ArtefaCToS. Revista de estudios de la ciencia y la tecnología
eISSN: 1989-3612
Vol. 9, No. 1 (2020), 2.ª Época, 13-45
DOI: https://dx.doi.org/10.14201/art2020911345

Teoría normativa y Psicología descriptiva en la comprensión del razonamiento causal: Papel de las intervenciones y la invarianza*

Normative Theory and Descriptive Psychology in Understanding Causal Reasoning: The Role of Interventions and Invariance

James F. WOODWARD

Universidad de Pittsburgh, Estados Unidos
jfw@pitt.edu

Recibido: 23/01/2020. Revisado: 29/01/2020. Aceptado: 03/02/2020

Resumen

Este artículo explora algunas relaciones entre [dos aspectos:] por un lado, las teorías filosóficas normativas de la causalidad y el razonamiento causal; y, por otro lado, las teorías descriptivas de la cognición causal del tipo generado en Psicología. Estos temas se tratan desde la perspectiva de una concepción intervencionista de la causalidad. La atención se centra en lo que llamo distinciones *entre* las relaciones causales, en términos de características tales como la invarianza, la especificidad y la proporcionalidad, y la relevancia psicológica de estas [relaciones]. Se argumenta que las teorías normativas y descriptivas sobre la causalidad tienen mucho que aprender unas de otras.

Palabras clave: causalidad; intervencionismo; invarianza; especificidad causal; proporcionalidad.

---

* Las ideas desarrolladas en este texto han recibido una gran influencia de los debates en curso con los miembros de la *Initiative on Causal Learning* (Iniciativa sobre aprendizaje causal), financiada por la Fundación James S. McDonnell, con Alison Gopnik como investigadora principal (nota del autor).

Abstract

This paper explores some relationships between, on the one hand, normative philosophical theories of causation and causal reasoning and, on the other hand, descriptive theories of causal cognition of the sort produced in psychology. These issues are discussed from the perspective of an interventionist account of causation. The focus is on what I call distinctions *among* causal relationships in terms of such features as invariance, specificity and proportionality and the psychological significance of these. It is argued that normative and descriptive theorizing about causation have a great deal to learn from each other.

Keywords: causation; interventionism; invariance; causal specificity; proportionality.

1. Introducción: Lo normativo y lo descriptivo

La creciente literatura sobre causalidad (*causation*) y conocimiento causal [*causal cognition*] (aprendizaje, razonamiento y juicio causales) abarca muchas disciplinas diferentes, que incluyen la Filosofía, la Estadística, la Inteligencia Artificial y el aprendizaje de máquina (*machine learning*), y la Psicología. Pero, a grandes rasgos, esta investigación puede dividirse en dos categorías, aun cuando la división esté lejos de ser nítida y haya una superposición (*overlap*) considerable. En primer lugar, está el trabajo que es de índole primordialmente *normativa*, en cuanto que se propone (*purport*) decir cómo *deberíamos* aprender y razonar acerca de las relaciones causales y [cómo deberíamos] hacer juicios causales. Este foco normativo es quizás más obvio en el caso de las teorías del aprendizaje y la inferencia causales a partir de diversos tipos de datos, que han sido desarrollados en Estadística y aprendizaje de máquina (*machine learning*). Estas teorías hacen propuestas sobre qué inferencias orientadas a conclusiones causales están *justificadas* (esto es, cuáles de esas inferencias llevan de manera fiable [*reliably*] al logro de alguna meta [*goal*] epistémica, como la verdad). En esta categoría de teorías normativas del aprendizaje causal incluyo a las técnicas convencionales de hacer modelos (*modeling*) causales basadas en modelos de ecuaciones estructurales (*structural equations modeling*); el enfoque de la inferencia causal sobre la base de limitaciones (*constraint-based*), desarrollado por Spirtes, Glymour y Scheines (2000); las ideas sobre razonamiento causal que se describen en Pearl (2000); las muchas variedades de enfoques bayesianos de la inferencia causal (véase, p. ej., Griffiths y Tenenbaum, 2009); y otras propuestas sobre inferencia causal que se basan en ideas del aprendizaje de máquina (*machine learning*), como las que se deben a Bernhard Scholkopf y sus colaboradores (Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniusis, Bastian Steudel y Bernhard Scholkopf, 2012). Quizá sea menos habitual pensar en los distintos enfoques de la causalidad y el razonamiento causal que se encuentran en la literatura filosófica como enfoques que tienen también una aspiración normativa (*nor-*

mative in aspiration), pero creo que podría ser útil verlos de esta manera, aunque también pueden tener otras metas (*goals*)[1]. Prácticamente todas esas propuestas filosóficas pueden verse como recomendaciones acerca de cómo *debemos* pensar sobre la causalidad (*causation*) o qué conceptos causales debemos utilizar; y, puesto que estas recomendaciones comportan, a su vez, juicios causales concretos, también [pueden verse] como propuestas normativas sobre los juicios causales que debemos realizar. Por ejemplo, sobre la base de este modo de ver las cosas, las propuestas filosóficas de causalidad según las cuales las denominadas relaciones de doble prevención (*double prevention relations*) [véase la sección 5 de este artículo] no son relaciones causales genuinas —debido a la ausencia de un proceso de conexión que relacione causa y efecto—, pueden pensarse como propuestas normativas acerca de cómo debemos conceptualizar la noción de causalidad y sobre qué juicios causales son correctos o están justificados (*warranted*) cuando están presentes relaciones de doble prevención. En concreto, la perspectiva que alguien puede adoptar, según la cual las relaciones de doble prevención sin procesos de conexión (*connecting processes*) no son causales, se puede pensar como [una perspectiva] vinculada a la propuesta de existir alguna base (*rationale*) o justificación normativa, [que está] conectada con metas (*goals*) relacionadas con el razonamiento casual, [como modo] para distinguir las relaciones de dependencia con procesos de conexión respecto de otras [relaciones de dependencia] que carecen de ese rasgo. Podemos entonces preguntar cuál es esta base (*rationale*) y si esto justifica la distinción en cuestión[2]. En esta manera de ver las cosas está implícita la idea según la cual realizamos *elecciones* (*choices*) sobre qué conceptos causales vamos a adoptar (y sobre qué compromisos deberían conllevar esos conceptos) y que esas elecciones pueden ser evaluadas en términos de si sirven bien a nuestras metas (*goals*) y fines (*purposes*).

---

[1] En mi opinión, muchos filósofos que trabajan sobre causalidad (*causation*) tienden a obviar (*efface*) las dimensiones normativas de lo que están haciendo, por ejemplo, a través de describir sus proyectos como [algo que] proporciona una exposición (*account*) acerca de lo que "es" la causalidad (*causation*), o [algo que] proporciona una "Metafísica" de la causalidad. Esto puede hacer que suene como si su papel fuera meramente informativo (*reportorial*) o descriptivo (de "realidad causal" [*causal reality*] o la Metafísica de la causalidad) y no tuviera ningún contenido normativo (más allá de esa descripción). Pero, por supuesto, hay muchas maneras diferentes de describir el mundo de manera correcta. (No hay nada incorrecto, por ejemplo, en informar solo acerca de correlaciones [*reporting correlations*]). Así, nosotros afrontamos la cuestión acerca de por qué debemos pensar de manera causal (*think causally*) y por qué, al hacerlo, debemos emplear alguna manera concreta de pensar sobre la causalidad (*causation*), en lugar de [emplear] cualquier otra [manera] entre varias alternativas posibles. A mi juicio, las respuestas a estas preguntas tendrán, de manera inevitable, una dimensión normativa y requerirán el hacer referencia a nuestras metas (*goals*) y finalidades (*purposes*).

[2] De nuevo, este modo de ver las cosas contrasta con la práctica más habitual (especialmente entre los metafísicos) de pensar en este asunto, en cuanto que [se trata de] simplemente una cuestión acerca de si las relaciones de doble prevención son "realmente" causales o no. Lo que yo estoy planteando (*urging*) es que debemos preguntar si *debemos*, dadas nuestras metas (*goals*) y fines (*purposes*), ver esas relaciones como causales, destacando así el aspecto normativo de esta cuestión.

Otro ejemplo es la conocida teoría contrafáctica de la causalidad de David Lewis (1973) —y la métrica de la semejanza (*similarity metric*) que caracteriza la cercanía de los mundos posibles, sobre cuya base descansa [la teoría]—; podría considerarse —entre otras cosas— como una propuesta normativa acerca de qué proposiciones causales deben ser evaluadas como verdaderas y las consideraciones que son relevantes para dictaminar (*assessing*) su verdad. A tenor de esta propuesta normativa, las afirmaciones causales deben juzgarse como verdaderas o falsas dependiendo de si están relacionadas con contrafácticos verdaderos, en la manera en que lo prescribe la teoría de Lewis. Más aún, los contrafácticos verdaderos son aquellos que se juzgan como tales a través de criterios particulares para juzgar la semejanza entre los mundos posibles que Lewis propone. Como una propuesta normativa alternativa, se podría imaginar un enfoque que también relacione las afirmaciones causales con contrafácticos, pero que haga uso de una métrica de la semejanza diferente, que de alguna manera se considere superior a la que utiliza Lewis (quizás porque esté de manera más clara o más estrechamente conectada a cualesquiera metas [*goals*] que pensemos que deben guiar la inferencia causal). En tal caso, llega a ser un asunto razonable el porqué debemos utilizar la métrica de Lewis en lugar de esta [métrica] alternativa[3]. Finalmente, la concepción "intervencionista" ("*interventionist*" *account*) de la causalidad, que yo apoyo, es también normativa, tanto en el sentido de hacer recomendaciones sobre qué proposiciones causales se deben evaluar como verdaderas o falsas, como también en otros aspectos —por ejemplo, como he tratado de explicar en el textos como Woodward (2014)—, la concepción intervencionista impone restricciones sobre los tipos de variables que se pueden tener en cuenta en proposiciones causales bien planteadas, incorpora ideas acerca de qué variables es adecuado "supervisar" ["*control for*"] al evaluar propuestas causales "multinivel" ["*multi-level" causal claims*], qué tipo de pruebas [*evidence*] es relevante para evaluar esas proposiciones, etc.). También se busca que se entiendan de manera normativa las distinciones, dentro de las afirmaciones causales —con respecto a rasgos como la estabilidad y la proporcionalidad—, que he investigado en Woodward (2010) y que se tratan con mayor detalle más abajo (afirmo que, dadas nuestras metas [*goals*] epistémicas, son distinciones que es apropiado o racional hacer).

