Explaining Image Classifiers

Hana Chockler[1] , **Joseph Y. Halpern**[2]
[1]King's College, London, U.K
[2]Cornell University, Ithaca, New York, USA
hana.chockler@kcl.ac.uk, halpern@cs.cornell.edu

Abstract

We focus on explaining image classifiers, taking the work of Mothilal et al. (2021) (MMTS) as our point of departure. We observe that, although MMTS claim to be using the definition of explanation proposed by Halpern (2016), they do not quite do so. Roughly speaking, Halpern's definition has a necessity clause and a sufficiency clause. MMTS replace the necessity clause by a requirement that, as we show, implies it. Halpern's definition also allows agents to restrict the set of options considered. While these difference may seem minor, as we show, they can have a nontrivial impact on explanations. We also show that, essentially without change, Halpern's definition can handle two issues that have proved difficult for other approaches: explanations of absence (when, for example, an image classifier for tumors outputs "no tumor") and explanations of rare events (such as tumors).

1 Introduction

Black-box AI systems and, in particular Deep Neural Networks (DNNs), are now a primary building block of many computer vision systems. DNNs are complex non-linear functions with algorithmically generated coefficients. In contrast to traditional image-processing pipelines, it is difficult to retrace how the pixel data are interpreted by the layers of a DNN. This "black box" nature of DNNs creates a demand for *explanations*. A good explanation should answer the question "Why did the neural network classify the input the way it did?" By doing so, it can increase a user's confidence in the result. Explanations are also useful for determining whether the system is working well; if the explanation does not make sense, it may indicate that there is a problem with the system.

Unfortunately, it is not clear how to define what an explanation is, let alone what a *good* explanation is. There have been a number of definitions of explanation given by researchers in various fields, particularly computer science (Chajewska and Halpern 1997; Halpern 2016; Halpern and Pearl 2005b; Pearl 1988) and philosophy (Gärdenfors 1988; Hempel 1965; Salmon 1970; Salmon 1989; Woodward 2014), and a number of attempts to provide explanations for the output of DNNs (Ribeiro, Singh, and Guestrin 2016; Selvaraju et al. 2017; Lundberg and Lee 2017; Sun et al. 2020; Chockler, Kroening, and Sun 2021) ((Molnar 2022) provides an overview). Here we focus on one particular definition of explanation, that was given by Halpern (2016), which is in turn based on a definition due to Halpern and Pearl (2005b). Mothilal et al. (2021) (MMTS from now on) already showed that this definition could be usefully applied to better understand and evaluate what MMTS called *attribution-based* explanation approaches, such as LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017), which provide a score or ranking over features, conveying the (relative) importance of each feature to the model's output, and contrast them with what they called *counterfactual-based* approaches, such as DICE (Mothilal, Tan, and Sharma 2020) and that of Wachter et al. (2017), which generate examples that yield a different model output with minimum changes in the input features.

In this paper, we take MMTS as our point of departure and focus on explaining image classifiers. We first observe that, although MMTS claim to be using Halpern's definition, they do not quite do so. Roughly speaking, Halpern's definition (which we discuss in detail in Section 2) has a necessity clause and a sufficiency clause. MMTS replace the necessity clause by a requirement that, as we show, implies it. Halpern's definition also allows agents to restrict the set of options considered. While these difference may seem minor, as we show, they can have a nontrivial impact on explanations. We also show that, essentially without change, Halpern's definition can handle two issues that have proved difficult for other approaches: explanations of absence (when, for example, an classifier for tumors outputs "no tumor") and explanations of rare events (again, a classifier for tumors can be viewed as an example; a tumor is a relatively rare event). The upshot of this discussion is that, while the analysis of MMTS shows that a simplification of Halpern's definition can go a long way to helping us understand notions of explanation used in the literature, we can go much further by using Halpern's actual definition, while still retaining the benefits of the MMTS analysis. MMTS conduct an empirical evaluation of the main approaches to causal explanations. Their results apply essentially without change to this paper, apart from the new work discussed in Section 4.

We note that the problem of actual causality, and hence also the problem of computing explanations, are intractable (Halpern 2016). For image classifiers, which are the topic of this paper, the models are large, so brute-

force computation is infeasible for all but very small images. However, efficient approximation algorithms compute explanations that are very close to the precise ones for all but very convoluted inputs; we mentioned these tools above (see also the analysis in MMTS). Of particular interest is the work of Chockler et al. 2021, who compute an efficient approximation of an explanation, treating the classifier as a black-box causal model. We thus believe that, in practice, complexity considerations will not be an obstacle to using these tools.

The rest of the paper is organized as follows: In Section 2, we review the relevant definitions of causal models and explanations; in Section 3, we discuss explanations of image classifiers and show their relation to explanations in actual causality; and in Section 4, we discuss explanations of absence and of rare events.

2 Causal Models and Relevant Definitions

In this section, we review the definition of causal models introduced by Halpern and Pearl (2005a) and relevant definitions of causes and explanations given by Halpern (2016). The material in this section is largely taken from (Halpern 2016).

We assume that the world is described in terms of variables and their values. Some variables may have a causal influence on others. This influence is modeled by a set of *structural equations*. It is conceptually useful to split the variables into two sets: the *exogenous* variables, whose values are determined by factors outside the model, and the *endogenous* variables, whose values are ultimately determined by the exogenous variables. The structural equations describe how these values are determined.

Formally, a *causal model* M is a pair $(\mathcal{S}, \mathcal{F})$, where $\mathcal{S}$ is a *signature*, which explicitly lists the endogenous and exogenous variables and characterizes their possible values, and $\mathcal{F}$ defines a set of *(modifiable) structural equations*, relating the values of the variables. A signature $\mathcal{S}$ is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is a set of exogenous variables, $\mathcal{V}$ is a set of endogenous variables, and $\mathcal{R}$ associates with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y (i.e., the set of values over which Y *ranges*). For simplicity, we assume here that $\mathcal{V}$ is finite, as is $\mathcal{R}(Y)$ for every endogenous variable $Y \in \mathcal{V}$. $\mathcal{F}$ associates with each endogenous variable $X \in \mathcal{V}$ a function denoted F_X (i.e., $F_X = \mathcal{F}(X)$) such that $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y)) \rightarrow \mathcal{R}(X)$. This mathematical notation just makes precise the fact that F_X determines the value of X, given the values of all the other variables in $\mathcal{U} \cup \mathcal{V}$. If there is one exogenous variable U and three endogenous variables, X, Y, and Z, then F_X defines the values of X in terms of the values of Y, Z, and U. For example, we might have $F_X(u, y, z) = u + y$, which is usually written as $X = U + Y$. Thus, if $Y = 3$ and $U = 2$, then $X = 5$, regardless of how Z is set.[1]

[1]The fact that X is assigned $U + Y$ (i.e., the value of X is the sum of the values of U and Y) does not imply that Y is assigned $X - U$; that is, $F_Y(U, X, Z) = X - U$ does not necessarily hold.

