's utility is equal to how much money they have.

The Romans decide not to gift the Phoenicians anything, and they are attacked by the Carthaginians and have all their money stolen. Intuitively, the Romans didn't harm the Phoenicians (any harm was caused by the Carthaginians)—instead the Romans failed to benefit them. However, by the BCH account the Roman's harmed the Phoenicians.

To see this, first note that if $d \leq 0$ then the Carthaginians actions do not constitute harm, as the factual utility is equal to the default and H1 cannot be satisfied. This would be a counter-intuitive result, so we assume that $d > 0$. The Phoenicians end up with no money, so H1 is satisfied as $U(y) < d$. H2 is also satisfied by a simple but-for counterfactual because if the Romans had given the Phoenicians money, the Carthaginians wouldn't have attacked and the Phoenicians would have \$4 which is more than their factual \$0. Finally, H3 is satisfied by the same argument as H2. Therefore the Romans harmed the Phoenicians. Applying our methods, it is sufficient to note that the implied default action should be $A = 0$ (By default we do not expect the Romans to give money to the Phoenicians, reflecting the ethical assumptions implicit in the 'failure to benefit' assertion) and this gives a counterfactual harm of zero because the factual and counterfactual actions are identical.

Analysis: Why do these issues arise? Firstly, the BCH account of harm proposes a single casual formula for harm that applies to all scenarios, allowing for any counterfactual action or contingency to establish harm much as is done in actual causality [34]. If H3 was not included, this could result in harm being attributed in cases of 'preventing worse' (as pointed out in [8] and described in Appendix B), but H3 is included to fix this by requiring that benefit does not occur in the case where no contingency is taken, which in these examples is the same as requiring that an action cannot be harmful if its total (path-independent) benefit is non-zero. But this is precisely the case in Example 1, where Alice prevents a worse outcome but, intuitively, we would want to ascribe harm to her actions. By separating direct and indirect harm, we can see that her actions were indirectly beneficial (she prevented a worse robbery), but directly harmful, and in this scenario the morally relevant (i.e. 'intutive') measure of harm is the direct harm. In the equivalent Case 2 example in Appendix B, the harm query we intuitively want is the total harm rather than the path-specific harm. This points to the conclusion that a one-size-fits-all harm query is not tenable, given that the intuitive measures of harm we desire are sometimes path-dependent and sometimes path-independent.

Secondly, Example 2 suggests that approaches using default utilities are not tenable, because they preclude the possibility of the user being harmed by any action or event that occurred previously if, in the end, the user obtains the default utility. Clearly this is not the case in general—users can achieve the expected or default outcome (e.g. leaving the market with as much money as they came in with) and still have been harmed. The BCH therefore cannot robustly detect harm in cases where both benefit and harm occur unless we choose large values of d (essentially removing d from the analysis). But it is not clear how these default utility values can be justified beyond fixing a problem that they cause—seeing as these large values of d do not correspond to any utility the user can expect to have (e.g. in Example 2 it would require that the user can expect to be richer than they initially were following a robbery).

Thirdly, in Example 3 we see that by allowing for any default action the BCH account can end up misattributing harm in cases where the agent has no ethical duty to act (or by extension, has a duty to perform specific actions). This is because the BCH allows the counterfactual action to take any value—in this case, the Romans gifting the Phoenicians money. This can be avoided by by not attributing harm to the Romans using counterfactual actions that they could never be expected to take from an ethical standpoint (i.e. using default actions). Just by allowing the Romans to (in theory) give any positive amount of money, we could even make the harm they cause by not giving the Phoenicians anything arbitrarily large. In conclusion, by evaluating over all possible actions that the agent could take the BCH doesn't allow normative assumptions about actions (e.g. to do with the ethical responsibility to act or not act) to be included in the harm query.

E

In this Appendix we prove Theorem 1. Noting that $\max\{0, U(a,x,y) - U(\bar{a},x,y)\} - \max\{0, U(\bar{a},x,y^*) - U(a,x,y)\} = U(a,x,y) - U(\bar{a},x,y^*)$, subtracting the expected harm from the expected benefit (Def 3) gives,

$$\mathbb{E}[b|a,x;\mathcal{M}] - \mathbb{E}[h|a,x;\mathcal{M}] \tag{15}$$

$$= \int_{y,y^*} P(y, Y_{\bar{a}} = y^*|a,x;\mathcal{M})(U(a,x,y) - U(\bar{a},x,y^*)) \tag{16}$$

$$= \int_{y} P(y|a,x)U(a,x,y) - \int_{y^*} P(Y_{\bar{a}} = y^*|x)U(\bar{a},x,y^*) \tag{17}$$

$$= \mathbb{E}[U|a,x] - \mathbb{E}[U|\bar{a},x] \tag{18}$$

F

In this Appendix we derive the SCM model for the treatment decision task in examples 1 and 2, and calculate the average treatment effect and counterfactual harm.

Patients who receive the default 'no treatment' $T = 0$ have a 50% survival rate. $T = 1$ has a 60% chance of curing a patient, and a 40% chance of having no effect, with the disease progressing as if $T = 0$, whereas $T = 2$ has a 80% chance of curing a patient as a 20% chance of killing them, due to some unforeseeable allergic reaction to the treatment.

Next we evaluate this expression for our two treatment by constructing an SCM for the decision task. The patient's response to treatment is described by three independent latent factors (for example genetic factors) that we model as exogenous variables. Firstly, half of the patients exhibit a robustness to the disease which means they will recover if not treated, which we encode as $E^1 \in \{0, 1\}$ where $e^1 = 1$ implies robustness with $P(e^1 = 1) = 0.5$. Secondly, the patients may exhibit a resistance to treatment 1 indicated by variable E^2, with $e^2 = 1$ implying resistance with $P(e^2 = 1) = 0.4$. Finally, the patients can be allergic to treatment 2, indicated by variable E^3 with $e^3 = 1$ and $P(e^3 = 1) = 0.2$. Given knowledge of these three factors the response of any patient is fully determined, and so we define the exogenous noise variable as $E^Y = E^1 \times E^2 \times E^3$ with $P(e^Y) = P(e^1)P(e^2)P(e^3)$.

Next we characterise the mechanism $y = f(t, e^Y) = f(t, e^1, e^2, e^3)$ where $f(0, e^Y) = [e^1 = 1]$ (untreated patients recover if they are robust), $f(1, e_Y) = [e^1 = 1] \vee [e^2 = 0]$ (patients with $T = 1$ recover if they are robust or non-resistant) and $f(2, e_Y) = [e^3 = 0]$ (patients with $T = 2$ recover if they are non-allergic), where $[X = x]$ are Iverson brackets which return 1 if $X = x$ and 0 otherwise, and $\vee$ is the Boolean OR.