Junto a estas ideas normativas, hay también una cantidad de investigación, muy rica y que crece rápidamente, que es más descriptiva en su aspiración: pretende describir, como un asunto de hechos empíricos, cómo varias poblaciones (de adultos humanos, niños de varias edades, animales no-humanos) aprenden y —en los casos apropiados— razonan y hacen juicios (*judge*) con respecto a relaciones causales. Buena parte de este trabajo lo dirigen psicólogos, pero también

[3] Por supuesto, asumo que decir solamente que la razón para adoptar la métrica de Lewis es que capta "nuestro" concepto de causalidad no es una respuesta adecuada. Esto simplemente plantea la cuestión acerca de por qué debemos utilizar (o por qué utilizamos) "nuestro" concepto, en lugar de alguno otro [concepto] alternativo. Lo que se necesita es una respuesta que muestre que "nuestro concepto" sirve a las metas y fines que tenemos. Cfr. Woodward (2003, 137 y ss).

puede encontrarse una parte importante [de estas investigaciones] en la literatura sobre aprendizaje animal, cognición animal comparativa y, en algunos casos, en disciplinas como la Antropología. Más aún, aunque muchos filósofos probablemente preferirían no pensar de este modo acerca de lo que están haciendo[4], la literatura filosófica sobre la causalidad está llena de lo que parecen proposiciones empíricas descriptivas, al menos en un primer análisis (por ejemplo, proposiciones que apelan a lo que la gente "diría" o a lo que podrían pensar o juzgar acerca de la verdad de varias propuestas causales). Así, uno se encuentra filósofos que afirman que la gente corriente juzga o no las relaciones entre dos eventos que son candidatos a ser causales, al menos en términos estrictos y adecuados, cuando hay una dependencia contra-fáctica, pero no [hay] un "proceso de conexión" (*connecting process*) que vincule los dos eventos (véase Dowe, 2000; Schaffer, 2000). Además, uno también encuentra en la literatura filosófica muchas afirmaciones que se entienden de manera natural (*naturally*) como afirmaciones descriptivas implícitas acerca de los tipos de prueba (*evidence*) y otras consideraciones en las cuales la gente confía para alcanzar conclusiones causales. Por ejemplo, muchos debates acerca de las denominadas teorías probabilísticas de la causalidad (como, p. ej., Suppes, 1970) tendrían poco sentido si no fuera verdad que la gente hace uso sistemáticamente de la información sobre relaciones probabilísticas entre eventos (y, al menos en algunas versiones de teorías probabilísticas, *solo* de esa información) para alcanzar conclusiones causales. De manera similar, el análisis de Lewis acerca de la asimetría causal parece descansar en la idea según la cual los juicios que hace la gente sobre esa asimetría están, de alguna manera manera, guiados por las creencias que mantienen acerca de cuestiones como el número de "milagros" que se necesitan para producir la convergencia y la divergencia

[4] Los filósofos que se inclinan a negar que están haciendo afirmaciones empíricas descriptivas sobre los juicios causales que la gente hace, de hecho, pero que, sin embargo, apelan a "lo que pensamos", describen con frecuencia lo que están haciendo en términos de informar sobre sus "intuiciones" (o las de otros) acerca de la causalidad (o algo similar, aun cuando no se utilice la palabra "intuición"). Una línea habitual (*standard line*) consiste en que las intuiciones se distinguen de las simples propuestas descriptivas sobre lo que piensa quien tiene la intuición (*intuiter*) u otros [agentes] en cuanto que las intuiciones (o, al menos, las intuiciones "correctas" o "verídicas") comportan algún tipo especial de acceso epistémico o percepción (*insight*) de un asunto no-psicológico (*non-psychological subject matter*) (por ejemplo, la "causalidad en sí misma" [*causation itself*]).

Aquellos que, como yo, piensan que esas concepciones sobre la intuición no son creíbles (*incredible*), tampoco se sentirán impresionados por esta supuesta distinción entre intuiciones genuinas y simples afirmaciones psicológicas empíricas, que describen lo que el filósofo u otros piensan. En cualquier caso, incluso los mayores partidarios de las intuiciones sin elaborar reconocen con frecuencia algunas limitaciones (*constraints*) de la Psicología empírica (por ejemplo, se preocupan con frecuencia [como deben hacerlo] por los resultados empíricos que muestran que muchas personas no comparten sus intuiciones). Así, incluso para los irredentos de la intuición, debería ser importante la investigación empírica acerca de cómo la gente, de hecho, aprende y razona de manera causal.

entre distintos mundos posibles (o están estrechamente relacionados con esas creencias).

Una cuestión natural, que se examinará en este artículo, tiene que ver con la relación entre estas dos formas de teorizar acerca de la causalidad y la cognición causal (la descriptiva y la normativa). ¿Debemos pensar en ellas como completamente independientes, quizás sobre la base de que hay una brecha fundamental e insalvable entre "es" [*is*] y "debe" [*ought*] y que cómo debería razonar [*ought to reason*] la gente sobre la causalidad —o sobre otro asunto— no tiene relación alguna con lo que, de hecho, razona, y viceversa? ¿O debemos pensar en esas dos empresas como relacionadas? Y, en ese caso, ¿de qué manera?

En este artículo —y en Woodward (2018)— se argumenta que, aunque hay por supuesto una diferencia entre cómo la gente, de hecho, razona de manera causal (*reason causally*) y cómo deberían razonar; sin embargo, esas dos empresas que se han descrito están relacionadas de muchas maneras y cada una puede estar influida de manera fructífera por la otra. Esta influencia puede adoptar varias formas distintas, que trataré de ilustrar después. En primer lugar, las teorías normativas de la causalidad pueden sugerir hipótesis descriptivas acerca de lo causal y la cognición, posibles experimentos para evaluar esas hipótesis y posibles interpretaciones de los resultados experimentales, de modo que sería improbable que los investigadores pensaran en ausencia de una teorización normativa. Una idea clave aquí consiste en que las teorías normativas pueden jugar su papel, en parte, porque proporcionan puntos clave (*benchmarks*) o ideales con los que se puede comparar, medir y comprender la actuación (*performance*) real. Esto es, al tratar de elaborar una teoría de la cognición causal de un sujeto (*subject's causal cognition*) que sea descriptivamente adecuada, es frecuentemente una buena estrategia el comenzar con una teoría normativa acerca de cómo debería conducirse (*conduct*) esa cognición y, después, preguntar —como un asunto descriptivo— en qué medida la cognición en cuestión cumple los requisitos de la teoría normativa. En otros ámbitos se han empleado de manera muy fructífera estrategias similares que utilizan teorías normativas como puntos clave para guiar la investigación empírica, incluyendo la investigación de la visión (con el uso del análisis del "observador ideal"), la teoría de la decisión y el estudio del aprendizaje no-causal[5]. Sugiero que, de manera semejante, [esas estrategias] son fructíferas cuando se aplican a la cognición causal.

Una idea estrechamente relacionada, que ayuda a proporcionar una motivación para esta estrategia de utilizar como guía una teoría normativa, consiste en [señalar] que, entre los hechos empíricos que queremos que explique una teoría

[5] Un ejemplo bien conocido, que comporta aprendizaje no-causal, lo proporcionan las teorías computacionales del aprendizaje por diferencias temporales (*temporal difference learning*), que se introdujeron en origen como propuestas normativas en Ciencias de la Computación. Se han utilizado con mucho éxito de manera descriptiva para ilustrar el comportamiento de las neuronas dopaminérgicas y su papel en el cálculo de recompensas (*reward computation*).

descriptiva, hay hechos acerca de cuándo las estrategias de cognición causal de varios sujetos (*subjects*) tienen *éxito* (*are successful*) o no en el aprendizaje acerca de la estructura causal del mundo. Para explicar ese éxito o fracaso (*failure*), necesitamos una teoría normativa que caracterice el éxito y el fracaso, y que nos diga qué procedimientos (*procedures*) de aprendizaje, razonamiento y demás llevarán al éxito (o a su carencia). En concreto, en comparación con otros animales, incluidos otros primates, los seres humanos tienen, de manera notoria, éxito en el aprendizaje sobre relaciones causales, y sus capacidades para el aprendizaje causal con éxito se desarrollan y mejoran a través del tiempo. De hecho, incluso los niños muy pequeños tienen mejor aprendizaje causal en diversas cuestiones que otros primates. Uno podría querer una teoría descriptiva adecuada de la cognición causal que, entre otras cosas, explique cómo se alcanza este éxito, del mismo modo que una teoría adecuada del procesamiento visual debería explicar cómo (a tenor de qué procedimientos de inferencia) el sistema visual es capaz de extraer información fiable acerca del entorno visual distante del estímulo visual[6].

Un segundo asunto acerca del papel de la teorización normativa es este: al hacer explícitas las relaciones lógicas entre varios tipos de proposiciones causales y otro tipo de proposiciones (por ejemplo, acerca de pautas de covariación [*patterns of covariation*]), las teorías normativas pueden limitar (*constrain*) de diversas maneras las conclusiones interpretativas que se extraen de los resultados empíricos. Supóngase, por ejemplo, que es correcto —como afirman los intervencionistas (véase la sección 2 de este artículo)— que los juicios causales deberían distinguirse de los juicios sobre pautas de covariación (la causalidad es distinta de la correlación) y que uno de los rasgos distintivos de los primeros, a diferencia de los segundos, es que tienen implicaciones acerca de lo que *sucedería* si *fueran* a realizarse varias intervenciones. Considérense las implicaciones de esta idea para la interpretación de los resultados experimentales que se basan en el diferencial de los tiempos al mirar, que son tenidos en cuenta por muchos psicólogos del desarrollo (*developmental psychologists*) para apoyar conclusiones fuertes acerca del conocimiento causal infantil. Acerca de esta propuesta, la pregunta que formulará alguien que mantenga una versión normativa del intervencionismo es esta: ¿Qué [sucedería] si algo en esos resultados apoyase la conclusión según la cual los niños tienen creencias o conocimientos propiamente (*distinctively*) *causales*,

[6] Por supuesto, nada de esto supone negar que, en ocasiones, la gente comete errores en la inferencia causal, mantiene ideas confusas (*confused*) o inadecuadas en términos normativos sobre la causalidad, etc. Pero el punto de vista general que se adopta en este artículo consiste en que muchos humanos (tanto adultos como niños) son más "racionales" y tienen un comportamiento (*behavior*) que se acerca más a lo que es un comportamiento "bueno" desde un punto de vista normativo en el aprendizaje y el juicio (*judgment*) de lo que, con frecuencia, se supone. De manera relacionada, hay más continuidad entre las estrategias cognitivas de los sujetos ordinarios, incluidos los niños, y las estrategias adoptadas por los investigadores científicos; y, a menudo, podemos utilizar nuestro conocimiento de estas últimas para arrojar luz sobre las primeras. Si se cumplen —y en qué aspectos— estas proposiciones es una compleja cuestión empírica, pero que tiene apoyo empírico en el caso de los niños. Véase, por ejemplo, Gopnik (2012).

como [algo] opuesto a las expectativas sobre la base de pautas de covariación que se han experimentado (*experienced patterns of covariation*)? Según el intervencionismo, en la medida en que los niños tienen lo primero, deben tener creencias o expectativas que tienen que ver con lo que sucedería respecto de acciones o manipulaciones, en el caso de que estas se llevasen a cabo. En algunas oportunidades, que esas pruebas (*evidence*) puedan existir o que se puedan obtener a partir de otras fuentes es algo que no parece que se pueda lograr solo de los estudios temporales, considerados por sí mismos. En la medida en que esas pruebas (*evidence*) no están disponibles, no hay bases para interpretar que los resultados temporales proporcionen el apoyo para creencias distintivamente causales.

Otro ejemplo atañe a los experimentos que se tratan en mayor detalle en Woodward (2018). Mantienen W. Ahn y otros autores (1995) que, como asunto descriptivo, las atribuciones causales de la gente descansan, con mucha frecuencia, en creencias sobre "mecanismos", en lugar de [hacerlo] en información acerca de covariaciones (*covariational information*), de modo que entienden que son alternativas mutuamente excluyentes. Pero, tanto muchas de las propuestas normativas acerca de la información de mecanismos (incluidas las propuestas intervencionistas, como [acontece] en *Making Things Happen* [Woodward 2003]) como los propios ejemplos de Ahn y otros autores (1995) sobre la información mecanicista (*mechanistic information*), dejan claro que esa información *implica* propuestas sobre pautas de covariación. En concreto, no parece que haya razones para dudar que los sujetos en los experimentos de Ahn y otros, que citan información mecanicista (*mechanistic information*), toman esto como si tuviera implicaciones sobre la información acerca de covariaciones [*covariational information*] (y como si se basase en ella). Si fuese así, uno no podría manejar la hipótesis según la cual los sujetos confían en la información mecanicista (*mechanistic information*) y en la información sobre covariación (*covariational information*) como alternativas rivales (*competing*) y que se excluyen mutuamente, y los resultados de Ahn y otros no pueden interpretarse como que muestran que los sujetos confían en la primera *en lugar de* [hacerlo] en la segunda.

De la misma manera que la teorización normativa puede ser de ayuda para la elaboración de una teorización descriptiva, también (voy a sugerir que) los resultados empíricos pueden, con frecuencia, ser indicativos (*suggestive*) para las teorías normativas. Un modo en que esto puede funcionar es este: si vemos que, en cuanto que es un asunto de un hecho empírico, el conocimiento y el aprendizaje causales de la gente muestra ciertos rasgos R; será, con frecuencia, una estrategia fructífera el considerar la posibilidad de que algunos de esos rasgos R contribuyan a ese éxito; y, entonces, [esos rasgos] serán candidatos a ser incorporados en la teoría normativa. Recurriré a esta estrategia, para proporcionar apoyo parcial para algunas de las afirmaciones normativas que se tratarán después. Y a la inversa, si durante la defensa de una propuesta normativa del juicio causal, un filósofo (o quien sea) recurre a afirmaciones empíricas sobre rasgos de los juicios que hace la mayoría de la gente y al modo en que estos contribuyen al éxito, y esas afirma-

ciones empíricas resultan ser falsas, entonces esto sugiere que la teoría normativa o bien está equivocada, o bien necesita ser apoyada de alguna otra manera.