The structural equations define what happens in the presence of external interventions. Setting the value of some variable X to x in a causal model $M = (\mathcal{S}, \mathcal{F})$ results in a new causal model, denoted $M_{X \leftarrow x}$, which is identical to M, except that the equation for X in $\mathcal{F}$ is replaced by $X = x$.

We can also consider *probabilistic causal models*; these are pairs $(M, \Pr)$, where M is a causal model and $\Pr$ is a probability on the contexts in M.

The dependencies between variables in a causal model $M = ((\mathcal{U}, \mathcal{V}, \mathcal{R}), \mathcal{F})$ can be described using a *causal network* (or *causal graph*), whose nodes are labeled by the endogenous and exogenous variables in M, with one node for each variable in $\mathcal{U} \cup \mathcal{V}$. The roots of the graph are (labeled by) the exogenous variables. There is a directed edge from variable X to Y if Y *depends on* X; this is the case if there is some setting of all the variables in $\mathcal{U} \cup \mathcal{V}$ other than X and Y such that varying the value of X in that setting results in a variation in the value of Y; that is, there is a setting $\vec{z}$ of the variables other than X and Y and values x and x' of X such that $F_Y(x, \vec{z}) \neq F_Y(x', \vec{z})$.

A causal model M is *recursive* (or *acyclic*) if its causal graph is acyclic. It should be clear that if M is an acyclic causal model, then given a *context*, that is, a setting $\vec{u}$ for the exogenous variables in $\mathcal{U}$, the values of all the other variables are determined (i.e., there is a unique solution to all the equations). In this paper, following the literature, we restrict to recursive models.

We call a pair $(M, \vec{u})$ consisting of a causal model M and a context $\vec{u}$ a *(causal) setting*. A causal formula ψ is true or false in a setting. We write $(M, \vec{u}) \models \psi$ if the causal formula ψ is true in the setting $(M, \vec{u})$. The $\models$ relation is defined inductively. $(M, \vec{u}) \models X = x$ if the variable X has value x in the unique (since we are dealing with acyclic models) solution to the equations in M in context $\vec{u}$ (i.e., the unique vector of values for the exogenous variables that simultaneously satisfies all equations in M with the variables in $\mathcal{U}$ set to $\vec{u}$). Finally, $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\varphi$ if $(M_{\vec{Y} = \vec{y}}, \vec{u}) \models \varphi$, where $M_{\vec{Y} \leftarrow \vec{y}}$ is the causal model that is identical to M, except that the equations for variables in $\vec{Y}$ in $\mathcal{F}$ are replaced by $Y = y$ for each $Y \in \vec{Y}$ and its corresponding value $y \in \vec{y}$.

A standard use of causal models is to define *actual causation*: that is, what it means for some particular event that occurred to cause another particular event. There have been a number of definitions of actual causation given for acyclic models (e.g., (Beckers 2021; Glymour and Wimberly 2007; Hall 2007; Halpern and Pearl 2005a; Halpern 2016; Hitchcock 2001; Hitchcock 2007; Weslake 2015; Woodward 2003)). In this paper, we focus on what has become known as the *modified* Halpern-Pearl definition and some related definitions introduced by Halpern (2016). We briefly review the relevant definitions below (see (Halpern 2016) for more intuition and motivation).

The events that can be causes are arbitrary conjunctions of primitive events (formulas of the form $X = x$); the events that can be caused are arbitrary Boolean combinations of primitive events.

Definition 1. *[Actual cause]* $\vec{X} = \vec{x}$ *is an* actual cause *of* φ *in* $(M, \vec{u})$ *if the following three conditions hold:*

AC1. $(M, \vec{u}) \models (\vec{X} = \vec{x})$ *and* $(M, \vec{u}) \models \varphi$.

AC2. *There is a a setting* $\vec{x}'$ *of the variables in* $\vec{X}$*, a (possibly empty) set* $\vec{W}$ *of variables in* $\mathcal{V} - \vec{X}'$*, and a setting* $\vec{w}$ *of the variables in* $\vec{W}$ *such that* $(M, \vec{u}) \models \vec{W} = \vec{w}$ *and* $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}]\neg\varphi$*, and moreover*

AC3. $\vec{X}$ *is minimal; there is no strict subset* $\vec{X}'$ *of* $\vec{X}$ *such that* $\vec{X}' = \vec{x}''$ *can replace* $\vec{X} = \vec{x}'$ *in AC2, where* $\vec{x}''$ *is the restriction of* $\vec{x}'$ *to the variables in* $\vec{X}'$.

AC1 just says that $\vec{X} = \vec{x}$ cannot be considered a cause of φ unless both $\vec{X} = \vec{x}$ and φ actually holds. AC3 is a minimality condition, which says that a cause has no irrelevant conjuncts. AC2 extends the standard but-for condition ($\vec{X} = \vec{x}$ is a cause of φ if, had $\vec{X}$ been $\vec{x}'$, φ would have been false) by allowing us to apply it while keeping some variables fixed to the value that they had in the actual setting $(M, \vec{u})$. In the special case that $\vec{W} = \emptyset$, we get the but-for definition.

To define explanation, we need the notion of *sufficient cause* in addition to that of actual cause.

Definition 2. *[Sufficient cause]* $\vec{X} = \vec{x}$ *is a* sufficient cause *of* φ *in* $(M, \vec{u})$ *if the following four conditions hold:*

SC1. $(M, \vec{u}) \models (\vec{X} = \vec{x})$ *and* $(M, \vec{u}) \models \varphi$.

SC2. *Some conjunct of* $\vec{X} = \vec{x}$ *is part of an actual cause of* φ *in* $(M, \vec{u})$*. More precisely, there exists a conjunct* $X = x$ *of* $\vec{X} = \vec{x}$ *and another (possibly empty) conjunction* $\vec{Y} = \vec{y}$ *such that* $X = x \wedge \vec{Y} = \vec{y}$ *is an actual cause of* φ *in* $(M, \vec{u})$.

SC3. $(M, \vec{u}') \models [\vec{X} = \vec{x}]\varphi$ *for all contexts* $\vec{u}' \in \mathcal{R}(\mathcal{U})$.

SC4. $\vec{X}$ *is minimal; there is no strict subset* $\vec{X}'$ *of* $\vec{X}$ *such that* $\vec{X}' = \vec{x}'$ *satisfies conditions SC1, SC2, and SC3, where* $\vec{x}'$ *is the restriction of* $\vec{x}$ *to the variables in* $\vec{X}'$.

Note that this definition of sufficient cause (which is taken from (Halpern 2016)) is quite different from that in (Halpern and Pearl 2005a). Like the necessity clause used by MMTS, the definition of (Halpern and Pearl 2005a) requires only that some subset of $\vec{X} = \vec{x}$ be an actual cause of φ, without allowing the subset to be extended by another conjunction $\vec{Y} = \vec{y}$ (and uses a different definition of actual cause—that of (Halpern and Pearl 2005b)), but (somewhat in the spirit of SC3) requires that this necessity condition hold in all contexts. It has no exact analogue of SC3 at all. An example might help clarify the definition.

