CDD: 121

EL *EFECTO KNOBE*: ASIMETRÍAS EN LA ATRIBUICIÓN DE INTENCIONALIDAD Y SUS CAUSAS [*]

ALEJANDRO ROSAS

MARÍA ALEJANDRA ARCINIEGAS

Departamento de Filosofía
Universidad Nacional de Colombia
Ciudad Universitária, Cra. 30 #45-03, Edificio 239
BOGOTÁ
COLOMBIA

arosasl@unal.edu.co
maarciniegasg@unal.edu.co

Received: 17.02.2013; Revised: 31/07/2013; Accepted: 08.11.2013

Resumen: En este artículo proponemos una explicación novedosa del *efecto Knobe*. El efecto Knobe es una asimetría peculiar en la atribución de intencionalidad a un agente con relación a los efectos colaterales previstos de su acción, dependiendo sólo de la valoración moral del efecto y sin que nada más cambie en la situación juzgada: los efectos colaterales malos, pero no los buenos, se consideran intencionalmente producidos. Nos enfocamos aquí en la pregunta por la explicación de esa peculiar asimetría: ¿basta la valencia moral del efecto colateral para explicarla? Hacemos un análisis sistemático de una gran variedad de viñetas presentes en los estudios experimentales y de sus resultados. Intentamos así aislar los factores explicativos. Proponemos que la asimetría se explica por concordancia o discordancia entre la valencia moral del efecto colateral y la actitud moral del agente, juzgada por los espectadores.

Palabras clave: Intencionalidad. Efecto Knobe. *Free-rider*. Juicio moral. Moralidad del daño. Bien público.

Abstract: In this article we discuss factors presumably responsible for the *Knobe effect* and offer a novel explanation. The *Knobe effect* refers to a peculiar asymmetry in attributions of intentionality to

[*] M.A.A agradece el apoyo de la Facultad de Ciencias Humanas de la Universidad Nacional a través de la Convocatoria Orlando Fals Borda; A.R. agradece el apoyo de la Vicerrectoría de Investigación, código Hermes 16012 y de la John Simon Guggenheim Memorial Foundation.

the foreseen side-effects of an action, depending only on their moral assessment and with no other changes in the circumstances: the bad effects, but not the good ones, are considered intentionally produced. We focus on the possible explanation: does the moral value of the effect explain the asymmetry? We analyze a variety of vignettes introduced in experimental studies and their results, trying to isolate the explanatory factors. We propose that a concordance or discordance between the moral valence of the side effect and the moral attitude of the agent, as judged by spectators, explains the asymmetry.

Keywords: Intentionality. Knobe-effect. Free-rider. Moral judgment. Harm-morality. Public goods.

1. EL *EFECTO KNOBE*

La discusión en torno al *efecto Knobe* comenzó cuando el filósofo estadounidense Joshua Knobe describió un extraño fenómeno: la gente atribuye intencionalidad a un agente en relación a un efecto colateral de una misma acción (el juicio se evoca mediante historias o viñetas en un experimento) dependiendo de la valencia moral binaria del efecto (Knobe 2003). Si el resultado colateral es negativo ("malo") tienden a calificarlo como intencional, pero no lo juzgan intencional si es positivo ("bueno"). Lo extraño del efecto es que todos los factores tradicionalmente considerados relevantes permanecen iguales en las historias. El efecto colateral es previsto y se distingue del propósito principal de la acción. Lo único que varía es su valencia moral. Aparentemente, este influjo de la valencia moral fue una novedad inesperada: no había sido contemplada en los análisis tradicionales del concepto de intencionalidad.

1.1 VIÑETAS ORIGINALES DEL *EFECTO KNOBE*

El *efecto Knobe* se comprobó originalmente en las dos historias siguientes:

> El vicepresidente de una compañía va a donde el gerente y le dice "Estamos pensando en comenzar un nuevo programa. Nos ayudará a generar más ganancias, pero también perjudicará al medio ambiente". El gerente responde: "No me importa en lo absoluto dañar al medio ambiente, sólo quiero producir la mayor cantidad de ganancias posibles. Implementemos el nuevo programa". Implementan el

> programa y efectivamente el medio ambiente se vio perjudicado. (Knobe, 2003, p. 191).[1]

Luego se pregunta a los encuestados: ¿Dañó el gerente intencionalmente al medio ambiente? La mayoría responde que sí. Para obtener más información, se les pregunta por qué. Mencionan el estado psicológico o mental del gerente, esto es, que éste decidió implementar el programa aunque sabía que iba a dañar el medio ambiente. Sin embargo, Knobe mostró que la previsión no es suficiente para determinar la intencionalidad aplicando una segunda historia. Ésta mantiene todos los elementos de la primera, con la única diferencia de que en este caso el efecto colateral previsto es ayudar al medio ambiente:

> El vicepresidente de una compañía va a donde el gerente y le dice "Estamos pensando en comenzar un nuevo programa. Nos ayudará a generar más ganancias, y también ayudará al medio ambiente". El gerente responde: "No me importa en lo absoluto ayudar al medio ambiente, sólo quiero producir la mayor cantidad de ganancias posibles. Implementemos el nuevo programa". Implementan el programa y efectivamente el medio ambiente se vio ayudado. (Knobe, 2003, p. 191)

La mayoría de los encuestados responde ahora que el empresario *no* ayudó intencionalmente al medio ambiente. Esta asimetría en la atribución de intencionalidad es conocida como el *efecto Knobe*. Sugiere que el concepto de intencionalidad no es una mera herramienta con usos predictivos, sino que es también sensible a consideraciones morales. El efecto se presenta con personas tomadas de grupos distintos: niños de 4 años (Leslie et al. 2006), personas con lesiones en la corteza prefrontral ventromedial (Young et al. 2006) y hablantes nativos de la lengua hindi (Knobe and Burra 2006).

[1] Todas las traducciones son de los autores.

1.2 DISCUSIÓN EN TORNO AL *EFECTO KNOBE*: DOS PROBLEMAS

Los estudios de filosofía experimental con viñetas – historias cortas que se presentan a los sujetos experimentales, seguidas de preguntas sobre la aplicabilidad de un concepto – tienen generalmente por objetivo sondear el uso común de conceptos de interés filosófico. En el caso del efecto Knobe, el concepto investigado es el de intencionalidad. Cuando Knobe descubrió inesperadamente que la valencia moral del efecto colateral previsto decide si se atribuye o no intencionalidad, algunos autores señalaron que la influencia del factor moral constituía una distorsión de los criterios o normas que rigen el uso del concepto de intencionalidad. La distorsión es provocada por la atribución de responsabilidad: los sujetos experimentales juzgan al gerente como responsable del daño al medio ambiente y lo expresan atribuyéndole la intención de dañarlo (Mele 2003; Nadelhofer 2004). Al parecer, desconocen que hay casos de responsabilidad sobre efectos no intencionales. Por ejemplo, quien conduce ebrio es responsable de los daños que cause, aunque no haya sido su intención producirlos. Knobe mostró experimentalmente que sujetos que juzgan la acción en estado de ebriedad como acción responsable pero no intencional, seguidamente continúan atribuyendo intencionalidad al gerente que daña al medio ambiente como efecto colateral. Knobe (2006) niega que el uso del concepto de intencionalidad en sus viñetas se pueda asimilar al caso del ebrio y defiende que la asimetría en cuestión refleja un uso correcto.

