人工智能发展至今，已经在许多领域和任务上取得了惊人的成就。在这背后，先进的统计模型、机器学习、深度学习算法以及大数据时代海量的训练数据起到了关键作用。强大的机器学习方法，如决策树、集成方法、深度神经网络能够准确地对数据进行拟合，更多的数据也有助于获得更精确的预测结果。但是，现代机器学习研究仅仅追求预测准确性是不够的，正确性和可解释性也是机器学习方法的目标[1]。许多人类能够轻易理解和掌握的概念，对于目前的人工智能来说仍然是难以理解的。如果检查现如今驱动机器学习的数据信息，我们会发现其几乎都是基于统计的[2]。而基于统计的方法只是学习到了数据中的关联或者相关性，无法像人类一样，解数据背后更深层次的逻辑和因果关系。然后复杂的世界是充满因果关系的，相关性并不能代替因果来推动人工智能做出决策。就像人类是世界的主宰者，善于通过变化万千的事物总结其中的因果关系来形成自己的知识，而目前的人工智能是不具备这种能力的。因此，越来越多的科学家意识到，机器人和人工智能需要像人类一样的因果推理能力，才能从弱人工智能走向强人工智能[3]。

2017 年图灵奖得主、人称“贝叶斯网络之父”的 Judea Pearl 是因果关系领域的重要人物。Pearl 写了一本关于因果关系的科普书籍《The Book of Why: The New Science of Cause and Effect》[4]。在本书中，他将因果关系分为三个层次（“因果关系之梯”）从上到下依次是关联（Association）、干预（Intervention）和反事实推理（Couterfactuals）。第一层是关联，也就是现在的机器学习模型尝试在做的事，即通过观察到的数据学习变量之间统计上的相关性。但只知道事件 A 和事件 B 相关是无法推出 A、B 之间是否存在因果关系的。回答的是“变量之间是怎样关联的？”、“观察到 A 会怎样改变我对 B 的看法？”这样的问题。第二层是干预，即通过一些手段干预数据，从而回答“如果我实施 X 行动，那么 Y 会怎么样？”、“如果我改变了事件 A，事件 B 是否会随之改变？”第三层是反事实，相当于通过结果来考虑原因，也就是我们希望知道，如果我们想让事件 B 发生某种变化，能否通过改变事件 A 来实现。反事实推断回答诸如“假如当时做了……会怎样？”、“是 X 引起了 Y 吗？假如 X 没有发生会如何？”之类的问题。三个层级中，相关显然不涉及因果，只有干预和反事实才有助于我们学习数据中的因果关系。

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in$Y$?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

表 1 因果关系的三个层次，关联、干预、反事实推理。

贝叶斯网络

贝叶斯网络（Bayesian Network），也被称为信念网络（Belief Network），是一种典型的“概率图模型”（Probabilistic Graphical Model, PGM），是一种通过有向无环图（Directed Acyclic Graph，DAG）表示一组变量及其条件依赖关系的方法[5]。贝叶斯网络的基本结构有节点和节点之间的单向箭头连线组成。如图 1 所示，节点（Nodes/Vertices/Variables）是图中的变量X、Y和Z。边（Link/Edge）是图中单方向的箭头A和B。路径（Path）是从一个变量沿着箭头的方向抵达另一个变量的经过，如$X \to Y \to Z$。从节点X指向节点Y的箭头表示变量Y依赖于变量X，且X是Y的父节点（Parent node），Y是X的子节点（Child node）。

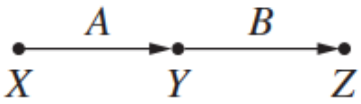

图 1 三个节点的有向无环图

贝叶斯网络可以表示为$G = (N, E)$，N、E分别代表有向无环图的节点集和边集。图中每一个节点都代表一个变量X_i。此外，贝叶斯网络不仅包含有向无环图的结构，还包括图中的参数θ，即各个节点之间的概率分布。给出n个节点的贝叶斯网络，其联合概率分布为

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | pa_i)$$

其中pa_i是节点X_i的父节点集。如果一个节点没有父节点，那么它的先验概率（边

缘概率）需要指定。如果一个节点有一个或多个父节点，那么它的条件概率分布需要被给出。给出观测数据和贝叶斯网络的结构，就可以估计每个节点的概率分布，这叫做贝叶斯网络的参数学习。如果网络结构未知，也可以通过观测数据进行贝叶斯网络的结构学习。虽然贝叶斯网络是一种有向图，但无法真正地表示因果关系。不过，贝叶斯网络中的向无环图仍是后面因果图的重要组成部分，贝叶斯网络中的概率公式也适用于因果图。

结构因果模型

为了能严格地处理因果关系，需要寻找到一种能够形式化表述数据背后因果假设地方法。为此，引入结构因果模型（Structural causal model，SCM），用于描述现实世界关联特征及其相互作用[6]。从形式上看，SCM 含有两个变量集U和V，以及一组函数：

$$f = \{f_X : W_X \to X | X \in V\}$$

其中$W_X \subseteq (U \cup V) - \{X\}$，即函数$f_X$根据模型中其他变量的值给变量$X$赋值。因果的定义：若$Y$存在于$f_X$的定义域中，则变量$Y$是变量$X$的直接原因；若$Y$是$X$的直接原因或原因的原因，则$Y$是$X$的原因。$U$中的变量称为外生变量，属于模型的外部，不必解释引起它们变化的原因。V中的变量称为内生变量，模型中的每一个内生变量都至少是一个外生变量的后代。外生变量没有祖先节点，因从不是任何其他变量的后代，外生变量也不能是内生变量的后代。如果知道外生变量的值，就可以利用$f_X \in f$完全确定每个内生变量的值。

