Facsímil 12

Lógica e incertidumbre

Capítulo 05

Razonamiento causal: de la correlación a la intervención

686f6c61 · 2026

Capítulo 05: Razonamiento causal: de la correlación a la intervención

Entrando en el tema

En el capítulo 2 una máquina aprendió a razonar con incertidumbre. Le dimos grados de creencia, el teorema de Bayes para actualizarlos y las redes bayesianas para organizar muchas creencias a la vez. Con esa maquinaria se responde a una pregunta muy concreta: dado que observo algo, ¿qué debo creer sobre lo demás? La red de la alarma calculaba $P(\text{Robo} \mid \text{Juan}, \text{María})$; el clasificador de spam calculaba $P(\text{clase} \mid \text{palabras})$. Todo eran probabilidades condicionales, todo era **observar**.

Pero hay una pregunta que esa maquinaria, por sí sola, no sabe contestar. No es “si veo que la alarma sonó, ¿qué creo del robo?”, sino “si yo **hago** sonar la alarma, ¿cambia algo el robo?”. La respuesta es, obviamente, no: forzar la alarma no provoca ladrones. Y sin embargo, en los datos, alarma y robo están correlacionados. Una probabilidad condicional no distingue las dos preguntas. Las dos se escriben casi igual, pero significan cosas radicalmente distintas, y confundirlas es la fuente de la mayoría de los errores cuando un sistema deja de predecir y empieza a **actuar**.

Este capítulo añade la pieza que faltaba: una semántica de la causa. Vamos a aprender a escribir, calcular y razonar sobre lo que ocurre cuando intervenimos en el mundo, no solo cuando lo miramos. La herramienta es la teoría causal de Judea Pearl, que no tira la probabilidad del capítulo 2 a la basura: la extiende con un operador nuevo, el operador `do`, y con una lectura causal de los mismos grafos dirigidos que ya conocemos.[1] El Facsímil 8, capítulo 7, ya recorrió el lado aplicado de esta historia: experimentos, A/B testing, estimadores como el ATE. Aquí damos el **formalismo** que sostiene aquello: por qué un experimento aleatorizado funciona, qué hace exactamente un ajuste por confusor y qué pregunta responde un contrafactual.

Empecemos por la trampa que justifica todo el capítulo.

La correlación miente: la paradoja de Simpson

Imaginemos un hospital que prueba un fármaco nuevo contra una enfermedad. Recoge datos de 200 pacientes: 100 reciben el fármaco, 100 reciben el tratamiento habitual (control). Mira la tasa de recuperación global y se lleva un disgusto:

GRUPO	PACIENTES	RECUPERAN	TASA
Fármaco nuevo	100	54	54 %
Control	100	76	76 %

Leído sin más, el veredicto es demoledor: el fármaco nuevo **empeora** las cosas, 54 % frente a 76 %. Cualquiera lo retiraría. Pero un médico cauto sospecha que los dos grupos no eran comparables y separa a los pacientes por gravedad de partida, la única variable que faltaba:

GRAVEDAD	TRATAMIENTO	PACIENTES	RECUPERAN	TASA
Leve	Fármaco nuevo	10	9	90 %
Leve	Control	90	72	80 %
Grave	Fármaco nuevo	90	45	50 %
Grave	Control	10	4	40 %

Mira las dos últimas columnas con calma. **Entre los casos leves**, el fármaco gana: 90 % frente a 80 %. **Entre los casos graves**, el fármaco también gana: 50 % frente a 40 %. El fármaco es mejor en cada subgrupo por separado y, a la vez, peor en el total. Esto no es un error de cálculo: las cifras encajan ($9 + 45 = 54$, $72 + 4 = 76$). Es la **paradoja de Simpson**: una tendencia presente en todos los grupos se invierte al juntarlos.[2]

¿Cómo es posible? Porque el fármaco nuevo se administró sobre todo a los pacientes **graves** (90 de los 100 que lo recibieron), y los graves se recuperan menos hagas lo que hagas. La gravedad es una variable que influye en dos sitios a la vez: en quién recibe el fármaco (los médicos lo reservaban para los casos difíciles) y en quién se recupera. Esa variable contamina la comparación global. El 54 % del fármaco no es bajo porque el fármaco sea malo, sino porque se probó en la población más enferma.

La pregunta del millón es: ¿cuál de las dos lecturas debo creer? ¿El fármaco ayuda o perjudica? Y aquí está la lección central del capítulo: **los datos solos no lo deciden**. Las mismas cifras serían compatibles con conclusiones opuestas según qué historia causal las generó. Si la gravedad es lo que empuja a recetar el fármaco, hay que mirar los subgrupos y el fármaco ayuda. Si fuera al revés —si el fármaco causara la gravedad, cosa absurda aquí pero no siempre—, habría que mirar el total. La estadística no contiene esa información; hay que traerla de fuera, en forma de un modelo de qué causa qué. Eso es exactamente lo que vamos a construir.

La escalera de la causalidad

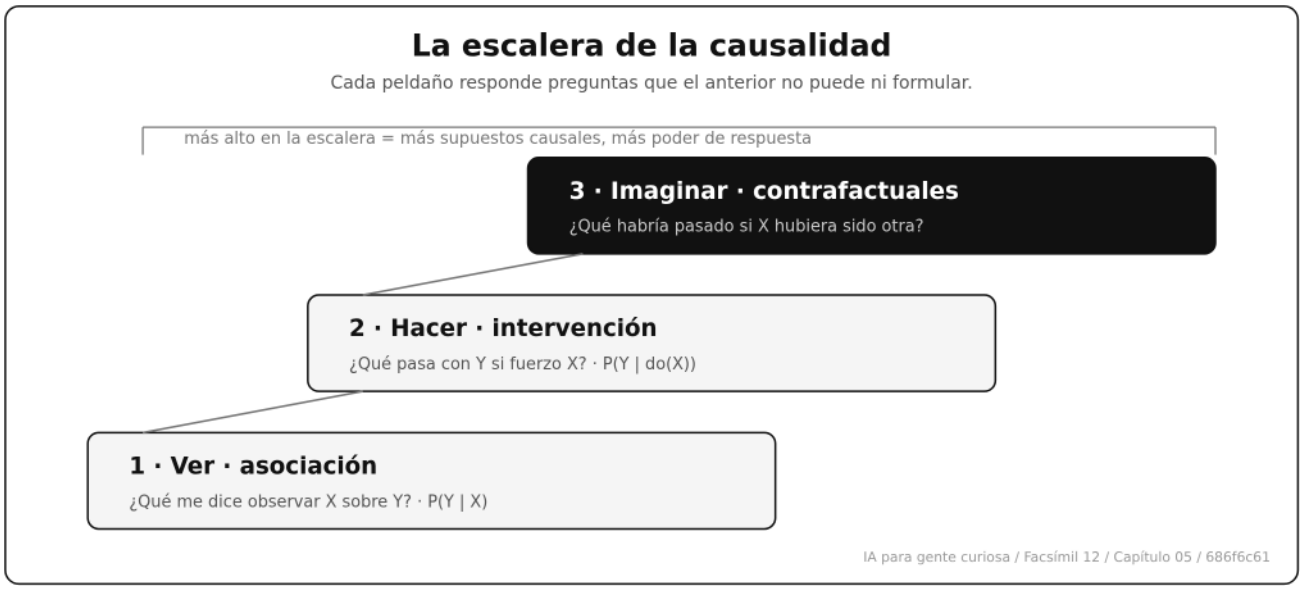
Pearl ordena las preguntas que podemos hacerle al mundo en tres peldaños, una **escalera de la causalidad** en la que cada nivel necesita más que el anterior y responde preguntas que el anterior no alcanza.[3]