No queremos entrar en el debate sobre si el *efecto Knobe* se produce por un error de desempeño (*performance error*), es decir, por una violación de las normas que rigen el uso del concepto de intencionalidad debido a la influencia impropia de consideraciones morales. Más bien, queremos distinguir dos problemas a discutir en relación con la asimetría. Por un lado, está la pregunta por el factor que la causa. ¿Es la valencia moral del efecto colateral la causante de la asimetría? ¿Es algún otro factor? ¿Es una combinación de factores? El

segundo problema consiste en decidir si esos factores que de hecho influyen, merecen formar parte de los criterios o normas que rigen el juicio de intencionalidad.

Obviamente, abordar la segunda pregunta requiere establecer previamente qué factor o factores producen el 'efecto Knobe'. Sobre este punto no hay aún consenso, y la literatura es extensa y crece. Nuestro propósito en este escrito es ofrecer una explicación novedosa, que se distingue, aunque sutilmente, de otras explicaciones en oferta. Nuestro método será analizar las variaciones sobre las viñetas originales que la literatura experimental ha introducido en la discusión. Ellas ofrecen la oportunidad de extraer, teniendo en cuenta los datos sobre la atribución, algunas pistas sobre los tipos de historia que producen el efecto y los que no lo producen. Eventualmente aislaremos los factores que explican la misteriosa asimetría en la atribución de intencionalidad a efectos colaterales previstos.

2. ASIMETRÍA CON VALENCIA MORAL

2.1 VALENCIA MORAL DE EFECTOS COLATERALES Y DE MEDIOS.

Según como Knobe mismo describió la asimetría (Knobe 2006), parece esencial que ella se presente en relación a un efecto colateral previsto que, en tanto colateral, no cae aparentemente bajo la intención del agente. Muchas viñetas ideadas por otros autores siguen esta misma estructura, tanto así que al 'efecto Knobe' también se le conoce como el 'Efecto del efecto colateral' (*The side-effect effect*) (Cova y Naar 2012a, Sripada 2012).

Sin embargo, algunos estudios experimentales sobre el efecto utilizan medios en sus viñetas, en lugar de efectos colaterales. Eso puede afectar su validez. Provisionalmente, se prodría decir que un efecto es colateral cuando no hubo intento de producirlo. Simplemente se produce en el curso de enfocarse en un objetivo diferente. Sin embargo, es posible que una acción tenga efectos previsibles que, si bien no se buscan directamente, son instrumentales y necesarios para el

objetivo y podrían confundirse con un medio y ser, entonces, intencionales. Así, lo primero que se necesita es una definición clara sobre qué es un efecto colateral, y cómo se distingue de un medio. Una buena definición es presentada por Cova y Naar (2012a p. 5):

> Entonces, ¿cómo debemos entender la diferencia entre medios y efectos colaterales? Una manera de trazarla es como sigue: un evento E es un medio para mi fin G sólo si E es un componente necesario de la explicación causal de cómo logré producir G. Si no es un componente de la explicación causal, pero es un efecto de lograr G, entonces es un efecto colateral.[2]

Cova y Naar aplican esta definición a los dos dilemas morales clásicos *Trolley* y *Footbridge*. En *Trolley*, un tren fuera de control se dirige a matar a cinco obreros desprevenidos en la vía. Cerca de la vía se encuentra un transeúnte junto a una palanca que cambia la dirección del tren hacia la vía secundaria, donde matará a una sola persona. En *Footbridge*, el tren amenaza con matar a las cinco personas y el transeúnte se encuentra en un puente sobre las vías. A su lado se encuentra una persona voluminosa y la única manera de salvar a las personas es arrojarlo sobre la vía del tren, antes de que alcance éste a los obreros. Sin embargo, esta persona voluminosa morirá.

En ambos casos el objetivo principal es salvar a las cinco personas. Sin embargo, es fácil ver que en *Trolley* matar a la persona es un efecto colateral, pues sería posible explicar la cadena causal de cómo se salvan a las personas tan sólo mencionando el cambio de carriles. Por el contrario, en *Footbridge* se sacrifica la vida de una persona como medio para salvar a otras. No es posible explicar causalmente la salvación de los obreros sin el sacrificio de la persona voluminosa.

---

[2] "So, how should we understand the difference between means and side effects? A way to draw the distinction is the following: an event E is a means to my goal G only if E is a necessary component of the causal explanation of my bringing about G. If it isn't a component of the causal explanation, but is still an effect of my reaching G, then it is a side effect."

Según esta definición, varios estudios sobre la asimetría sustituyen los efectos colaterales por medios. También hay casos ambiguos. El primer caso de ambigüedad es nada menos que la viñeta original del 'Gerente y el medio ambiente' (*Chairman*). Hay buenas razones para afirmar que el daño al medio ambiente no es un mero efecto colateral, en particular porque es instrumental para obtener las ganancias. Si la nueva política que genera ganancias implica producir desechos tóxicos, botarlos al medio ambiente es un recurso que ahorra costos a la empresa a expensas del bien público. Si la empresa tuviese que procesar los desechos, se generarían costos que podrían anular las ganancias. Siendo así, arrojar desechos sin procesar es un medio para generar ganancias. En el caso contrario de que se produjese un desecho que ayuda al medio ambiente, por ejemplo, eliminar CO_2 del aire, ese efecto no generaría las ganancias; se darían igual si no hubiera desecho alguno. Esto puede ser parte de las intuiciones que tienen los sujetos al responder.

Esta reflexión sugiere una explicación de la asimetría en el estudio original: el daño al medio ambiente es un medio intencionalmente escogido, mientras que la ayuda es un genuino efecto colateral que no cae bajo la intención del agente. La asimetría en la atribución de intencionalidad no tendría nada de particular, o al menos no sería un efecto sorprendente en busca de una explicación inusual.

Otro ejemplo aún más patente del uso de medios en una viñeta y de un efecto colateral en la otra, es el estudio de Machery (2008). Según Machery, la asimetría en la atribución de intencionalidad se produce sólo en los casos que constituyen un '*trade-off*', y pueden ser casos sin importe moral: el agente incurre en un costo para obtener lo que considera un beneficio. Para mostrar esto, Machery presenta dos viñetas, una de las cuales es un *trade-off* – *Extra dollar* – mientras que la otra, similar en todo lo demás, no lo es – *Free cup*. En ambas viñetas el agente quiere comprar la bebida más grande que ofrece el local. En una el cajero le informa que la bebida subió de precio en un dólar.

Obviamente el dólar extra es un *medio* para obtener la bebida deseada. En la otra le informan que por su compra le obsequian un vaso conmemorativo, lo que al agente le tiene sin cuidado. En este caso, obtener el vaso conmemorativo no es una condición causal del fin principal (bien podría omitirse y obtendría igualmente su bebida ofreciendo el pago). En *Extra dollar*, 95% de los participantes atribuyeron intencionalidad, mientras que en *Free cup* sólo 45% lo hicieron. Pero las viñetas de Machery no son simétricas: en una el efecto en cuestión es un medio, y en la otra es un genuino efecto colateral. De nuevo, la explicación de la asimetría en el juicio de intencionalidad podría apelar a esta diferencia.