类似地，一个 SCM 与图形化的因果模型相关联，图模型中的节点表示U和V中的变量，节点之间的边表示f中的函数。由于 SCM 和图模型之间的这种关系，可以给出因果关系的形式化定义：在图模型中，如果变量X是另一个变量Y的子节点，那么Y是X的直接原因；如果X是Y的后代，那么Y是X的一个潜在原因。例如，图 2 中的 SCM 展示了“学历”X、“工龄”Y和“工资”Z之间的因果关系：

$$U = \{X, Y\}, V = \{Z\}, F = \{f_z\}$$

$$f_z : Z = 2X + 3Y$$

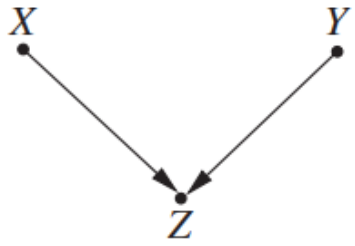

图 2 “学历”X、“工龄”Y和“工资”Z之间的 SCM 对应的图模型

在这个模型中，X和Y都出现在f_z，因此X和Y都是Z的直接原因，f_z基于X和Y对Z进行赋值。如果X和Y有祖先，则他们将是Z的潜在原因。在很多情况下，我们无法测量每一个变量的值，但这些变量的值却影响其他我们希望了解的变量。下面是一个部分细化的 SCM，展示了助学金X、考试成绩Y和考研录取概率Z的关系。

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$

$$f_X: X = U_1$$

$$f_Y: Y = \frac{X}{3} + U_Y$$

$$f_Z: Z = \frac{Y}{16} + U_Z$$

在这个模型中，外生变量$U = \{U_X, U_Y, U_Z\}$有时称为“误差项”或“省略因素”，代表观测变量的未知或随机的外生原因。

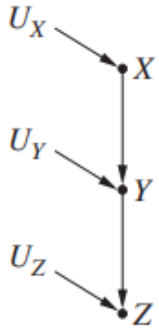

图 3 助学金X、考试成绩Y和考研录取概率Z之间的 SCM 对应的图模型

三种接合结构

在上述图模型中，根据边方向的不同，可以将三个相邻节点的路径结构分为：链式（Chain）、叉式（Fork）、反叉式/对撞（Inverted fork/Collider）。Pearl 将这中三个节点的网络结构称为接合（Junction）。所有的贝叶斯网络（或因果图）都可以被拆解成这三种接合结构的组合。

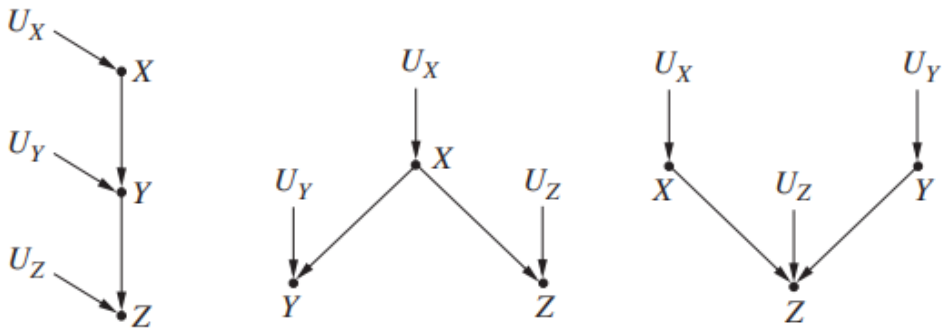

图 4 三种接合结构，依次是链式（Chain）、叉式（Fork）、对撞（Collider）。

在链式结构中，变量X、Y、Z之间的独立性和依赖关系如下：

（1） Z和Y可能是相互依赖的：对于某些z，y，有

$$P(Z = z|Y = y) \neq P(Z = z)$$

（2） Y和X可能是相互依赖的：对于某些y，x，有

$$P(Y = y|X = x) \neq P(Y = y)$$

（3） Z和X可能是相互依赖的：对于某些z，x，有

$$P(Z = z|X = x) \neq P(Z = z)$$

（4） Z和X在Y的条件下是独立的：对于所有的x，y，z，有

$$P(Z = z|X = x, Y = y) = P(Z = z|Y = y)$$

在一个典型的因果模型中，由边相连的两个变量是依赖的。从变量Y指向变量Z的箭头表示前者是后者的原因，也就是说，变量Y是确定后变量Z的函数的一部分，变量Z的值依赖于变量Y的值。如果Z依赖于Y，而Y又依赖于X，那么Z和X很可能是依赖的（在某些特殊情况下并非如此）。对于上述（4），考虑$Y = a$的情况，即选择具有不同的X值时，U_Y的值需随之变化以使得Y的值为a。因为Z值仅取决于Y和U_Z，不依赖于U_Y，所以U_Y的变化不会影响Z值。因此，不同的X值也不会带来Z值的改变，因此，当$Y = a$，Z和X是独立的。