PELDAÑO	VERBO	PREGUNTA TÍPICA	QUÉ NECESITA
1. Asociación	ver	¿Qué me dice observar X sobre Y?	Datos y probabilidad condicional.
2. Intervención	hacer	¿Qué pasa con Y si yo fuerzo X?	Un modelo causal o un experimento.
3. Contrafactuales	imaginar	¿Qué habría pasado con Y si X hubiera sido otra, en este caso concreto?	El modelo causal completo.

El **primer peldaño** es donde vive todo el capítulo 2 y casi todo el aprendizaje automático actual. Se calcula con $P(Y \mid X)$: observo un valor y actualizo mi creencia. Un modelo de este nivel puede ser excelente prediciendo y aun así no saber nada de causas. Sabe que los gallos cantan antes del amanecer; no sabe que silenciarlos no retrasa el sol.

El **segundo peldaño** pregunta por intervenciones. Ya no observo X, lo fijo yo. Se escribe con el operador `do`: $P(Y \mid do(X{=}x))$. Aquí es donde el ejemplo del fármaco vive de verdad, porque la decisión clínica es una intervención: «si receto este fármaco, ¿se recuperan más?». Ningún volumen de datos puramente observacionales sube por sí solo a este peldaño; hace falta un experimento o un modelo causal que diga qué ajustar.

El **tercer peldaño** es el más alto y el más humano. Pregunta por mundos que no ocurrieron: «este paciente concreto tomó el fármaco y se recuperó; ¿se habría recuperado igualmente sin él?». Es una pregunta sobre un individuo y sobre un pasado alternativo. De ahí salen las nociones de responsabilidad («¿fue su acción la causa del daño?»), de explicación («¿por qué ocurrió?») y de equidad («¿le habrían dado el crédito si hubiera sido de otro grupo?»). La probabilidad condicional ni siquiera sabe formular estas preguntas.



La moraleja operativa es dura: un modelo entrenado solo con datos observacionales vive en el peldaño 1. Si lo usas para decidir una intervención (peldaño 2), estás extrapolando fuera de lo que el modelo realmente sabe, y la paradoja de Simpson es el aviso de lo mal que puede salir. Subir de peldaño no es cuestión de más datos ni de un modelo más grande: es cuestión de añadir información causal que los datos, por su naturaleza, no contienen.

Grafos causales: las flechas significan causas

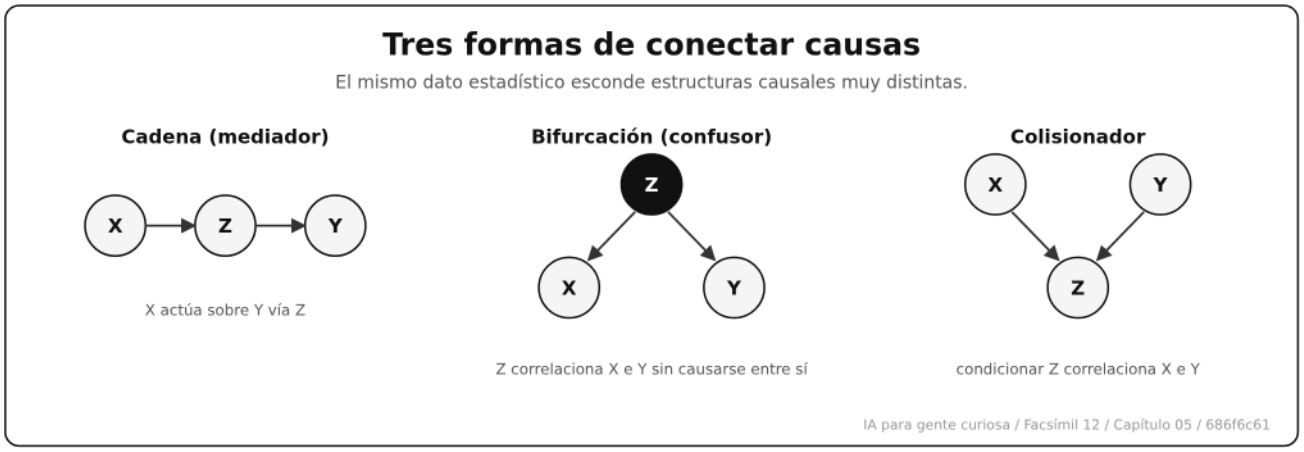
Para subir al segundo peldaño necesitamos escribir nuestras hipótesis sobre qué causa qué. La herramienta es la misma del capítulo 2 con un cambio de lectura: un **grafo dirigido acíclico**, pero ahora cada flecha $X \to Y$ afirma que X es una **causa directa** de Y, no solo que hay dependencia estadística entre ellas.[4] Esta es la diferencia profunda con una red bayesiana corriente: una red bayesiana solo promete factorizar bien la distribución conjunta, y la dirección de sus flechas puede elegirse de varias formas equivalentes. Un grafo causal promete algo más fuerte: que si intervienes en un nodo, el efecto se propaga **siguiendo el sentido de las flechas** y nunca hacia atrás.

Como en el capítulo 2, hay tres formas básicas en que tres variables se conectan, y su comportamiento causal es lo que organiza todo lo demás. Conviene tenerlas grabadas, porque decidir bien una intervención es, en el fondo, saber reconocerlas.

PATRÓN	ESTRUCTURA	LECTURA CAUSAL	QUÉ PASA AL CONDICIONAR EL NODO CENTRAL
Cadena	$X \to Z \to Y$	X causa Y a través de Z (mediador).	Se bloquea: X e Y se vuelven independientes.
Bifurcación	$X \leftarrow Z \to Y$	Z es causa común de X e Y (confusor).	Se bloquea: desaparece la correlación espuria.
Colisionador	$X \to Z \leftarrow Y$	X e Y son causas de Z.	Se abre: aparece correlación falsa entre X e Y.

La **bifurcación** es la villana de este capítulo. Cuando una variable Z causa a la vez a X y a Y, se la llama **confusor**, y es la responsable de la paradoja de Simpson: la gravedad causaba tanto recibir el fármaco como recuperarse. Una bifurcación produce correlación entre X e Y **sin que exista ninguna flecha de X a Y**. Helados y ahogamientos suben juntos no porque el helado ahogue, sino porque el calor (la Z) provoca ambas cosas. Quien confunde esa correlación con causa actuaría prohibiendo los helados para salvar vidas.