Este cuadro de la interacción entre lo descriptivo y lo normativo encaja (*fits*) de manera natural con otra idea, que está implícita en lo que he expuesto hasta ahora. Esta [idea] tiene que ver con la importancia y fecundidad (*fruitfulness*), cuando se llevan a cabo tanto proyectos descriptivos como proyectos normativos, cuando se intenta comprender el aprendizaje, razonamiento y juicio causales en (lo que llamaré) términos *funcionales*[7]. Esto comporta pensar en la cognición causal en términos de las *metas* (*goals*) u *objetivos* (*aims*) que trata de cumplir (sus funciones) y evaluar las propuestas normativas sobre el aprendizaje causal, el juicio (*judgment*) causal, etc., desde la perspectiva de si alcanzan bien esas metas (*goals*). En otras palabras, al intentar comprender el razonamiento causal, debemos preguntar qué queremos *hacer* con ese razonamiento y qué objetivos (*aims*) y fines (*purposes*) estamos intentando alcanzar. El razonamiento causal debería ser concebido como orientación *hacia* diversas metas (no-triviales)[8] y no simplemente como un fin en sí mismo. Así, llega a ser de una importancia crucial el especificar cuáles son esas metas (o cuáles podrían ser).

Como se explica en la sección 2, los intervencionistas piensan en esas metas como si tuvieran que ver, de manera directa, con la manipulación y el control, pero puede entenderse que un enfoque funcional de la causalidad abarca otras metas posibles (por ejemplo, otra meta posible de la cognición causal podría ser la representación de información correlacional de un modo condensado [*compressed*] y uniforme o, quizás, la representación de varios hechos sobre la dependencia e independencia de la información [*informational dependence and independence*], como [lo hace] Janzig y otros autores [2012]). Sin embargo, en lo que sigue me centraré únicamente en las metas intervencionistas.

El resto de este artículo se organiza de la manera siguiente: En la sección 2 se presenta una visión de conjunto del enfoque intervencionista de la causalidad y se describe su ampliación (*extension*) a las distinciones entre diferentes tipos de juicio causal, las cuales sostengo que tienen sentido a la luz de las metas (*goals*) relacionadas con la intervención. Después, en la sección 3, se examinan algunas posibles implicaciones psicológicas de la idea intervencionista básica, según la cual las afirmaciones causales pueden entenderse como propuestas sobre los resultados de experimentos hipotéticos (*hypothetical experiments*). Más tarde, en las secciones 4 y 5, se analiza el papel normativo de la invarianza en el juicio causal y sus implicaciones empíricas, con especial atención en la sección 5 a los juicios en

[7] Para una defensa más amplia de esta idea, véase Woodward (2015).

[8] Por supuesto, es posible trivializar esta idea sobre la comprensión de la causalidad en términos funcionales (por ejemplo, al insistir en que la meta [*goal*] del pensamiento causal es simplemente enunciar [*stating*] la verdad sobre los hechos causales y nada más). Considero que este es un planteamiento que carece por completo de aportación alguna, puesto que es una meta compartida por prácticamente todas las concepciones de causalidad.

casos de doble prevención (*double prevention*), en los cuales no hay una conexión física entre la causa y el efecto. En Woodward (2018) se desarrolla el examen de la importancia (*significance*) normativa y descriptiva de la invarianza, centrándose en casos en los cuales hay una conexión física entre la causa y el efecto, en ejemplos que ilustran la importancia de la proporcionalidad y en lo que sucede cuando, de manera contraria a la estrategia que yo recomiendo, las investigaciones descriptivas no están guiadas por la teoría normativa.

2. Intervencionismo y distinciones entre relaciones causales

He descrito mi versión de la propuesta intervencionista en detalle en otro lugar (Woodward, 2003). Me limitaré aquí a [hacer] un pequeño resumen. En concreto, mi meta en lo que sigue no es proporcionar una defensa detallada del intervencionismo como una teoría normativa, sino ilustrar cómo los componentes normativos de la teoría podrían estar relacionados con diversos tipos de investigaciones descriptivas. El punto de partida de la propuesta intervencionista es la idea según la cual las relaciones causales son relaciones que pueden utilizarse (*exploitable*), en principio, para la manipulación y el control. Podemos precisar esto algo más mediante el siguiente principio:

> (M) C causa E, si y solo si es posible intervenir para cambiar el valor de C, de tal modo que, si esta intervención ocurriera, el valor de E o la distribución de probabilidad de E cambiaría.

Aquí, una "intervención" es una manipulación idealizada no confundible de C, que cambia E, si es que lo hace, solamente a través de C, y no de alguna otra manera. Desde el punto de visto de la Psicología descriptiva del aprendizaje causal, es un hecho importante el que algunas acciones humanas se califiquen como intervenciones y puedan ser reconocidas como tales; pero la propia noción de intervención puede especificarse sin ninguna referencia a los seres humanos o sus actividades. Sin embargo, la noción de intervención no requiere conceptos causales para su especificación (de manera más obvia, porque una intervención sobre C *causa* un cambio en esta variable[9]). Nótese también que (M) *no* dice que la proposición según la cual C causa E sea verdadera solo cuando C *es* cambiada mediante una intervención, o que uno solo pueda aprender acerca de las proposiciones causales realizando intervenciones. Según (M), lo que importa respecto de si C causa E es si son posibles las intervenciones sobre C y si es verdadera una proposición contrafáctica acerca de qué le *sucedería* a E, si C *fuese* cambiada mediante una intervención, no si C *es* cambiada mediante una intervención. (M) tiene en cuenta la posibilidad de que uno pueda aprender sobre las relaciones causales a partir de muchas fuentes (*sources*) distintas, incluidas las observaciones pasivas que no comportan intervenciones; pero implica que, en todos esos casos,

[9] Para un análisis más detallado, véase Woodward (2003).

(M) capta el contenido de lo que se aprende. En otras palabras, cuando se aprende que C causa E, a partir de una observación pasiva (o a partir de alguna otra fuente que no comporta la realización de una intervención), se debería pensar acerca de uno mismo que ha aprendido que E cambiaría bajo alguna intervención sobre C, pero sin realizar realmente la intervención en cuestión. Si las pruebas que uno tiene no son suficientes para establecer esas proposiciones acerca de lo que pasaría respecto de las intervenciones sobre C, [entonces] las pruebas que uno tiene no son suficientes para establecer que C causa E[10].

(M) fue pensada inicialmente como un criterio normativo para distinguir *entre* relaciones causales y relaciones que son no-causales (porque, por ejemplo, son meramente correlacionales o fallan para satisfacer alguna otra condición necesaria para la causación) y, de manera relacionada, [pretendía ser] un principio que ayudase a aclarar el contenido de las proposiciones causales (véase más adelante en este artículo). La mayoría de las teorías filosóficas de la causalidad (ya sean contra-fácticas, probabilísticas o basadas en la regularidad) tenían como objetivo (*aim*) proporcionar criterios de distinción (*distinguishing criteria*) de este tipo. Hay, sin embargo, otro proyecto normativo con respecto a la causalidad que ha recibido mucha menos atención y que también vale la pena atender. Este segundo proyecto tiene también una amplia motivación intervencionista y es potencialmente fructífero desde el punto de vista de la Psicología empírica. Este proyecto comporta trazar distinciones *entre* relaciones causales que son importantes desde un punto de vista normativo o desde una perspectiva descriptiva. Este proyecto plantea la siguiente cuestión: supóngase que tenemos varias relaciones que son causales, en el sentido de que satisfacen (M). ¿Qué distinciones ulteriores —si es que hay alguna— podría ser útil o importante trazar (para finalidades normativas o descriptivas) *entre* estas relaciones? Aquí están algunas distinciones posibles de este tipo, de las cuales trataré en este artículo y también lo hago en Woodward (2018). En cada caso, las relaciones causales que satisfacen (M) pueden diferir en el grado en que poseen el rasgo en cuestión, y esto puede tener importancia normativa y descriptiva.

> Invarianza: Las relaciones causales que satisfacen (M) pueden [ser] más o menos *invariantes* o estables. Esto es, pueden diferir en el grado en el cual continúan manteniéndose cuando se producen cambios en las circunstancias del entorno (*background circumstances*).
>
> Proporcionalidad: Una causa puede satisfacer (M) y ser más o menos *proporcional* o estar al nivel adecuado para su efecto. Una causa es proporcional a su efecto, en la medida en que todos los cambios posibles en la va-

[10] Se sigue [de eso] que podemos evaluar las inferencias causales propuestas, sobre la base de si proporcionan pruebas (*evidence*) que permitan extraer conclusiones de manera fiable (*reliably*) acerca de los resultados de experimentos hipotéticos adecuados (*appropriate*). Ciertos procedimientos inferenciales, como el uso de variables instrumentales o los diseños de regresión discontinua (*regression discontinuity designs*), cumplen especialmente bien este criterio.

riable de causa (*cause variable*) están asociados a cambios en la variable de efecto (*effect variable*), y todos los cambios posibles en la variable de efecto están asociados a cambios en el valor de la variable de causa.

Especificidad (*specificity*): Las causas que satisfacen (M) pueden ser más o menos *específicas*. Una relación causa-efecto es específica, en la medida en que la causa tiene únicamente un tipo de efecto (entre algún rango de tipos posibles de efectos) y el efecto es el efecto de únicamente un tipo de causa (entre algún rango de tipos posibles de causas).

Como ya se ha sugerido, veo el proyecto de examinar estas distinciones en cuanto que tiene tanto un componente normativo como un ingrediente empírico. En la vertiente normativa, podemos preguntar si, para la gente, tiene sentido normativo hacer estas distinciones entre afirmaciones causales, dadas las metas (*goals*) que tienen o las funciones asignadas al pensamiento causal. En el ámbito empírico/descriptivo, podemos preguntar si la gente, de hecho, hace distinciones entre sus juicios causales en términos de estos rasgos y si estas distinciones sirven a esas metas y cómo [lo hacen]. Mi planteamiento, que defenderé después, consiste en que hay una base normativa (*normative rationale*) para cada una de estas distinciones y que, en cada caso, esta base (*rationale*) está estrechamente ligada a la idea intervencionista, según la cual el razonamiento causal está netamente conectado a nuestro interés (*concern*) por la manipulación y el control. En otras palabras, nuestro interés por la manipulación y el control no solo ayuda a estructurar las distinciones que hacemos entre afirmaciones causales y no-causales (reflejado, como yo propongo, por [M]), sino que también influye en las distinciones mencionadas, que hacemos entre relaciones causales.

Antes de entrar en detalles, hay, sin embargo, otro asunto que quiero poner sobre la mesa. La mayoría de los enfoques de aprendizaje y juicio causales, ya sean normativos o descriptivos, empiezan con una serie de variables (o propiedades o términos descriptivos), que se supone que, de alguna manera, están disponibles con anterioridad para la caracterización de las relaciones causales o de los datos a partir de los cuales se aprenden [esas relaciones]. Sin embargo, por lo general, no ofrecen ninguna propuesta acerca de dónde provienen esas variables. Más aún, es algo común que su elección inicial de variables influya fuertemente los resultados de cualquier análisis causal. Por tomar únicamente una de las posibilidades más simples, dadas dos variables aleatorias (*random*) X e Y, que caracterizan un conjunto de datos, siendo X e Y independientes desde una perspectiva probabilística (*probabilistically independent*), uno siempre puede transformar estas [variables] en variables diferentes (por ejemplo, Z = X+Y, W = X-Y), que son dependientes. Así, las conclusiones causales que se alcanzan mediante cualquier procedimiento inferencial, que infiere relaciones causales a partir de información sobre relaciones de dependencia e independencia (y esto es verdad para prácticamente todos estos procedimientos inferenciales), dependerá de las variables que se utilicen (por ejemplo, la conclusión puede ser que o bien X e Y no están relacionadas

causalmente o bien que Z causa W, dependiendo de qué variables se utilicen). De manera semejante, diferentes elecciones de procedimientos para agregar más micro-variables muy detalladas (*fine-grained micro variables*) en macro-variables más generales (*coarse grained macro variables*) influirán también en las conclusiones acerca de las relaciones entre esas variables macro. Y, básicamente, por la misma razón cualquier propuesta filosófica que trate de comprender la causalidad en términos de regularidades, relaciones estadísticas o contrafácticos también dará lugar a resultados que dependen de las variables utilizadas. Por ejemplo, dada una situación que se describe correctamente como una [situación] en la que Y depende contrafácticamente de X, uno siempre puede re-describir (*re-describe*) la misma situación en términos de las variables W y Z, que se definen a partir de X e Y, donde estas variables son independientes de modo contrafáctico (*counterfactually independent*).