Suppose that we have a dry forest and three arsonists. There are three contexts: in u_1, it takes just one dropped match to burn the forest down, but arsonist 1 and arsonist 2 drop matches; in u_2, it takes just one dropped match to burn the forest down, and all three arsonists drop a match; finally, in u_3, it takes two dropped matches to burn the forest down, and arsonists 1 and 3 drop matches. To model this, we have binary variables ML_1, ML_2, and ML_3, denoting which arsonist drops a match, and FB denoting, whether the forest burns down. (Note that we use the structural equation for FB to capture these differences; for example, if x_i is the value of ML_i, then $F_{FB}(x_1, x_2, x_3, u_1) = 1$ iff at least one of x_1, x_2, and x_3 is 1, and $F_{FB}(x_1, x_2, x_3, u_3) = 1$ iff at least two of x_1, x_2, and x_3 are 1.) We claim that $ML_1 = 1 \wedge ML_2 = 1$ is a sufficient cause of $FB = 1$ in (M, u_1) and (M, u_2), but not (M, u_3). To see that it is not a sufficient cause in (M, u_3), note that SC1 does not hold: arsonist 2 does not drop a match. It is also easy to see that $(M, u_i) \models [ML_1 = 1 \wedge ML_2 = 1](FB = 1)$ for $i = 1, 2, 3$, so SC3 holds. Now in (M, u_1), $ML_1 \wedge ML_2 = 1$ is an actual cause of $FB = 1$ (since the values of both ML_1 and ML_2 have to be changed in order to change the value of FB). Similarly, in (M, u_2), $ML_1 \wedge ML_2 = 1 \wedge ML_3 = 1$ is an actual cause of $FB = 1$ (so we get SC2 by taking $Y = ML_3$). Thus, SC2 holds in both (M, u_1) and (M, u_2).

While SC3 is typically taken to be the core of the sufficiency requirement, to show causality, we also need SC2, since it requires that $\neg\varphi$ hold for some setting of the variables. We might hope that if there were a setting where φ was false, SC3 would imply SC2. As the following example shows, this is not the case in general.

Example 1. *Consider a causal model* M *with three binary variables* A*,* B*, and* C*, and the structural equations* $A = B$ *and* $C = A \vee (\neg A \wedge B)$*. Let* $\vec{u}$ *be a context in which all variables are set to* 1*, and let* φ *be the formula* $C = 1$*. In context* $\vec{u}$*,* $A = 1$ *satisfies SC1, SC3, and SC4 for* φ*. There is also some setting of the variables for which* $C = 0$ *:* $(M, \vec{u}) \models [A \leftarrow 0, B \leftarrow 0](C = 0)$*. However,* $A = 1$ *does not satisfy SC2. Indeed,* $A = 1$ *by itself is not an actual cause of* $C = 1$*, as* B*=1 holds as well, nor is* $A = 1$ *part of an actual cause, as* $B = 1$ *is already an actual cause of* $C = 1$.

On the other hand, if there are no dependencies between the variables and some other assumptions made by MMTS in their analysis of image classifiers hold, then, roughly speaking, SC1, SC3, and SC4 do imply SC2. To make this precise, we need two definitions.

Definition 3. *The variables in a set* $\vec{X}$ *of endogenous variables are* causally independent *in a causal model* M *if, for all contexts* $\vec{u}$*, all strict subsets* $\vec{Y}$ *of* $\vec{X}$*, all assignments* $\vec{y}$ *to* $\vec{Y}$*, and all* $Z \in \vec{X} - \vec{Y}$*,* $(M, \vec{u}) \models Z = z$ *iff* $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}](Z = z)$.

Intuitively, the variables in $\vec{X}$ are causally independent if setting the values of some subset of the variables in $\vec{X}$ has no impact on the values of the other variables in $\vec{X}$.

Definition 4. $\vec{X}$ *is* determined by the context *if for each setting* $\vec{x}$ *of* $\vec{X}$*, there is a context* $\vec{u}_{\vec{x}}$ *such that* $(M, \vec{u}_{\vec{x}}) \models X = x$.

Theorem 1. *Given a set* $\vec{X}'$ *of endogenous variables in a causal model* M *such that (a) the variables in* $\vec{X}'$ *are causally independent, (b)* $\vec{X}'$ *is determined by the context, (c)* $\vec{X}'$ *includes all the parents of the variables in* φ*, (d) there is some setting* $\vec{x}'$ *of the variables in* $\vec{X}'$ *that makes* φ *false in*

context $\vec{u}$ (i.e., $(M, \vec{u}) \models [\vec{X}' \leftarrow \vec{x}']\neg\varphi$), and (e) $\vec{X} \subseteq \vec{X}'$, then $\vec{X} = \vec{x}$ is a sufficient cause of φ in $(M, \vec{u})$ iff it satisfies SC1, SC3, and SC4 (i.e., SC2 follows).

Proof. By assumption, there is some setting $\vec{x}'$ such that $\vec{X}'$ that makes φ false in context $\vec{u}$ (i.e., $(M, \vec{u}) \models [\vec{X}' \leftarrow \vec{x}']\neg\varphi$). Choose $\vec{x}'$ to be such a setting that differs minimally from the values that variables get in $\vec{u}$; that is, the set of variables $Y \in \vec{X}$ such that $(M, \vec{u}) \models Y = y$, and the value of Y in $\vec{x}'$ is different from Y is minimal. Let $\vec{Y}$ be the set of variables in this minimal set. Let $\vec{u}_{\vec{x}'}$ be a context such that $(M, \vec{u}_{\vec{x}'}) \models \vec{X}' = \vec{x}'$; by assumption, such a context exists. Since $\vec{X}'$ includes all the parents of the variables in φ, we must have $(M, \vec{u}_{\vec{x}'}) \models [\vec{X}' \leftarrow \vec{x}']\neg\varphi$. Since $(M, \vec{u}_{\vec{x}'}) \models \vec{X}' \leftarrow \vec{x}'$, it follows that $(M, \vec{u}_{\vec{x}'}) \models \neg\varphi$.

$\vec{Y}$ must contain a variable in $\vec{X}$. For suppose not. By SC1, $(M, \vec{u}) \models \vec{X} = \vec{x}$, so if Y does not contain a variable in X, the values that the variables in $\vec{X}$ get in the setting $\vec{X}' = \vec{x}'$ is the same as their value in $(M, \vec{u})$. Thus, $(M, \vec{u}_{\vec{x}'}) \models \vec{X} = \vec{x}$. By SC3, $(M, \vec{u}_{\vec{x}'}) \models [\vec{X} = \vec{x}]\varphi$, from which it would follow that $(M, \vec{u}_{\vec{x}'}) \models \varphi$, a contradiction.