El **colisionador** es la trampa contraria y la más sutil. Aquí X e Y son independientes de partida, pero condicionar sobre su efecto común Z las correlaciona artificialmente. Es el *explaining away* del capítulo 2 leído en clave causal: si dos causas independientes pueden producir un mismo efecto y observamos el efecto, saber que una causa estuvo presente reduce la sospecha de la otra. El peligro práctico es feroz: ajustar (condicionar) por un colisionador, lejos de limpiar la comparación, la **ensucia**. No todo lo que se puede medir se debe controlar; controlar la variable equivocada introduce sesgo en lugar de quitarlo.



El operador do: observar no es intervenir

Volvamos al fármaco y escribamos las dos preguntas que tanto se parecen. El grafo causal del experimento es una bifurcación: la gravedad Z causa el tratamiento X (los médicos recetaban según la gravedad) y causa la recuperación Y; y el tratamiento puede causar la recuperación. Es decir, $Z \to X$, $Z \to Y$ y $X \to Y$.

La pregunta de **observación** es $P(Y \mid X{=}x)$: entre los pacientes que de hecho tomaron el fármaco, ¿cuántos se recuperaron? Esa cifra mezcla dos cosas: el efecto real del fármaco y el hecho de que el fármaco fue a parar a los pacientes más graves. Es el 54 % engañoso.

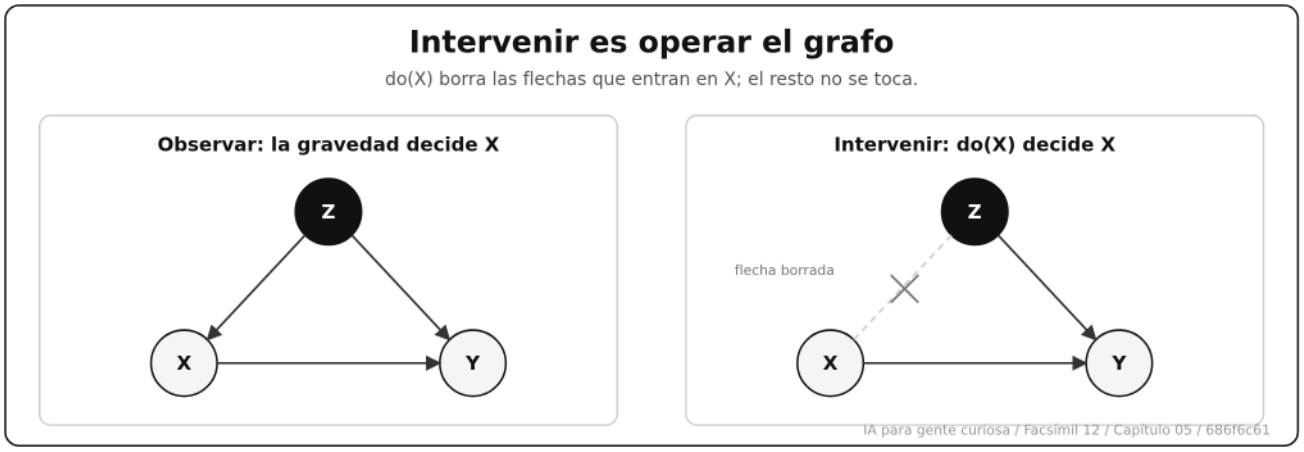
La pregunta de **intervención** es $P(Y \mid do(X{=}x))$: si yo administrara el fármaco a un paciente **sin que la gravedad influya en esa decisión**, ¿cuántos se recuperarían? El operador `do` representa precisamente eso: agarrar la variable X y fijarla por la fuerza, ignorando sus causas habituales. Gráficamente, intervenir con `do(X=x)` equivale a **borrar todas las flechas que entran en** X: la gravedad ya no decide el tratamiento, lo decido yo. Esa amputación quirúrgica del grafo es la definición de Pearl, y es lo que un experimento aleatorizado consigue físicamente: al echar a suertes quién recibe el fármaco, rompemos el vínculo entre la gravedad y el tratamiento. Por eso la aleatorización es la regla de oro, y por eso el Facsímil 8, capítulo 7, dedica tanto cuidado a que la asignación esté de verdad rota respecto a todo lo demás.

$$\underbrace{P(Y \mid X{=}x)}_{\text{observar: el confusor sigue actuando}} \quad \neq \quad \underbrace{P(Y \mid do(X{=}x))}_{\text{intervenir: el confusor ya no decide } X}$$

SÍMBOLO	SIGNIFICADO	EJEMPLO
$P(Y \mid X{=}x)$	Probabilidad de Y **entre quienes se observa** que tienen $X{=}x$.	Recuperación entre los que de hecho tomaron el fármaco (el 54 %).
$P(Y \mid do(X{=}x))$	Probabilidad de Y si **fijamos** $X{=}x$ por intervención.	Recuperación si administramos el fármaco al margen de la gravedad (el 70 %).
$do(X{=}x)$	Operador de intervención: borra las flechas que entran en X.	Repartir el fármaco a suertes, no según el criterio del médico.
$\neq$	Los dos lados difieren mientras quede un confusor sin cerrar.	54 % observado frente a 70 % intervenido.

En palabras: lo que ves entre quienes ya tienen un valor de la causa **no** es lo mismo que lo que pasaría si tú impusieras ese valor a todos. La izquierda mezcla el efecto de la causa con la huella de quién acabó teniéndola; la derecha aísla el efecto, porque al intervenir el confusor deja de elegir quién recibe el tratamiento. Solo coinciden cuando no hay nada que confunda.

Esa amputación de las flechas que entran en X es literal en el grafo. A la izquierda, el mundo observado: la gravedad decide el tratamiento y el resultado a la vez. A la derecha, el mundo intervenido con $do(X{=}x)$: la flecha de la gravedad al tratamiento desaparece, porque ahora quien decide el tratamiento eres tú, no la gravedad. El resto del grafo se queda intacto.



La diferencia entre los dos lados de esa desigualdad es el **sesgo del confusor**. Cuando no hay confusores (cuando ninguna variable causa a la vez X e Y), los dos lados coinciden y observar equivale a intervenir; ese es el mundo cómodo de un experimento bien aleatorizado. Cuando sí los hay, como en el fármaco, la observación está sesgada y hay que corregirla. La pregunta es cómo, y la respuesta tiene nombre.

El criterio de la puerta trasera y la fórmula de ajuste

El sesgo del confusor entra por un camino concreto del grafo. Entre el tratamiento X y el resultado Y hay dos tipos de caminos: el **camino causal** que va de X hacia Y siguiendo las flechas ($X \to Y$), que es el que queremos medir, y los **caminos de puerta trasera**, que conectan X e Y pero empiezan con una flecha que entra en X. En el fármaco, ese camino traicionero es $X \leftarrow Z \to Y$: sale de X por detrás, pasa por la gravedad y desemboca en la recuperación. Por ahí se cuela la correlación espuria.