A nivel normativo, esto parece implicar que los enfoques acerca del aprendizaje y el juicio causales son incompletos, a no ser que vayan acompañados por propuestas acerca de la elección de variables; esto es, por propuestas que proporcionen una guía normativa sobre cómo elegir variables de manera apropiada (o sobre cómo "mejorar" la elección de variables que uno hace, dado algún punto inicial de partida, o cómo aprender variables nuevas y mejores para los fines [*purposes*] del conocimiento causal). De manera semejante, en el nivel de la Psicología descriptiva, un proyecto importante, si buscamos comprender por completo el aprendizaje y el razonamiento causal, es comprender de dónde proceden las variables o las descripciones que utilizan los sujetos (por qué los sujetos utilizan un conjunto de variables en lugar de otro, cómo aprenden variables nuevas o más apropiadas, etc.). Parece justo decir que, a pesar de la importancia de estos asuntos, en este momento sabemos muy poco sobre los aspectos normativo y descriptivo de este problema.

Proporcionar una propuesta satisfactoria acerca de elección de variables está más lejos del alcance de este artículo. Solo quiero sugerir que las distinciones entre propuestas causales que examinaré —distinciones con respecto a la invarianza, la proporcionalidad, etc.— tienen que jugar algún papel en la comprensión de la elección de variables (*variable choice*). Algunas elecciones de variables serán de tal manera que podemos expresar relaciones entre esas variables que han de situarse relativamente altas en lo que atañe a las dimensiones de invarianza, proporcionalidad y demás; mientras que otras elecciones de variables llevarán a la expresión de relaciones que se sitúan más bajas en cuanto a esas dimensiones. En la medida en que hay un racional normativo (*normative rationale*) para valorar (*valuing*) rasgos como la invarianza y la proporcionalidad en las relaciones causales, tendrá sentido, de un modo normativo, elegir variables que faciliten estas metas (*goals*). En la medida en que, en cuanto asunto empírico, el razonamiento causal de la gente está guiado por un interés (*concern*) por estas metas, también podemos esperar, en tanto que asunto empírico, que su conocimiento causal refleje estos rasgos. Así, una razón adicional para prestar atención a nociones

como la invarianza y la estabilidad consiste en que esto puede ofrecernos alguna ayuda con los problemas de la elección de variables. La sección 4, en particular, ilustrará esta idea.

3. Interpretaciones intervencionistas de la causación: Consideraciones normativas y descriptivas

En esta sección mi atención está en las propuestas intervencionistas de la causalidad, en cuanto que representadas por el principio (M), y algunas afirmaciones empíricas psicológicas en las que uno puede pensar como "sugeridas" o "motivadas" por este principio. Para evitar malentendidos, permítaseme destacar que es ciertamente posible considerar que (M) es solamente un principio normativo, y rechazar considerar si, como asunto descriptivo, el comportamiento (*behavior*) y cognición de los sujetos se adecua a algo parecido a (M). No hay nada en (M), en cuanto que afirmación normativa, que nos obligue a conectarlo con la Psicología empírica[11]. Por otro lado, por todas las razones que se han sugerido en la sección 1, el análisis de esas conexiones parece un empeño que, potencialmente, vale la pena, y en lo que sigue procederé sobre la base de este supuesto.

Permítasenos considerar, entonces, las posibilidades (*prospects*) de interpretar (M) como parte de una propuesta descriptiva acerca de la cognición causal de humanos adultos y quizá de otros sujetos (esto es, como si caracterizara aspectos de cómo la gente piensa y razona acerca de afirmaciones causales). Comienzo destacando que, aunque (M) —como se ha señalado antes— no propone que uno *solamente* pueda aprender sobre propuestas causales mediante la realización de intervenciones, es natural suponer —si (M) tiene alguna plausibilidad (*plausibility*) como una teoría descriptiva— que la gente aprenderá sobre las relaciones causales de manera relativamente rápida (*readely*) y fiable (*reliably*), si son capaces de llevar a cabo, realmente, intervenciones adecuadas (*appropriate*) y observar los resultados. En otras palabras, si (M) es plausible desde una perspectiva psicológica (*psychologically plausible*), cabe esperar (*expect*) que el realizar intervenciones ha de facilitar el aprendizaje causal del sujeto en diversos contextos. Además, si la teoría normativa asociada con el intervencionismo describe aspectos acerca de cómo los sujetos de hecho razonan, [entonces] se podría esperar que los sujetos serán sensibles (*sensitive*) en su aprendizaje y juicio causales a las diferentes distinciones que son importantes de manera normativa, dentro de un marco intervencionista. Por ejemplo, uno puede preguntar si varios sujetos (humanos adultos, niños de varias edades, animales no-humanos) son sensibles de una manera adecuada a la diferencia normativa entre intervenir (*intervening*) y condi-

[11] Por la misma razón, no hay inconsistencia lógica en mantener que (M) es correcta en cuanto que teoría normativa, pero que, habitualmente, las inferencias de la gente fallan en adecuarse a algo parecido a (M) (por tanto, que la gente es mayoritariamente irracional [*mostly irrational*]). Por supuesto, esta no es la perspectiva que se adopta en este artículo.

cionar (*conditioning*)[12] y si responden de modos normativamente adecuados a la información que sugiere que sus manipulaciones están confundidas (*confounded*) y, por tanto, no son verdaderas intervenciones. La investigación empírica que se ha realizado hasta ahora acerca del aprendizaje causal humano parece sugerir respuestas afirmativas a todas estas cuestiones[13].

Otro asunto empírico que (M) sugiere es este: los adultos humanos rápidamente ponen en común lo que aprenden acerca de las relaciones causales sobre la base de la observación pasiva y lo que aprenden sobre la base de intervenciones (de hecho, el concepto de causalidad de los seres humanos adultos es un concepto según el cual uno piensa que la misma relación causal pude estar presente entre X e Y, con independencia de si X se ha producido a través de una intervención o si se ha observado de manera pasiva). Porque los seres humanos adultos piensan en las relaciones causales de esta manera, pueden utilizar lo que han aprendido acerca de las relaciones causales a partir de la observación pasiva para diseñar intervenciones novedosas (*novel*): podemos establecer que X causa Y sobre la base de una observación pasiva y, después, utilizar esta información para dar lugar a Y mediante la intervención en X. Una cuestión que surge es si hay una etapa en el desarrollo de la cognición humana causal en la que los niños pequeños todavía no son capaces de hacer esto. La respuesta a esta pregunto de nuevo parece ser "sí"; aunque los niños de tres años aprenden rápidamente las relaciones causales a partir de la observación de los resultados de sus propias intervenciones (y las de otros) cuando están ausentes diversas condiciones que lo facilitan, son incapaces de utilizar información correlacional a partir de la observación pasiva para diseñar sus propias intervenciones. En cambio, los niños de cinco años a los que se presentan los mismos estímulos experimentales (*experimetnal stimuli*) sí son capaces de hacer esto (cfr. Bonawitz, et al). Esto representa uno de los muchos aspectos en los cuales los conceptos humanos causales cambian a través del tiempo en el curso del desarrollo y el aprendizaje.

Hay también una serie de críticas filosóficas que se han dirigido a (M), que plantean tanto problemas empíricos interesantes como normativos, que merecen un mayor análisis. Una de esas críticas filosóficas consiste en que (M) es viciosamente "circular", porque [M] sostiene que elucida (*elucidate*) la noción de X que causa Y mediante la apelación a una noción (la de intervención) que es, obviamente, en sí misma de índole causal. Un enfoque adecuado de la causalidad, dicen los críticos, debe ser no-circular; debe ser "reductivo" (*reductive*), en el sen-

[12] Para un debate acerca de la diferencia entre condicionar e intervenir, véase Woodward (2007). Dadas las variables aleatorias X, Y y Z, uno puede preguntar, por ejemplo: (i) si X e Y son independientes de modo condicional respecto de Z. Esta es una cuestión respecto de condicionalización. Uno también puede preguntar (ii) si, por ejemplo, X e Y serían independientes bajo una intervención en X. Esta es una cuestión acerca de intervenir. (i) y (ii) son no-equivalentes y esto es lo que sucede en las intervenciones, que es diagnóstico de la causalidad.

[13] Para más detalles, véase Woodward (2007).

tido de que explica (*explicates*) qué es para X causar Y en términos de conceptos (como "regularidad", "correlación", etc.), que son en sí mismos completamente no-causales. Según los críticos, (M) intenta, en cambio, de hecho "explicar la causalidad en términos de causalidad" y se preguntan cómo esto puede ser esclarecedor (*illuminating*). Una segunda preocupación, señalada por los críticos filosóficos, se puede expresar de esta manera: cuando se puede realizar, en efecto, una manipulación experimental de X, es quizá plausible (dice el crítico) que (M) proporcione un criterio para determinar si X causa Y. Sin embargo, incluso en este caso —afirma el crítico— este criterio es (en el mejor de los casos) de importancia (*significance*) meramente epistemológica. No nos dice nada acerca de la Semántica o la Metafísica de la causalidad (qué significan las afirmaciones causales, qué es la causalidad u otras cosas parecidas). Más aún, cuando una manipulación experimental adecuada de X no se va a realizar o no puede realizarse, todavía está menos claro cómo (M) podría ser esclarecedor (*illuminating*): (M) conecta "X causa Y" con un contrafáctico sobre qué sucedería si se produjese una intervención en X, pero ¿cómo puede ese contrafáctico ser de alguna utilidad, si no podemos llevar a cabo la intervención en cuestión?

Se trata de cuestiones complejas. No intentaré [proporcionar] nada parecido a una respuesta completa aquí[14]; pero, en vez de eso, me centraré en una faceta o aspecto de ellas, que ilustra mi tema (*theme*) de la interacción entre lo descriptivo y lo normativo. Comienzo con una observación empírica, aunque es una [observación] que expresa una afirmación normativa que se hace con frecuencia. Muchos investigadores de una serie de disciplinas diferentes dicen que es útil o esclarecedor (*illuminating*) conectar las proposiciones causales y los resultados de los experimentos hipotéticos (*hypothetical experiments*) de un modo algo cercano a lo descrito por (M). Por ejemplo, el marco de la respuesta potencial (*the potential response framework*), que desarrollaron Rubin, Holland y otros (véase, por ejemplo, Rubin, 1974) y que ahora se utiliza mucho en Estadística, Econometría y en otros ámbitos en Ciencias Sociales, se organiza en torno a la elaboración de proposiciones causales precisamente de esta manera. Otro ejemplo es que muchos historiadores (al menos oficialmente) son cautelosos respecto de hacer proposiciones causales de cualquier tipo; pero aquellos historiadores que hacen esas propuestas insisten de nuevo con frecuencia en conectar esas proposiciones a contrafácticos, donde estos tienen una interpretación ampliamente intervencionista. Los investigadores que adoptan este enfoque, tanto en Historia como en Ciencias Sociales, proponen habitualmente que eso ayuda a esclarecer el contenido de las proposiciones causales y a comprender qué tipo de pruebas (*evidence*) son relevantes para su evaluación (*assessment*), si se las relaciona (*associate*) con experimentos hipotéticos adecuados; y esto es así, tanto cuando uno puede, en efecto, realizar el experimento en cuestión como cuando —incluso de

[14] Para una exposición más amplia de cómo una propuesta no-reductiva de la causalidad puede, sin embargo, ser explicativa, véase Woodward (2003).

manera más interesante—cuando no es posible realizarlo (Davis, 1968; Angrist y Pischke, 2009). Por supuesto, es posible para los críticos responder que estos investigadores están, simplemente, confundidos —piensan que esta conexión con experimentos meramente hipotéticos es útil y esclarecedora, cuando no lo es—, pero yo defendería un enfoque más comprensivo (*charitable*), en el cual tratamos de comprender, tanto a nivel psicológico como a nivel metodológico, cómo es posible que sea informativa esta conexión.