The recovery rate for $T = 1$ and $T = 2$ can be calculated with (1) to give $P(Y_1 = 1) = P(e^1 = 1 \vee e^2 = 0) = 1 - P(e^1 = 0)P(e^2 = 1) = 0.8$, and likewise $P(Y_2 = 1) = P(e^3 = 0) = 0.8$. Hence the two treatments have identical outcome statistics (recovery/mortality rates), and all observational and interventional statistical measures are identical, such as risk, expected utility and the effect of treatment on the treated. Note as there are no unobserved confounders the recovery rate for action $A = a$ is equal to $\mathbb{E}[Y_a]$.

We compute the counterfactual expected harm by evaluating (4), noting that $Y_0^*(e) = 1$ if $e^1 = 1$, $Y_1^*(e) = 0$ if $e^1 = 0$ and $e^2 = 1$, and $Y_2^*(e) = 0$ if $e^3 = 1$. This gives $P(Y_1 = 0, Y_0^* = 1) = 0$, i.e. there are no values of e^Y that satisfy both $Y_1(e) = 0$ and $Y_0(e) = 1$, and therefore do$(T_1 = 1)$ causes zero harm. However, $P(Y_2 = 0, Y_0^* = 1) = P(e^1 = 1)P(e^3 = 1) = 0.1$, and so do$(T_2 = 2)$

causes non-zero harm. This is due to the existence of allergic patients who are also robust, and will die if treated with $T = 2$ but would have lived had $T = 0$.

G

In this Appendix we derive the policies of agents 1-3 in Example 3. We note that outcome Y is described by a heteroskedastic additive noise model with the default action $\bar{a}$ (no action) corresponding to $A = 1, K = 1$. The expected harm is given by Theorem 5 with $\sigma(\bar{a}) = 100$, $\sigma(A = 2) = 100$, $\sigma(A = 3) = 0$ and $\sigma(A = 1, K) = 100K$. $\mathbb{E}[U|\bar{a}] = 100$ $\mathbb{E}[U|A = 2] = 110$, $\mathbb{E}[U|A = 3] = 80$ and $\mathbb{E}[U|A = 1] = 100K$, where we have used $\text{Var}(KY) = K^2\text{Var}(Y)$ and $\text{Var}(Y + 10) = \text{Var}(Y)$

Agent 1 takes action 1 and the maximum value $K = 20$ as this extremizes $\mathbb{E}[U|a]$.

Agent 2 chooses $a = \arg\max_a\{\mathbb{E}[Y|a] - \text{Var}(Y|a)\}$ which for each action is given by,

$$E[Y|A=1] - \lambda\text{Var}(Y|A=1) = 100K - 100^2K^2\lambda \tag{19}$$

$$E[Y|A=2] - \lambda\text{Var}(Y|A=2) = 110 - 100^2\lambda \tag{20}$$

$$E[Y|A=3] - \lambda\text{Var}(Y|A=3) = 80 \tag{21}$$

For action 1 the optimal $K = 1/200\lambda$, which gives $E[Y|A = 1] - \text{Var}(Y|A = 1) = 1/4\lambda$. Note that $1/4\lambda > 110 - 100^2\lambda$ for $\lambda < 0.0032$, which $80 > 1/4\lambda$ for $\lambda > 0.003125$. Therefore there is no value of λ for which agent 2 selects action 2, choosing action 1 for $\lambda < 0.003125$ and action 3 otherwise.

For agent 3 applying Theorem 5 gives,

$$\mathbb{E}[Y|A=1] - \lambda\mathbb{E}[h|A=1,K] = 100K - \lambda\left[\frac{|100(K-1)|}{\sqrt{2\pi}}e^{-\frac{1}{2}} + \frac{100(K-1)}{2}\left(\text{erf}\left(\frac{\text{sign}(K-1)}{\sqrt{2}}\right) - 1\right)\right] \tag{22}$$

$$= \begin{cases} 100K - 8.332(K-1)\lambda, & K \geq 1 \\ 100K - 59.937(1-K)\lambda, & K < 1 \end{cases} \tag{23}$$

$$E[Y|A=2] - \lambda\mathbb{E}[h|A=2] = 110 \tag{24}$$

$$E[Y|A=3] - \lambda\mathbb{E}[h|A=3] = 80 - \lambda\left[\frac{100}{\sqrt{2\pi}}e^{-\frac{20^2}{2\times100^2}} + \frac{20}{2}\left(\text{erf}\left(\frac{20}{\sqrt{2}\times100}\right) - 1\right)\right] \tag{25}$$

Clearly, the agent will never take action 3 as its expected HPU is smaller than that for action 2 for all λ. For action 1, for $K < 1$ the expected HPU is also smaller that that for action 2, for all λ. For action 1 with $K > 1$, if $\lambda < 12.002$ the optimal $K = 20$, otherwise it is 0. As a result, for $\lambda < 11.93$ the agent chooses action 1 with $K = 20$, and otherwise chooses action 2.

H

In this Appendix we derive an expression for the expected counterfactual harm in generalized additive models. To calculate the expected counterfactual harm we derive a solution for a broad class of SCMs, heteroskedastic additive noise models, which includes our GAM (11),

Definition 12 (Heteroskedastic additive noise models)**.** *For Y, $Pa(Y) = A \cup X$, the mechanism $y = f_Y(a, x)$ is a heteroskedastic additive noise model if Y is normally distributed with a mean and variance that are functions of a, x,*

$$y = \mu(a,x) + e^Y\sigma(a,x), \quad e^Y \sim \mathcal{N}(0,1) \tag{26}$$

In Appendix I we show that the dose response model (11) can be parameterised as a heteroskedastic additive noise model and calculate the expected counterfactual harm using the following theorem,

Theorem 5 (Expected harm for heteroskedastic additive noise model)**.** *For $Y = f_Y(a, x, e^Y)$ where f_Y is a heteroskedastic additive noise model (Definition 12) and default action $A = \bar{a}$, the expected harm is*

$$\mathbb{E}[h|a,x] = \frac{|\Delta\sigma|}{\sqrt{2\pi}} e^{-\frac{\Delta U^2}{2\Delta\sigma^2}} + \frac{\Delta U}{2}\left(\text{erf}\left(\frac{\Delta U}{\sqrt{2}|\Delta\sigma|}\right) - 1\right) \tag{27}$$

where $\text{erf}(\cdot)$ *is the error function,* $\Delta U = \mathbb{E}[U|a,x] - \mathbb{E}[U|\bar{a},x]$, $\Delta\sigma = \sigma(a,x) - \sigma(\bar{a},x)$.