Let $\vec{y}'$ be the restriction of $\vec{x}'$ to the variables in $\vec{Y}$, and let $\vec{y}$ be the values of the variables in $\vec{Y}$ in $(M, \vec{u})$, that is $(M, \vec{u}) \models \vec{Y} = \vec{y}$. We claim that $\vec{Y} = \vec{y}$ is a cause of φ in $(M, \vec{u})$. By assumption, $(M, \vec{u}) \models \vec{Y} = \vec{y}$; by SC1, $(M, \vec{u}) \models \varphi$. Thus, AC1 holds. By construction, $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}']\neg\varphi$, so AC2 holds (with $\vec{W} = \emptyset$). Finally, suppose that $\vec{W}$ is such that $(M, \vec{u}) \models \vec{W} = \vec{w}$ and for some subset $\vec{Y}'$ of $\vec{Y}$, we have that $(M, \vec{u}) \models [\vec{Y}' \leftarrow \vec{y}'', \vec{W} \leftarrow \vec{w}]\neg\varphi$, where $\vec{y}''$ is the restriction of $\vec{y}'$ to the variables in $\vec{Y}'$. We can assume that $\vec{W}$ is a subset of $\vec{X}'$, since $\vec{X}'$ includes all the parents of the variables in φ. By causal independence, $(M, \vec{u}) \models [\vec{Y}' \leftarrow \vec{y}'](\vec{W} = \vec{w})$. Thus, $(M, \vec{u}) \models [\vec{Y}' \leftarrow \vec{y}'']\neg\varphi$. But this contradicts the minimality of $\vec{Y}$. It follows that AC3 holds.

We have shown that $\vec{Y} = \vec{y}$ is a cause of φ in $(M, \vec{u})$. Since $\vec{Y}$ includes a variable in $\vec{X}$, it follows that some conjunct of $\vec{X} = \vec{x}$ is part of a cause of φ in $(M, \vec{u})$, so SC2 holds, as desired. ∎

The notion of explanation builds on the notion of sufficient causality, and is relative to a set of contexts.

Definition 5. *[Explanation] $\vec{X} = \vec{x}$ is an* explanation *of φ relative to a set $\mathcal{K}$ of contexts in a causal model M if the following conditions hold:*

EX1. *$\vec{X} = \vec{x}$ is a sufficient cause of φ in all contexts in $\mathcal{K}$ satisfying $(\vec{X} = \vec{x}) \wedge \varphi$. More precisely,*

- *If $\vec{u} \in \mathcal{K}$ and $(M, \vec{u}) \models (\vec{X} = \vec{x}) \wedge \varphi$, then there exists a conjunct $X = x$ of $\vec{X} = \vec{x}$ and a (possibly empty) conjunction $\vec{Y} = \vec{y}$ such that $X = x \wedge \vec{Y} = \vec{y}$ is an actual cause of φ in $(M, \vec{u})$. (This is SC2 applied to all contexts $\vec{u} \in \mathcal{K}$ where $(\vec{X} = \vec{x}) \wedge \varphi$ holds.)*
- *$(M, \vec{u}') \models [\vec{X} = \vec{x}]\varphi$ for all contexts $\vec{u}' \in \mathcal{K}$. (This is SC3 restricted to the contexts in $\mathcal{K}$.)*

EX2. *$\vec{X}$ is minimal; there is no strict subset $\vec{X}'$ of $\vec{X}$ such that $\vec{X}' = \vec{x}'$ satisfies EX1, where $\vec{x}'$ is the restriction of $\vec{x}$ to the variables in $\vec{X}'$. (This is SC4).*

EX3. *$(M, u) \models \vec{X} = \vec{x} \wedge \varphi$ for some $u \in \mathcal{K}$.*

Note that this definition of explanation (which is taken from (Halpern 2016)) is quite different from that in (Halpern and Pearl 2005a). What is called EX2 in (Halpern and Pearl 2005a) is actually an analogue of the first part of EX1 here; although it is called "sufficient causality", it is closer to the necessity condition. But, like the necessity clause used by MMTS, it requires only that some subset of $\vec{X} = \vec{x}$ be an actual cause of φ (without allowing the subset to be extended by another conjunction $\vec{Y} = \vec{y}$) (and uses a different definition of actual cause—that of (Halpern and Pearl 2005b)). The definition of (Halpern and Pearl 2005a) has no analogue to the second part of EX1 here.

Of course, if the assumptions of Theorem 1 hold, then we can drop the requirement that the first part of EX1 holds. (Note that if the assumptions of Theorem 1 hold for some context, they hold for all contexts; in particular, assumption (d) holds, since $\vec{X}'$ includes all the parents of the variables in φ.) Although the changes made by MMTS to Halpern's definition seem minor, they are enough to prevent Theorem 1 from holding. We show this in Example 3 in the next section, after we have discussed the assumptions made by MMTS in more detail.

The requirement that the first part of condition EX1 as given here holds in all contexts in $\mathcal{K}$ that satisfy $\vec{X} = \vec{x} \wedge \varphi$ and that the second part holds in all contexts in $\mathcal{K}$ is quite strong, and often does not hold in practice. We are often willing to accept $\vec{X} = \vec{x}$ as an explanation if these requirements hold with high probability. Given a set $\mathcal{K}$ of contexts in a causal model M, let $\mathcal{K}_\psi$ consist of all contexts $\vec{u}$ in $\mathcal{K}$ such that $(M, \vec{u}) \models \psi$, and let $\mathcal{K}(\vec{X} = \vec{x}, \varphi, \text{SC2})$ consist of all contexts $\vec{u} \in \mathcal{K}$ that satisfy $\vec{X} = \vec{x} \wedge \varphi$ and the first condition in EX1 (i.e., the analogue of SC2).

Definition 6. *[Partial Explanation] $\vec{X} = \vec{x}$ is a* partial explanation *of φ with goodness (α, β) relative to $\mathcal{K}$ in a probabilistic causal model $(M, \Pr)$ if*

EX1′. *$\alpha \leq Pr(\mathcal{K}(\vec{X} = \vec{x}, \varphi, SC2) \mid \mathcal{K}_{\vec{X}=\vec{x}\wedge\varphi})$ and $\beta \leq Pr(\mathcal{K}_{[\vec{X}=\vec{x}]\varphi})$.*

EX2′. *$\vec{X}$ is minimal; there is no strict subset $\vec{X}'$ of $\vec{X}$ such that $\alpha \leq Pr(\mathcal{K}(\vec{X}' = \vec{x}', \varphi, SC2) \mid \mathcal{K}_{\vec{X}'=\vec{x}'\wedge\varphi})$ and $\beta \leq Pr(\mathcal{K}_{[\vec{X}'=\vec{x}']\varphi})$, where x' is the restriction of $\vec{x}$ to the variables in $\vec{X}$.*

EX3′. *$(M, u) \models \vec{X} = \vec{x} \wedge \varphi$ for some $u \in \mathcal{K}$.*

Theorem 1 has no obvious counterpart for partial explanations. The problem is that the conjuncts in EX1′ can be satisfied by different contexts in the set $\mathcal{K}$, while still satisfying the probabilistic constraint. (The theorem would hold with $\alpha = \beta$ if both conjuncts were satisfied by the same subset of $\mathcal{K}$.)