El **criterio de la puerta trasera** de Pearl dice cómo cerrar esos caminos: hay que encontrar un conjunto de variables Z que, al condicionar sobre ellas, **bloquee todos los caminos de puerta trasera** sin bloquear el camino causal y sin abrir colisionadores. Cuando ese conjunto existe y lo observamos, el efecto causal se vuelve calculable a partir de datos observacionales mediante la **fórmula de ajuste**:

$$P\big(Y \mid do(X{=}x)\big) = \sum_{z} P\big(Y \mid X{=}x,\, Z{=}z\big)\, P(Z{=}z)$$

SÍMBOLO	SIGNIFICADO	EJEMPLO
$P(Y \mid do(X{=}x))$	Probabilidad de Y si intervenimos fijando $X{=}x$.	Recuperación si administramos el fármaco.
$X{=}x$	La causa, fijada al valor que evaluamos.	`fármaco = sí`.
$Z{=}z$	Cada valor del confusor que ajustamos.	Gravedad leve o grave.
$P(Y \mid X{=}x,\, Z{=}z)$	Tasa observada de Y dentro de cada subgrupo.	Recuperación de tratados leves, de tratados graves.
$P(Z{=}z)$	Peso de cada subgrupo en la población **entera**.	Qué fracción de pacientes es leve y cuál grave.
$\sum_z$	Suma sobre todos los valores del confusor.	Recombinar los subgrupos.

En palabras: mide el efecto del tratamiento **dentro de cada subgrupo del confusor**, donde la comparación sí es justa, y luego promedia esos efectos usando como pesos lo que pesa cada subgrupo en la población total, no en el grupo tratado. Esa última coletilla es la clave: el confusor sesgaba porque el tratamiento se repartía mal entre subgrupos; al promediar con los pesos de la población completa, deshacemos ese reparto torcido.

Hagamos el cálculo con las cifras del hospital. De los 200 pacientes, 100 son leves y 100 graves, así que $P(Z{=}\text{leve}) = P(Z{=}\text{grave}) = 0{,}5$. Las tasas de recuperación por subgrupo ya las tenemos en la tabla. Para el fármaco:

$$P(Y \mid do(\text{fármaco})) = \underset{\text{leves}}{0{,}90 \cdot 0{,}5} + \underset{\text{graves}}{0{,}50 \cdot 0{,}5} = 0{,}45 + 0{,}25 = 0{,}70$$

Y para el control, repitiendo la operación con sus tasas:

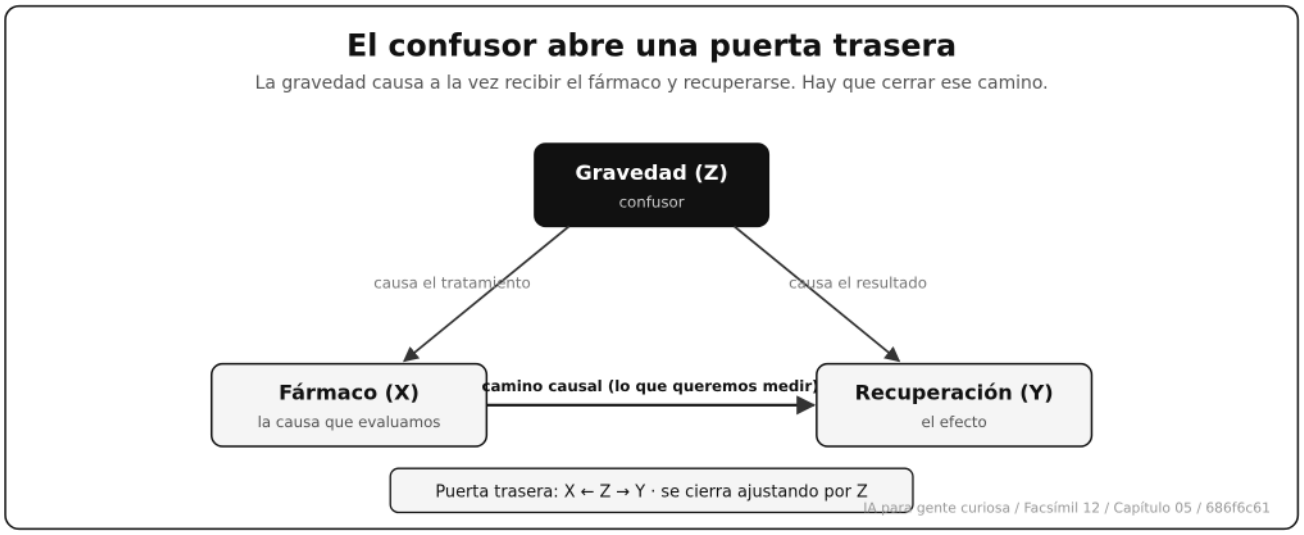
$$P(Y \mid do(\text{control})) = 0{,}80 \cdot 0{,}5 + 0{,}40 \cdot 0{,}5 = 0{,}40 + 0{,}20 = 0{,}60$$

En palabras: en lugar de comparar el grupo tratado con el de control tal como quedaron (uno lleno de graves, el otro de leves), cogemos la tasa de cada subgrupo y la pesamos por lo que ese subgrupo representa en **toda** la población, mitad y mitad. Al usar los mismos pesos para los dos brazos, ambos se evalúan sobre una población idéntica y la comparación deja de estar contaminada por el reparto torcido del tratamiento.

El efecto causal del fármaco es $0{,}70 - 0{,}60 = +0{,}10$: **diez puntos porcentuales más de recuperación**. El fármaco ayuda, justo lo contrario de lo que gritaba la cifra global. Compara las dos lecturas en una tabla:

CANTIDAD	CÁLCULO	RESULTADO	CONCLUSIÓN
Observación ingenua	$P(Y \mid \text{fármaco}) - P(Y \mid \text{control})$	$0{,}54 - 0{,}76 = -0{,}22$	El fármaco parece dañino.
Intervención ajustada	$P(Y \mid do(\text{fármaco})) - P(Y \mid do(\text{control}))$	$0{,}70 - 0{,}60 = +0{,}10$	El fármaco ayuda.

El ajuste por el confusor **invierte la conclusión**, de -22 puntos a $+10$. Y esto no es magia estadística: es la consecuencia matemática de haber escrito el grafo causal correcto y de haber bloqueado el camino de puerta trasera $X \leftarrow Z \rightarrow Y$. Si la gravedad no hubiera sido un confusor, ajustar por ella no habría cambiado nada. El grafo es lo que nos dice **qué ajustar**; sin él, ajustar por una variable es tan probable que ayude como que estropee.



Conviene subrayar lo que la fórmula de ajuste **no** es. No es «controlar por todas las variables que tengas». Recuerda el colisionador: ajustar por la variable equivocada introduce sesgo. El criterio de la puerta trasera es preciso justamente para evitar ese error: ajusta por lo que cierra puertas traseras, deja en paz mediadores y colisionadores. Y tiene un límite honesto: solo funciona si el confusor está **medido**. Si la gravedad no se hubiera registrado, ningún cálculo la habría recuperado; haría falta un experimento. Esa es la diferencia entre tener datos y poder concluir causas.