El primer asunto a señalar en esta conexión es que las propuestas causales, con frecuencia, se exponen de maneras que no son claras (*unclear*) o están indeterminadas (*indeterminate*). Una de las cosas que se pueden lograr al conectar las proposiciones causales con experimentos hipotéticos —en la manera descrita en (M)— es esclarecer (*clarify*) qué comportan (*mean*) esas afirmaciones ("comportar" [*mean*] en el sentido de aquello a lo que nos comprometen [*commit*][15]) y hacerlas más determinadas y precisas; lo hacemos al hacer explícito que han de ser entendidas en términos de algún experimento hipotético concreto —que especificamos—, en vez de algún otro experimento [posible]. Hacer esto requiere, entre otras cosas, que las variables, que entendemos que están relacionadas de manera causal, deben de ser el tipo de factores que, en principio, pueden ser manipulados o cambiados mediante intervenciones: deben ser factores tales que "tenga sentido" pensar en términos de manipularlos. También requiere que hagamos explícito cuáles son los valores posibles de esas variables, y asimismo ha de hacerse explícito cómo cambiar la variable causa (*cause-variable*) de uno de sus valores a otros [valores] conlleva cambios en el valor de la variable efecto (*effect-variable*). En el caso de muchas proposiciones causales de la forma "X causa Y", habrá un número de maneras posibles no-equivalentes de hacer esto (esto es, distintas proposiciones posibles sobre los resultados de experimentos hipotéticos que podrían asociarse con la proposición causal original). Así, indicar qué experimento posible se ha planeado pude aclarar (*clarify*) o desambiguar la proposición causal original.

Al modo de una sencilla ilustración, considérese la proposición según la cual (Glymour, 1986),

> (3.1) Fumar cinco paquetes de cigarrillos al día causa un aumento sustancial en la probabilidad de [contraer] cáncer de pulmón.

Una manera (muy poco comprensiva [*uncharitable*]) de asociar esto con un experimento hipotético es interpretar que (3.1) propone

---

[15] O quizá sería mejor decir que (M) nos dice lo que se *debería* querer decir (*mean*) mediante la proposición causal o cómo deberían interpretarse esas proposiciones, si van a hacerse de manera clara y precisa. Nótese que, cuando se entiende de esta manera, (M) cuenta como una propuesta *semántica*, incluso si no proporciona una reducción de las proposiciones causales a proposiciones no-causales. A este respecto, parece desacertada (*misguided*) la crítica según la cual (M) no tiene implicaciones para la Semántica o el contenido de las proposiciones causales.

> (3.1*) Cualquier intervención que cambie el que alguien fume cinco cajetillas a algún número inferior de cajetillas (por ejemplo, 4.9 cajetillas) cambiará sustancialmente la probabilidad de que la persona desarrolle cáncer de pulmón.

Según esta interpretación, (3.1) es probablemente falsa. Otra interpretación más comprensiva (*charitable*) de (3.1) —probablemente, más próxima a la intención del hablante— es interpretar (3.1) como si propusiese que

> (3.1**) Una intervención que cambie el que alguien fume cinco cajetillas [al día] a que esa persona no fume nada cambiará, de manera sustancial, la probabilidad de que esa persona desarrolle cáncer de pulmón.

Bien podría ser (3.1**) verdadera y es, en cualquier caso, obviamente una proposición distinta a (3.1*). Alguien que afirme (3.1) puede aclarar qué es lo que quiere decir, indicando si su interpretación es (3.1*) o (3.1**).

Otro ejemplo, considerablemente más controvertido, considera la proposición [siguiente:]

> (3.2) Ser mujer (*being a woman*) es causa de discriminación en las contrataciones (*hiring*).

De nuevo, los intervencionistas se inclinan a considerar que esta proposición es poco clara (*unclear*) y a pensar que puede hacerse más clara o menos ambigua haciendo explícito qué se pretende proponer respecto del resultado de un experimento hipotético. Desde una perspectiva intervencionista, el problema básico con (3.2), según se plantea, consiste en que la noción de una manipulación o una intervención sobre "género" (*gender*) o "ser mujer" (*being a woman*) no es clara (*unclear*). Una manera posible de manipular el género es cambiar la estructura de los cromosomas del sexo de un individuo inmediatamente después de la concepción (substituyendo un cromosoma X por un cromosoma Y o viceversa). Interpretada con esta intervención en concreto en mente, se podría entender (3.2) como una propuesta según la cual (3.2*), como intervención, cambiaría la probabilidad que alguien tiene de ser contratado para determinados trabajos. Mientras que esta proposición (asumo) es verdadera, una construcción alternativa (*alternative construal*) de (3.2), que yo supongo que está más próxima a captar lo que pretende decir la mayoría de los que afirman (3.2), es esta[16]:

---

[16] Para ver la diferencia entre (3.2*) y (3.2**), nótese que (3.2*) sería verdadera dentro de un régimen donde la contratación se basase por completo en los méritos y las cualificaciones de los aspirantes, en tanto que [las personas de] diferentes géneros desarrollan intereses y capacidades distintas, lo que causa que estén cualificados de distinta manera (*differentially*) para diversos trabajos. Supongo que lo que se pretende a través de (3.2) es [hacer] alguna proposición en el sentido según el cual las decisiones respecto de los contratos que afectan a las mujeres no se hacen sobre la base del mérito y las cualificaciones de los aspirantes, y que hay mujeres cualificadas que no son contratadas debido a su género. (3.2**) está más próxima a captar esto.

(3.2**) Intervenir, para cambiar las creencias sobre el género de un aspirante a un puesto de trabajo que tiene quien contrata (*employer*), cambiará la probabilidad de esa persona para ser contratada.

Aquí —a diferencia de en [lo expuesto] 3.2*— lo que se implica (*implication*) es que, dados dos aspirantes idénticos en todo (por ejemplo, idénticos en sus calificaciones), pero distintos en el género y en rasgos como las características sexuales externas, es más probable que sean contratados los aspirantes varones. Nótese que, en este caso, la variable que se contempla como el objeto de la intervención (y la causa) es "las creencias acerca del género que tiene quien emplea", en lugar de serlo el género mismo. (Esto ejemplifica cómo el marco intervencionista fuerza a uno a ser más preciso sobre qué variables se entiende que están relacionadas causalmente). Es (3.2**) una proposición que podría ser contrastada [*tested*] (y, de hecho, lo ha sido) mediante, por ejemplo, presentar currículos, que sean en lo demás idénticos, donde solo se ha alterado el género de los aspirantes al puesto de trabajo. En cualquier caso, la cuestión importante para nuestros fines (*purposes*) es que (3.2) y (3.2**) son propuestas no-equivalentes, que bien podrían tener valores de verdad diferentes. No parece ser controvertido que merece la pena para alguien que afirme (3.2) el pensar cuál de estas posibilidades es la que tiene en mente.

Estas consideraciones sugieren que al menos parte de (o un componente de) una narrativa (*story*) sobre cómo es posible que una afirmación como (M) sea útil o esclarecedora para dilucidar el contenido de las proposiciones causales, a pesar de su aparente "circularidad": Pensar en términos de (M) —y, de manera más general, interpretar las propuestas causales como proposiciones sobre los resultados de experimentos hipotéticos— fuerza a uno a ser más preciso y explícito sobre con qué proposiciones causales nos comprometemos (*commit*) y cómo podrían contrastarse (*tested*). Caso de ser correcto, esto explicaría la observación hecha antes (que los investigadores en muchas disciplinas distintas encuentran que es útil [*helpful*] asociar proposiciones causales con experimentos hipotéticos). Nótese también que nada en esta estrategia requiere que se sea capaz de llevar a cabo una reducción de las proposiciones causales a proposiciones no-causales.

Si algo como esta sugerencia/especulación (*suggestion/spectulation*) fuese correcto, [entonces] se atisban cuestiones empíricas adicionales. Por ejemplo, valdría la pena investigar en qué medida es verdadero (como sugiere el relato que se ha hecho antes) que diversos sujetos (tanto expertos como de los otros) desempeñan mejor el aprendizaje causal y las tareas de razonamiento (*reasoning tasks*), si se les incita a asociar las proposiciones causales con experimentos hipotéticos de la manera descrita[17].

[17] Hay algunas pruebas (*evidence*) en cuanto a los expertos en relaciones internacionales cuando hacen juicios causales: quienes contemplan de manera sistemática contrafácticos son más fiables (*reliable*) que los expertos que no lo hacen. Cfr. Tetlock (2005).

También sería interesante aprender más acerca de los mecanismos psicológicos que subyacen a cualesquiera capacidades que están utilizando cuando asociamos proposiciones causales con experimentos hipotéticos. En una interesante serie de artículos (véase, por ejemplo, Gendler, 2007), Gendler recurre a la teoría de sistemas duales, para explicar el posible hecho de ser capaces de aprender a partir de experimentos mentales (*thought experiments*). A grandes rasgos, la teoría de los sistemas duales afirma que el procesamiento psicológico humano se organiza en dos sistemas. El sistema 1 es rápido, automático, "intuitivo", y con frecuencia opera de manera inconsciente; mientras que el sistema 2 es más lento, más deliberativo y dependiente de un procesamiento explícito, consciente. La idea de Gendler consiste en que, a menudo, cuando tomamos parte en un experimento de pensamiento (*thought experiment*) nosotros "corremos" o bien hacemos uso de la información que proporciona el sistema 1 y la conectamos o la hacemos disponible para su procesamiento de una manera más explícita a través del sistema 2. Uno puede pensar, como yo hago, que la dicotomía sistema 1 frente a sistema 2 está enormemente simplificada y, aún así, cabe pensar que hay algo correcto en la idea básica de Gendler y que también se puede utilizar alguna versión de ella para ayudar a dilucidar cómo el asociar proposiciones causales con experimentos causales puede ser esclarecedor. La idea sería que, cuando se considera una proposición causal como "X causa Y", no todo lo que es relevante para razonar con la propuesta o contrastarla (*testing*) está (al menos, en un comienzo) explícito y disponible para la valoración crítica. Se puede contemplar la proposición "X causa Y" sin pensar, al menos de manera muy clara y explícita, qué comportaría cambiar o manipular X o cómo uno espera que Y cambie bajo varias manipulaciones posibles de X. Asociar "X causa Y" con un experimento hipotético fuerza a uno a ser explícito sobre estos asuntos y, con frecuencia, esto se hace al hacer inferencias a partir de la información que, de alguna manera, se "tiene" (está presente en nuestro sistema 1), pero que uno no ha integrado previamente de manera explícita en su juicio causal (*causal judgment*). Expresado de esta manera, la idea que subyace a (M) no es tanto que siempre que se contempla una proposición causal se está, necesariamente, pensando en ella o representándola como una proposición sobre dos estados posibles de la causa, el resultado de un experimento hipotético, etc., sino más bien que se puede aclarar (*clarify*) o hacer preciso lo que está pensando en un principio mediante la expansión de la propuesta causal en la línea de lo indicado por (M). De nuevo, esta es una sugerencia que, en principio, podría contrastarse (*tested*) empíricamente.

4. Invarianza/estabilidad: Consideraciones normativas y descriptivas

Pasaré ahora a un examen más detallado de las distinciones entre juicios causales descritas en la sección 2, centrándose esta sección y la sección 5 en la noción de invarianza. Supóngase que dos variables están relacionadas en la manera descrita por (M): E es dependiente de modo contrafáctico de C en algún conjunto

concreto de circunstancias de entorno (*background circumstances*) Bi, donde la dependencia en cuestión tiene una interpretación intervencionista (esto es, E cambiaría en el marco de alguna intervención que podría llevarse a cabo sobre C en Bi). Lo que llamaré la *invarianza* (estabilidad, robustez [*robustneess*], ausencia de sensibilidad [*insensitivity*]) de esta relación tiene que ver con el grado en que se seguiría satisfaciendo (M) en otras circunstancias de entorno (*background circumstances*) distintas de Bi. En otras palabras, la invarianza de la relación C→E tiene que ver con el grado en que se sigue manteniendo la dependencia de E respecto de C, cuando las condiciones de entorno cambian respecto de aquellas [que hay] en la situación presente[18].