3 Using Causality to Explain Image Classification

Following MMTS, we view an image classifier (a neural network) as a probabilistic causal model. MMTS make a number of additional assumptions in their analysis. Specifically, MMTS take the endogenous variables to be the set $\vec{V}$ of pixels that the image classifier gets as input, together with an output variable that we call O. The variable $V_i \in \vec{V}$ describes the color and intensity of pixel i; its value is determined by the exogenous variables. The equation for O determines the output of the neural network as a function of the pixel values. Thus, the causal network has depth 2, with the exogenous variables determining the feature variables, and the feature variables determining the output variable. Following MMTS, we assume that there are no dependencies between the feature variables. This is a nontrivial assumption and, in general, is not true in practice, where we expect the color and intensity of a pixel to be causally related to color and intensity of other pixels: if a group of pixels captures, say, a cat's ear, then a group of pixels below it should capture a cat's eye. That said, assuming independence greatly simplifies explanation extraction, hence we adopt the assumption. Moreover, for each setting $\vec{v}$ of the feature variables, there is a setting of the exogenous variables such that $\vec{V} = \vec{v}$. That is, the variables in $\vec{V}$ are causally independent and determined by the context, in the sense of Theorem 1. Moreover, all the parents of the output variable O are contained in $\vec{V}$. So for any explanation of the output involving the pixels, conditions (a), (b), (c), and (e) of Theorem 1 hold if we take φ to be some setting of O. While a neural network often outputs several labels, MMTS assume that the output is unique (and a deterministic function of the pixel values). Given these assumptions, the probability on contexts directly corresponds to the probability on seeing various images (which the neural network presumably learns during training).

Although they claimed to be using Halpern's definition, the definition of (partial) explanation given in MMTS differs from that of Halpern in three respects. The first is that, rather than requiring that a conjunct of the explanation can be extended to an actual cause (i.e., the first part of EX1), they require that in each context, some subset of the explanation be an actual cause. Second, they do not use Halpern's definition of actual cause; in particular, in their analogue of AC2, they do not require that $\vec{W} = \vec{w}$ be true in the context being considered. This appears to be an oversight on their part, and is mitigated by the fact that they focus on but-for causality (for which $\vec{W} = \emptyset$, as we observed, so that the requirement that $\vec{W} = \vec{w}$ in the context being considered has no bite). Finally, they take $\mathcal{K} = \mathcal{R}(\mathcal{U})$. Since we have identified contexts and images, this amounts to considering all images possible. As we shall see in Section 4, there are some benefits in allowing the greater generality of having $\mathcal{K}$ be an arbitrary subset of $\mathcal{R}(\mathcal{U})$. We now show that the conditions considered by MMTS suffice to give Halpern's notion.

Theorem 2. *For the causal model $(M, \Pr)$ corresponding to the image classifier, $\vec{X} = \vec{x}$ is a partial explanation of $O = o$ with goodness (α, β), where $\alpha, \beta > 0$, if the following conditions hold:*

- $\beta \leq Pr(\mathcal{K}_{[\vec{X}=\vec{x}](O=o)})$;
- *there is no strict subset $\vec{X}'$ of $\vec{X}$ such that $\beta \leq Pr(\mathcal{K}_{[\vec{X}'=\vec{x}'](O=o)})$, where x' is the restriction of $\vec{x}$ to the variables in X';*
- $\alpha \leq \Pr(\{\vec{u} : \exists \vec{x}'((M, \vec{u}) \models [\vec{X} = \vec{x}'](O \neq o))\} \mid \mathcal{K}_{\vec{X}=\vec{x} \wedge O=o})$.

Proof. The first condition in the theorem clearly guarantees that the second part of EX1′ holds; the second condition guarantees that EX2′ holds. The fact that $\beta > 0$ means that for some $\vec{u}$, we must have $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}](O = o)$. Let $\vec{u}'$ be a context that gives the pixels in $\vec{X}$ the value $\vec{x}$ and agrees with $\vec{u}$ on the pixels in $\vec{Y}$. Clearly $(M, \vec{u}') \models \vec{X} = \vec{x} \wedge O = o$, so EX3′ holds.

It remains to show that the first part of EX1′ holds. Suppose that $\vec{u}$ is a context for which there exists a setting $\vec{x}'$ of the pixels in $\vec{X}$ such that $(M, \vec{u}) \models [\vec{X} = \vec{x}'](O \neq o)$. (Since $\alpha > 0$, there must exist such a context.) Let $\vec{X}'$ be a minimal subset of $\vec{X}$ for which there exists a setting $\vec{x}''$ such that $(M, \vec{u}) \models [\vec{X}' = \vec{x}''](O \neq o)$. It is easy to see that $\vec{X}' = \vec{x}^*$ is a cause of $O = o$, where $\vec{x}^*$ is the restriction of $\vec{x}$ to $\vec{X}'$. Thus, the first part of EX1′ holds for $\vec{u}$. The desired result follows. ∎

The following example illustrates Theorem 2.

Example 2. *Consider an image, defined by a context $\vec{u}_I$, that is classified as "cat" by the image classifier. Let $\vec{X}$ be a set of pixels corresponding to the cat's head, with their values $\vec{x}$ determined by $\vec{u}$. Intuitively, $\vec{X} = \vec{x}$ is a small picture of a cat's head. In the absence of other information, we assume a uniform probability distribution over all images. Then β is bounded by the fraction of the images that would be classified as "cat" if the cat's head is superimposed on top of the image. If $\vec{X}$ is small, this might happen only with the original image, in which case β is going to be quite small. If $\vec{X}$ occupies a large area of the image, superimposing it on top of another image is more likely to change the classification to "cat", hence β would be higher. The probability α is bounded by the fraction of images in which changing the values of $\vec{X}$ to $\vec{x}'$ leads to a change in classification. As $\mathcal{K}$ contains all images, in particular it contains images that are classified as cats, but would not be classified as cats if not for the subset $\vec{X} = \vec{x}$ (e.g., only a cat's head is visible, and the rest is hidden in the image). For those images, setting the values of $\vec{X}$ to $\vec{x}'$ changes the classification to "not cat".*

We note that it is crucial that $\mathcal{K}$ contains all images, otherwise it is possible that there is no $\vec{u}$ for which setting $\vec{X}$ to $\vec{x}'$ leads to a "not cat" classification. Indeed, consider, for example, a set of images of cats where, in all images, the whole cat is clearly visible. Let $\vec{u}_I$ depict one such image, and assume that the cat in this image is quite small.

It might be that $\vec{X}$, the cat's head, while for a human a perfectly plausible explanation of the classification, is too small to lead to a change in the classification for any image in $\mathcal{K}$ if these pixels' values are changed. Note that it would not even lead to a change in the classification of the original image $\vec{u}_I$, as it is quite possible that the image would be classified as "cat" due to its body shape, even if its head were not visible. Hence, $\alpha = 0$.

How reasonable are the assumptions made by MMTS? The following example shows that their assumption that, in each context, some subset of the explanation be an actual cause (as opposed to some conjunct of the explanation being extendable to an actual cause, as required to EX1) leads to arguably unreasonable explanations.