El colisionador: cuando condicionar inventa correlación

Si el confusor es la trampa que nos empuja a ajustar, el **colisionador** es la trampa que nos castiga por ajustar de más. Su criterio es la imagen invertida del de la puerta trasera, y conviene enunciarlo aparte porque es responsable de algunos de los errores más caros y menos evidentes del análisis de datos: **condicionar sobre un colisionador (o sobre**

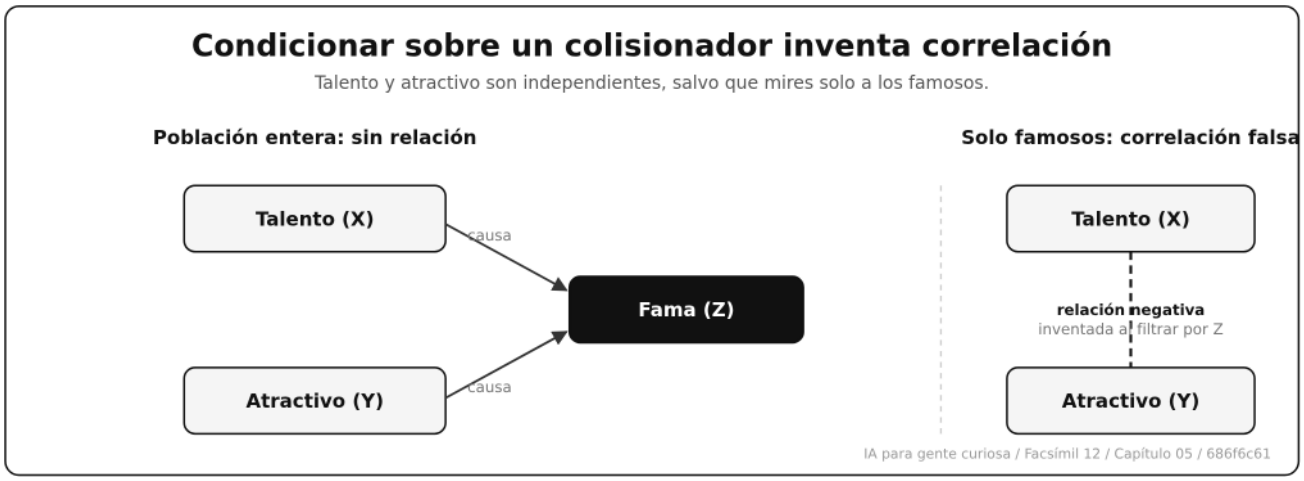
cualquier descendiente suyo) abre un camino que estaba cerrado y crea correlación donde no había ninguna. No hace falta ninguna flecha entre X e Y para que, tras filtrar por su efecto común, parezcan relacionados.

El mecanismo es el *explaining away* del capítulo 2 visto en otra clave. Tomemos un ejemplo de calle. Imagina dos rasgos de los actores que llegan a ser famosos: el **talento** X y el **atractivo físico** Y. En la población general no tienen por qué ir juntos: hay gente con mucho talento y poco atractivo, y al revés. Pero para hacerse famoso ayuda cualquiera de los dos, así que la **fama** Z es un colisionador, $X \rightarrow Z \leftarrow Y$. Ahora haz lo que hace cualquiera sin darse cuenta: mira **solo a los famosos**. Dentro de ese grupo aparece una correlación negativa entre talento y atractivo que en la población entera no existía. ¿Por qué? Porque si alguien es famoso y resulta que tiene poco talento, algo lo explica: seguramente es muy atractivo. Filtrar por el efecto común hace que sus dos causas compitan por explicarlo, y eso las correlaciona artificialmente. Quien observara solo a los famosos concluiría, falsamente, que el talento «estropea» el atractivo.

El mismo patrón, con otro disfraz, aparece por todas partes:

- **Admisiones y becas.** Una universidad admite por nota o por mérito deportivo. Entre los admitidos, ambas cosas salen correlacionadas a la baja, aunque en los aspirantes fueran independientes. Estudiar solo a los admitidos engaña.
- **Sesgo de selección en sanidad.** Si analizas únicamente a pacientes **hospitalizados**, dos enfermedades que llevan al hospital pueden parecer asociadas (la *paradoja de Berkson*) sin que una tenga nada que ver con la otra.
- **Contratación.** Entre los empleados ya contratados, la destreza técnica y la soltura en la entrevista pueden correlacionar negativamente: si te contrataron pese a una entrevista floja, tu perfil técnico debía de ser muy fuerte.

La regla práctica es incómoda porque contradice el instinto de «cuantas más variables controle, más limpio el análisis»: **no todo lo que se puede medir se debe controlar**. Condicionar (ajustar, estratificar, o simplemente seleccionar la muestra) por un colisionador no quita sesgo, lo fabrica. El grafo causal es, de nuevo, lo único que distingue un confusor (que sí hay que ajustar) de un colisionador (que hay que dejar en paz), porque en los datos brutos ambos se ven igual de correlacionados.



Hay un tercer error de ajuste, simétrico del anterior, que conviene nombrar para cerrar el cuadro: ajustar por un **mediador**. Si la causa actúa **a través** de una variable intermedia, $X \to M \to Y$, esa M no es un confusor ni un colisionador, sino el propio canal por el que viaja el efecto. Imagina que el fármaco cura reduciendo la inflamación, y que la inflamación reducida es lo que mejora la recuperación. Si controlas por la inflamación, **borras** justo la vía por la que el fármaco hace su trabajo, y el efecto medido se desploma hacia cero aunque el fármaco funcione. Esto se llama **sesgo de sobrecontrol**, y es la cara opuesta del colisionador: el colisionador inventa correlación, el mediador la suprime. La conclusión une las tres figuras en una sola regla: ajusta por confusores, deja en paz mediadores y colisionadores. Cuál es cuál no lo dicen los datos, lo dice el grafo.

Contrafactuales: qué habría pasado si

El tercer peldaño pregunta por mundos alternativos sobre casos concretos. «María tomó el fármaco y se recuperó. ¿Se habría recuperado igualmente sin tomarlo?» Esta no es una pregunta sobre promedios de población, sino sobre **un individuo y un pasado que no ocurrió**. Si la respuesta es «sí, se habría recuperado igual», entonces el fármaco no fue la causa de su recuperación, aunque la tomara y se curara. Si la respuesta es «no, sin el fármaco no se habría recuperado», entonces el fármaco fue, para ella, decisivo.

Formalizar contrafactuales requiere maquinaria que va más allá de lo que cubre este capítulo: hace falta el modelo causal completo, con sus ecuaciones, para «rebobinar» el caso, cambiar el valor de la causa y volver a simular el desenlace. Pearl describe ese rebobinado en tres pasos, y aunque la mecánica completa quede fuera de aquí, el esquema mental es sencillo y vale la pena tenerlo:

1. **Mirar el caso real y deducir lo que no se ve.** María tomó el fármaco y se recuperó. De ese hecho, y del modelo, inferimos las circunstancias individuales que el promedio de la población esconde: su gravedad, su respuesta al tratamiento, todo lo que la hace ser ella y no «un paciente medio».