Primero, ilustraré la idea básica mediante un par de ejemplos de David Lewis (1986) y, después, intentaré precisar más [esa idea]. Supóngase, en primer lugar, que Lewis escribe una carta de recomendación para un aspirante N a un puesto de trabajo, lo que lleva a N a obtener un trabajo que, de otro modo, no hubiera conseguido; lo cual, a su vez, lleva a N a conocer y a casarse con alguien con quien, de otra manera, no se hubiera casado; lo que lleva después a la existencia de unos niños (y, finalmente, a su muerte) que, de otro modo, no hubieran existido. El que esos niños existan y mueran o no (D) es contrafácticamente dependiente de si (W) Lewis escribe la carta de recomendación, donde la dependencia en cuestión es de tipo no-retroactivo (*non-backtracking*), la cual —según Lewis— es suficiente para la causalidad. Más aún, la relación W→D satisface la condición (M) y, así, cuenta como causal dentro del marco intervencionista. No obstante, una investigación no formal sugiere que, para mucha gente, la afirmación según la cual W *causa* D es, de alguna manera, extraña, engañosa o deficiente[19]. Un diagnóstico (*diagnosis*) plausible para esta reacción es que, aunque la relación W→D satisface (M), es relativamente no-invariante [*non-invariant*] (relativamente inestable [*unstable*] o relativamente sensible [*sensitive*]). Es relativamente no-invariante en el sentido de que, si una de las muchas condiciones de entorno realmente predominantes (*actually obtaining background conditions*) hubiera sido

[18] Para un estudio adicional, véase Woodward (2006). Lo que llamo aquí invarianza se denomina sensibilidad (*sensitivity*) en ese artículo; hay algunas diferencias de detalle en comparación con el presente texto.

[19] Asumir que la proposición según la cual W causa D es defectuosa de algún modo, aunque sea pequeño, llevará a considerar, en lo que sigue, en qué pensamos exactamente que consiste ese defecto. Si (M) es correcto, la propuesta según la cual W causa D es literalmente verdadera y, por tanto, su deficiencia (*defectiveness*) debe comportar algo más, aparte de su falsedad literal [*literal falsity*] (podría ser o bien engañosa [*misleading*] —porque está muy lejos de ser causal de una manera paradigmática— o bien no-informativa [*uninformative*] o no-explicativa [*unexplanatory*] o inadecuada desde un punto de vista pragmático sobre la base de su no-invarianza). Otro posible planteamiento consiste en que la afirmación es literalmente falsa, lo cual, por supuesto, requeriría la revisión de (M), de modo que algún requisito de invarianza se incorpore a su cláusula de suficiencia. Yo considero que el primer planteamiento es más simple, pero creo que no necesito insistir en que el segundo planteamiento está equivocado en la medida en que la deficiencia (*defectiveness*) está relacionada con una relativa no-invarianza.

diferente, incluso en formas relativamente pequeñas (si, digamos, N hubiera recibido una oferta todavía mejor de otra escuela o si se hubiera quedado un poco menos en el bar donde conoció a su futuro cónyuge) la dependencia contrafáctica de D en W no se hubiera obtenido por más tiempo.

Compárese esto con un segundo ejemplo: A dispara a B en el corazón con un arma de gran calibre a corta distancia y B muere. No solo la muerte de B depende de modo contrafáctico de que A le dispare en el corazón —su relación satisface (M)—, sino que esta relación de dependencia es relativamente invariante en el sentido según el cual —dentro de muchas variaciones no de ficción científica (*non-science fictionistish*) en cuanto a las condiciones de entorno (consistente con el disparo de A)— se seguiría manteniendo esta dependencia. Puesto que alguien ha sido disparado al corazón, no hay mucho que nosotros o la Naturaleza podamos hacer, en circunstancias ordinarias, para evitar su muerte.

Como ilustran estos ejemplos, mencionar relaciones que satisfacen (M) que sean relativamente no-invariantes o inestables tiende a parecernos —al menos de manera habitual— de alguna manera menos satisfactorio que mencionar relaciones más invariantes que también satisfacen (M). Las primeras se ven como ejemplos menos "buenos" (*less "good"*), menos paradigmáticos o menos satisfactorios de relaciones causales. Por razones que he relegado a una nota al pie (la número 17), no dedicaré tiempo a tratar de decir de manera más exacta en qué consiste la índole insatisfactoria (*unsatisfactory*) de las primeras (quizá se trate de un asunto acerca de que las relaciones relativamente no-invariantes [*non-invariant*] son defectuosas [*defective*], desde el punto de vista de la explicación causal, o son engañosas o no-informativas (*uninformative*), de alguna otra manera). Desde mi punto de vista, el hecho importante es que nuestros juicios causales parecen (al menos en cuanto a lo que atañe a las pruebas incidentales citadas antes) sensibles a la diferencia entre las relaciones relativamente invariantes y relativamente no-invariantes que satisfacen (M). Más aún, esto no es solo un hecho del pensamiento de la gente corriente acerca de la causalidad, sino que parece que impregna una gran cantidad de pensamiento causal, también en contextos científicos: los científicos también parecen valorar relativamente más las relaciones invariantes, considerándolas como más explicativas o bien más satisfactorias.

Para ejemplificar esta última cuestión, cabe considerar la preferencia habitual [que hay] en las Ciencias Biológicas por las explicaciones "mecanicistas" (*mechanistic*). Qué hace que una explicación sea "mecanicista" es una cuestión interesante[20], pero un rasgo que comparten muchas explicaciones mecanicistas es este: Se comienza con alguna relación general entre una entrada (*input*) y una salida [*output*] (una relación I→O), que satisface (M) (por ejemplo, un estímulo y una respuesta, una presión en el acelerador y la aceleración de un coche, la presencia

[20] Para un examen más detallado de lo que hace que una explicación sea mecanicista, véase Woodward (2013).

de lactosa en el entorno de la bacteria *E. coli* y su síntesis de una enzima que digiere la lactosa). Entonces, se explica por qué esta relación I→O se mantiene al descomponerla en sus conexiones (*links*) intermedias o componentes. En el caso más simple, esto podría comportar una cadena de conexiones individuales [*single links*] (I→C1→Cn→O); aunque, en los casos más complejos, las causas intermedias podrían comportar ramificaciones y convergencias. Si se ejecuta de manera adecuada, este proceso proporciona el sentimiento de iluminación o de una experiencia de "¡ajá!", un sentimiento de que la "caja negra" asociada con la relación I→O original, que parecía bastante arbitraria y misteriosa, se ha abierto (*open up*) y se ha hecho más inteligible.

Con frecuencia, los enfoques filosóficos acerca de la explicación mecanicista paran en este punto; pretenden decirnos qué es un mecanismo o una descripción mecanicista, pero no por qué es explicativa o por qué es algo bueno el tener una explicación de ese tipo. Me parece que esto no es satisfactorio. Lo que hace falta es un tratamiento [de esta cuestión] que sitúe el enfoque propuesto acerca de la descripción mecanicista en el contexto de un enfoque más general de la causalidad y la explicación, y que nos ayude a comprender por qué el completar (*filling in*) las conexiones intermedias proporcionan una comprensión causal más profunda. Me parece que la observación que hice antes acerca de la invarianza contribuye a proporcionar esto, al menos de alguna manera. Con mucha frecuencia (al menos en contextos en los cuales la explicación mecanicista parece apropiada), las conexiones intermedias descubiertas en la descripción mecanicista son más invariantes que la relación original O→I. Este rasgo contribuye —creo— a nuestro sentido de que proporciona una explicación más profunda la información mecanicista sobre las conexiones intermedias; esta información nos llama la atención como más profunda o más satisfactoria, en parte, *porque* comporta relaciones causales que son más invariantes.

Aunque creo que ejemplos como los que se han analizado proporcionan alguna motivación intuitiva para la idea según la cual valoramos la invarianza en las relaciones causales, tanto la dimensión normativa como la vertiente descriptiva de la idea merecen mucho más análisis. A nivel normativo, he hablado de relaciones que son o no relativamente invariantes, o que son más o menos invariantes, pero he dicho muy poco en esta presentación que haga esto preciso o para explicar por qué, desde una perspectiva normativa, la gente debería preocuparse (*care*) por la invarianza. A nivel empírico, está la siguiente cuestión: suponiendo que la gente, de hecho, se preocupe de alguna manera por sobre el grado en que las relaciones causales son invariantes, ¿qué tipo de invarianza les preocupa y por qué?

Permítaseme comenzar con la cuestión normativa. En primer lugar, como debería estar claro por mis comentarios anteriores, creo que es mejor pensar en la invarianza como en una cuestión de grado y como [algo] relativo a alguna clase particular de cambios (en las condiciones de entorno [*background conditions*]), en lugar de simplemente pensar en términos de relaciones que son (de manera

absoluta) invariantes o no. En concreto, quiero plasmar la idea según la cual una relación R podría ser invariante con respecto a un conjunto de cambios en las condiciones de entorno, pero no con respecto a algún otro conjunto de cambios. Así, *no* propongo algún criterio único para clasificar (*sorting*) las relaciones en términos de una simple dicotomía (por ejemplo, relaciones altamente invariantes frente a relaciones no-invariantes). En segundo término, hay un tipo especial de situación donde las comparaciones de grado acerca de la invarianza son por completo no-problemáticas: cuando un conjunto de cambios respecto de los cuales la generalización R es invariante es un subconjunto propio del conjunto de cambios respecto de los cuales una segunda generalización R' es invariante, entonces R' es, por supuesto, más invariante que R. De manera mucho más habitual, sin embargo, este tipo de base para [establecer] un orden de grados de invarianza no estará disponible. En esas situaciones, sugiero otras dos consideraciones que parecen, de hecho, afectar a los juicios sobre la invarianza (y se podría decir que, como un asunto normativo, deben [afectar a esos juicios][21]). (Dejo abierta la posibilidad de que otras consideraciones también puedan ser relevantes).

La primera consideración tiene que ver con la tipicidad (*typicality*): ¿Con qué frecuencia se producen (o se piensa que se producen) diversos tipos de cambios en las condiciones de entorno o, de manera más general, en qué medida esos cambios parecen habituales (*typical*) frente a poco probables (*unlikely*) o [parecen] inverosímiles o de ficción científica (*science-fictionish*)[22]? Si, digamos, una relación de interés R es invariante respecto de cambios en las condiciones de entorno, [que son cambios] de un tipo que ocurre con frecuencia[23], pero R es no-invariante bajo otro tipo de cambios que son más raros o más inverosímiles (*far-fectched*), consideraremos R como relativamente invariante; el veredicto opuesto se obtendrá si R es no-invariante bajo cambios comunes o habituales, pero [es] estable bajo cambios que ocurren de manera poco frecuente. Como ejemplo, considérese que la dependencia contrafáctica que hay entre ingerir una cantidad sustancial de cianuro y la muerte fuese alterada si el individuo en cuestión recibiera de manera inmediata un antídoto y el tratamiento médico propio del momento, pero que el recibir este tratamiento fuese, en la actualidad, un suceso poco común; la mayoría de la gente que ingiriese cianuro, dentro de un ámbito de condiciones de entorno que se dieran de manera habitual, moriría.

---

[21] Véase más adelante [en este artículo].

[22] Se puede especular que, en la medida en que estamos interesados en contextos científicos, la tipicidad (*typicality*) puede importar más en algunas Ciencias que en otras (por ejemplo, puede importar más en Biología y en las Ciencias Sociales y del Comportamiento [*behavioral*] que en Física).

[23] Por supuesto, si las condiciones ocurren de manera frecuente o no dependen de la clase de referencia o del rango de ambientes que se consideren. Lo que sucede de manera frecuente ahora mismo en la superficie terrestre podría suceder únicamente de manera muy poco frecuente en el Universo tomado en su conjunto. Parece que los biólogos están más interesados e influidos por lo primero; [mientras que] los físicos [lo están] por lo segundo.