Proof. Note that if $e^Y \sim \mathcal{N}(\mu, V)$ we can replace $e^Y \to e'^Y = e^Y/\sqrt{V} - \mu$ and absorb these terms into $f(a,x)$ and $\sigma(a,x)$. Hence we need only consider zero-mean univariate noise. In the following we use $e^Y = \varepsilon \sim \mathcal{N}(0,1)$ to denote the fact the the exogenous noise term is univariate normally distributed. We also use the fact that there are no unobserved confounders between A and Y to give $P(y|a,x) = P(y_a|x)$. Calculating the expected counterfactual harm using gives

$$\mathbb{E}[h|a,x] = \int_y dy \int_{y^*} dy^* P(y, Y_{\bar{a}} = y^*, |a,x) \max(0, U(\bar{a},x,y^*) - U(a,x,y)) \tag{28}$$

$$= \int_y dy \int_{y^*} dy^* P(Y_a = y, Y_{\bar{a}} = y^*|x) \max(0, U(\bar{a},x,y^*) - U(a,x,y)) \tag{29}$$

$$= \int_\varepsilon P(\varepsilon)d\varepsilon \int_y dy \int_{y^*} dy^* P(Y_a = y, Y_{\bar{a}} = y^*|\varepsilon,a,x) \max(0, U(\bar{a},x,y^*) - U(a,x,y)) \tag{30}$$

$$= \int_\varepsilon P(\varepsilon)d\varepsilon \int_y dy \int_{y^*} dy^* P(Y_a = y|\varepsilon,a,x) P(Y_{\bar{a}} = y^*, |\varepsilon,a,x) \max(0, U(\bar{a},x,y^*) - U(a,x,y)) \tag{31}$$

Substituting in $U(a,x,y) = y$ and $P(y|\varepsilon,a,x) = \delta(y - f(a,x) - \varepsilon\sigma(a,x))$ gives,

$$\mathbb{E}[h|a,x] = \int d\varepsilon P(\varepsilon) \max\{0, f(\bar{a},x) - f(a,x) + \varepsilon(\sigma(\bar{a},x) - \sigma(a,x))\} \tag{32}$$

$$= \int d\varepsilon P(\varepsilon) \max(0, -(\mathbb{E}[U|a,x] - \mathbb{E}[U|\bar{a},x]) - \varepsilon(\sigma(a,x) - \sigma(\bar{a},x))) \tag{33}$$

where we have used the fact that $\mathbb{E}[U|a,x] = \int d\varepsilon P(\varepsilon)(f(a,x) + \varepsilon\sigma(a,x)) = f(a,x)$. For ease of notation we use $\Delta U = E[U|a,x] - E[U|\bar{a},x]$, $\Delta\sigma = \sigma(a,x) - \sigma(\bar{a},x)$. Next, we remove the $\max()$ by incorporating it into the bounds for the integral. If $\Delta U > 0$ and $\Delta\sigma > 0$, this is equivalent to $\varepsilon < -\Delta U/\Delta\sigma$ and hence,

$$\mathbb{E}[h|a,x] = \int_{\varepsilon < -\Delta U/\Delta\sigma} d\varepsilon P(\varepsilon)(-\Delta U - \varepsilon\Delta\sigma) \tag{34}$$

$$= -\Delta U \int_{-\infty}^{-\Delta U/\Delta\sigma} P(\varepsilon)d\varepsilon - \Delta\sigma \int_{-\infty}^{-\Delta U/\Delta\sigma} \varepsilon P(\varepsilon)d\varepsilon \tag{35}$$

(36)

Using the standard Gaussian integrals

$$\int_a^b P(\varepsilon)d\varepsilon = \frac{1}{2}\left[\text{erf}(\frac{b}{\sqrt{2}}) - \text{erf}(\frac{a}{\sqrt{2}})\right] \tag{37}$$

$$\int_a^b \varepsilon P(\varepsilon)d\varepsilon = P(a) - P(b) \tag{38}$$

where $P(\varepsilon) = e^{-\varepsilon^2/2}/\sqrt{2\pi}$ and erf(z) is the error function, we recover

$$\mathbb{E}[h|a,x] = \frac{-\Delta U}{2}\left[\text{erf}(\frac{-\Delta U}{\sqrt{2}\Delta\sigma}) - \text{erf}(-\infty)\right] - \Delta\sigma\left[P(-\infty) - P(-\Delta U/\Delta\sigma)\right] \tag{39}$$

$$= \frac{\Delta U}{2}\left[\text{erf}\left(\frac{\Delta U}{\sqrt{2}\Delta\sigma}\right) - 1\right] + \frac{\Delta\sigma}{\sqrt{2\pi}}e^{-\frac{\Delta U^2}{2\Delta\sigma^2}} \tag{40}$$

where we have used erf$(-z) = -$erf(z) and $P(-z) = P(z)$. Similarly, if $\Delta U > 0$, $\Delta\sigma < 0$ then the $\max()$ in (33) can be replaced with a definite intergral over $\varepsilon > \Delta U/\Delta\sigma$ giving,

$$\mathbb{E}[h|a,x] = \int\limits_{\varepsilon > \Delta U/\Delta\sigma} d\varepsilon P(\varepsilon)\left(-\Delta U - \varepsilon\Delta\sigma\right) \tag{41}$$

$$= -\Delta U \int\limits_{-\Delta U/\Delta\sigma}^{\infty} P(\varepsilon)d\varepsilon - \Delta\sigma \int\limits_{-\Delta U/\Delta\sigma}^{\infty} \varepsilon P(\varepsilon)d\varepsilon \tag{42}$$

$$= -\frac{\Delta U}{2}\left[\text{erf}(\infty) - \text{erf}(\frac{-\Delta U}{\sqrt{2}\Delta\sigma})\right] - \Delta\sigma\left[P(\frac{-\Delta U}{\sqrt{2}\Delta\sigma}) - P(\infty)\right] \tag{43}$$

$$= \frac{\Delta U}{2}\left[\text{erf}\left(\frac{\Delta U}{\sqrt{2}|\Delta\sigma|}\right) - 1\right] + \frac{|\Delta\sigma|}{\sqrt{2\pi}}e^{-\frac{\Delta U^2}{2\Delta\sigma^2}} \tag{44}$$

Next, if $\Delta U < 0$ and $\Delta\sigma > 0$ we recover the same integral as (35), and if $\Delta U < 0$ and $\Delta\sigma < 0$ we recover the same integral as (41). Hence the general solution for all $\Delta\sigma$ is (44).

□

I

In this Appendix we present the GAM dose response model including parameter values, and show that it corresponds to a heteroskedastic additive noise model and calculate the expected harm for a given dose.