Para este punto, es relevante un estudio experimental de Cova y Naar (2012a). Su resultados muestran que la asimetría también se presenta cuando se pregunta por la intencionalidad de los medios utilizados por un agente. Sin embargo, también muestran que hay una tendencia clara: más participantes atribuyen intencionalidad a los medios que a los efectos colaterales. Los resultados de Cova y Naar sugieren que no es grave que a veces se utilicen medios en las viñetas. Sin embargo, es preciso señalar que debería haber al menos consistencia intra-estudios: Por la tendencia a atribuir intencionalidad a los medios más que a los efectos colaterales, no es apropiado demostrar asimetría utilizando medios en una viñeta y efectos colaterales en la otra, cosa que, como vimos, no siempre se tiene en cuenta. Esto va a dificultar nuestro análisis, pues no siempre podremos analizar estudios que sólo usan efectos colaterales o sólo medios. Trataremos de comparar viñetas que usan efectos colaterales. Aunque no descartamos que las hipótesis y los principios que propondremos se puedan aplicar también a las viñetas de medios, no haremos aquí ese examen.

2.2 VALENCIA MORAL DEL EFECTO Y DEL AGENTE

La valencia moral (positivo o negativo) del efecto colateral (o eventualmente del medio) no produce por sí sola el efecto, como creía Knobe (2006). Estudios posteriores (Phelan y Sarkissian 2009) han mostrado, a veces sin proponérselo (viñeta *City Planner* en Phelan y Sarkissian 2008), que la actitud del agente de la historia respecto a la valencia moral del efecto producido tiene una influencia en el juicio de intencionalidad. En *City Planner*, un funcionario distrital produce un efecto colateral previsible y moralmente negativo (desempleo) en su intento por solucionar el problema de la polución en la cuidad. Acabar con la polución es su fin principal, mientras que el desempleo es un efecto previsible no deseado. A diferencia de la viñeta original, sólo un 29% de los sujetos experimentales le atribuyó la producción intencional del desempleo. ¿Por qué? Porque a diferencia de la viñeta original, el agente dijo: "Me siento terrible por incrementar el desempleo."[3] Pero lo importante en realidad, es la coherencia entre lo que declaró y lo que hizo. A diferencia de la viñeta original, el efecto colateral malo no fue consecuencia de un fin egoísta, sino de luchar contra la polución, es decir, de hacer algo por el bien público. Es instructivo comparar aquí *City Planner* con *Reluctant Chairman* de Cova y Naar (2012b). Esta última es similar a la viñeta original en cuanto que el gerente daña al medio ambiente con el ánimo de ganar más dinero; y se distingue de ella porque el gerente afirma que le importa el medio ambiente, como en *City Planner*. Pero a diferencia de lo que sucede en esta última, en *Reluctant Chairman* el 76% de los participantes consideraron intencional el efecto negativo. ¿Por qué? Porque en vistas de su conducta orientada a las ganancias, es probable que muchos participantes no hayan creído que *Reluctant Chairman* tuviera un interés real en el medio ambiente. Entre *City Planner* y *Reluctant Chairman* la única diferencia es que el funcionario distrital quiere combatir la polución, mientras que el

[3] "I feel terrible about increasing joblessness."

gerente quiere ganar más dinero. Y es plausible que la valencia moral del *fin principal* confirme o anule lo que el agente dice con relación al efecto negativo. Las palabras no pesan si se acompañan con acciones incongruentes y se confirmaría así el dicho "los actos dicen más que las palabras". El fin principal de la acción podría influir de manera decisiva en la opinión que uno se forma de la actitud del agente. De estos casos podríamos inferir que no basta que el efecto colateral sea malo para producir la asimetría en la atribución de intencionalidad. Hace falta establecer la actitud del agente, ya sea por sus declaraciones y acciones concordantes, o solo por sus acciones. Esta conjetura es un componente central de la hipótesis que formularemos más adelante.

Cova y Naar defienden que los participantes juzgaron al *Reluctant Chairman* ligeramente pro-ambiente. En una escala de -3 a 3, donde -3 significa tener una actitud anti-ambiente (*anti-environment*) y +3 una actitud pro-ambiente (*pro-environment*), el *Reluctant Chairman* obtuvo 0.64 (la indiferencia es el punto 0). Aunque el propósito de Cova y Naar (2012b) es criticar a Sripada y su modelo de concordancia con el yo profundo (*Deep-Self Concordance Model,* ver sección 4 de este artículo), adoptan una tesis de ese modelo: que la concordancia entre los valores anti o pro X del agente y el efecto colateral previsto, es fundamental para atribuir intencionalidad. Creemos, sin embargo, que no es posible inferir de esas viñetas los valores del agente en una escala descriptiva de *anti* a *pro*. Mostraremos en la sección 4 que es más natural tomar a los agentes que dañan al medio ambiente como egoístas y gorrones, que como despreciando el medio ambiente.

En todo caso, si la viñeta declara la actitud del agente y la acompaña de una acción principal congruente, eso influye en la opinión de los participantes. Hay viñetas con la misma estructura que *City Planner*, pero donde se omite una alusión a la actitud del agente. Se trata de las viñetas clásicas del *Trolley*, en donde una acción principal beneficiosa para terceros acarrea un efecto colateral moralmente indeseable. Machery (2008) recogió datos sobre la atribución de

intencionalidad en *Trolley* y obtuvo un 56% de atribución positiva sobre el daño colateral (la muerte de un trabajador). Esto contrasta con *City Planner* (29%) cuya estructura es similar: un fin bueno para otros acarrea un efecto colateral malo. Aunque los sujetos no tienen información explícita sobre la actitud del agente en Trolley, el fin bueno sugiere que es averso al efecto colateral moralmente indeseable. Lanteri (2009) obtuvo con la misma viñeta *Trolley* un 42% de atribución de intencionalidad al daño colateral. Machery (2008) también recogió datos sobre una variación de *Trolley* – *Dog* – con un efecto colateral bueno: salvar adicionalmente a un perro que está delante de los cinco trabajadores en la vía del tren. En *Dog* agregó también una frase en donde el agente declara explícitamente su actitud: "no me importa el perro". Obtuvo un 23% de respuestas afirmativas sobre la intencionalidad de salvar al perro, lo cual concuerda con la viñeta original de Knobe en la que se utiliza una expresión similar. En contraste con *Dog*, en *Nice Chairman* de Cova y Naar (2012b el gerente dice que le importa el ambiente y escoge un programa que ayuda al medio ambiente entre dos programas que generan iguales ganancias. A diferencia de *Dog*, el fin es egoísta – ganar dinero. Pero se distingue tambien por la actitud del agente: al agente le importa un fin bueno y escoge deliberadamente el programa lucrativo que conduce a ese fin. El 80% de los participantes le atribuyó intencionalidad a su ayuda. Es de notar que aquí la elección entre dos programas convierte el efecto colateral de la viñeta original del gerente y el medio ambiente en un fin escogido.

Phelan y Sarkissian (2009) utilizan explícitamente la actitud del agente como variable en su estudio con *Lieutenant*. Se trata de viñetas en donde un teniente da la orden de tomar una colina que implica el sacrificio de las vidas de muchos de sus soldados. Las viñetas varían la importancia militar de la colina (ninguna o mucha) y la declaración del teniente sobre si le importa o no la vida de sus soldados. Allí es donde descubren que la actitud no se fija únicamente por la declaración

explicita "me importa" o "no me importa". Pues al medir únicamente esa variable, no había efecto sobre la atribución de intencionalidad. Los autores descubren que en las viñetas *Lieutenant*, cuando el teniente dice interesarse (*care*) por sus soldados, y aun así los manda a morir por una causa sin importancia, los sujetos experimentales le atribuyen intencionalidad sobre daño colateral (71%), como si hubiese dicho "no me importa". Es decir, juzgan como si tomasen su declaración explícita por una mentira. Los autores dicen que la opinión de un sujeto experimental sobre la actitud del agente se forma más por el fin principal de la acción que por la actitud expresada. Esto concuerda con lo que acabamos de confirmar en las viñetas analizadas. En *Luitenant*, la muerte de los soldados sólo se atribuye intencionalmente al teniente, si éstas se sacrifican por fines que no están militarmente justificados.