同样，对于叉式结构，变量X、Y、Z之间的独立性和依赖关系可以总结为：

（1） X和Y可能是相互依赖的：对于某些x，y，有

$$P(X = x|Y = y) \neq P(X = x)$$

（2） X和Z可能是相互依赖的：对于某些x，z，有

$$P(X = x|Z = z) \neq P(X = x)$$

（3） Z和Y可能是相互依赖的：对于某些z，y，有

$$P(Z = z|Y = y) \neq P(Z = z)$$

（4） Y和Z在X的条件下是独立的：对于所有的x，y，z，有

$$P(Y=y|Z=z,X=x)=P(Y=y|X=x)$$

在叉式结构中，由于Y和Z都与X直接相连，因此X与Y、Z是相互依赖的。进一步，当X改变时，Y和Z可能会一起发生变化，由于可以从Y值的变化中得到Z值变化的信息，所以Y和Z可能是相互依赖的。对于（4），考虑$X=a$的情况，由于此时X的值是固定的，Y和Z的值不会随着X值而变化，只会随着U_Y和U_Z而变化，因为已经假设U_Y和U_Z是独立的，因此Y和Z在X的条件下是独立的。

对于对撞结构，变量X、Y、Z之间的独立性和依赖关系可以总结为：

（1） X和Z可能是相互依赖的：对于某些x，z，有

$$P(X=x|Z=z)\neq P(X=x)$$

（2） Y和Z可能是相互依赖的：对于某些y，z，有

$$P(Y=y|Z=z)\neq P(Y=y)$$

（3） X和Y是独立的：对于所有的x，y，有

$$P(X=x|Y=y)=P(X=x)$$

（4） X和Y在Z的条件下可能是相互依赖的：对于某些x，y，z，有

$$P(X=x|Y=y,Z=z)\neq P(X=x|Z=z)$$

前两点和链式、叉式结构是一样的道理。因为X和Y不是彼此的后代或者祖先，也没有依赖于同一个变量的值，仅分别依赖于U_X和U_Y，所以X和Y是独立的。对于（4），当以Z为条件时，Z的值是固定的，又由于Z同时依赖于X和Y，因此，当X和Y中的一个变化时，必须通过另一个值的变化来进行“补偿”以使得Z是不变的。所以，在对撞结构中，当以共同效应Z作为条件时，两个独立的变量X和Y就变得相互依赖了。

条件独立性与阻断

更一般地，我们可以将三种接合结构中的条件独立性概括为：

（1） 链式结构中的条件独立性：如果变量X和Z之间只有一条单向路径，Y是截断这条路径的任何一组变量，则在Y的条件下，X和Z是独立的。在Y的条件下，也即给定Y的值，那么X和Z之间的路径被阻断（Blocking）。

（2） 叉式结构中的条件独立性：如果变量X是变量Y和变量Z的共同原因，并且

Y和Z之间只有一条路径，则Y和Z在X的条件下独立，Y和Z之间的路径被阻断。

（3） 对撞结构中的条件独立性：如果变量Z是变量X和Y之间的对撞节点，并且X与Y之间只有一条路径，那么X与Y是无条件独立的，但是在Z或Z的任何子孙条件下是相互依赖的。

d-分离

真实的因果模型是复杂的，变量之间通常不会只有一条路径，而是多条路径相连，且每个路径上包含多个链式、叉式和对撞结构。因此，对于任意复杂的因果图模型，需要一个准则来判断任何一对节点之间的相关性和独立性。d-分离法则的全称是有向分离（directional separation）法则，是一种判断变量是否条件独立的方法。通过该法则，我们能够确定任何一对节点是d-连通的还是d-分离的。一对节点是d-连通的，指这两个变量可能是相互依赖的；一对节点是d-分离的，指这两个变量是独立的或者条件独立的。具体来说，如果两个节点X和Y之间存在的任何路径都被阻断，则它们是（关于阻断变量）d-分离的；如果X和Y之间存在一条路径没有被阻断，则它们是d-连通的。

d-分离的定义：一条路径会被一组节点Z阻断，当且仅当：

a） 路径p包含链式结构$A \to B \to C$或分叉结构$A \leftarrow B \to C$，且中间节点B在Z中（即以B为条件）；或者

b） 路径p包含一个对撞结构$A \leftarrow B \to C$，且对撞节点B及其子孙节点都不在Z中。

如果Z阻断了X和Y之间的每一条路径，则X和Y在Z的条件下是d-分离的。d-分离分为两种情况：以某些节点为条件和不以某些节点为条件。前者是指当以非对撞结构的中间节点为条件时，路径的两端点会变得条件独立；后者是指当路径中存在对撞结构时，对撞节点会阻断该路径。

辛普森悖论与混杂

辛普森悖论（Simpson’s paradox）以第一个论及该问题的统计学家 Edward Simpson 命名。该悖论说明了这样一个事实：存在这样的数据，总体上的统计结果与其每一个子部分的统计结果相反。以下是一个经典的例子：假设一种新药被