We follow the set-up described in [18], where outcome Y denotes the level of improvement in the symptoms of schizoaffective patients following treatment and compared to pre-treatment levels, measured in terms of the Positive and Negative Syndrome Scale (PANSS) [44]. The response of Y w.r.t dose A (Aripiprazole mg/day) is determined using a generalized additive model fit with a cubic splines regression and random effects,

$$y = \theta_1 a + \theta_2 f(a) + \varepsilon_0 \tag{45}$$

where the parameters θ_i are random variables $\theta_i \sim \mathcal{N}(\hat{\theta}_i, V_i)$, $\varepsilon_0 \sim \mathcal{N}(0, V_0)$ is the sample noise, and the spline function $f(a)$ is given by,

$$f(a) = \frac{(a-k_1)_+^3 - \frac{k_3-k_1}{k_3-k_2}(a-k_2)_+^3 + \frac{k_2-k_1}{k_3-k_2}(a-k_3)_+^3}{(k_3-k_1)^2} \tag{46}$$

where k_1, k_2, k_3 are the knots at $a = 0, 10$ and 30 respectively, with $(u)_+ = \max\{0, u\}$. In the following we assume for simplicity that θ_1 and θ_2 are independent. This hierarchical model can be expressed as an SCM with the mechanism for Y given by,

$$y = \left(\hat{\theta}_1 a + \hat{\theta}_2 f(a)\right) + \varepsilon_1 a + \varepsilon_2 f(a) + \varepsilon_0 \tag{47}$$

where $\varepsilon_i \sim N(0, V_i)$. We will now reparameterise this as an equivalent SCM that is an additive heteroskedastic noise model. Using the identifies $Z = kY$, $Y \sim \mathcal{N}(0,1) \implies Z \sim \mathcal{N}(0, k^2)$, and $Z = X + Y$, $X \sim \mathcal{N}(0, V_X)$, $Y \sim \mathcal{N}(0, V_Y) \implies Z \sim \mathcal{N}(0, V_X + V_Y)$ (where V_X is the variance of X and likewise for V_Y, Y), we can replace $\varepsilon_1 a + \varepsilon_2 f(a) \to \varepsilon g(a)$ where $\varepsilon \sim \mathcal{N}(0, 1)$ and $g(a) = \sqrt{a^2 V_1 + f(a)^2 V_2}$. We can therefore reparameterise the mechanism for Y as

$$y = \mathbb{E}[U|a] + g(a)\varepsilon + \varepsilon_0 \tag{48}$$

where we have used $U(a, x, y) = U(a, y) = y$ and the fact that ε, ε_0 are mean zero to give $\mathbb{E}[U|a] = \theta_1 a + \theta_2 f(a)$. Finally, we note that the sample noise term ε_0 cancels in the expression for the harm,

$$\mathbb{E}[h|a] = \int\limits_y dy \int\limits_{y^*} dy^* P(y, Y_{\bar{a}} = y^*, |a) \max\left(0, U(\bar{a}, y^*) - U(a, y)\right) \tag{49}$$

$$= \int\limits_y dy \int\limits_{y^*} dy^* P(Y_a = y, Y_{\bar{a}} = y^*) \max\left(0, U(\bar{a}, y^*) - U(a, y)\right) \tag{50}$$

$$= \int\limits_\varepsilon P(\varepsilon) d\varepsilon \int\limits_{\varepsilon_0} P(\varepsilon_0) d\varepsilon_0 \int\limits_y dy \int\limits_y dy^* P(Y_a = y, Y_{\bar{a}} = y^* | \varepsilon, \varepsilon_0, a) \max\left(0, U(\bar{a}, y^*_{\bar{a}}) - U(a, y)\right) \tag{51}$$

$$= \int\limits_\varepsilon P(\varepsilon) d\varepsilon \int\limits_{\varepsilon_0} P(\varepsilon_0) d\varepsilon_0 \int\limits_y dy \int\limits_{y^*} dy^* P(y, |\varepsilon, \varepsilon_0, a) P(Y_{\bar{a}} = y^*, |\varepsilon, \varepsilon_0) \max\left(0, U(\bar{a}, y^*) - U(a, y)\right) \tag{52}$$

Substituting in $P(Y_a = y | \varepsilon, \varepsilon_0) = \delta(y - f(a) + g(a)\varepsilon + \varepsilon_0)$ gives,

$$\mathbb{E}[h|a] = \int\limits_\varepsilon P(\varepsilon) d\varepsilon \int\limits_{\varepsilon_0} P(\varepsilon_0) d\varepsilon_0 \max\left(0, f(\bar{a}) + g(\bar{a})\varepsilon + \varepsilon_0 - f(a) - g(a)\varepsilon - \varepsilon_0\right) \tag{53}$$

$$= \int\limits_\varepsilon P(\varepsilon) d\varepsilon \int\limits_{\varepsilon_0} P(\varepsilon_0) d\varepsilon_0 \max\left(0, f(\bar{a}) - f(a) + (g(\bar{a}) - g(a))\varepsilon\right) \tag{54}$$

$$= \int\limits_\varepsilon P(\varepsilon) d\varepsilon \max\left(0, f(\bar{a}) - f(a) + (g(\bar{a}) - g(a))\varepsilon\right) \tag{55}$$

Therefore we can ignore the sample noise term when calculating the expected harm, instead calculating the expected harm for the model $Y = f(a) + g(a)\varepsilon$. This is a heteroskedastic additive noise model, and therefore by Theorem 5 the expected harm is,

$$\mathbb{E}[h|a] = \frac{\Delta U}{2}\left[\text{erf}\left(\frac{\Delta U}{\sqrt{2}\Delta\sigma}\right) - 1\right] + \frac{\Delta\sigma}{\sqrt{2\pi}} e^{-\Delta U^2 / 2\Delta\sigma^2} \tag{56}$$

where $\Delta U = \mathbb{E}[U|a] - \mathbb{E}[U|\bar{a}]$, $\Delta\sigma = g(a) - g(\bar{a})$ and $g(a) = \sqrt{a^2 V_1 + f(a)^2 V_2}$

The resulting curves prefented in Figure 2 are calculated using (56) and the parameter values taken from [18] (Table 1), which are fitted in a meta-analysis of the dose-responses reported in [19, 43, 57, 70, 86].

J

In this Appendix we present proofs of Theorems 2, 3 and 4. First, we prove Theorem 2.

Theorem 2: For any utility functions U, environment $\mathcal{M}$ and default action $A = \bar{a}$ the expected HPU is never a harmful objective for $\lambda > 0$.

Table 1: Parameters for the hierarchical generalized additive dose-response model reported in [18]

Parameter	Value
$\hat{\theta}_1$	0.937
$\hat{\theta}_2$	-1.156
V_1	0.03
V_2	0.10

Proof. Let $a_{\text{max}} = \arg\max_a \{\mathbb{E}[U|a,x] - \lambda\mathbb{E}[h|a,x;\mathcal{M}]\}$. If $\exists\, a' \neq a_{\text{max}}$ such that $\mathbb{E}[U|a',x] \geq \mathbb{E}[U|a_{\text{max}},x]$ and $\mathbb{E}[h|a',x;\mathcal{M}] < \mathbb{E}[h|a_{\text{max}},x;\mathcal{M}]$, then $\mathbb{E}[U|a_{\text{max}},x] + \lambda\mathbb{E}[h|a_{\text{max}},x;\mathcal{M}] < \mathbb{E}[U|a',x] + \lambda\mathbb{E}[h|a',x;\mathcal{M}]\ \forall\, \lambda > 0$ and so $a_{\text{max}} \neq \arg\max_a \{\mathbb{E}[U|a,x] - \lambda\mathbb{E}[h|a,x;\mathcal{M}]\}$. □

Next, we prove theorems 3 and 4 by example, constructing distributional shifts that reveal if an objective function is harmful. To do this we make use of a specific family of structural causal models—counterfactually independent models.