En todos estos casos hay un patrón claro. Cuando las viñetas introducen efectos colaterales previstos con valencia moral binaria – negativa/mala (μ), o positiva/buena (β) –, es preciso inferir la actitud del agente hacia ese efecto antes de atribuir intencionalidad: ¿denota una inclinación a β y una aversión a μ, o indiferencia a ambos? Si el efecto colateral es β, su producción tiende a ser juzgada intencional cuando el agente muestra (en conducta y expresión) que le importa (+), y tiende a no ser juzgada intencional si muestra que no le importa (-). Si el efecto colateral es μ, su producción se juzga intencional siempre y cuando el agente muestre con su declaración o con un fin principal que no le importa (-) producirlo; en cambio no se juzga intencional si el agente persigue un fin principal que moralmente compensa el efecto negativo y muestra que desearía no haber producido dicho efecto (+). En suma, +β y -μ son intencionales; -β y +μ no son intencionales. Llamemos a esta combinación de afirmaciones *el principio M*, por "moral".

El caso clásico del gerente y el medio ambiente combina -β, -μ, mostrando una asimetría en la atribución de intencionalidad. Del mismo modo, según *el principio M*, si combinamos +β con +μ,

obtenemos un 'efecto Knobe invertido': un agente con un compromiso moral (+) – y en eso opuesto al *Chairman* de la viñeta original – produce un efecto β intencionalmente y uno μ no-intencionalmente, reversando la asimetría de las historias originales. En resumen, en los casos en los que el efecto colateral tiene una valencia moral, la actitud moral del agente es condición para juzgar si el efecto colateral fue intencional. Los sujetos que juzgan establecen la actitud del agente de la historia en dos pasos: (i) primero detectan si hay un aspecto moral (bueno o malo) en el efecto colateral (o medio), y (ii) luego establecen si el agente es moralmente indiferente o, al contrario, si tiene un compromiso moral que lo lleva a tener en cuenta los efectos colaterales de su acción. El sujeto experimental hace (implícitamente) un juicio moral sobre la actitud del agente ante el efecto colateral previsto, y ese juicio influye, como veremos más adelante, en la asimetría típica del efecto.

3. EFECTOS COLATERALES SIN VALENCIA MORAL

En casos moralmente neutros, se plantea la pregunta de si la actitud del agente es o no es relevante. Uno podría pensar que sólo lo es cuando el efecto colateral (o medio) tiene valencia moral. Pero no es así. Los filósofos han puesto ejemplos de acciones con efectos colaterales previstos sin valencia moral en donde la actitud del agente influye en el juicio de intencionalidad del espectador. Nado (2008, p. 725-26) nos recuerda un ejemplo de Harman (1976, p. 433) y otro de Bratman (1984, p. 400). Un francotirador decide disparar al blanco aunque alerte así al enemigo de su presencia (efecto colateral); un corredor participa en una maratón con el fin de competir y ganar aunque sus zapatos se desgastarán como efecto colateral. Ambos ejemplos sugieren que, cuando al agente le importa el efecto colateral previsto, ya sea que lo vea como costo o como beneficio, el efecto colateral se ve como intencional. Llamemos a esto *el principio nM*, por "no-moral". Si el efecto colateral es un costo compensado es asumido

intencionalmente por el agente, y si es un beneficio, es una razón más para ejecutar la acción principal y es razonable que el agente la incluya en la intención de su acción. Nótese que las viñetas de Machery (2008) – *extra dollar* y *free-cup* – no desconfirman este principio: cuando el agente dice en *extra dollar* que no le interesa pagar un dólar extra, la declaración no significa que el dólar no es un costo, sino que es un costo compensado. El *principio nM* predice intencionalidad y el resultado es 95%. Cuando en *free-cup* el agente declara que no le interesa el "beneficio", i.e., la *free-cup*, la declaración significa que el agente no ve la copa como un beneficio para él; el *principio nM* predice que no se juzgará intencional y el resultado es 45%. Aquí se produce algo similar a la asimetría, pero solo porque en una viñeta el efecto es un costo y en la otra no es un beneficio. En un par de viñetas que sólo varían porque en una el efecto colateral es un costo y en la otra un beneficio, no se produce un efecto Knobe, pues ambos efectos se juzgarían como intencionales: en una el costo es asumido por mor del fin principal y en la otra el beneficio agrega una razón para perseguir ese fin. En resumen, el efecto colateral no-moral no se juzga intencional si no es un costo ni un beneficio para al agente, y se juzga intencional si le representa un costo o un beneficio sólo a él (valencia no-moral). Pero esto no parece ser algo paradójico o extraño, que hubiera que explicar con alguna teoría especial.

Algunos estudios traen ejemplos de viñetas con efectos colaterales moralmente neutros y que explicitan, en algunos casos, la actitud del agente hacia aquellos. Están, por ejemplo, dos versiones de *Apple Tree* diseñados por Cova y Naar (2012b). En ambas viñetas usadas no hay valencia moral, pero si hay valencia para el gerente de la empresa, quien decide: pues se trata de expandir el edificio de la empresa (fin principal) cortando un árbol, lo cual acarrea un efecto colateral en relación a las emociones del gerente: el árbol protagoniza agradables recuerdos de su niñez en la viñeta 1; el árbol ha sido un fastidio constante, presuntamente por obstruirle la vista desde su

ventana en la viñeta 2. Ambos son casos de efectos colaterales que importan al agente, porque tocan sus emociones; y como según la viñeta nadie más es afectado por esa acción, los consideramos aquí como efectos moralmente neutros, a los que aplica *el principio nM*. El principio dice que si el agente tiene razones a favor o en contra del efecto colateral y consiente en producirlo, éste se ve como intencional. Sin embargo, los resultados no lo confirman: el 78% de los encuestados le atribuyó producción intencional cuando tenía emociones en contra de cortarlo, pero sólo 29% cuando tenía emociones a favor. ¿Debemos concluir que el principio queda experimentalmente refutado? No es así. Lo que sucede en estas viñetas, en ambas, es que hay una contradicción entre dos informaciones sobre el agente: por un lado dice que tiene razones afectivas para no cortar (version 1) o para cortar (versión 2) y por otro lado se dice en ambas que no le importa cortarlo. El principio que sugirió Nado vale para agentes para quienes es claro que el efecto colateral importa, positiva o negativamente. Pero en estas viñetas el agente se revela patentemente inconsistente, razón por la cual los resultados no pueden invalidar el principio.

Otros casos a los que aplica *el principio nM* son 2 pares de viñetas para estudiar la intencionalidad de efectos colaterales vs. medios en Cova y Naar (2012ª): "*Mice" y "Burglar*". Son 4 viñetas con efectos moralmente neutros: un agente cuyo fin principal es desactivar una bomba en el sótano de un edificio perturba ratones como efecto colateral o como medio (*Mice*); o un ladrón cuyo fin es robar una residencia proyecta una sombra, visible desde fuera, como efecto colateral o como medio de comunicación con su cómplice (*Burglar*). Las versiones de efecto colateral obtuvieron 48% (*Burglar*) y 45% (*Mice*) de atribución de intencionalidad. Las versiones de medio obtuvieron 71% y 100% respectivamente, lo que muestra que los medios se juzgan más intencionales que los efectos colaterales. Pero concentrémonos aquí en los efectos colaterales. En ambos casos el efecto colateral no tiene valencia moral; y tampoco se puede decir que represente para el agente

un costo o un beneficio que debiera tener en cuenta. En este caso, si aplicamos el *principio nM*, la predicción es que los sujetos experimentales no van a juzgar intencional ninguno de los dos efectos. En efecto, los resultados son 48% (*Burglar*) y 45% (*Mice*) que, si bien están por debajo del 50%, están demasiado cerca de esa divisoria como para decir que confirman rotundamente el *principio nM*. Pero en todo caso tampoco constituyen una refutación del mismo.