研发出了，一组患者可以选择是否尝试这种新药。辛普森发现，当对患者按性别进行划分后，服药的男性患者的痊愈率比不服药的男性患者的痊愈率高。同样的，服药的女性患者的痊愈率比不服药的女性的痊愈率高。看来这种药不论对男性患者还是女性患者，都是有益的。然而，根据总体统计，服用该药的患者的痊愈率却低于未服药的患者。也就是说，从男性患者和女性患者所构成的全体来看，这种药是没有效果的。这个问题无法简单地从统计学中找到答案，因为数据本身不足以确定药物到底是有没有效果的。

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

表 1 一种新药在男性和女性患者群体上的痊愈率

如果我们分析这个问题中的三个变量，即服用药物X、痊愈Y和性别Z之间的关系，不难发现，男性和女性服药与不服药的比例是大不一样的，男性服药的人数远小于不服药的人数，而女性服药的人数远多于不服药的人数。其实，在这个问题中，性别同时影响服药和痊愈情况，是二者的共因，也被称为“混杂因子”（Confounder），如下图所示。混杂因子会造成伪相关关系，和真正的因果关系混合在一起，带来“混杂”（Confounding）。因果推理的一大目标就是尽量消除混杂带来的影响，找出真正的因果关系[3]。

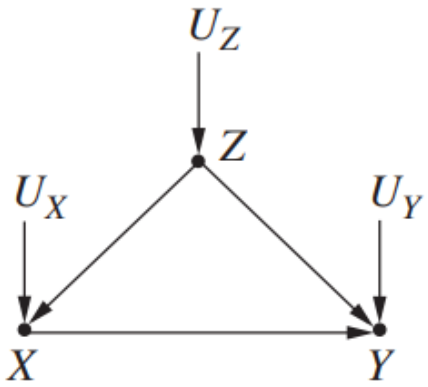

图 5 辛普森悖论中服用药物X、痊愈Y和性别Z之间的因果关系

干预与矫正公式

就目前为止，我们提到的操作仍然只属于因果关系之梯的第一层级——关联。

即便是“以某个变量为条件”的操作，也只是依据现有观测到的数据进行统计，并没有改变数据的分布。回到辛普森悖论中服用药物的例子，我们实施了观察性研究，从数据中观察到服用药物和痊愈之间的相关性或者关联，但这种研究方法的问题在于很难将因果关系从相关关系中提取出来。我们仅仅记录数据，而没有控制数据。而要进行干预，我们就得改变现有的数据分布。为了确定药物的有效性， 可以设想一种干预措施，即让整个人群统一地服用这种药物（让原来没有服药的人服药），与阻止整个人群服用这种药物（让原来服药的人不服药），然后对两种情况下人群的痊愈率进行比较。

在实际应用中，干预是非常重要的。比如，当我们对一种新的抗癌药物进行研究时，我们试图确定当我们对病人进行药物干预时，病人的病情如何变化。当我们研究暴力电视节目和儿童的攻击行为之间的关系时，我们希望知道，干预减少儿童接触暴力电视节目是否会减少他们的攻击性。

干预和以变量为条件有着本质的区别。当我们在模型中对一个变量进行干预时，我们将固定这个变量的值。其他变量的值也随之改变。当我们以一个变量为条件时，我们什么也不会改变；我们只是将关注的范围缩小到样本的子集，选取其中我们感兴趣的变量的值。因此，以变量为条件改变的是我们看世界的角度，而干预则改变了世界本身[6]。

在符号表达式上，干预用do算子来表示，并将其与条件概率区分开来。$P(Y=y|X=x)$表示在$X=x$的条件下$Y=y$的概率；$P(Y=y|do(X=x))$表示通过干预使$X=x$时$Y=y$的概率。前者反映了在X取值都为x的个体上Y的总体分布，后者反映了如果将群体中每个个体的X值都固定为x时Y的总体分布。注意两者的区别，以变量为条件不改变数据的分布，而干预则改变了数据的分布。

以上述的辛普森悖论问题为例，我们介绍一种通过观察数据进行干预并分解因果关系的方法。假设这种干预措施是：给整个群体都服用药物，即$do(X=1)$；以及给整个群体都不服用药物，即$do(X=0)$。然后，计算它们之间的差异，称为“因果效应差异”或“平均因果效应”（average causal effect, ACE），以此来判断服用药物与痊愈之间的因果关系。

$$P\big(Y=1\big|do(X=1)\big)-P(Y=1|do(X=0))$$

首先，对图5进行处理，删除$Z\rightarrow X$这条边，得到图6的因果图。因为$do(X=x)$

操作将数据中所有个体的X值固定为x，阻断了所有指向该节点的联系。我们用P_m表示修改后模型的概率分布，也被称为操纵概率，那么$P(Y=y|do(X=x))$与$P_m(Y=y|X=x)$是相等的。