Definition 13 (counterfactual independence (CFI))**.** *Y is counterfactually independent in with respect to A in $\mathcal{M}$ if,*

$$P(y^*_{a^*}, y_a|x) = \begin{cases} P(y_a|x)\delta(y_a - y^*_{a^*}) & a = a^* \\ P(y^*_{a^*}|x)P(y_a|x) & \textit{otherwise} \end{cases} \tag{57}$$

Counterfactually independent models (CFI models) are those for which the outcome Y_a is independent to any counterfactual outcome $Y_{a'}$. Next we show that there is always a CFI model that can induce any factual outcome statistics.

Lemma 1. *For any desired outcome distribution $P(y|a,x)$ there is a choice of exogenous noise distribution $P(e^Y)$ and causal mechanism $f_Y(a,x,e^Y)$ such that Y is counterfactually independent with respect to A*

Proof. Consider the causal mechanism $y = f_Y(a,x,e^Y)$ for some fixed $X = x$, and exogenous noise distribution $P(E^Y = e^Y)$. Let the noise term by described by the random field $E^Y = \{E^Y(a,x) : a \in A, x \in X\}$, with $P(E^Y = e^y) = \times_{a \in A, x \in X} P(E^y(a,x) = e^y(a,x))$ and with $\text{dom}(E^Y(a,x)) = \text{dom}(Y)\ \forall\, A = a, X = x$. I.e. we choose the noise distribution to be joint state over mutually independent noise variables, one for every action $A = a$ and context $X = x$, and where each of these variables has the same domain as Y. Next, we choose the causal mechanism,

$$f_Y(a,x,e^Y) = e^Y(a,x) \tag{58}$$

i.e. the value of Y for action $A = a$ and context $X = x$ is the state of the independent noise variable $E^Y(a,x)$. By construction this is a valid SCM, and we note that the factual distributions (calculated with (4)) are given simply by,

$$P(y|a,x) = P(E^Y(a,x) = y) \tag{59}$$

Likewise applying our choice of mechanism and noise distribution to (4) gives (for $a \neq a'$) the counterfactual distribution,

$$\begin{aligned} P(Y_a = y, Y_{a'} = y'|x) &= P(E^Y(a,x) = y)P(E^Y(a',x) = y') && (60) \\ &= P(Y_a = y|x)P(Y_{a'} = y'|x) && (61) \end{aligned}$$

and likewise gives $P(y_a|x)\delta(y_a - y'_{a'})$ for $a = a'$. Finally, we note that we can choose any $P(y_a|x) = P(E^Y(a,x) = y)$, hence there is a CFI model that induces any factual outcome distribution we desire. □

Next, we show that in counterfactually independent models there are outcome distributional shifts that only change the expected harm of individual actions, without changing any other factual or counterfactual statistics.

Lemma 2. *For $\mathcal{M}$ and (context-dependent) default action $A = \bar{a}(x)$, if U is outcome dependent for the default action $\bar{a}(x)$ and some other action $a \neq \bar{a}(x)$, then there are three outcome distributionally shifted environments $\mathcal{M}_0$, $\mathcal{M}_+$ and $\mathcal{M}_-$ such that;*

1. $\mathbb{E}[h|a, x; \mathcal{M}_-] < \mathbb{E}[h|a, x; \mathcal{M}_0] < \mathbb{E}[h|a, x; \mathcal{M}_+]$
2. $\mathbb{E}[h|b, x; \mathcal{M}_-] = \mathbb{E}[h|b, x; \mathcal{M}_0] = \mathbb{E}[h|b, x; \mathcal{M}_+] \;\forall\, b \neq a$
3. $P(y|a', x; \mathcal{M}_0) = P(y|a', x; \mathcal{M}_+) = P(y|a', x; \mathcal{M}_-) \;\forall\, a' \in A$*, including* $a, \bar{a}(x)$

Proof. In the following we suppress the notation $\bar{a}(x) = \bar{a}$. To construct the environment $\mathcal{M}_0$ we restrict to a binary outcome distribution for each action such that $P(y_a|x)$ is completely concentrated on the highest and lowest utility outcomes,

$$Y_a = 1 \implies Y_a = \arg\max_y U(a, x, y) \tag{62}$$
$$Y_a = 0 \implies Y_a = \arg\min_y U(a, x, y) \tag{63}$$
$$1 = P(Y_a = 1|x; \mathcal{M}_0) + P(Y_a = 0|x; \mathcal{M}_0) \tag{64}$$

Note that we abuse notation as the variables $Y_a = 1$ and $Y_b = 1$ will not be in the same state in general, and the states $1, 0$ denote the max/min utility states under any given action, rather than a fixed state of Y. By Lemma 1 we can choose Y_a to be counterfactually independent with respect to A. Recalling our parameterization of CFI models in Lemma 1, with noise distribuiton $P(E^Y = e^Y) = \times_{a \in A, x \in X} P(E^Y(a, x) = e^Y(a, x))$, $\text{dom}(E^Y(a, x)) = \text{dom}(Y)$, and causal mechanism $f_Y(a, x, e^Y) = e^Y(a, x)$, therefore $E^Y(a, x) \in \{0, 1\} \;\forall\, a, x$. The expected harm for action $\text{do}(A = a)$ is,

$$\mathbb{E}[h|a, x; \mathcal{M}_0] = \sum_{y_a=0}^{1} \sum_{y_{\bar{a}}=0}^{1} P(y_{\bar{a}}|x) P(y_a|x) \max\{0, U(\bar{a}, x, y_{\bar{a}}) - U(a, x, y_a)\} \tag{65}$$

where we have used the fact that $P(y^*_{\bar{a}}, y|a, x) = P(y^*_{\bar{a}}, y_a|x)$ and used counterfactual independence. $U(a, x, 0) < U(\bar{a}, x, 1)$ and so if we choose non-deterministic outcome distributions for $P(y_a|x)$ and $P(y_{\bar{a}}|x)$ then (65) is strictly greater than 0.