Por último, Knobe y Mendlow (2004) presentaron una viñeta con efectos colaterales sin valencia moral: la gerente de una empresa implementa una nueva política que baja las ventas en la ciudad M y las aumenta en N, con un efecto total positivo, que es el que interesa a la gerente. El 75% de los sujetos experimentales respondió que la gerente bajó intencionalmente las ventas en M. Se presume que bajar las ventas en M importa a la gerente. Sin embargo Phelan & Sarkissian (2008) replicaron el experimento y preguntaron a los participantes si consideraban malo bajar las ventas en M. Sólo un 14% respondió afirmativamente. Pero en este caso, como en otro similar que ponen a prueba, es obvio en el texto de la viñeta que la gerente evalúa el efecto colateral: sopesa si es un costo compensado o no. La pregunta sobre si el efecto es "malo" no es suficientemente precisa para sondear si los participantes piensan que el efecto colateral entra como costo en su deliberaciones, que es lo que importa para la atribución de intencionalidad.

Hasta aquí hemos aislado dos principios que parecen predecir los datos de las diversas viñetas. Ambos afirman que la atribución de intencionalidad de efectos colaterales previstos requiere que previamente se establezca la actitud del agente ante el efecto. Queremos aquí limitarnos al caso clásico en donde la valencia del efecto es moral y binaria (bueno o malo). La actitud del agente también es binaria: o es indiferente a la valencia moral del efecto colateral o revela el compromiso moral de tenerla en cuenta. La primera la simbolizamos como (-) y la segunda como (+). Así vemos que cada una de ellas,

combinada con la valencia moral del efecto, da un juicio de intencionalidad consistente a lo largo de las viñetas investigadas: +β y -μ son intencionales; y -β y +μ no son intencionales (ver tabla 1, apéndice 1). En cierto modo, podemos decir que cuando las valencias *concuerdan* se atribuye intencionalidad, y cuando no, se niega la intencionalidad. Esto tiene un parecido con un modelo basado en un principio de concordancia (Sripada y Konrath 2011). En la sección que sigue mostramos las diferencias con ese modelo y lo criticamos.

4. EL MODELO '*DEEP SELF*'

Sripada y Konrath (2011) (S&K) y Sripada (2012) introdujeron un modelo explicativo en donde lo que realmente cuenta en la atribución de intencionalidad es la concordancia entre el efecto colateral previsto y los valores del agente. La concordancia se establece sobre los valores del agente en tanto estados mentales descritos y no valorados por el sujeto experimental; no se emite (implícitamente) un juicio moral sobre el agente, sino un juicio descriptivo. S&K señalan que las normas morales del agente de la historia y las del sujeto experimental pueden ser distintas. Por tanto, pueden divergir en su valoración del efecto colateral: uno podría verlo malo y el otro no. El sujeto experimental se limita, entonces, a constatar descriptivamente los valores del agente y su concordancia con el efecto producido, sin juzgar moralmente al agente. La divergencia moral se presentaría en casos de normas controversiales, es decir, de normas que tienen defensores y detractores. Un caso así es una norma en favor del aborto, usada en la viñeta *Clinic* (Sripada 2012). Aunque este caso parece especial, en realidad nos alerta, según S&K, sobre lo que sucede en la mente de los sujetos experimentales en todos los casos. Ellos estarían juzgando, descriptivamente, la concordancia del efecto colateral con los valores del agente de la historia.

Para establecer que ese juicio descriptivo es el factor decisivo para atribur intencionalidad de manera asimétrica, utilizan una

metodología estadística conocida como modelo de ecuaciones estructurales (*structural equation modelling*) que evalúa las relaciones causales posibles entre múltiples variables. Su aplicación de la metodología ha sido cuestionada del modo siguiente (Rose et al 2012): El número de modelos posibles sobre las relaciones causales entre las múltiples variables es muy alto, y sus autores no hicieron una búsqueda exhaustiva. Partieron de su modelo preferido y luego realizaron pruebas de ajuste. Esto no garantiza encontrar el mejor modelo. Por otro lado, el índice de ajuste para su modelo es bueno, pero eso se debe a que incluyeron en "su modelo" no sólo su hipótesis positiva, sino también la exclusión de otros modelos como el de Knobe (2006). La exclusión de esas hipótesis tiene mejor puntaje que sus propias hipótesis, y ese puntaje es el que explica el buen índice de ajuste. No queremos entrar en este debate técnico. Pero apoyamos las conclusiones de Rose et al. en la medida en que el modelo preferido de S&K es conceptualmente débil. La justificación conceptual sobre el papel del juicio descriptivo en la atribución de intencionalidad suena plausible y convincente, pero bien mirada, adolece de un defecto fundamental. Básicamente, sucede que el tipo de historias que han usado los estudios sobre el efecto Knobe – y que usa el mismo Sripada (2012) – no permite sacar inferencias sólidas sobre los valores del agente respecto de asuntos controversiales como el aborto o menos controversiales como el medio ambiente. Veamos esto con detenimiento.

Las viñetas de Sripada (2012) siguen las viñetas originales de Knobe. En todas la viñetas el agente es el presidente de una empresa que colateralmente daña o ayuda a otra empresa llamada Beta expresando explícitamente: "no me importa dañar/ayudar a…" mientras persigue un fin principal egoísta, es decir, de ganancia para él y su empresa. Lo que varía es la identidad de Beta y su estatus moral: Beta es una empresa que limpia el medio ambiente (*Charity*); o una empresa que lo ensucia (*Chemical*) o una clínica que practica abortos en etapas avanzadas de gestación (*Clinic*). Es importante que en las tres

viñetas el agente que ayuda /daña busca ganar más dinero y declara explícitamente que no le importa dañar o ayudar a *Charity/Chemical/Clinic*. Como en la viñeta original de Knobe, la mayoría de los participantes le atribuyeron intencionalidad al daño y se la negaron a la ayuda. Según Sripada (2012), el hecho de que dañe o ayude a *Charity/Chemical/Clinic* permite inferir sus valores por una vía directa (2012, p. 233; 237):

> En particular, es plausible que una persona que declara que 'no le importa en absoluto dañar (ayudar) al medio ambiente' proporciona evidencia fuerte de poseer una orientación anti-ambientalista. La persona, puede uno inferir con razón, siente desprecio por el medio ambiente, le otorga un valor bajo al medio ambiente, o evidencia una disposición para dañar el ambiente en una diversidad de situaciones y contextos. (p. 233)[4]

Estas actitudes de su "yo profundo", según Sripada, pueden inferirse de la viñeta. Pero esa inferencia nos parece a todas luces injustificada. Por ejemplo, en la viñeta en donde el presidente daña a la empresa que limpia el medio ambiente, es mucho más natural interpretarlo como una instancia del típico gorrón (*free-rider*) que, en lo que se refiere a recursos comunes o bienes públicos, prefiere vivir de las contribuciones de los demás. El *free rider* no desprecia los recursos comunes ni los considera de bajo valor; simplemente sabe que con algo de astucia y algo de suerte puede disfrutarlos sin contribuir. "Que los demás trabajen por un medio ambiente limpio; si puedo, gozaré los beneficios sin incurrir en costos" podría ser la máxima de ese agente. Pero incluso esto, siendo más natural, es demasiado inferir. En las tres