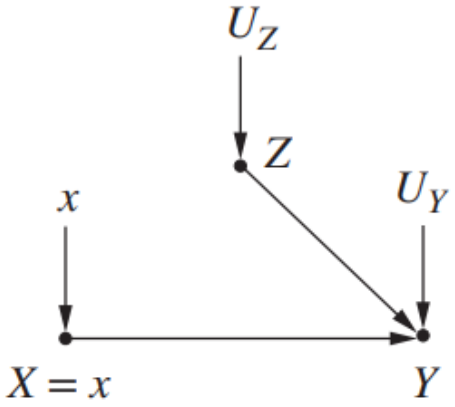

图 6 修改之后的辛普森悖论因果图

在这里，我们假设对于一个变量进行干预不会造成其他影响，即不会改变其他变量的值，也不会对其他变量的关系产生影响。计算因果效应的关键在于观察操纵概率P_m。在原模型和修改后的模型之间，边缘概率$P(Z=z)$在干预前后保持不变，即$P_m(Z=z)=P(Z=z)$，这是因为删除$Z\to X$不会改变Z的确定过程，因此干预前后数据中男性患者和女性患者的比例保持不变。此外，条件概率$P(Y=y|Z=z,X=x)$也是不变的，即$P_m(Y=y|Z=z,X=x)=P(Y=y|Z=z,X=x)$，因为$Y$对$X$和$Z$的响应函数$f_Y=f(x,z,u_Y)$是不会因为$X$的改变而发生变化的。进一步，可以得到：

$$P(Y=y|do(X=x))=P_m(Y=y|X=x)\text{（由定义）}$$

$$=\sum_z P_m(Y=y|X=x,Z=z)P_m(Z=z|X=x)\text{（由全概率公式）}$$

$$=\sum_z P_m(Y=y|X=x,Z=z)P_m(Z=z)\text{（由}Z\text{和}X\text{的独立性）}$$

最后，由前面的不变性关系，得到下面一个以干预前概率表示的因果效应公式，也被称为“矫正公式”（adjustment formula），这个运算过程叫做“对Z的矫正”或者“对Z的控制”。

$$P(Y=y|do(X=x))=\sum_z P(Y=y|X=x,Z=z)P(Z=z)$$

矫正公式的右边可以直接从数据中估算得到，这就意味着我们可以用观测数据来进行干预。接下来我们将矫正公式应用于辛普森悖论，有：

$$\begin{aligned}P(Y=1|do(X=1)) &\\ &= P(Y=1|X=1,Z=1)P(Z=1)\\ &+ P(Y=1|X=1,Z=0)P(Z=0)\end{aligned}$$

$$\begin{aligned}P(Y=1|do(X=0)) &\\ &= P(Y=1|X=0,Z=1)P(Z=1)\\ &+ P(Y=1|X=0,Z=0)P(Z=0)\end{aligned}$$

根据表 1 中的数据，可以计算得到：

$$P(Y=1|do(X=1)) = 0.93\times\frac{87+270}{700}+0.73\times\frac{263+80}{700}=0.832$$

$$P(Y=1|do(X=0)) = 0.87\times\frac{87+270}{700}+0.69\times\frac{263+80}{700}=0.7818$$

因此，让整个群体都服用药物$do(X=1)$和让整体群体都不服用药物$do(X=0)$之间的平均因果效应为

$$\begin{aligned}ACE &= P(Y=1|do(X=1))-P(Y=1|do(X=0)) = 0.832-0.7818\\ &= 0.00502\end{aligned}$$

这表明服药对痊愈是具有积极作用的。而如果只看条件概率，有

$$P(Y=1|X=1)) = 0.78$$

$$P(Y=1|X=0)) = 0.83$$

可见，如果不加干预，仅计算原始数据中的条件概率是无法得到正确的因果关联的。

通过以上的例子，我们爬上了因果关系之梯的第二层级——干预。通过观察矫正公式的导出过程，我们发现，当通过外部操作固定X值时，X的父节点的影响被消除了。进一步，我们得到如下更一般化的矫正公式，总结为以下这条规则：

给定一个图G，设变量X的父节点集合为PA，则X对Y的因果效应为：

$$P(Y=y|do(X=x)) = \sum_{z} P(Y=y|X=x,PA=z)P(PA=z)$$

其中，z的取值范围是PA中变量所有可能的取值集合。因此，如果我们要研究两个变量X和Y之间的因果关系，那么就可以对X的父节点集进行调整。然而，在很多情况下，变量通常有无法观察到的父节点，即虽然该父节点在图中有所体现，

但节点的值却无法得到。所以，我们需要找到可替代的变量集合用于矫正。后门准则就是一种用来识别需要矫正的变量集合的方法。对于不同的因果效应模型，还可以通过前门准则、逆概率加权、工具变量、中介等概念或干预工具来求解。相关内容可以参考[6]第三章。此外，有一个被称为“do-演算”的强大的工具可以揭示给定图模型中能够被识别的所有因果效应[7-9]。