Estas consideraciones nos llevan a pensar la relación de la ingesta de cianuro → muerte como relativamente invariante/estable. En cambio, es probable que se juzgue como relativamente no-invariante la relación entre que escriba Lewis una carta de recomendación y la existencia y la muerte de los niños, no solo porque hay muchos cambios posibles a partir del estado actual de las condiciones de entorno, que podrían alterar la relación, sino porque esos cambios son de clases extremadamente comunes o típicas.

Un segundo conjunto de consideraciones (quizá no muy distinto del primero) que son relevantes para las valoraciones (*assessment*) de invarianza son consideraciones específicas del objeto de estudio; consideraciones que tienen que ver con el tema [*subject-matter*] o contenido de la generalización de interés. El papel de estas consideraciones es especialmente obvio en los contextos científicos, aunque mi conjetura es que algo análogo funciona (*is operative*) también en otros contextos, más ordinarios[24]. Para ilustrar lo que tengo en mente, considérese el papel de los juicios de invarianza en Economía. Economistas serios y con principios se preocupan mucho acerca del grado en el cual las generalizaciones en las que se basan sus modelos son relativamente invariantes [*relatively invariant*] (este es el tema central de la así llamada crítica de [Robert] Lucas respecto de muchas generalizaciones macro-económicas), pero lo que más les preocupa es la invarianza bajo tipos particulares de cambios que se consideran especialmente importantes o relevantes para la Economía (cambios que atañen a ciertos tipos de variables "de Economía"). Así, por ejemplo, los economistas se preocuparán especialmente acerca de si las generalizaciones en las que confían son invariantes respecto de cambios en los incentivos, los precios relativos o la información que está disponible para los agentes económicos. Las generalizaciones que no son invariantes respecto de ese tipo de cambios serán consideradas como defectuosas (*defective*) para finalidades explicativas o de los modelos. Por otro lado, la mayoría de los economistas no se perturbarían al saber que las generalizaciones que utilizan no son invariantes respecto, digamos, de la ingesta de drogas que alteran la mente de la mayoría de los miembros de la población sobre la que versa el modelo (*population modeled*); estos cambios no se consideran relevantes económicamente, en parte quizá porque se consideran poco probables, pero también porque no se considera que la preocupación por ellos sea parte del objeto de estudio (*subjet-matter*) de la Economía.

Como segundo ejemplo, que se ilustra en mayor detalle después, en muchos contextos biológicos el tipo de invarianza que se considera como más importante es la invarianza respecto de cambios en condiciones que son "biológicamente normales" (*biologically normal*); "normal" en el sentido de condiciones internas

[24] Así, por ejemplo, se puede conjeturar que las condiciones de fondo, [aquellas] que más importan para las evaluaciones de invarianza en contextos que incluyen juicios ordinarios de causalidad psicológica, difieren de las condiciones de fondo que se consideran como más importantes en los juicios ordinarios acerca de la causalidad física.

al organismo o en cuanto al ambiente externo. Así, dentro de una célula u organismo, las relaciones que caracterizan la regulación genética y la síntesis de proteínas, o las que caracterizan el funcionamiento del sistema inmunológico, se juzgarán como relativamente invariantes en la medida en que sean estables respecto de un rango de condiciones fisiológicas que son normales para el organismo; que estas se rompan bajo condiciones fisiológicas extremas (al aumentar la temperatura del organismo hasta 100º) se considera como menos significativo desde el punto de vista de la evaluación de la invarianza. Por supuesto, los juicios acerca de la normalidad (*normality*) están influidos por los que sucede habitualmente, pero también están influidos por ideas acerca de la función biológica, si una condición es o no patológica, etc.

Finalmente, debemos señalar que el grado en que una relación es invariante depende de cuáles sean los factores que se incluyen o se excluyen de esa relación. En concreto, una relación entre factores que se caracteriza con un nivel de descripción muy detallado y que llega al detalle más fino, pero que es incompleta en el sentido de que omite factores relevantes, podría ser relativamente no-invariante en comparación con una relación que se expresa atendiendo a factores más genéricos (*coarse-grained*), pero que no omite factores relevantes a ese nivel genérico (*coarse-grained*). Si, digamos, se le da a alguien (C1) el momento y la posición exactos del 90 por ciento de las moléculas en un mol de gas en un tiempo 1 y se le pide que formule una relación entre C1 y C2 (la posición y el momento de esas mismas moléculas un minuto más tarde), no habrá ninguna relación de dependencia estable interesante entre C1 y C2. (Dificultades de cálculo aparte, C2 dependerá de los estados exactos de *todas* las moléculas de gas en el tiempo 1; la omisión de información sobre el restante 10 por ciento de C1 hará a C2, de hecho, independiente de C1). Por otro lado, si a uno se le dan los valores de únicamente unas pocas variables macro (*macro variables*) —por ejemplo, los valores para variables termodinámicas tales como la presión, temperatura y volumen—, uno puede expresar una generalización relativamente invariante en términos de estas variables. Como ilustra este ejemplo, las relaciones de nivel superior, relaciones macro o genéricas (*coarse-grained*) pueden ser más invariantes que las relaciones expresadas de acuerdo a variables de nivel más micro, si (como habitualmente —quizá casi siempre— es el caso) las últimas son incompletas en el sentido de que omiten algunos factores relevantes. Esta es solamente una de las muchas maneras en las cuales una preocupación por encontrar relaciones relativamente invariantes puede, a veces (pero de ninguna manera siempre) llevar a una preferencia por relaciones formuladas a un nivel más macro, en cuanto que opuesto a un nivel más micro o con mayor detalle. De manera más general, parece suceder que es posible formular relaciones relativamente invariantes entre un conjunto de variables, pero no entre algún otro conjunto, aun cuando ambos conjuntos pueden utilizarse para describir algún sistema de interés. En este sentido, una preocupación (*concern*) por la invarianza puede ayudar a guiar la elección de variables. De un modo similar, supóngase, para sopesar el argumento, que si

uno tuviese caracterizaciones suficientemente exactas de un conjunto completo de variables neuronales que describen cerebros individuales y [tuviese] capacidades de cálculo ilimitadas (*unlimited*), uno podría formular relaciones altamente invariantes en términos de ellas [estas variables]. No sería menos verdadero que, si uno no conociera todas esas micro-variables relevantes y/o fuera incapaz de medirlas con precisión arbitraria (*with arbitrary precision*), cualquier relación que uno pudiera formular entre ellas sería muy no-invariante (*very non-invariant*). Como en el caso del gas, se estaría en mejores circunstancias, desde el punto de vista de encontrar relaciones invariantes, si se utilizase un conjunto de variables mucho más general (*coarse-grained*) y una representación con niveles de libertad muy reducidos.

Pero ¿por qué la invarianza relativa en las proposiciones causales es, metodológicamente hablando, una virtud? Entre otras consideraciones, la identificación de una relación causal relativamente invariante (en comparación con una menos invariante) proporciona mejores oportunidades para la manipulación y la predicción. Una relación relativamente invariante es más generalizable o exportable (*exportable*) a situaciones nuevas o diferentes y podemos tener más confianza en que no se vendrá abajo cuando intentemos utilizarla[25]. Consideraciones similares se aplican a los factores que se han identificado antes como influyentes para las valoraciones (*assessments*) de la invarianza. Si, por ejemplo, quiero usar alguna relación supuesta para controlar y predecir, entonces es obvio por qué me debe importar en qué medida es probable que suceda que los cambios en las circunstancias de entorno alteren la relación (en este entorno, en las presentes circunstancias). Esto explica, al menos en parte, por qué las consideraciones respecto de la tipicidad (*typicality considerations*) deben importar en la evaluación de la invarianza.

Hasta aquí en cuanto a la idea general de invarianza y su base (*rationale*) metodológica. Dada esta idea, surgen una serie de cuestiones empíricas [que son] de un tipo que puede ser de interés para los psicólogos. En primer lugar —y de manera más obvia—, aunque los ejemplos de Lewis y otros [ejemplos] descritos más arriba son sugerentes, son por completo anecdóticos. Valdría la pena investigar de un modo más sistemático si los sujetos de varias poblaciones distinguen entre varios tipos de propuestas causales con respecto a su invarianza de un algún modo parecido a la manera que se ha sugerido antes. Reconociendo que no soy un psicólogo y no tengo formación (*training*) en el diseño experimental, hay aquí algunas sugerencias sobre posibles maneras en las que esto podría hacerse. Primero, dada alguna escala para valorar (*rating*) el grado en que una relación candidata es un ejemplo bueno o paradigmático de una relación causal, o en qué medida es "adecuado" (*appropriate*) describir la relación como causal (la escala de evaluación [*rating*] utilizada por Lombrozo para experimentos, descrita más adelante), parece razonable esperar (si lo que he señalado anteriormente está en el camino

[25] Para un análisis adicional, véase Woodward (2010) y Lombrozo (2010).

correcto) que, permaneciendo las demás cosas iguales, los sujetos valorarán (*rate*) las proposiciones relativamente invariantes como más causales de modo paradigmático[26]. Una tarea algo más ambiciosa sería una investigación acerca de si, cuando se les ofrecen pruebas (*evidence*) a los sujetos de relaciones de covariación o de dependencia a diferentes "niveles" o descritas en términos de distinta variables, prefieren el nivel donde hay relaciones [que son] más invariantes, en lugar de menos invariantes; "prefieren" en el sentido de que aprenden más relaciones invariantes de manera más rápida (*readily*), las utilizan preferentemente en explicaciones y para la predicción, etc. (Los resultados de este tipo serían análogos a los resultados de los experimentos respecto de la proporcionalidad dirigidos por Lien y Cheng (2000) y analizados en Woodward (2018), pero centrándose, en cambio, en el modo en que las consideraciones basadas en la invarianza [*invariance-based considerations*] influyen en la elección de variables [*variable choice*] o la elección de nivel).

Además de la cuestión general acerca de si los sujetos, de hecho, distinguen entre proposiciones causales más o menos invariantes en la manera en que se ha sugerido, está la cuestión empírica ulterior (suponiendo que no hagan, en absoluto, esas distinciones sobre la base de la invarianza) de qué factores influyen en esos juicios. Antes he descrito dos candidatos interrelacionados para esos factores: uno tiene que ver con la frecuencia o tipicidad (*frequency or typicality*) con la que se producen cambios en las condiciones de entorno (*background conditions*); y el otro [tiene que ver] con limitaciones más específicas del objeto de estudio (*subject-matter specific constraints*). Pero estas sugerencias se derivan únicamente de mi empirismo casual (*casual empiricism*) de sillón. De nuevo, valdría la pena —en mi opinión— investigar todo esto de una manera más cuidadosa y sistemática. ¿Es verdad, por ejemplo, como implica mi sugerencia, que dada una relación causal R que se mantiene en condiciones Bi, pero que se alteraría a través de un cambio en las condiciones de entorno Bk, que el manipular la información sobre la frecuencia en la que sucede Bk afecta los juicios de los sujetos sobre la invarianza de R, evaluada (*assessed*) de la manera en que se ha descrito antes?[27]

---

[26] Esto, de hecho, es lo que se halla en la clase concreta de casos investigados por Lombrozo, que se tratan después.

[27] Nótese que la frecuencia de la existencia de alguna condición de entorno posible (diferente de las condiciones de entorno existentes), respecto la cual una relación causal C podría venirse abajo es un factor que parece natural considerar como altamente "extrínseco" a la propia C. Según varias propuestas filosóficas acerca de la causalidad, la relación causal entre dos eventos debe depender solamente de factores que son "intrínsecos" a esa relación. Mi suposición es que, como una cuestión de hecho empírico, los juicios causales de la gente fallan de diversas maneras en cuanto a respetar este requisito de lo intrínseco (*intrinsicness*); pero esto, de nuevo, es un asunto que merece investigación experimental. En la medida en que es adecuado, desde un punto de vista normativo, que consideraciones acerca de la invarianza influyan en el juicio causal, también habrá razones normativas por las cuales el juicio causal no debe estar basado por completo en consideraciones "intrínsecas".