[4] "Specifically, it is plausible that a person declaring that he **'**doesn't care at all about harming [helping] the environment**'** provides strong evidence that the person has an anti-environment orientation. The person, one might reasonably infer, has contempt for the environment, places a low value on the environment, or evidences a trait-like readiness to harm the environment across a range of situations and contexts."

viñetas el agente declara que le tiene sin cuidado (*I don't care*) si daña o ayuda a esos colectivos, y que sólo le interesan las ganancias. ¿Es eso tener, en *Charity* y *Chemical*, un valor anti- o pro-ambiente? Los sujetos experimentales así lo juzgaron. Pero hay buenas razones para preguntarse qué entendieron por "anti-ambiente". En todo caso, estar en contra de valores ambientales sería como propugnar por la destrucción de la naturaleza. Un valor como este, por extravagante que sea, puede ser un valor; pero nadie que lo tenga declararía al mismo tiempo que le tiene sin cuidado ayudar a sus "enemigos ideológicos" o dañar a los que piensan igual que él, sin ser un traidor a ese valor. El valor debe, al menos algunas veces, influir en sus deseos y acciones. Tampoco puede inferirse de la viñeta que el agente detenta un valor anti-ambiente que traiciona por dinero cuando daña a *Chemical*; pues con la misma razón (es decir, con ninguna) podríamos afirmar que el agente detenta valores pro-ambientalistas que traiciona por dinero cuando daña a *Charity*. Lo más sensato es inferir que el agente manifiesta indiferencia hacia valores morales ambientalistas ampliamente compartidos, sin adherir a valores opuestos. Aunque una mayoría lo clasificó como anti-ambiente en ambas viñetas, nos parece que no hay ningún patrón de inferencia que conduzca confiablemente a esa conclusión a partir de los datos de la historia. Es importante señalar que la pregunta por los valores subyacentes del agente se formuló utilizando el nombre de las compañías dañadas o ayudadas, no el de los valores que plausiblemente representan (Sripada 2012, p. 234). Esto permitiría interpretar que los participantes juzgaron que el agente tenía actitudes a favor o en contra de las compañias, más que a favor en contra de los valores expresados por ellas. Por otro lado, ese juicio sobre la posición del agente no predice, según Rose et al. (2012), la atribución de intencionalidad. El buen ajuste del modelo de S&K no se debe a esa predicción en particular, sino a las predicciones que excluye.

En las viñetas que recorrimos en la sección 2.2, los efectos colaterales previstos de daño/ayuda tenían valores morales binarios

indiscutibles. Si el agente es indiferente hacia esos valores, se puede clasificar sin escrúpulos como reprochable. Las viñetas de S&K (2012) son aparentemente más complejas: los valores binarios de daño o ayuda se modifican al interactuar con el estatus moral de los agentes dañados/ayudados. Pero ya vimos que esa complejidad, aunque real, no se puede evaluar en este tipo de viñetas. Cuando al agente de esas viñetas en particular no le importa si daña o ayuda a X y sólo le interesa el dinero, su actitud no es ni pro-X ni anti-X, sino indiferente a X; y por ello no genera la complejidad que podría generarse cuando un espectador debe evaluar las actitudes profundas del agente (pro-X o anti-X) a partir de sus interacciones con representantes de X (las empresas que daña o ayuda en las viñetas). La conclusión sobre su actitud sigue siendo la misma que en las viñetas originales: el agente es indiferente a los valores morales; y cuando estos son de amplio consenso o no son controversiales, es reprochable. El *principio M* mantiene aquí su validez. Los datos recogidos por S&K confirman este punto de vista: en todas las viñetas, los sujetos experimentales juzgaron al agente como reprochable (*blameworthy*), tanto en la condición de daño como en la de ayuda. Cuando los valores son controversiales, como en la viñeta *Clinic*, en donde el agente, por afán de ganancia, daña o ayuda colateralmente a una clínica que practica abortos en etapas avanzadas del embarazo, es de esperarse que la clínica así descrita tenga defensores y detractores. El veredicto mayoritario de reprochable, que cae sobre el agente que la daña/ayuda colateralmente mientras persigue un beneficio monetario, se puede interpretar aquí del siguiente modo. Probablemente, dañar a *Clinic* por un beneficio personal no fue interpretado por los participantes a favor del aborto como el acto de un enemigo ideológico, ni por sus detractores como el acto de un amigo, sino simplemente como daño con el fin de obtener un beneficio personal, es decir, algo indiscutiblemente malo en ambos casos. Esto nos lleva a proponer nuestra hipótesis en torno al controversial *efecto Knobe*.

5. LA ASIMETRÍA Y LA CONCORDANCIA ENTRE LOS EFECTOS Y EL CARÁCTER MORAL: DIFERENCIAS CON OTROS MODELOS

El modelo que proponemos se deriva del *principio M*. El *principio M* simplemente recoge las características binarias del agente y las características binarias del efecto colateral previsto con valencia moral que generan asimetría en la atribución de intencionalidad. Para generar asimetrías podemos variar la valencia del efecto y mantener constante la actitud del agente hacia él, o podemos mantener constante la valencia del efecto y variar la actitud del agente hacia él. En ambos procedimientos las características binarias morales ya sea del efecto, ya sea del agente, son los únicos factores que varían y generan la asimetría en la atribución de intencionalidad. Esta formulación completa el elemento faltante en la hipótesis original de Knobe, como está formulada en Knobe (2006), que sostenía que un efecto colateral previsto es intencional si es moralmente malo; y no lo es, si es bueno. Ahí faltaba explicitar que la actitud del agente debía mantenerse constante en ambos casos y que debía ser de indiferencia hacia ese valor moral binario. Con esa formulación por ejemplo, no era posible predecir la existencia del "*efecto Knobe invertido*": un agente manifiestamente comprometido con fines morales puede protagonizar un par de historias, produciendo intencionalmente un efecto colateral bueno en una, y un efecto colateral malo sin intención en la otra; ambos previstos. Tampoco era posible predecir que obtenemos una asimetría si mantenemos la valencia moral del efecto colateral constante y variamos solamente la actitud del agente. Pero estas últimas son más difíciles de formular. La dificultad estriba en que el mejor método para transmitir en una historia breve una actitud moral comprometida (+) es describir un agente que actúa para producir un bien público; mientras que transmitir una actitud moral indiferente (-) se logra mejor con un agente que actúa por beneficio propio contra el bien público. Las viñetas serían diferentes en el fin principal y se perdería la simplicidad

de las viñetas originales. La viñetas del *efecto Knobe invertido* son más fáciles de formular y ponemos un ejemplo en el apéndice 2.

Las actitudes binarias del agente son el compromiso con valores morales (+) o indiferencia hacia esos valores y prioridad para los fines egoístas (-). El efecto colateral previsto puede ser de valencia moral positiva (β) o negativa (μ). De ese modo tenemos 4 combinaciones posibles que generan asimetría. En el cuadro que sigue, las dos filas representan el caso original y el invertido: (-β,−μ) y (+β,+μ); y las dos columnas representan la asimetría con efecto de valencia moral constante: (+β, −β) y (−μ,+μ).