We can construct the desired $\mathcal{M}_\pm$ by keeping the causal mechanism but changing the factorized exogenous noise distribution in $\mathcal{M}$ to be,

$$P'(E^Y = e^Y; \mathcal{M}_+) = P(E^Y = e^Y; \mathcal{M}_0) + (-1)^{e^Y(a,x) - e^Y(\bar{a},x)} \phi_+ \tag{66}$$
$$P'(E^Y = e^Y; \mathcal{M}_-) = P(E^Y = e^Y; \mathcal{M}_0) + (-1)^{e^Y(a,x) - e^Y(\bar{a},x)} \phi_- \tag{67}$$

where $\phi_\pm \in \mathbb{R}$ are constants that satisfy the bounds $\max\{-P(Y_{\bar{a}} = 1|x)P(Y_a = 1|x), -P(Y_{\bar{a}} = 0|x)P(Y_a = 0|x)\} \leq \phi_\pm \leq \min\{P(Y_{\bar{a}} = 1|x)P(Y_a = 0|x), P(Y_{\bar{a}} = 0|x)P(Y_a = 1|x)\}$. It is simple to check that for any ϕ that satisfies these bounds we recover $\sum_{e^Y} P'(E^Y = e^Y) = 1$, $P'(E^Y = e^Y) \geq 0 \;\forall\, e^Y$, and therefore P' is a valid noise distribution. Keeping the same causal mechanism f_Y is $\mathcal{M}_\pm$ as in $\mathcal{M}_0$ gives $P(y_a|x; \mathcal{M}_0) = P(y_a|x; \mathcal{M}_+) = P(y_a|x; \mathcal{M}_-)$ as,

$$P'(y_i|x) = \sum_{e^Y(0,x)=0}^{1} \cdots \sum_{e^Y(i-1,x)=0}^{1} \sum_{e^Y(i+1,x)=0}^{1} \cdots \sum_{e^Y(|A|,x)=0} \left[\prod_{j=1}^{|A|} P(e^Y(j, x)) + (-1)^{e^Y(i,x) - e^Y(\bar{a},x)} \phi_\pm \right] \tag{68}$$

$$= P(e^Y(i, x)) + (-1)^{e^Y(i,x) - 0} \phi_\pm + (-1)^{e^Y(i,x) - 1} \phi_\pm \tag{69}$$
$$= P(e^Y(i, x)) = P(y_i|x) \tag{70}$$

and likewise for $i = \bar{a}$. This implies that for any desired outcome statistics $P(y_a|x)$ there is a model where $Y_a \perp Y_{a'} \;\forall\, (a, a')$ where $a \neq a'$ except for the pair $a, \bar{a}$, so long as $P(y_{\bar{a}}|x)$ and $P(y_a|x)$ are non-deterministic (if they are deterministic, $\phi_\pm = 0$ and $\mathcal{M}_0 = \mathcal{M}_\pm$). Because $Y_{a'} \perp Y_{\bar{a}} \;\forall\, a' \neq a$, then $H(a', x; \mathcal{M}_0) = H(a', x; \mathcal{M}_\pm) \;\forall\, a' \neq a$. Also note that $H(\bar{a}, x; \mathcal{M}) = 0$ for any U or $\mathcal{M}$ if

$P(a|x) = \delta(a - \bar{a})$, as $P(Y_{\bar{a}} = i, , Y_{\bar{a}} = k) = 0$ if $i \neq k$ and if $i = k$ (factual and counterfactual outcomes are identical) then the expected harm is zero. The only difference between $\mathcal{M}_0$ and $\mathcal{M}_\pm$ is $P(y_{\bar{a}}, y_a|x; \mathcal{M}_+) \neq P(y_{\bar{a}}, y_a|x; \mathcal{M}_-) \neq P(y_{\bar{a}}, y_a|x; \mathcal{M}_0)$, which differ for $\phi_+ \neq 0$, $\phi_- \neq 0$ and $\phi_+ \neq \phi_-$. Substituting (66) and (67) into our expression for the expected harm as using the notation $\Delta_{y,y'} = \max\{0, U(\bar{a}, x, y) - U(a, x, y')\}$ gives,

$$\mathbb{E}[h|a, x; \mathcal{M}_\pm] = \mathbb{E}[h|a, x; \mathcal{M}_0] + \phi_\pm \left[\Delta_{00} + \Delta_{11} - \Delta_{10} - \Delta_{01}\right] \tag{71}$$

$$\mathbb{E}[h|a', x; \mathcal{M}_\pm] = \mathbb{E}[h|a', x; \mathcal{M}_0]\,, \quad a' \neq a \tag{72}$$

Now, as $\max_y U(a, x, y) > \min_y U(\bar{a}, x, y)$ then $\Delta_{01} = 0$. For the coefficient of $\phi_\pm$ in (71) to be zero, we would therefore require that $\Delta_{00} + \Delta_{11} = \Delta_{10}$. We know $\Delta_{10} > 0$ because otherwise $\min_y U(a, x, y) > \max_y U(\bar{a}, x, y)$, therefore the minimial value of Δ_{10} is $\max_y U(\bar{a}, x, 1) - \min_y U(a, x, y)$. If $\Delta_{00} \neq 0$ and $\Delta_{11} \neq 0$ then $\Delta_{00} + \Delta_{11} \geq \Delta_{10}$ implies $\min_y U(\bar{a}, x, y) \geq \max_y U(a, x, y)$ which violates our assumptions, therefore $\Delta_{00} + \Delta_{11} < \Delta_{10}$. If $\Delta_{00} = 0$ clearly we cannot have $\Delta_{11} = \Delta_{10}$ as $\min_y U(a, x, y) < \max_y U(a, x, y)$ by our assumptions, and likewise if $\Delta_{11} = 0$ we cannot have $\Delta_{00} = \Delta_{10}$ as this would imply $\min_y U(\bar{a}, x, y) = \max_y U(\bar{a}, x, y)$ which violates our assumptions. Therefore we can conclude that the coefficient in (71) in greater than zero.

Therefore if we choose any $0 < \phi_+ < \min\{P(Y_{\bar{a}} = 1|x)P(Y_a = 0|x), P(Y_{\bar{a}} = 0|x)P(Y_a = 1|x)\}$ we get $\mathbb{E}[h|a, x; \mathcal{M}_+] > \mathbb{E}[h|a, x; \mathcal{M}_0]$, and any $\max\{P(Y_{\bar{a}} = 1|x)P(Y_a = 1|x), P(Y_{\bar{a}} = 0|x)P(Y_a = 0|x)\} < \phi_- < 0$, we get $\mathbb{E}[h|a, x; \mathcal{M}_-] < \mathbb{E}[h|a, x; \mathcal{M}_0]$. □

Lemma 3. *For (context dependent) default action $A = \bar{a}(x)$, $\mathbb{E}[h|\bar{a}(x), x; \mathcal{M}] = 0 \,\forall\, \mathcal{M}$*

Proof. In the following we suppress the notation $\bar{a}(x) = \bar{a}$.

$$\begin{aligned}
\mathbb{E}[h|\bar{a}, x; \mathcal{M}] &= \int\limits_{y^*, y} P(Y_{\bar{a}} = y^*, Y = y|\bar{a}, x; \mathcal{M}) \max\{0, U(\bar{a}, x, y^*) - U(\bar{a}, x, y)\} && (73)\\
&= \int\limits_{y^*, y} P(Y_{\bar{a}} = y^*, Y_{\bar{a}} = y|x; \mathcal{M}) \max\{0, U(\bar{a}, x, y^*) - U(\bar{a}, x, y)\} && (74)\\
&= \int\limits_{y^*, y} P(Y_{\bar{a}} = y|x; \mathcal{M}) \delta(y^* - y) \max\{0, U(\bar{a}, x, y^*) - U(\bar{a}, x, y)\} && (75)\\
&= \int\limits_{y} P(Y_{\bar{a}} = y|x; \mathcal{M}) \max\{0, U(\bar{a}, x, y) - U(\bar{a}, x, y)\} && (76)\\
&= 0 && (77)
\end{aligned}$$

$P(y|a, x) = P(y_a|x)$.