反事实推理

到这里，我们来到了因果关系之梯的第三层级——反事实（推理）。第二层级——干预的目标是，找到研究中处理的某个总效应或者在某些典型个体或子总体中的效应（平均因果效应）。但到目前为止，我们仍不具备在特定事件或个体层面上谈论个性化的因果关系的能力[4]。举例来说，“吸烟致癌”和“我的叔叔 30 年来每天都抽一包烟，假如他不曾抽烟的话，那他现在可能还活着”完全是两回事。这种“如果”、“假如”（不真实的或者未能实现的）的称述形式被称为反事实。反事实的“如果”部分称为假设条件，或者前件。使用反事实来着重强调，想要在完全一致的现实条件下比较不同的前件的结果。而通过未发生的前件来推理可能出现的结果，就称作为反事实推理。在上面的例子中，“我的叔叔 30 年来每天都抽一包烟，并且他已经去世了”是事实，而“他不曾抽烟”是反事实的假设条件或前件，对应于“他每天都抽烟”。通过“如果他不曾抽烟”来推测“他现在可能还活着吗？”就是一个反事实的推理。与干预不同的是，反事实试图在给定叔叔已经死亡的条件下，推测在选择不抽烟的世界中，叔叔是否可能会活着。而干预是在选择不抽烟的世界中估计其对死亡的平均因果效应。前者参考了叔叔死亡的世界的信息，而后者不参考任何一个世界中的信息。正如哲学家大卫·路易斯所言，我们是通过比较我们的世界和在其他方面与现实世界最相似的那个假如世界来评估反事实陈述的[4]。

下面介绍反事实的形式化定义和计算。我们用M表示结构因果模型，U表示外生变量集，V表示内生变量集，F表示M中的函数集。$U = u$表示对外生变量进行赋值，每一个赋值$U = u$唯一确定了V中所有变量的值。比如，如果$U = u$代表某个人的属性，X代表变量“薪资”，那么$X(u)$就代表了他的薪资。反事实陈述“在$U = u$的情况下，如果X当初取值为x，则Y会取值为y”可以表示为$Y_x(u) =$

y，其中X和Y是V中的任意两个变量。“如果X当初取值为x”建立了前置条件$X = x$，可能与X的实际观测值$X(u)$冲突。因为前置条件的引入相当于将模型中的X赋值为了常量x，所以可以将其看作是一个外部干预，即$do(X = x)$。

下面是一个简单的线性结构因果模型M，它只包含三个变量X、Y、U，定义如下：

$$X = aU$$

$$Y = bX + U$$

（1）计算反事实$Y_x(u)$，即在$U = u$的情况下，如果X当初取x，Y的取值应该是多少。用$X = x$替换第一个方程，得到修改后的结构因果模型M_x：

$$X = x$$

$$Y = bX + U$$

将$U = u$带入，得到：

$$Y_x(u) = bx + u$$

假设$a = b = 1$，$u = 1$，$x = 2$，那么在原模型M中，$X(u) = 1$，$Y(u) = 2$；在修改后的模型M_x中，$Y_1(u) = 3$，即在$U = 1$的情况下，如果X当初取2，Y的取值会从原来的 2 变为 3。

（2）计算反事实$X_y(u)$，即在$U = u$的情况下，如果Y当初取y，X的取值应该是多少。这里，用$Y = y$替换第二个方程，并求解得到$X_y(u) = au$，这意味着在假设条件为“Y当初取y”的情况下，X保持不变，因为Y不会影响X，可以被理解为未来事件并不会改变过去。假设$a = b = 1$，U可以取 1，2，3，下表给出了该结构因果模型中不同反事实陈述所对应的取值情况。

u	$X(u)$	$Y(u)$	$Y_1(u)$	$Y_2(u)$	$Y_3(u)$	$X_1(u)$	$X_2(u)$	$X_3(u)$
1	1	2	2	3	4	1	1	1
2	2	4	3	4	5	2	2	2
3	3	6	4	5	6	3	3	3

表 2 上述模型的反事实陈述取值情况

可以将反事实概念推广到任何的结构因果模型M。考虑任何两个变量X和Y，M_x表示用$X = x$替换X后得到的修改后的模型，反事实$Y_x(u)$定义为：

$$Y_x(u) = Y_{M_x}(u)$$

即模型M中的反事实$Y_x(u)$的值等于修改后的模型M_x中Y的值。当X和Y是变量集合、M_x的X集合中所有成员均被常数值替换后的时候，该定义同样适用。另外，关于反事实陈述的一致性原则是，如果观测到$X = x$，那么$Y_x = Y$，因为此时反事实中的假设条件就是事实，没有“反”事实。如果X是二值的，那么一致性原则可以写成：

$$Y = XY_1 + (1 - X)Y_0$$

Y_1为X取值为 1 时Y的观测值，如果$X = x = 1$，那么$Y = Y_1$。X取值为 0 时同理。

反事实推理实例

考虑以下的结构因果模型，其中X表示课外补习的时间，H表示家庭作业量，Y表示考试成绩：

$$X = U_X$$

$$H = aX + U_H$$

$$Y = bX + cH + U_Y$$

$$\forall i, j \in \{X, H, Y\}, \sigma_{U_i U_j} = 0$$

假设所有的因子U都是独立的，从数据中估计得到系数的值为：

$$a = 0.5, b = 0.7, c = 0.4$$

图 7 为对应的因果图结构：

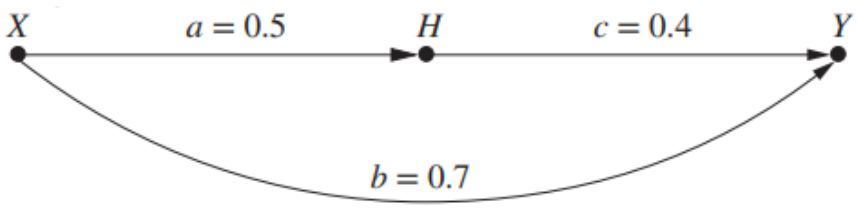

图 7 课外补习时间X、家庭作业量H和考试成绩Y之间的因果模型图

假设我们观测到一个叫小明的学生的各项数据：$X = 0.5$，$H = 1$，$Y = 1.5$。那么请问，如果小明当初的家庭作业量加倍，他的考试成绩会怎么变化？