	Valencia del efecto	
Actitud del Agente	**Malo (μ)**	**Bueno (β)**
Negativa (-) (Indiferencia moral)	**(-μ)** Intencional	**(-β)** No intencional
Positiva (+) (Compromiso moral)	**(+μ)**No Intencional	**(+β)** Intencional

Cuadro 1: Intencionalidad por tipos de Viñetas (ValAgente x ValEfecto)

Las diferencias con Knobe (2006) son claras, pues Knobe no había considerado la actitud del agente. Nos parece importante destacar ahora las diferencias con otros dos modelos: el modelo de Nadelhofer (2004) y otros, refinado en Nado (2008); y el modelo de Sripada y Konrath (2011).

Según Nadelhofer (2004), cuando el efecto colateral tiene una valencia moral negativa, los sujetos atribuyen primero responsabilidad al agente y luego lo juzgan intencional. Nado (2008) agrega que este proceso sucede automáticamente a un nivel sub-personal. En el modelo que defendemos esta explicación falla en vistas de que los sujetos experimentales juzgan que los efectos colaterales de valencia moral negativa no son producidos intencionalmente cuando el agente muestra compromiso con los valores morales binarios. Nuestro análisis sugiere

que hay un paso previo a la atribución de intencionalidad y de responsabilidad, que es la evaluación de la actitud moral del agente. El fin principal de su acción y su declaración al ser confrontado con la valencia del efecto colateral dan información sobre su actitud moral. Una vez se establece ésta, se puede atribuir o negar intencionalidad y responsabilidad; y estas dos atribuciones pueden ser simultáneas. Lo que sí tiene que preceder es un juicio sobre la actitud moral del agente.

Entre el modelo aquí sugerido y el modelo de Sripada y Konrath (2011) hay un punto de acuerdo fundamental: las cualidades de carácter y disposiciones presumiblemente profundas (*Deep Self*) del agente son el criterio que los sujetos experimentales utilizan para atribuir intencionalidad. La diferencia con S&K es que el juicio sobre las cualidades de carácter es al mismo tiempo una descripción de las disposiciones del agente y una evaluación de esas disposiciones desde un punto de vista moral. Los participantes no juzgan si el agente es 'anti' o 'pro' con relación a valores controversiales que admiten valores opuestos, como sucedería en relación a una norma que favorece el aborto. El sujeto que juzga no toma un punto de vista objetivo y desapegado sobre los valores en cuestión. Más bien juzga si el agente tiene compromiso con, o indiferencia hacia, valores morales binarios que no son ni se consideran controversiales. Juzga si el agente se compromete con valores morales binarios que no están en discusión.

Quizás quiera verse como un punto débil de nuestra propuesta que afirmemos tajantemente que la actitud del agente se evalua con valores morales incuestionables. Pero examinemos la viñeta original de Knobe, *Chairman,* y establezcamos qué asunto moral está ahí en juego. El *Chairman* decide ensuciar el medio ambiente con un curso de acción que lo lleva a ganar más dinero. Ensuciar el medio ambiente es hacer un daño a toda una comunidad. Y aunque su objetivo principal es ganar dinero y no dañar, no deja de ser cierto que decide ganar dinero a costas de dañar a toda una comunidad. Esta forma de proceder es típica de un *free-rider*. Es bien conocida a los seres humanos desde tiempos

inmemoriales (Trivers 1971), y es también desde siempre considerada una falta moral incontestable. Su aplicación al medio ambiente limpio, como recurso común, es reciente, y quizas sólo por eso es todavía posible ser, impunemente, un *free-rider* en ese contexto. *Chairman* se comporta como un *free-rider*, sea o no consciente de ello. La ignorancia no lo disculparía, porque cualquiera puede entender que quiere obtener beneficios a costas de un bien público. Hacia esa actitud todos tenemos una aversión ancestral y profunda. Del mismo modo, tenemos respeto por quienes dan muestras claras de compromiso con la norma contra esas conductas. Esto da respaldo a la idea de que hay, en efecto, valores morales incontestables y que están presentes en la asimetría.

De aquí se derivan algunas sugerencias para un problema que no pretendimos tratar aquí: si ser influenciado por factores morales es parte, o no, de la competencia linguística en relación al concepto de intencionalidad. Nos limitamos a un breve apunte. Es posible que el concepto de intencionalidad sea un instrumento para lidiar también con estructuras de interacción social, probablemente ancestrales, que los científicos sociales conocen como "dilemas sociales" (Trivers 1971; Kollock 1998). De ahí su estrecha conexión con el concepto de responsabilidad, pues los dilemas sociales se resuelven comunmente mediante normas de comportamiento públicamente acordadas. Pero la conexión normativa no se riñe con su aspecto predictivo. Este se expresa en la exigencia de una concordancia entre el estado mental del agente y los efectos de su acción, de manera que la causa sea connatural al efecto. El interés predictivo es natural en relación con los rasgos de carácter que facilitan un comportamiento adecuado en el grupo. La concordancia que interesa aquí se refiere a disposiciones fundamentales hacia el cumplimiento de normas de conducta que se instituyen cuando están en juego costos y beneficios, y por tanto conflictos, que ponen en peligro la estabilidad social.

Apéndice 1. Tabla de viñetas con contenido moral y % de atribución de intencionalidad

Autor, año	Viñeta	Versión	Tipo (ValAct. x ValEfecto)	% Si Intencional	Cumple predicción? (S/N)
Knobe (2003)	Chairman and Enviroment	Harm Case	(–) **μ**	82%	S
Knobe (2003)	Chairman and Enviroment	Help Case	(–) **β**	23%	S
Machery (2008)	Trolley /Lever	Harm/ Good Goal	(+?) **μ**	56%	N
Machery (2008)	Dog (Trolley/Lever)	Help /Uncaring	(–) **β**	23%	S
Mallon (2008)	Terrorist	Harmful Terrorist	(–) **μ**	92%	S
Mallon (2008)	Terrorist	Helpful Terrorist	(–) **β**	12%	S
Mallon (2008)	Gang Leader	Harmful Gang Leader	(–) **μ**	62%	S
Mallon (2008)	Gang Leader	Helpful Gang Leader	(–) **β**	28%	S
Phelan y Sarkissian (2009)	Lieutenant	Caring / Important	(+) **μ**	45%	S
Phelan y Sarkissian (2009)	Lieutenant	Caring / Unimportant	(–) **μ**	71%	S
Phelan y Sarkissian (2009)	Lieutenant	Uncaring / Important	(+) **μ**	50%	S?
Phelan y Sarkissian (2009)	Lieutenant	Uncaring / Unimportant	(–) **μ**	76%	S
Lanteri (2009)	Trolley/ Lever	Harm/Good Goal	(+?)**μ**	42%	S
Lanteri (2009)	Trolley/ Footbridge	Harm/ means/ Good G	(+?)**μ**	90%	
Cova y Naar (2012a)	Terrorist /side-effect	Harmful Terrorist	(–) **μ**	62%	S
Cova y Naar (2012a)	Terrorist /side-effect	Helpful Terrorist	(–) **β**	10%	S
Cova y Naar (2012a)	Productivity/means	Bad Means	(–) **μ**	82%	S
Cova y Naar (2012a)	Productivity/means	Good Means	(–) **β**	42%	S
Cova y Naar (2012b)	Chairman and Enviroment	Reluctant Chairman	(–) **μ**	76%	S
Cova y Naar (2012b)	Chairman and Enviroment	Very Nice chairman	(+) **β**	80%	S
Phelan y Sarkissian (2008)	City Planner	Harm/Good Goal	(+) **μ**	29%	S
Autor, año	**Viñeta**	**Versión**	**Tipo (ValAct. x ValEfecto)**	**% SI (resp. > 0)**	**Cumple predicción? (S/N)**
Sripada (2012)	Charity	Harm version	(–) **μ**	68%	S
Sripada (2012)	Charity	Help version	(–) **β**	9%	S
Sripada (2012)	Chemical	Harm version	(–) **μ / β**	57%	S?
Sripada (2012)	Chemical	Help version	(–) **β / μ**	19%	S?
Sripada (2012)	Clinic	Harm version	(–) **μ / μ,β**	59%	S?
Sripada (2012)	Clinic	Help version	(–) **β / β,μ**	15%	S?