2. **Cambiar solo la causa, dejando lo demás fijo.** Imaginamos a esa misma María, con esas mismas circunstancias, pero sin el fármaco. No es otra paciente parecida: es ella, con un único cambio.
3. **Volver a calcular el desenlace.** Con la causa cambiada y todo lo demás congelado, preguntamos si se habría recuperado.

Pongamos números a la intuición. Supongamos que el modelo dice que, en su subgrupo, el fármaco solo es decisivo para una parte de los pacientes; para el resto, se habrían curado igual o no se habrían curado de ninguna manera. Si las circunstancias de María la sitúan entre los primeros, el contrafactual «sin fármaco, no se recupera» es verdadero **para ella**, aunque a escala de población el fármaco solo suba la tasa diez puntos. Aquí está la diferencia crucial: el efecto promedio de la intervención (peldaño 2) responde «de cada cien tratados, diez se curan que no se habrían curado»; el contrafactual (peldaño 3) responde «**esta** persona concreta, ¿es una de esas diez?». Son preguntas distintas, y la segunda exige más supuestos que la primera, porque mezcla lo observado con un pasado que no ocurrió.

La intuición de los dos mundos posibles para un mismo caso (lo que pasa con tratamiento y lo que pasaría sin él) la formalizó Rubin con los resultados potenciales, el lenguaje hermano del de Pearl para hablar de efectos individuales que solo observamos a medias.[5] De cada persona solo vemos uno de los dos mundos (tomó el fármaco o no lo tomó), nunca los dos a la vez; ese es el **problema fundamental de la inferencia causal**, y es la razón de que un contrafactual individual nunca se compruebe directamente, solo se estime con un modelo. Pero la **idea** es accesible y es la que importa para entender por qué la causalidad es indispensable. Tres lugares donde un contrafactual hace todo el trabajo:

- **Responsabilidad.** Decir que una acción «causó» un daño es, en el fondo, afirmar un contrafactual: sin esa acción, el daño no habría ocurrido. Los tribunales razonan así («but for» en el derecho anglosajón). Un sistema de IA que reparte culpas o méritos está, lo sepa o no, haciendo afirmaciones contrafactuales.
- **Equidad.** Una decisión es injusta por discriminación si **habría sido distinta** caso de cambiar solo el atributo protegido (género, etnia) manteniendo todo lo demás. Eso es un contrafactual puro, y es la base de las definiciones causales de equidad que veremos asomar en el Facsímil 9.
- **Explicación.** «¿Por qué me denegaron el crédito?» se responde mejor con un contrafactual accionable: «si tus ingresos hubieran sido 3 000 en lugar de 2 000, te lo habrían concedido». Eso le dice a la persona qué cambiar; una mera correlación, no.

El caso de la equidad merece un segundo vistazo porque enlaza directamente con el Facsímil 9. Un modelo puede no usar el género como entrada y aun así discriminar, si se apoya en una variable que **es** consecuencia del género (el tipo de empleo, el historial, el barrio). Distinguir cuándo eso es injusto exige un contrafactual: «¿la decisión sería la misma si esta persona hubiera nacido de otro género, arrastrando los cambios que ello implica?». Esa pregunta es la que separa una correlación legítima con el resultado de un camino de discriminación, y es justo el tipo de distinción que una métrica puramente estadística no sabe hacer; por eso muchas definiciones de equidad numéricas se contradicen entre sí. Aquí también acecha el

colisionador: condicionar sobre una variable de selección (quién fue contratado, quién fue admitido) puede inventar o esconder disparidades entre grupos, y solo el grafo causal dice si la cifra que miras es real o un artefacto de a quién estás mirando.

La conexión con el capítulo 2 es elegante: una red bayesiana corriente vive en el peldaño 1, un grafo causal con el operador `do` sube al peldaño 2, y un modelo causal estructural completo alcanza el peldaño 3. Es la misma familia de grafos ganando, en cada nivel, más capacidad de respuesta a cambio de más compromiso con cómo funciona el mundo.[6]

En el día a día

Aunque suene a manual, esta distinción decide cosas todos los días. Un panel de marketing muestra que los usuarios que reciben la newsletter compran más; ¿conviene mandarla a todos? Solo si el efecto es causal y no un confusor (quizá quien se suscribe ya era más fiel). Un hospital ve que los pacientes ingresados en cierta unidad mueren más; ¿es peor la unidad, o es que allí mandan los casos más graves? Es la paradoja de Simpson otra vez, con vidas de por medio. Un equipo de producto observa que los usuarios que usan una función avanzada retienen mejor; ¿forzar la función mejorará la retención, o solo los usuarios ya comprometidos la descubren?

La misma trampa asoma en la calle comercial. **Precios y demanda:** los datos muestran que los meses en que más coches vende un concesionario son los de precio más alto; ¿subir el precio venderá más? Por supuesto que no: el confusor es la temporada (en primavera sube la demanda y suben los precios a la vez). Quien lea esa correlación como una palanca se arruina. **Publicidad y ventas:** una marca gasta más en anuncios justo en campaña de rebajas, cuando las ventas suben por sí solas; el panel atribuye al anuncio un efecto que era del calendario. **Ejercicio y salud:** quien hace deporte vive más, sí, pero parte de esa correlación la sostiene un confusor (quien ya está sano hace más deporte), y solo un diseño cuidadoso separa el efecto real del ejercicio de la salud previa que lo permite. En todos, la pregunta es la misma: ¿la cifra mide el efecto de mover la palanca, o solo refleja quién acabó moviéndola?

Por eso el patrón oro sigue siendo intervenir de verdad. El **A/B testing** es exactamente el operador do hecho producto: al repartir a los usuarios a suertes entre la versión A y la B, rompemos por diseño cualquier vínculo entre la asignación y los rasgos del usuario, de modo que la diferencia observada **sí** es $P(Y \mid do(X))$ y no una correlación contaminada. Cuando no se puede experimentar (por coste, ética o tiempo), queda la otra vía: ajustar por los confusores correctos guiados por un grafo causal. Lo que nunca es válido es tratar una correlación observada como si fuera el resultado de una intervención.

La pregunta higiénica que conviene hacerse siempre es: «¿esta cifra viene de **mirar** o de **hacer**?». Si viene de mirar y vas a actuar, falta un paso. Ese paso es, según los casos, un experimento (Facsímil 8, capítulo 7) o un ajuste por los confusores correctos guiado por un grafo causal.

Por qué debería importarte

Porque la IA moderna está, casi toda, atrapada en el primer peldaño. Un modelo entrenado por máxima verosimilitud aprende correlaciones: $P(Y \mid X)$ a gran escala.[7] Eso basta para predecir, y a veces deslumbra. Pero en cuanto el modelo **actúa** (recomienda, prioriza, decide a quién se concede un crédito, qué prueba se ordena, qué caso se escala), está tomando decisiones de intervención con una herramienta de observación. Y ahí aparecen tres fallos caros.