Example 3. *Consider the following voting scenario. There are three voters, A, B, and C, who can vote for the candidate or abstain; just one vote is needed for the candidate to win the election. The voters make their decisions independently. By Definition 5, $A = 1$ (the fact that A voted for the candidate) is an explanation of the outcome (as well as $B = 1$ and $C = 1$, separately). By assumption, $A = 1$ is sufficient for the candidate to win in all contexts. Now consider any context $\vec{u}$ where the candidate wins and $A = 1$. It is easy to see that the conjunction of voters for the candidate in $\vec{u}$ is a cause of the candidate winning. So, for example, if A and C voted for the candidate in context $\vec{u}$, but B did not, then $A = 1 \wedge C = 1$ is a (but-for) cause of the candidate winning; if both votes change, the candidate loses.*

On the other hand, $A = 1$ is not *an explanation of the candidate winning according to the MMTS definition. To see why, note that in the context $\vec{u}$ above, no subset of $A = 1$ is an actual cause of the candidate winning. Rather, according to the MMTS definition, $A = 1 \wedge B = 1 \wedge C = 1$ is the only explanation (since there are contexts—namely, ones where all three voters voted for the candidate)—where $A = 1 \wedge B = 1 \wedge C = 1$ is the only actual cause). This does not match our intuition. (Of course, we can easily convert this to a story about image classification, where the output is 1 if any of the pixels A, B, and C fire.)*

As we observed, the assumptions of MMTS imply that conditions (a), (b), (c), and (e) of Theorem 1 hold, if we take $\vec{X}'$ to be the pixels, $\vec{X}$ to be some subset of pixels, and φ to be some setting of the output variable O. They also hold in this example, taking $\vec{X}'$ to be the set of voters, $\vec{X}$ to be any subset of voters, and φ to be the outcome of the election. Moreover, condition (d) holds. So in this example, we do not need to check the first part of EX1 to show that $A = a$ is an explanation of the candidate winning, given that the second part of EX1 and EX2 hold. But this is not the case for the MMTS definition. Although $A = 1$ is a sufficient cause of the candidate winning and is certainly minimal, as we observed, it is not an explanation of the outcome according to the MMTS definition. This is because MMTS require a subset of the conjuncts in the explanation to be an actual cause, which is not the case for SC2.

Example 4. *For another, perhaps more realistic, example of this phenomenon in image classification, as observed by Shitole et al. (2021) and Chockler et al. (2023), images usually have more than one explanation. Assume, for example, that the input image I is an image of a cat, labeled as a cat by the image classifier. An explainability tool might find several explanations for this label, such as the cat's ears and nose, the cat's tail and hind paws, or the cat's front paws and fur. All those are perfectly acceptable as explanations of why this image was labeled a cat, but only their conjunction is an explanation according to MMTS. Most of the existing explainability tools for image classifiers output a* saliency map *as an explanation. This saliency map can be used to isolate a part of the input image that is sufficient for the classification—in other words, one explanation (i.e., a single conjunct). Thus, these explanations match Definition 5 (rather than that of MMTS) for the set of contexts that includes the original image and all partial coverings of this image.*

MMTS's restriction to but-for causality is reasonable if we assume that there are no causal connections between the setting of various pixels. However, there are many examples in the literature showing that but-for causality does not suffice if we have a richer causal structure. Consider the following well-known example due to Hall (2004):

Example 5. *Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's rock would have shattered the bottle had it not been preempted by Suzy's throw. The standard causal model for this story (see (Halpern and Pearl 2005a)) has endogenous binary variables ST (Suzy throws), BT (Billy throws), SH (Suzy's rock hits the bottle), BH (Billy's rock hits the bottle), and BS (the bottle shatters). The values of ST and BT are determined by the exogenous variable(s). The remaining equations are $SH = ST$ (Suzy's rock hits the bottle if Suzy throws), $BH = BT \wedge \neg BH$ (Billy's rock hits the bottle if Billy throws and Suzy's rock does not hit the bottle—this is how we capture the fact that Suzy's rock hits the bottle first), and $BS = SH \vee BH$ (the bottle shatters if it is hit by either rock).*

In Example 5, suppose that $\mathcal{K}$ consists of four contexts, corresponding to the four possible combinations of Suzy throwing/not throwing and Billy throwing/not throwing, and $\Pr$ is such that all context have positive probability. In this model, Suzy's throw (i.e., $ST = 1$) is an explanation for the bottle shattering (with goodness (1,1)). Clearly the second part of EX1 holds; if Suzy throws, the bottle shatters, independent of what Billy does. The first part of EX1 holds because $ST = 1$ is a cause of the bottle shattering; if we hold BT fixed at 0 (its actual value), then switching ST from 1 to 0 results in the bottle not shattering. But ST=1 is not a but-for cause of $BS = 1$. Switching ST to 0 still results in $BS = 1$, because if Suzy doesn't throw, Billy's rock will hit the bottle.

We can easily convert this story to a story more appropriate for image classification, where we have an isomorphic model. Suppose that we have two coupled pixels. ST and BT correspond to whether we turn on the power to the pixels. However, if we turn the first one (which corresponds to

SH) on, the second (which corresponds to BH) is turned off, even if the power is on. We classify the image as "active" if either pixel is on. Since this model is isomorphic to the Suzy-Billy story, turning on the first pixel is an explanation for the classification, but again, the second condition of Theorem 2 does not hold. Given a context (an input image) in which the first pixel is on, the explainability tools for image classifiers would output the first pixel being on as an explanation for the classification. This is because they do not take the dependencies between pixels into account. These systems would not call the second pixel part of an explanation, as it is off in the input image.

The following example shows that even without assuming causal structure between the pixels, the second condition in Theorem 2 may not hold.