Tabla 1. Esta tabla muestra las viñetas clasificadas por tipos según la tipología de la sección 5 y muestra si confirman la predicción del modelo defendido en el texto en cuanto a la atribución de intencionalidad sobre los efectos colaterales (o medios, en algunos casos). Todas las viñetas confirman salvo Trolley/Lever (Machery 2008). Se muestran por pares que varían, ya sea la valencia del efecto colateral (β, μ), ya sea el compromiso moral del agente (+, -). Hay una viñeta sin pareja (*City Planner*). En el caso de las viñetas de Sripada *Chemical* y *Clinic*, hay dificultades en establecer la valencia moral del efecto colateral, según si se da primacía al hecho de que el efecto es daño o ayuda a un agente, o si se tiene en cuenta el estatus moral del agente como bueno o malo. La predicción se cumple sólo si se describe el efecto como daño o ayuda y ponemos por tanto un signo de interrogación en la predicción. Sripada publicó sus datos como promedios de puntaje de intencionalidad en una escala de -3 a 3. Nosotros registramos las respuestas > 0 como respuestas SI a intencionalidad. Agradecemos a Sripada habernos compartido sus datos.

Apéndice 2. Viñetas de *efecto Knobe invertido* (+β,+μ)

V1(**+β)**: El asistente va donde el alcalde y le dice: "Su larga y esforzada búsqueda de una estrategia efectiva para acabar con la polución y los residuos tóxicos dio resultados: por medio de simulaciones hemos encontrado una estrategia exitosa. Sé además lo mucho que le preocupan las tasas de empleo. Por eso tenga en cuenta que la estrategia genera empleo." El alcalde responde: "Me siento feliz por generar empleo, y tenemos que acabar con la polución: implementemos la estrategia. La implementan, eliminan la polución y generan empleo."

¿Generó el alcalde empleo intencionalmente?

V2(**+μ)**: El asistente va donde el alcalde y le dice: "Su larga y esforzada búsqueda de una estrategia efectiva para acabar con la polución y los residuos tóxicos dio resultados: por medio de simulaciones hemos encontrado una estrategia exitosa. Sé además lo mucho que le preocupan las tasas de empleo. Por eso tenga en cuenta que la estrategia genera desempleo." El alcalde responde: "Me siento muy mal por generar desempleo, pero tenemos que acabar con la polución: implementemos la estrategia". La implementan, eliminan la polución y generan desempleo.

¿Generó el alcalde desempleo intencionalmente?

7. BIBLIOGRAFÍA

BRATMAN, M. "Two faces of intention". *Philosophical Review,* 93 (3) pp. 375-405, 1984.

COVA, F., NAAR, H. "Side-Effect effect without side effects: The pervasive impact of moral considerations on judgments of intentionality". *Philosophical Psychology* 25 (6) pp. 837-854, 2012a.

————. "Testing Sripada's Deep Self model". *Philosophical Psychology* 25 (5) pp. 647 – 659, 2012b.

HARMAN, G. "Practical Reasoning". *Review of Metaphysics*, 29 pp. 431–63, 1976.

HINDRIKS, F. "Intentional action and the praise-blame asymmetry". *Philosophical Quarterly* 58 (233) pp. 630-641, 2008.

HOLTON, R. "Norms and the Knobe Effect". *Analysis* 70 (3) pp. 1-8, 2010.

KNOBE, J. "Intentional action and side effects in ordinary language". *Analysis* 63 (3) pp. 190–194, 2003.

————. "The concept of intentional action: A case study in the uses of folk psychology". *Philosophical Studies* 130 (2) pp. 203-231, 2006.

KNOBE, J., BURRA, A. "The folk concepts of intention and intentional action: A cross-cultural study". *Journal of Cognition and Culture* 6 (1-2) pp. 113-132, 2006.

KNOBE, J., MENDLOW G. "The good, the bad, and the blameworthy: Understanding the role of evaluative considerations in folk psychology". *Journal of Theoretical and Philosophical Psychology* 24 pp. 252–258, 2004.

KOLLOCK, P. "Social Dilemmas: The Anatomy of Cooperation". *Annual Review of Sociology* 24 pp. 183-214, 1998.

LANTERI, A. "Judgments of intentionality and moral worth: Experimental challenges to Hindriks". *Philosophical Quarterly* 59 (237) pp. 713-720, 2009.

LESLIE, A., KNOBE, J., COHEN, A. "Acting intentionally and the side-effect effect: 'Theory of mind' and moral judgment". *Psychological Science* 17 pp. 421-427, 2006.

MACHERY, E. "The folk concept of intentional action: Philosophical and experimental issues". *Mind and Language* 23 (2) pp. 165–189, 2006.

MALLON, R. "Knobe vs. Machery: Testing the trade-off hypothesis". *Mind and Language* 23 (2), pp. 247-255, 2008.

McCANN, H. "Intentional Action and Intending: Recent Empirical Studies". *Philosophical Psychology*, 18 (6), pp. 737-748, 2005.

MELE, A. "Intentional action: Controversies, data, and core hypotheses". *Philosophical Psychology,*16 (2), pp. 325-340, 2003.

NADELHOFFER, T. "On praise, side effects, and folk ascriptions of intentionality". *Journal of Theoretical and Philosophical Psychology,* 24 (2), pp. 196-213, 2004,.)

NADO, J. "Effects of moral cognition on judgments of intentionality". *British Journal for the Philosophy of Science,* 59 (4), pp. 709-731, 2008.

PHELAN, M., SARKISSIAN, H. "The folk strike back; or, why you didn't do it intentionally, though it was bad and you knew it". *Philosophical Studies,* 138 (2), pp. 291 – 298, 2008.

————. "Is the 'trade-off hypothesis' worth trading for?" *Mind and Language,* 24 (2), pp. 164-180, 2009.

ROSE, D., LIVENGOOD, J., SYTSMA, J., MACHERY, E. "Deep trouble for the deep self". *Philosophical Psychology,* 25 (5), pp. 629 – 646, 2011.

SRIPADA, C. "Mental State Attributions and the Side-Effect Effect". *Journal of Experimental Social Psychology,* 48 (1), pp. 232-238, 2012.

SRIPADA, C., KONRATH, S. "Telling More Than We Can Know About Intentional Action". *Mind and Language,* 26 (3), pp. 353-380, 2011.

TRIVERS, R., "The evolution of reciprocal altruism". *Quarterly Review of Biology,* 46 (1), pp. 35–57, 1971.

YOUNG, L., CUSHMAN, F., ADOLPHS, R., TRANEL, D., HAUSER, M. "Does emotion mediate the effect of an action's moral status on its intentional status? Neuropsychological evidence". *Journal of Cognition and Culture,* 6, pp. 291-304, 2006.