□

Theorem 3: For any (context dependent) default action $A = \bar{a}(x)$, if there is a context $X = x$ where the user's utility function is outcome dependent for $\bar{a}(x)$ amd some other action $a \neq \bar{a}(x)$, then there is an outcome distributional shift such that U is harmful in the shifted environment.

Proof. For the expected utility to not be harmful by Definition 6, it must be that $\mathbb{E}[h|a, x] > \mathbb{E}[h|b, x] \implies \mathbb{E}[U|a, x] < \mathbb{E}[U|b, x]$. Given our assumption of outcome dependence, we know there is a context $X = x$ such that the utility functions for $\bar{a}(x)$ and $a \neq \bar{a}(x)$ overlap, that is $\min_y U(a, x, y) < \max_y U(\bar{a}(x), x, y)$ and $\max_y U(a, x, y) > \min_y U(\bar{a}(x), x, y)$. In the following we drop the notation $\bar{a}(x) = \bar{a}$. We can restrict our agent to choose between these two actions and construct an outcome distributional shift such that; i) The outcomes Y_a and $Y_{\bar{a}}$ are binary with one outcome maximizing the utility for that action and the other minimizing the utility, i.e. $Y_a \in \{\max_y U(a, x, y), \min_y U(a, x, y)\}$ and $Y_{\bar{a}} \in \{\max_y U(\bar{a}, x, y), \min_y U(\bar{a}, x, y)\}$, ii) $\mathbb{E}[U|a, x] = \mathbb{E}[U|\bar{a}, x]$, iii) $P(y_a|x)$ and $P(y_{\bar{a}}|x)$ are non-deterministic. This follows from the fact that the set of possible expected utility values for an action a is the set of mixtures over $U(a, x, y)$

with respect to y, and as $Y_a = 0, 1$ are the extremal points of this convex set, the expected utility for action a in context x can be written as $P(Y_a = 0|x)U(a, x, 0) + P(Y_a = 1|x)U(a, x, 1)$. Then, as the utility functions for a and $\bar{a}$ overlap there is point in the intersection of these convex sets that is non-extremal (and hence, a non-deterministic mixture).

By Lemma 3 the default action causes zero expected harm. By Lemma 2 we can construct a shifted environment $\mathcal{M}_0$ where the non-default action $a \neq \bar{a}$ has non-zero harm for any non-deterministic $P(y_a|x)$. We can therefore construct $\mathcal{M}_0$ such that i) $\mathbb{E}[Y_a|x] = \mathbb{E}[Y_{\bar{a}}|x]$, and ii) $\mathbb{E}[h|a, x] > \mathbb{E}[h|\bar{a}, x]$, violating our requirement that $\mathbb{E}[h|a, x] > \mathbb{E}[h|b, x] \implies \mathbb{E}[U|a, x] < \mathbb{E}[U|b, x]$.

□

Theorem 4: For any (context dependent) default action $A = \bar{a}(x)$, if there is a context $X = x$ where the user's utility function is outcome dependent for $\bar{a}(x)$ and two other actions $a_1, a_2 \neq \bar{a}(x)$, then for any factual objective function J there is an outcome distributional shift such that maximizing the J is harmful in the shifted environment.

Proof. By assumption there is a context $X = x$ for which the utility functions for a_1, a_2 and $\bar{a}(x)$ overlap. In the following we drop the notation $\bar{a}(x) = \bar{a}$. There is a choice of non-deterministic outcome distributions $P(y_{\bar{a}}|x)$, $P(y_{a_1}|x)$ and $P(y_{a_2}|x)$ such that all three actions have the same expected utility. By Lemma 2 for any non-deterministic outcome distribution we can choose $\mathcal{M}_0$ such that $\mathbb{E}[h|a_1, x; \mathcal{M}_0] > 0$, and $\mathbb{E}[h|a_2, x; \mathcal{M}_0] > 0$, and by Lemma 3 $E[h|\bar{a}, x; \mathcal{M}] = 0 \; \forall \; \mathcal{M}$. Therefore $\exists \; \mathcal{M}_0$ that is an outcome distributional shift of the original environment $\mathcal{M}$ such that $\bar{a}, a_1, a_2$ have the same expected utility, $\bar{a}$ has zero expected harm and a_1, a_2 have non-zero expected harm.

If $\mathbb{E}[h|a_1, x\mathcal{M}_0] = \mathbb{E}[h|a_2, x; \mathcal{M}_0]$ then by Lemma 2 there are outcome-shifted environments $\mathcal{M}_\pm$ such that $\bar{a}, a_1$ and a_2 have the same factual statistics as in $\mathcal{M}_0$ and $\mathbb{E}[h|a_2, x\mathcal{M}_0] = \mathbb{E}[h|a_2, x\mathcal{M}_\pm]$, but the harm caused by a_1 is increased(descreased) by some non-zero amount. Therefore in $\mathcal{M}_+$ a_1 and a_2 have the same expected utility but a has a strictly higher expected harm, and in order to be non-harmful it must be that $\mathbb{E}[J|a_1, x; \mathcal{M}_+] < \mathbb{E}[J|a_2, x; \mathcal{M}_+]$. Likewise in $\mathcal{M}_-$ a_1 and a_2 have the same expected utility but the expected harm for a_1 is strictly lower than for a_2, therefore in order to be non-harmful it must be that $\mathbb{E}[J|a_1, x; \mathcal{M}_-] > \mathbb{E}[J|a_2, x; \mathcal{M}_-]$. Finally we note that $\mathbb{E}[J|a, x; \mathcal{M}_+] = \mathbb{E}[J|a, x; \mathcal{M}_-] = \mathbb{E}[J|a, x; \mathcal{M}_0] \; \forall \; a \in A$ as the factual statistics are identical in $\mathcal{M}_0, \mathcal{M}_\pm$, i.e. $P(y_a|x; \mathcal{M}_+) = P(y_a|x; \mathcal{M}_-) = P(y_a|x; \mathcal{M}_0)$. Therefore any J must be harmful in either $\mathcal{M}_+$ and $\mathcal{M}_-$, and therefore there is an outcome distributional shift $\mathcal{M} \to \mathcal{M}_+$ or $\mathcal{M} \to \mathcal{M}_-$ such that J is harmful in the shifted environment.