Example 6. *Suppose that pixels have values in* $\{0, 1\}$. *Let an image be a* $(2n + 1)$*-tuple of pixels. Suppose that an image is labeled 0 if the first pixel is a 0 and the number of 0s in the remaining* $2n$ *pixels is even (possibly 0), or the first pixel is a 1, and the number of 0s in the remaining pixels is even and positive. Suppose that the probability distribution is such that the set of images where there is an even number of 0s in the final* $2n$ *pixels has probability* .9*, Moreover, suppose that* $X_1 = 0$ *with probability* $1/2$*, where* X_1 *denotes the first pixel. Now the question is whether* $X_1 = 0$ *is a partial explanation of* $O = 0$ *with goodness* $(\alpha, .9)$ *for some* $\alpha > 1/2^{2n-1}$. *Clearly the probability that* $O = 0$ *conditional on* $X_1 = 0$ *is .9, while conditional on* $X_1 = 1$ *and unconditionally, the probability that* $O = 0$ *is less than* .9. *Thus, the second part of EX*$1'$ *and EX2*$'$ *both hold. Now consider the first part of EX1*$'$. *Suppose that* $(M, \vec{u}) \models X_1 = 0 \wedge O = 0$*, so in* $\vec{u}$*, an even number of the last* $2n$ *pixels are* 0. *Suppose in fact that a positive number of the last* $2n$ *pixels are* 0. *Then we claim that there is no* $\vec{Y}$ *and* $\vec{y}$ *such that* $X_1 = 0 \wedge \vec{Y} = \vec{y}$ *is a cause of* $O = 0$ *in* $(M, \vec{u})$. *Clearly,* $X_1 = 0$ *is not a cause (since setting it to* 1 *does not affect the labeling). Moreover, if* $\vec{Y}$ *is nonempty, let* $Y \in \vec{Y}$ *and let* y *be the value of* Y *in* $\vec{Y}$. *Then it is easy to see that* $Y = y$ *is a cause of* $O = 0$ *(flipping the value of* Y *results in changing the value of* O *to* 1*), so* $X_1 = 0 \wedge \vec{Y} = \vec{y}$ *is not a cause of* $O = o$ *(AC3 is violated). Thus,* $X_1 = 0$ *is a cause of* $O = 0$ *only in the context* $\vec{u}$ *where* $X_1 = 0$ *and all the last* $2n$ *pixels are* 1. *Using Pascal's triangle, it is easy to show that this context has probability* $1/2^{2n-1}$ *conditional on* $X_1 = 0 \wedge O = 0$. *Note that this is also the only context where changing the value of* X_1 *affects the value of* O.

Example 6 is admittedly somewhat contrived; it does not seem that there are that many interesting examples of problems that arise if there is really no causal structure among the pixels. But, as Example 5 suggests, there may well be some causal connection between pixels in an image. Unfortunately, none of the current approaches to explanation seems to deal with this causal structure. We propose this as an exciting area for future research; good explanations will need to take this causal structure into account.

4 Beyond Basic Explanations: Rare Events and Explanations of Absence

So far we have considered classifiers that output only positive labels, that is, labels that describe the image. However, there are classifiers that output negative answers. These are especially common in healthcare, where image classifiers are used as a part of the diagnostic procedure for MRI images, X-rays, and mammograms (Amisha, Pathania, and Rathaur 2019; Payrovnaziri et al. 2020). In these cases, a diagnosis of absence of abnormalities is a possible output of a classifier: a brain tumor detector based on an MRI outputs either "tumor" or "no tumor".

There are many papers in the medical domain pointing out the importance of explanations in terms of justifying clinical decisions to patients and colleagues (see, e.g., (Amann et al. 2022)). A "right to explanation" is also defined in the EU AI Act. Hence, explanations, and in particular, explanations of absence, are essential, in particular, for clinical diagnosis. However, they have not been addressed up to now.

While the discussion above shows what would it would take for $\vec{X} = \vec{x}$ to be an explanation of "tumor", what would count for $\vec{X} = \vec{x}$ to be an explanation of "no tumor"? We claim that actually the same ideas apply to explanations of "no tumor" as to tumor. For the second part of EX1, $\vec{X} = \vec{x}$ would have to be such that, with high probability, setting $\vec{X}$ to $\vec{x}$ would result in an output of "no tumor". For the first part of EX1, we need to find a minimal subset of pixels that includes a pixel in $\vec{X}$ such that changing the values of these pixels would result in a label of "tumor".

There is a subtlety here though. A tumor is a rare event. Overall, the probability that someone develops a brain tumor in their lifetime is less than 1%, and the probability that a random person on the street has a brain tumor at this moment is much lower than that. Suppose that the classifier derived its probability using MRI images from a typical population. Given an image I that is (correctly) labeled "no tumor", let X be a single pixel whose value in I is x such that X is part of an explanation $\vec{X} = \vec{x}$ of the label "tumor" in a different image I', and in I', $X = x' \neq x$. Then $X = x$ is an explanation of "no tumor" with goodness (α, β) for quite high values of α and β: in most images where $X = x$, the output is "no tumor" (because the output is "no tumor" with overwhelming probability), and changing X to x', as well as the values of other pixels in $\vec{X}$, we typically get an output of "tumor". But $X = x$ does not seem like a good explanation of why we believe there is no tumor!

To deal with this problem (which arises whenever we are classifying a rare event), we (a) assume that the probability is derived from training on MRI images of patients who doctors suspect of having a tumor; thus, the probability of "tumor" would be significantly higher than it is in a typical sample of images; and (b) expect an explanation of "no tumor" to have goodness (α, β) for α and β very close to 1. With these requirements, an explanation $\vec{X} = \vec{x}$ of "no tumor" would include pixels from all the most likely tumor sites, and these pixels would have values that would allow us to preclude there being a tumor at those sites (with high proba-

bility). This is an instance of a situation where $\mathcal{K}$ would not consist of all contexts.

The bottom line here is that we do not have to change the definitions to accommodate explanations of absence and rare events, although we have to modify the probability distribution that we consider. That said, finding an explanation for "no tumor" seems more complicated than finding an explanation for "tumor". The standard approaches that have been used do not seem to work. In fact, none of the existing image classifiers is able to output explanations of absence. The reason for this is that, due to the intractability of the exact computation of explanations, all existing black-box image classifiers construct some sort of *ranking* of pixels of an input image, which is then used as a (partially) ordered list from which an approximate explanation is constructed greedily. Unfortunately, for explaining absence, there is no obvious ranking of pixels of the image: since a brain tumor can appear in any part of the brain, all pixels are equally important for the negative classification.

In general, people find it difficult to explain absences. Nevertheless, they can and do do it. For example, an expert radiologist might explain their decision of "no tumor" to another expert by pointing out the most suspicious region(s) of the brain and explaining why they did not think there was a tumor there, by indicating why the suspicious features did in fact not indicate a tumor. But this is exactly what Definition 6 provides, for appropriate values of the probabilistic bounds (α, β).

Indeed, note that we can get an explanation to focus on the most suspicious regions by making β sufficiently small. Since explanations must be minimal, the explanation will then return a smallest set of regions that has total probability (at least) β.[2] Alternatively, we can just restrict $\mathcal{K}$ to the most suspicious regions, by considering only contexts where non-suspicious regions have all their pixels set to white, or some other neutral color (this is yet another advantage of considering $\mathcal{K}$ rather than all of $\mathcal{R}(\mathcal{U})$). In addition, to explain why there is no tumor in a particular region, the expert would likely focus on certain pixels and say "there's no way that those pixels can form part of a tumor"; that's exactly what the "sufficiency" part of the explanation does. The expert might also point out pixels that would have to be different in order for there to be a tumor; that's exactly what the "necessity" part of the explanation does. Of course, this still must be done for a number of regions, where the number is controlled by β or the choice of $\mathcal{K}$. Thus, it seems to us that the definition really is doing a reasonable job of providing an explanation in the spirit of what an expert would provide.

We can get simpler (although perhaps not as natural) explanations for the absence of tumors by taking advantage of domain knowledge. For example, if we know the minimal size of a tumor, explaining the absence of a tumor can be done by covering the image with a "net" of pixels, none of which can be part of a tumor, such that the distance between neighboring pixels in the net is smaller than the size of a minimal tumor.