5. La doble prevención (*double prevention*) y el papel de la invarianza

Un último conjunto de problemas empíricos que suscita el papel de la invarianza en el juicio causal es este: ¿Podemos utilizar esta noción para explicar pautas específicas en los juicios causales que hace la gente? Aquí ha habido algún trabajo experimental muy interesante y sugerente sobre la importancia (*significance*) de la invarianza para la causalidad a través de la doble prevención (un trabajo del que trataré ahora[28]). En la doble prevención, si algún evento *d* fuera a ocurrir, evitaría (*prevent*) el acaecimiento de un segundo evento *e* (que, de otra manera, hubiera ocurrido en ausencia de *d*) y, además, el acaecimiento de un tercer evento *c* evitaría el acaecimiento de *d*, con el resultado de que *e* tendría lugar. En el conocido ejemplo de Ned Hall (2004), el avión de Suzy bombardería un objetivo (*e*), si no se evita que lo haga. Un piloto enemigo *p* derribaría el avión de Suzy (*d*), a no ser que se evite que lo haga. Billy, que pilota otro avión, derriba *p* (*c*) y Suzy bombardea el objetivo.

En casos como este hay una dependencia contrafáctica de *e* en *c* (donde la dependencia en cuestión satisface el requerimiento (M) y es también el tipo de dependencia no-retroactiva (*non-backtracking*), que se asocia con la relación causal (*causal relatedeness*) en las teorías contrafácticas de la causalidad). Sin embargo, una reacción habitual de muchos filósofos consiste en que los casos en los cuales *e* está relacionada con *c* a través de la doble prevención, o bien no son casos en los cuales *c* causa *e*, o bien, cuando menos, son casos que carecen de algún rasgo que es central para muchos otros casos de causación. Esto se refleja en que muchos se resisten a decir que [el hecho de que] Billy derribe el avión enemigo *causa* el bombardeo de Suzy. En algún momento, el propio Hall (2004) utiliza este y otros ejemplos, para motivar la afirmación según la cual trabajamos con (al menos) dos conceptos distintos de causalidad: uno que descansa en la dependencia contrafáctica no retroactiva [*non-backtracking counterfactual dependence*] ("dependencia") y otro que comporta lo que él llama "producción". No intentaré reproducir en detalle la caracterización de Hall de esta segunda noción, pero la idea básica es que, a diferencia de la dependencia, la producción satisfará algún tipo de restricción local y, al menos en muchos casos paradigmáticos, estará presente un proceso de conexión [*a connecting process*] (por ejemplo, uno que comporte una transferencia energía/momento) que conecte causa y efecto. Así, la producción está presente cuando, por ejemplo, una piedra que se arroja golpea una botella y causa que se haga añicos. Hall sugiere que las relaciones de dependencia que comportan producción tienden a parecernos como causales de manera paradigmática y que es la ausencia de los rasgos asociados a la producción en la relación entre la acción de Billy y el bombardeo de Suzy aquello que explica

[28] Como se ha señalado antes, estos experimentos parecen proporcionar algún apoyo para la propuesta según la cual las consideraciones basadas en la invarianza influyen en el juicio causal, pero solo en un contexto muy específico.

por qué pensamos en la relación de doble prevención como no-causal o, al menos, no completa o paradigmáticamente causal.

Si pensamos en estas ideas sobre la doble prevención desde la perspectiva "funcional", que se recomienda en este artículo, surgen de manera natural varias cuestiones. Primero, como un asunto de Psicología descriptiva, ¿en qué medida se comparten ampliamente estos juicios intuitivos que presentan Hall y otros sobre el estatus causal de la doble prevención?

Segundo, en la medida en que la gente distingue entre relaciones de doble prevención y relaciones de dependencia que comportan producción, ¿podemos decir algo sobre *por qué* la gente hace esta distinción (esto es, qué base normativa, si es que hay alguna, habría para esta distinción)? ¿Qué tiene de especial la producción, de manera que la gente la distingue de otros tipos de relaciones de dependencia? Puesto que he analizado en mayor detalle en otro lugar la investigación empírica que se ha realizado sobre estas cuestiones (Woodward, 2012), me limitaré aquí a hacer una breve síntesis (*summary*). Primero, hay de hecho pruebas (*evidence*) que provienen de los experimentos llevados a cabo por Walsh y Sloman (2011) y Lombrozo (2010) acerca de que los sujetos adultos distinguen, al menos, entre algunos casos de relaciones causales que comportan producción y algunos casos que comportan doble prevención; y están más dispuestos (*ready*) a juzgar los primeros como causales o más paradigmáticamente causales. Segundo, y de manera más interesante, las pruebas de Lombrozo muestran que los sujetos distinguen *entre* relaciones de doble prevención, considerando algunas (en particular, aquellas que comportan acciones intencionales, artefactos [*artifacts*] con funciones diseñadas y funciones biológicas; que tienen, por tanto, "estatus fucional") como más paradigmáticamente causales que otras relaciones de doble prevención, por lo demás semejantes, que no comportan estos rasgos.

Siguiendo una reflexión de Woodward (2006), Lombrozo sugiere una explicación para esta última pauta (*pattern*): las relaciones de doble prevención que comportan un estatus funcional son típicamente (*typically*) más invariantes que otras relaciones de doble prevención, por lo demás similares, que carecen de estos rasgos; y esta diferencia explica por qué las primeras se juzgan como más paradigmáticamente causales. Por ejemplo, cuando una relación de doble prevención comporta una adaptación biológica (como en el caso de ejemplos de regulación genética que comportan doble prevención, que son bastante habituales), entonces esta relación de dependencia se seguirá manteniendo ante estos cambios. En cambio, en el supuesto acerca de Billy y Suzy que se ha descrito antes, la dependencia del bombardeo de Suzy respecto de que Billy derribe al enemigo es muy inestable/no-invariante (*unstable/non-invariant*) respecto de muchos cambios muy frecuentes en el entorno. Por ejemplo, esta relación de dependencia no se mantendría si el piloto enemigo recibiera un mensaje para volver a la base antes de tener la oportunidad de atacar a Suzy, si un segundo combatiente enemigo hubiera podido atacar a Suzy, etc. Desde esta perspectiva, la dependencia que está

presente en el supuesto acerca de Billy y Suzy es relativamente similar a la dependencia presente en el ejemplo de la carta de recomendación de Lewis. Ambas son relativamente no-invariantes y las dos son juzgadas como menos paradigmáticamente causales —al menos, en parte— por esta razón.

Sin embargo, aun cuando fuera correcto este análisis, todavía deja abierta la cuestión de por qué los sujetos tienden a juzgar que las relaciones de dependencia que satisfacen (M), en las cuales están presentes conexiones físicas, son más paradigmáticamente causales que las relaciones de dependencia en las cuales no están presentes conexiones físicas. ¿Podríamos utilizar también la noción de invarianza para explicar esta pauta (*pattern*) en el juicio? En Woodward (2018) se examina esta cuestión y otras más.

Referencias bibliográficas

Ahn, Woo-kyoung, Kalish, Charles W., Medin, Douglas M. y Gelman, Susan A. (1995). The Role of Covariation versus Mechanism Information in Causal Attribution. *Cognition*, *54*, 299-352.

Angrist, Joshua, and Pischke, Jörn-Steffen (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Baillargeon, Renée (2002). The Acquisition of Physical Knowledge in Iknfancy. En Usha Goshwami (ed.), *Blackwell Handbook of Childhood Cognitive Development* (pp. 47-83). Oxford: Blackwell Publishing.

Bonawitz, Elizabeth B., Ferranti, Darlene, Saxe, Rebecca, Gopnik, Alison, Meltzoff, Andrew N., Woodward, James y Schulz, Laura (2010). Just Do It? Investigating the Gap between Prediction and Action in Toddlers' Causal Inferences. *Cognition*, *115*, 104-117.

Cheng, Patricia W. (1997). From Covariation to Causation: A Causal Power Theory. *Psychological Review*, *104*(2), 367-405.

Davis, Lance E. (1968). And It Will Never Be Literature— The New Economic History: A Critique. *Explorations in Enterpreneurial History*, *6*(1), 75-92.

Dowe, Phil (2000). *Physical Causation*. N. York: Cambridge University Press.

Gendler, T. (2007). Philosophical Thought Experiments, Intuitions and Cognitive Equilibrium. *Midwest Studies in Philosophy*, *23*, 68-89.

Glymour, Clark (1986). Comment. *Journal of the American Statistical Association*, 81, 964-966.

Gopnik, Alison y Schulz, Laura (2007). *Causal Learning: Psychology, Philosophy and Computation*. N. York: Oxford University Press.

Gopnik, Alison (2012). Scientific Thinking in Young Children: Theoretical Advances, Empirical Research, and Policy Implications. *Science*, *337*, 1623-1627.

Griffiths, Thomas L. y Tenenbaum, Joshua Brett (2009). Theory Based Causal Induction. *Psychological Review*, *116*, 661–716.

Hall, Ned (2004). Two Concepts of Causation. En John Collins, Ned Hall y L. A. Paul (eds.), *Causation and Counterfactuals* (pp. 225-276). Cambridge, MA: The MIT Press.

Janzing, Dominik, Mooij, Joris, Zhang, Kun, Lemeire, Jan, Zscheischler, Jakob, Daniusis, Povilas, Steudel, Bastian y Scholkopf, Bernhard (2012). Information-geometric Approach to Inferring Causal Directions. *Artificial Intelligence*, *182-183*, 1-3.

Lewis, David (1973). Causation. *Journal of Philosophy*, *70*, 556-567.

Lewis, David (1986). Postscripts to 'Causation'. En David Lewis, *Philosophical Papers, vol. 2* (pp. 172-213). Oxford: Oxford University Press.

Lien, Yunnwen y Cheng, Patricia W. (2000). Distinguishing Genuine from Spurious Causes: A Coherence Hypothesis. *Cognitive Psychology*, *40*(2), 87-137.

Lombrozo, Tania (2010). Causal-Explanatory Pluralism: How Intentions, Functions, and Mechanisms Influence Causal Ascriptions. *Cognitive Psychology*, *61*, 303-332.

Pearl, Judea (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Rubin, Don B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, *66*(5), 688-701.

Schaffer, Jonathan (2000). Causation by Disconnection. *Philosophy of Science*, *67*(2), 285-300.

Spirtes, Peter, Glymour, Clark y Scheines, Richard (2000). *Causation, Prediction and Search*. Cambridge, MA: The MIT Press.

Suppes, Patrick (1970). *A Probabilistic Theory of Causality*. Ámsterdam: North-Holland.

Tetlock, Philip E. (2005). *Expert Political Judgment*. Princeton University Press: Princeton, NJ.

Walsh, Clare R., and Sloman, Steven A. (2011). The Meaning of Cause and Prevent: The Role of Causal Mechanism. *Mind and Language*, *26*(1), 21-52.

Woodward, James (2003). *Making Things Happen: A Theory of Causal Explanation*. N. York: Oxford University Press.

Woodward, James (2006). Sensitive and Insensitive Causation. *Philosophical Review*, *115*, 1-50.

Woodward, James (2007). Interventionist Theories of Causation in Psychological Perspective. En Alison Gopnik y Laura Schulz (eds.), *Causal Learning: Psychology, Philosophy and Computation* (pp. 19-36). N. York: Oxford University Press.

Woodward, James (2010). Causation in Biology: Stability, Specificity, and the Choice of the Levels of Explanation. *Biology and Philosophy*, *25*, 287-318.

Woodward, James (2012). Causation: Interactions between Philosophical Theories and Psychological Research. *Philosophy of Science*, *79*(5), 961-972.

Woodward, James (2013). Mechanistic Explanation: Its Scope and Limits. *Aristotelian Society Supplementary Volume*, *87*(1), 39-65.

Woodward, James (2014). A functional account of causation; or a defense of the legitimacy of causal thinking by reference to the only standard that matters — usefulness (as opposed to metaphysics and agreement with intuitive judgment). *Philosophy of Science*, *81*(5), 691-713.

Woodward, James (2015). Methodology, Ontology, and Interventionism. *Synthese*, *192*(11), 3577-3599.

Woodward, James (2018). Causal Cognition: Physical Connections, Proportionality, and the Role of Normative Theory. En Wenceslao J. Gonzalez (ed.) (2018), *Philosophy of Psychology: Causality and Psychological Subject. New Reflections on James Woodward's Contribution* (pp. 105-137). Boston/Berlín: Walter de Gruyter.