我们可以从观测数据中计算出小明的U值：

$$U_X = X = 0.5$$

$$U_H = H - aX = 1 - 0.5 \times 0.5 = 0.75$$

$$U_Y = Y - bX - cH = 1.5 - 0.7 \times 0.5 - 0.4 \times 1 = 0.75$$

“作业量加倍”是反事实的假设条件，即$H=2$。接下来，用$H=2$进行对应方程的替换，得到修改之后的模型$M_{\hat{h}}$：

$$X = U_X$$

$$H = 2$$

$$Y = bX + cH + U_Y$$

$$\forall i,j \in \{X,H,Y\}, \sigma_{U_i U_j} = 0$$

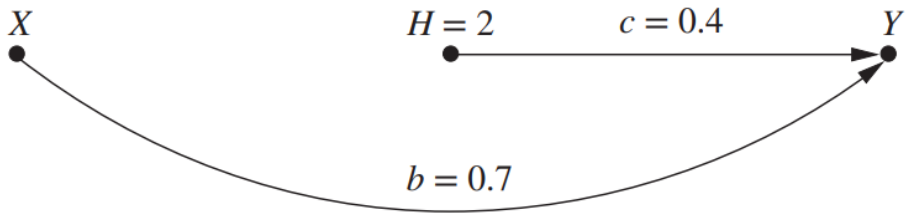

图 8 将家庭作业量固定为 2 后，得到的因果模型图

在修改后的模型中，计算Y的值，得到：

$$Y_{H=2}(U_X = 0.5, U_H = 0.75, U_Y = 0.75) = 0.7 \times 0.5 + 0.4 \times 2 + 0.75 = 1.90$$

因此，如果小明当初的作业量加倍，那么他的考试成绩将变为1.90。

计算反事实的三个步骤

根据以上的例子，可以进一步概括任何确定性模型中反事实值的确定方法，分为以下三步：

（1） 归因（Abduction）：通过观测证据$E=e$计算外生变量U的值；

（2） 行动（Action）：修改模型M，将变量X出现在左边的方程移除，用$X=x$来替代，获得修改之后的模型M_x；

（3） 预测（Prediction）：使用计算得到的外生变量U的值和修改后的模型M_x，计算Y的值，得到反事实的结果。

以上方法可以解决任何确定性的反事实问题，但是反事实问题也可能是一个概率问题。对于概率性模型，计算反事实$E(Y_{X=x}|E=e)$的方法也分为三步：

（1） 归因（Abduction）：通过观测证据$E=e$更新$P(U)$,得到$P(U|E=e)$；

（2） 行动（Action）：修改模型，将变量X出现在左边的方程移除，用$X=x$来替代，获得修改之后的模型M_x；

（3） 预测（Prediction）：使用计算得到的$P(U|E=e)$和修改后的模型M_x，计算Y的期望，得到反事实的结果。

反事实的概率

考虑以下的结构因果模型，其中$X=1$表示上过大学，Z表示技能水平，$U_2=1$表示有工作经验，Y表示薪水：

$$X=U_1$$

$$Z=aX+U_2$$

$$Y=bZ$$

对应的因果图如下：

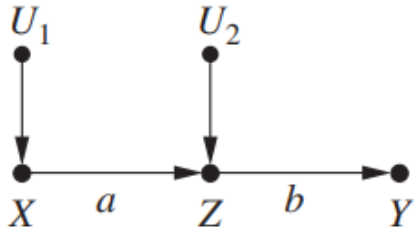

图 9 大学X、技能水平Z和薪水Y之间的因果模型图

假如我们要回答以下问题：假如一个工作技能$Z=1$的个体当初接受了大学教育的话，他的工资的期望值应该是多少？即计算$E(Y_{X=1}|Z=1)$。在这里，事实是$Z=1$，反事实是$X=1$，是当$Z=1$发生的情况下假象的条件。如果假设u_1和u_2都只能取 0 或者 1，那么可能情况有 4 种，对于每一种情况，可以求出对应的$X(u)$、$Z(u)$、$Y(u)$，即可能发生的事实，以及其对应的反事实$Y_0(u)$和$Y_1(u)$。例如，当$u_1=0$，$u_2=1$时，可以得到：