The upshot of this discussion is that the techniques we have presented can be used to find explanations of absence.

Computing explanations of absence as defined in this section is quite nontrivial, and is quite domain-dependent. For example, we can explain "no tumor"' by providing a grid of sufficiently small dark grey patches (as tumors appear on the MRI as light in color) such that the distance between the patches is too small to contain the smallest tumor recognised by the model. But this approach would not work for other domains, as there is no clear distinction between the colors of the absent object and the colors of other possible objects in the image. (For example, it would not work to explain why there is no cat in an image, as images of cats contain many colors.) A full implementation of an algorithm to compute explanations of absence (even approximately) is beyond the scope of the paper, and is the subject of ongoing work.

5 Conclusions

We conclude by repeating the point that we made in the introduction (now with perhaps more evidence): while the analysis of MMTS shows that a simplification of Halpern's definition can go a long way to helping us understand notions of explanation used in the literature, we can go much further by using Halpern's actual definition, while still retaining the benefits of the MMTS analysis. In particular, we can use the definition to provide explanations of absence and explanations of rare events, both of which arise frequently in practice.

However, there is still more to be done. For one thing, dealing with the full definition may involve added computational difficulties. We believe that using domain knowledge may well make things more tractable, although this too will need to checked. For example, We showed above how domain knowledge could be used to provide simpler explanations of the absence of tumors. For another example, if we are explaining the absence of cats in an image of a seascape, knowledge of zoology allows to eliminate the sea part of the image as a possible area where cats can be located, as cats are not marine animals. This seems doable in practice.

Acknowledgments: We thank Ilse van der Linden for pointing us to the work of MMTS. Hana Chockler was supported in part by the UKRI Trust-worthy Autonomous Systems Hub (EP/V00784X/1), the UKRI Strategic Priorities Fund to the UKRI Research Node on Trustworthy Autonomous Systems Governance and Regulation (EP/V026607/1), and CHAI – EPSRC AI Hub for Causality in Healthcare AI with Real Data (EP/Y028856/1). Joe Halpern was supported in part by NSF grant FMitF-2319186, ARO grant W911NF-22-1-0061, MURI grant W911NF-19-1-0217, and a grant from the Cooperative AI Foundation.

[2]There will be may be many choices for this; we can add code to get the choice that involves the smallest set of pixels. These will be the ones of highest probability.

References

Amann, J.; Vetter, D.; Blomberg, S. N.; Christensen, H. C.; Coffee, M.; Gerke, S.; Gilbert, T. K.; Hagendorff, T.; Holm, S.; Livne, M.; Spezzatti, A.; Str´umke, I.; Zicari, R. V.; and Madai, V. I. 2022. To explain or not to explain?—artificial intelligence explainability in clinical decision support systems. *PLOS Digital Health* 1(2).

Amisha, P. M.; Pathania, M.; and Rathaur, V. K. 2019. Overview of artificial intelligence in medicine. *Journal of Family Medicine and Primary Care* 8(7):2328–2331.

Beckers, S. 2021. Causal sufficiency and actual causation. *Journal of Philosophical Logic* 50:1341–1374.

Chajewska, U., and Halpern, J. Y. 1997. Defining explanation in probabilistic systems. In *Proc. Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI '97)*, 62–71.

Chockler, H.; Kelly, D. A.; and Kroening, D. 2023. Multiple different explanations for image classifiers. Available at https://arxiv.org/pdf/2309.14309.pdf.

Chockler, H.; Kroening, D.; and Sun, Y. 2021. Explanations for occluded images. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV 2021)*, 1214–1223.

Gärdenfors, P. 1988. *Knowledge in Flux*. Cambridge, Mass.: MIT Press.

Glymour, C., and Wimberly, F. 2007. Actual causes and thought experiments. In Campbell, J.; O'Rourke, M.; and Silverstein, H., eds., *Causation and Explanation*. Cambridge, MA: MIT Press. 43–67.

Hall, N. 2004. Two concepts of causation. In Collins, J.; Hall, N.; and Paul, L. A., eds., *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

Hall, N. 2007. Structural equations and causation. *Philosophical Studies* 132:109–136.

Halpern, J. Y., and Pearl, J. 2005a. Causes and explanations: a structural-model approach. Part I: causes. *British Journal for Philosophy of Science* 56(4):843–887.

Halpern, J. Y., and Pearl, J. 2005b. Causes and explanations: a structural-model approach. Part II: explanations. *British Journal for Philosophy of Science* 56(4):889–911.

Halpern, J. Y. 2016. *Actual Causality*. Cambridge, MA: MIT Press.

Hempel, C. G. 1965. *Aspects of Scientific Explanation*. New York: Free Press.

Hitchcock, C. 2001. The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy* XCVIII(6):273–299.

Hitchcock, C. 2007. Prevention, preemption, and the principle of sufficient reason. *Philosophical Review* 116:495–532.

Lundberg, S. M., and Lee, S. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 4765–4774.

Molnar, C. 2022. *Interpretable Machine Learning*. 2 edition.

Mothilal, R. K.; Mahajan, D.; Tan, C.; and Sharma, A. 2021. Towards unifying feature attribution and counterfactual explanations: different means to the same end. In *Proc. of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*, 652–663.

Mothilal, R. K.; Tan, C.; and Sharma, A. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proc. of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT 2020)*, 607–616.

Payrovnaziri, S. N.; Chen, Z.; Rengifo-Moreno, P.; Miller, T.; Bian, J.; Chen, J. H.; Liu, X.; and He, Z. 2020. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *Journal of the American Medical Informatics Association* 27(7):1173–1185.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. San Francisco: Morgan Kaufmann.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Knowledge Discovery and Data Mining (KDD)*, 1135–1144. ACM.

Salmon, W. C. 1970. Statistical explanation. In Kolodny, R., ed., *The Nature and Function of Scientific Theories*. Pittsburgh, PA: University of Pittsburgh Press. 173–231.

Salmon, W. C. 1989. *Four Decades of Scientific Explanation*. Minneapolis: University of Minnesota Press.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV 2017)*, 618–626.

Shitole, V.; Li, F.; Kahng, M.; Tadepalli, P.; and Fern, A. 2021. One explanation is not enough: structured attention graphs for image classification. In *Proc. Advances in Neural Information Processing Systems (NeurIPS 2021)*, 11352–11363.

Sun, Y.; Chockler, H.; Huang, X.; and Kroening, D. 2020. Explaining image classifiers using statistical fault localization. In *Proc. 18th Conference on computer vision (ECCV 2020), Part XXVIII*, 391–406.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard Journal of Law and Technology* (31):841–887.

Weslake, B. 2015. A partial theory of actual causation. *British Journal for the Philosophy of Science*. To appear.

Woodward, J. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford, U.K.: Oxford University Press.

Woodward, J. 2014. Scientific explanation. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy* (Winter 2014 edition). Available at http://plato.stanford.edu/archives/win2014/entries/scientific-explanation/.