If $\mathbb{E}[h|a_1, x\mathcal{M}_0] \neq \mathbb{E}[h|a_2, x; \mathcal{M}_0]$, assume without loss of generality that $\mathbb{E}[h|a_1, x\mathcal{M}_0] > \mathbb{E}[h|a_2, x; \mathcal{M}_0]$. As $\bar{a}, a_1$ and a_2 have the equal expected utilities then so does any mixture of these actions, in $\mathcal{M}_0$ and $\mathcal{M}_\pm$. Restrict the agent to choose between action a_2 and a mixture of actions $\bar{a}$ and a_1—i.e. a stochastic or 'soft' intervention [17, 66], which involves replacing the causal mechanism for A with a mixture $\tau := q[A = a_1] + (1 - q)[A = a_0]$ where q is an independent binary noise term. By linearity the expected utility for this mixed action is $\mathbb{E}[U_\tau|x] = q\mathbb{E}[U_{a_1}|x] + (1-q)E[U_{\bar{a}}|x] = \mathbb{E}[U_{a_1}|x]$ as all three actions have the same expected utility, and has an expected harm $\mathbb{E}[h|\tau, x; \mathcal{M}_0] = q\mathbb{E}[h|a_1, x; \mathcal{M}] + (1-q)\mathbb{E}[h|\bar{a}, x; \mathcal{M}] = q\mathbb{E}[h|a_1, x; \mathcal{M}]$ as $\mathbb{E}[h|\bar{a}, x; \mathcal{M}] = 0 \; \forall \; \mathcal{M}$. Therefore as $\mathbb{E}[h|a_1, x; \mathcal{M}] > 0$ and $\mathbb{E}[h|a_2, x; \mathcal{M}] > 0$ we can choose $p > 0$ such that $\mathbb{E}[h|\tau, x; \mathcal{M}_0] = \mathbb{E}[h|a_2, x; \mathcal{M}_0]$. Therefore in $\mathcal{M}_0$, a_2 and τ have the same expected harm and utility, and in $\mathcal{M}_+$ they have the same expected utility but τ is more harmful than a_2 as $\mathbb{E}[h|a_1, x; \mathcal{M}_+] > \mathbb{E}[h|a_1, x; \mathcal{M}_0]$ and $p > 0$, and in $\mathcal{M}_-$ they have the same expected utility but a_2 is more harmful that τ. As the factual statistics $P(y_a|x)$ are identical for $\mathcal{M}_0$ and $\mathcal{M}_\pm$, so is the value of any factual objective function across all three environments. Hence, any factual objective function must be harmful in either $\mathcal{M}_+$ or $\mathcal{M}_-$. □

K

In this appendix we discuss related works; counterfactual fairness [48] and path-specific objectives [26], as well as discussing some deep learning implementations that are capable of supporting counterfactual inferences of the type used to estimate counterfactual harm. For the sake of generality our results are derived in the SCM framework, and so taken at face value they assume knowledge of the SCM for the data generating process. Often in complex domains we will not have access to the true SCM that describes the data generating process but some approximation. However, there have been several recent proposals for performing counterfactual inference using deep learning methods, with promising results in diverse complex domains including learning deep structural causal models for medical imaging [65], visual question answering [61], and vision-and-language navigation in robotics [64] and text generation [54]. These studies evidence that deep learning algorithms can learn to make good counterfactual inferences that can be used to support decision making. This is achieved often without perfect knowledge of the underlying SCM (one notable exception being when the environment is simulated). This is somewhat analogous to the fact that human decision making often utilizes counterfactual reasoning for various cognitive tasks [23] (for example, it is important for legal and ethical reasoning [50]). This is in spite of the fact that humans clearly do not having access to perfect structural causal models of their environments, but have to learn good enough approximations through heuristics and inductive biases. While it is known that counterfactuals cannot be identified from data alone [79] but are only defined up to a structural causal models of the environment, clearly humans [32] and increasingly AI systems are capable of learning good structural causal models of real-world environments and using these to make counterfactual inferences capable of guiding actions and reasoning about harm.

K.1 Related work

Counterfactual fairness deals with prediction tasks $\hat{Y} : X \to Y$ where the desire is to have a predictor $\hat{Y}$ that is not unfairly influenced by a protected attribute A such as gender or race. Note A is a feature that typically cannot be intervened on, whereas is our setup A denotes an agent's action. Counterfactual fairness quantifies this unfair influence causally, using the counterfactual constraint,

$$P(\hat{Y}_a = y | X = x, A = a) = P(\hat{Y}_{a'} = y | X = x, A = a) \quad \forall\, a' \in A,\, y \in \hat{Y} \tag{78}$$

which states that the probability of predicting any given outcome should not be caused on average by the protected attribute A, where type causation is established using the counterfactual $P(\hat{Y}_{a'} = y | X = x, A = a)$ which is the probability of $\hat{Y}$ given $A = a$ if A had been equal to a'. Note that the counterfactual in (78) does not deal with the joint statistics of the factual outcome $\hat{Y}_a$ and the counterfactual outcome $\hat{Y}_{a'}$, as so is an example of type causality compared to harm which an example of actual causality [34]. Harm is conceptually distinct from fairness—for example, it is possible to apply a needlessly harmful action fairly—but the two measures can be used in tandem. For example, one could quantify if a action or decision was unfair, and whether or not the user was harmed due to this unfair action.

Another perhaps more related use of counterfactual inference for ethical AI is path-specific objectives [26]. This work similarly refines expected utility theory in the CID framework to take into account the fact that we often want to maximize utility via specific causal pathways due to ethical constraints. For example we can consider a simple model where the agent's action A influences user feedback Y (and utility $U(y)$) but also effects the users preferences H where $A \to Y$, $A \to H$ and $H \to Y$. To maximize utility without intentionally manipulating the user we must maximize along the causal pathway $(1) : A \to Y$ without including contributions to the expected utility from the mediator pathway $(2) : A \to H \to Y$. This involves replacing the expected utility with its path-specific equivalent, much as our path-specific harm (Definition 9) generalizes our path-independent definition of harm (Definition 3). As such the path-specific expected utility is still agnostic to harm just as the expected utility is, although it could be combined with the path specific harm in [26] to give a path-specific variant of the HPU (Definition 4). This would allow for harm averse decision making where the necessary degree of harm-aversion $\lambda_{(i)}$ differs depending on the causal path (i)—for example, if we desire agents that have a high aversion for being directly harmful, but a lower degree of harm-aversion for indirect harm mediated by the actions of other agents (as described in Appendix B).