$$X(u)=u_1=0$$

$$Z(u)=a\times 0+u_2=1$$

$$Y(u)=b\times 1=b$$

$Z(u)=1$表明该个体的技能水平等于 1。接着，计算反事实推理结果：

$$Z_0(u)=a\times 0+u_2=1$$

$$Z_1(u)=a\times 1+u_2=a+1$$

$$Y_0(u)=b\times 1=b$$

$$Y_1(u)=b\times(a+1)=(a+1)b$$

用同样的方式可以计算出其他情况下的反事实推理结果，如表 3 所示：

$X = u_1 \quad Z = aX + u_2 \quad Y = bZ$								
u_1	u_2	$X(u)$	$Z(u)$	$Y(u)$	$Y_0(u)$	$Y_1(u)$	$Z_0(u)$	$Z_1(u)$
0	0	0	0	0	0	ab	0	a
0	1	0	1	b	b	$(a+1)b$	1	$a+1$
1	0	1	a	ab	0	ab	0	a
1	1	1	$a+1$	$(a+1)b$	b	$(a+1)b$	1	$a+1$

表 3 u_1和u_1不同取值下，图 9 的模型取值结果

通过观察，表中第二行就是一个工作技能$Z = 1$的个体，我们找到其反事实$Y_0(u)$和$Y_1(u)$：

$$Y_0(u) = b$$

$$Y_1(u) = (a+1)b$$

由于$Y_1(u) - Y_0(u) = ab \neq 0$，因此对于那些工作技能$Z = 1$的个体，$X$对$Y$是有效应的，即如果他们没有接受大学教育，那么假如他们当初接受了大学教育的话，他们的薪水会有所提高。

在上面的计算中，我们没有考虑u_1和u_1发生的概率$P(u_1)$和$P(u_1)$，因为$Z = 1$仅在$u_1 = 0$，$u_2 = 1$的情况下发生（假定$a \neq 0$且$a \neq 1$），此时$P(u_1 = 0, u_2 = 1|Z = 1) = 1$。如果假设模型中$a = 1$，那么$Z = 1$的情况有两种：$u_1 = 0, u_2 = 1$和$u_1 = 1, u_2 = 0$。根据反事实计算的第三步，需要计算条件概率$P(u_1 = 0, u_2 = 1|Z = 1)$和$P(u_1 = 1, u_2 = 0|Z = 1)$，有：

$$P(u_1 = 0, u_2 = 1|Z = 1) = \frac{P(u_1 = 0)P(u_2 = 1)}{P(u_1 = 0)P(u_2 = 1) + P(u_1 = 1)P(u_2 = 0)}$$

$$P(u_1 = 1, u_2 = 0|Z = 1) = \frac{P(u_1 = 1)P(u_2 = 0)}{P(u_1 = 0)P(u_2 = 1) + P(u_1 = 1)P(u_2 = 0)}$$

因此，反事实推理结果，即Y的期望为：

$$\begin{aligned} E(Y_{X=0}|Z = 1) &= b \cdot \frac{P(u_1 = 0)P(u_2 = 1)}{P(u_1 = 0)P(u_2 = 1) + P(u_1 = 1)P(u_2 = 0)} + 0 \\ &\cdot \frac{P(u_1 = 1)P(u_2 = 0)}{P(u_1 = 0)P(u_2 = 1) + P(u_1 = 1)P(u_2 = 0)} \\ &= b \cdot \frac{P(u_1 = 0)P(u_2 = 1)}{P(u_1 = 0)P(u_2 = 1) + P(u_1 = 1)P(u_2 = 0)} \end{aligned}$$

$$E(Y_{X=1}|Z=1)$$
$$= 2b \cdot \frac{P(u_1=0)P(u_2=1)}{P(u_1=0)P(u_2=1)+P(u_1=1)P(u_2=0)} + b \cdot \frac{P(u_1=1)P(u_2=0)}{P(u_1=0)P(u_2=1)+P(u_1=1)P(u_2=0)}$$
$$= b(1+\frac{P(u_1=0)P(u_2=1)}{P(u_1=0)P(u_2=1)+P(u_1=1)P(u_2=0)})$$

我们发现$E(Y_{X=1}|Z=1) > E(Y_{X=0}|Z=1)$，同样说明接受大学教育对薪水的多少是有效应的。

关于反事实推理的更多内容可以参考[6]中第四章。

参考文献

[1] YAO L, CHU Z, LI S, et al. A survey on causal inference [J]. arXiv preprint arXiv:200202770, 2020.

[2] PEARL J. Theoretical impediments to machine learning with seven sparks from the causal revolution [J]. arXiv preprint arXiv:180104016, 2018.

[3] 望止洋. 因果推理初探系列 [Z]. 2020

[4] PEARL J, MACKENZIE D. The book of why: the new science of cause and effect [M]. Basic books, 2018.

[5] PEARL J. Bayesian networks [J]. 2011.

[6] PEARL J, GLYMOUR M, JEWELL N P. Causal inference in statistics: A primer [M]. John Wiley & Sons, 2016.

[7] BAREINBOIM E, PEARL J. Causal inference by surrogate experiments: z-identifiability [J]. arXiv preprint arXiv:12104842, 2012.

[8] SHPITSER I, PEARL J. What counterfactuals can be tested; proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence, UAI 2007, F, 2007 [C].

[9] TIAN J, PEARL J. A general identification condition for causal effects; proceedings of the Aaai/iaai, F, 2002 [C].