El primero son las **correlaciones espurias y los atajos** (*shortcut learning*): un clasificador que distingue lobos de huskies fijándose en la nieve del fondo, o un modelo médico que detecta neumonía por el tipo de aparato de rayos X en lugar de por la patología. Funcionan en los datos de entrenamiento porque la correlación está ahí, y fallan estrepitosamente cuando el mundo cambia o cuando intervienes. Son confusores no controlados disfrazados de rasgos predictivos. Visto con la lente de este capítulo, el atajo es exactamente el camino de puerta trasera del clasificador: en el aparato de rayos X concreto influye el tipo de hospital, que a la vez se asocia con la gravedad de los pacientes que recibe; el modelo aprende ese rodeo en lugar del camino causal que va de la patología al diagnóstico. Por eso un atajo no se arregla con más datos del mismo sitio, igual que el sesgo del confusor no se arreglaba con más filas: lo que falta es saber qué rasgo es causa y cuál es solo compañía de viaje.

El segundo es la **falta de robustez**. Un modelo que solo correlaciona se hunde cuando la distribución se mueve, porque las correlaciones espurias no sobreviven al cambio de entorno mientras que las relaciones causales sí tienden a hacerlo. Aprender la causa, y no el atajo, es una de las apuestas más serias para construir IA que no se rompa fuera del laboratorio.

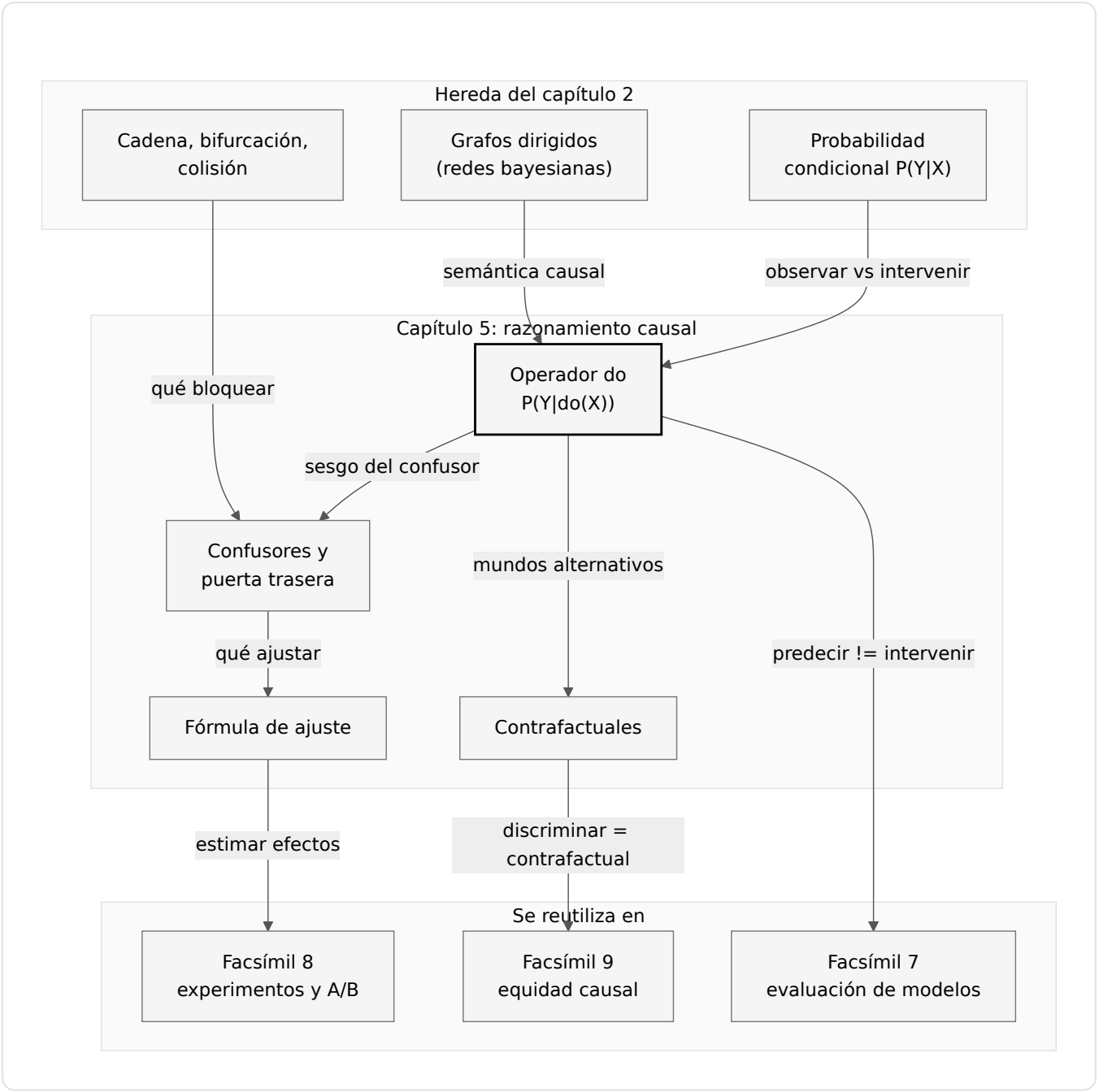
El tercero es la **equidad**. Como vimos, discriminar es un concepto contrafactual, y muchas métricas de equidad puramente estadísticas se contradicen entre sí precisamente porque ignoran la estructura causal de cómo se generan los datos. El Facsímil 9 retoma este hilo: sin causalidad, «justo» es una palabra que cada cual rellena a su gusto. La causalidad no resuelve sola el problema ético, pero da el lenguaje para plantearlo sin trampas.

Dónde solía tropezar yo

ERROR	POR QUÉ ES UN ERROR	ANTÍDOTO
Leer $P(Y \mid X)$ como si fuera $P(Y \mid do(X))$	Observar e intervenir coinciden solo si no hay confusores; con confusores, la observación está sesgada.	Pregúntate si la cifra viene de mirar o de hacer. Si vas a actuar, ajusta o experimenta.
Ajustar por todo lo que tengo medido	Condicionar sobre un colisionador o un mediador introduce sesgo en lugar de quitarlo.	Dibuja el grafo y aplica el criterio de la puerta trasera: ajusta solo lo que cierra puertas traseras.
Creer que más datos resuelven la confusión	El sesgo del confusor no desaparece con el tamaño de la muestra; un millón de filas confundidas siguen confundidas.	Lo que falta es información causal (qué causa qué), no más filas.
Mirar solo la cifra agregada	La paradoja de Simpson invierte conclusiones al juntar grupos con distinto reparto del tratamiento.	Desagrega por el confusor sospechoso y compara con el agregado antes de concluir.
Confundir un buen predictor con una buena palanca	Una variable puede predecir Y sin causarla; forzarla no mueve Y.	Antes de intervenir sobre una variable, comprueba que hay una flecha causal hacia el resultado, no solo correlación.

Cómo encaja todo

Este capítulo cierra el facsímil dando el salto que faltaba: de la creencia a la causa. Hereda directamente del capítulo 2 los grafos dirigidos y la probabilidad condicional, pero les añade una semántica nueva (el operador `do`) que distingue observar de intervenir. Lo que aquí se aprende (grafos causales, confusores, puerta trasera, ajuste y contrafactuales) es el cimiento formal de tres lugares del facsímil: los experimentos y la causalidad aplicada del Facsímil 8, las definiciones causales de equidad del Facsímil 9, y la evaluación honesta de modelos del Facsímil 7, que solo tiene sentido si distinguimos lo que un modelo predice de lo que provoca al actuar.



Leído en palabras: las redes bayesianas del capítulo 2 nos dieron los grafos y la probabilidad condicional, pero se quedaban en el peldaño de la asociación. Al cambiar la lectura de las flechas a «causa directa» y añadir el operador `do`, subimos al peldaño de la intervención; los patrones de conexión que ya conocíamos (cadena, bifurcación, colisión) nos dicen ahora qué variables bloquear, el criterio de la puerta trasera elige el conjunto correcto y la fórmula de ajuste lo convierte en un cálculo. De ahí salen, hacia delante, la estimación de efectos del Facsímil 8, las definiciones contrafactuales de equidad del Facsímil 9 y la disciplina del Facsímil 7 de no confundir un buen predictor con una buena palanca.

Vocabulario aprendido

TÉRMINO	DEFINICIÓN
Correlación frente a causalidad	Que dos cosas varíen juntas no implica que una cause la otra.
Grafo causal	Grafo dirigido acíclico cuyas flechas representan relaciones de causa a efecto.
Operador do	Notación de Pearl para una intervención: fijar una variable por la fuerza, no observarla.
Confusor	Variable que causa a la vez la supuesta causa y el efecto, y crea correlación espuria.
Criterio de la puerta trasera	Regla gráfica para elegir qué variables ajustar y estimar un efecto causal.
Fórmula de ajuste	Cálculo del efecto de una intervención sumando sobre los confusores controlados.
Colisionador	Variable causada por otras dos; condicionar sobre ella crea correlación falsa entre sus causas.
Escalera de la causalidad	Los tres niveles de Pearl: asociación (ver), intervención (hacer) y contrafactuales (imaginar).
Contrafactual	Afirmación sobre qué habría pasado si algo hubiera sido distinto.
Paradoja de Simpson	Una tendencia que aparece en grupos separados se invierte al juntarlos.

Antes de pasar página

- [ ] ¿Sé explicar con un ejemplo por qué $P(Y \mid X)$ y $P(Y \mid do(X))$ son distintas? (Si no, vuelve a «El operador do: observar no es intervenir».)
- [ ] ¿Entiendo cómo la paradoja de Simpson invierte una conclusión y qué papel juega el confusor? (Si no, vuelve a «La correlación miente: la paradoja de Simpson».)
- [ ] ¿Distingo los tres peldaños de la escalera de la causalidad y qué pregunta responde cada uno? (Si no, vuelve a «La escalera de la causalidad».)
- [ ] ¿Reconozco una cadena, una bifurcación y un colisionador, y sé qué pasa al condicionar el nodo central de cada uno? (Si no, vuelve a «Grafos causales: las flechas significan causas».)

- [ ] ¿Sé aplicar la fórmula de ajuste por puerta trasera y por qué uso los pesos de la población entera? (Si no, vuelve a «El criterio de la puerta trasera y la fórmula de ajuste».)
- [ ] ¿Entiendo por qué ajustar (o seleccionar la muestra) por un colisionador mete sesgo en lugar de quitarlo? (Si no, vuelve a «El colisionador: cuando condicionar inventa correlación».)
- [ ] ¿Sé formular un contrafactual y conectarlo con responsabilidad, equidad o explicación? (Si no, vuelve a «Contrafactuales: qué habría pasado si».)
- [ ] ¿Sé decir qué lectura (efecto ingenuo o ajustado) usaría para decidir y por qué? (Si no, repasa la sección correspondiente del capítulo.)

En resumen

IDEA FUERZA	DETALLE
La correlación no basta para decidir.	$P(Y \mid X)$ responde qué observo; para actuar necesito $P(Y \mid do(X))$, que puede ser muy distinta.
La paradoja de Simpson es el aviso.	Una tendencia presente en cada subgrupo puede invertirse al agregar, por culpa de un confusor.
La causalidad sube una escalera de tres peldaños.	Asociación (ver), intervención (hacer) y contrafactuales (imaginar); cada uno exige más que el anterior.
El operador do borra las flechas que entran en la causa.	Intervenir es fijar X por la fuerza; un experimento aleatorizado lo consigue físicamente.
La fórmula de ajuste mide efectos a partir de datos.	Si el confusor está medido y cierra la puerta trasera, se promedian los efectos por subgrupo con los pesos de la población.
No todo lo medido se debe controlar.	Ajustar por un colisionador inventa correlación y ajustar por un mediador la borra; solo el grafo dice qué es un confusor.
La IA que solo correlaciona falla al intervenir.	Atajos, falta de robustez e injusticia nacen de confundir el primer peldaño con el segundo.

Para saber más

Koller, D. y Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2.ª ed.). Cambridge University Press.

Pearl, J., Glymour, M. y Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. Wiley.

Pearl, J. y Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5), 688-701. https://doi.org/10.1037/h0037350

Russell, S. y Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4.ª ed.). Pearson.

Spirtes, P., Glymour, C. y Scheines, R. (2000). *Causation, Prediction, and Search* (2.ª ed.). MIT Press.

Notas

1. Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2.ª ed.). Cambridge University Press. Pearl formaliza la diferencia entre observar e intervenir mediante el operador `do`, demuestra el criterio de la puerta trasera y construye toda una jerarquía de preguntas causales que la teoría de la probabilidad clásica no sabía distinguir. ↩

2. Pearl, J., Glymour, M. y Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. Wiley. El manual dedica un tratamiento detallado a la paradoja de Simpson y muestra que la pregunta «¿qué cifra es la correcta, la agregada o la de los subgrupos?» no se resuelve con más estadística, sino con un modelo causal que diga cuál es el confusor. ↩

3. Pearl, J. y Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books. El libro divulga la escalera de la causalidad (asociación, intervención, contrafactuales) como metáfora central y argumenta que el aprendizaje automático basado solo en correlaciones se queda atrapado en el primer peldaño. ↩

4. Spirtes, P., Glymour, C. y Scheines, R. (2000). *Causation, Prediction, and Search* (2.ª ed.). MIT Press. Los autores desarrollan la teoría de los grafos causales y los algoritmos que intentan recuperar la estructura causal a partir de datos observacionales, estableciendo qué se puede y qué no se puede aprender sin intervenir. ↩

5. Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5), 688-701. https://doi.org/10.1037/h0037350. Rubin escribe, para cada unidad, el resultado bajo tratamiento y bajo control, de los que solo uno se observa; el otro es el contrafactual, y de ahí surge el problema fundamental de la inferencia causal. ↩

6. Koller, D. y Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press. El tratado conecta las redes bayesianas con los modelos causales y muestra que la intervención y los contrafactuales exigen información estructural que la distribución conjunta,

por sí sola, no determina. ↩

7. Russell, S. y Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4.ª ed.). Pearson. El capítulo sobre redes causales presenta el operador `do`, la fórmula de ajuste y los modelos causales estructurales como la extensión natural de las redes bayesianas para razonar sobre acciones, y advierte de los riesgos de tratar un modelo predictivo como si entendiese causas. ↩