Senci, M. 2019. Castigo y normas sociales. Acerca de la validez ecológica del castigo en la economía experimental, parte I
Cinta moebio 66: 321-332
https://doi.org/10.4067/S0717-554X2019000300321

Castigo y normas sociales. Acerca de la validez ecológica del castigo en la economía experimental, parte I

Punishment and social norms. On the ecological validity of punishment in experimental economics, part I

Maximiliano Senci (maximiliano.senci@uns.edu.ar) Instituto de Investigaciones Económicas y Sociales del Sur, Universidad Nacional del Sur - CONICET (Bahía Blanca, Argentina) ORCID: 0000-0001-9131-3843

Abstract

This article focuses on the connection between punishment and norms. Experimental studies in economics assume that transgression of norms of justice and equity is the cause of the application of punishment. While such a connection is evident, there are reasons to believe that external validity of such inference may be undermined due to certain experimental artefacts. These experimental artefacts refer to two sets of problems linked to ecological validity: first, the "normative irrelevance" of punishment, that is, that participants in experimental economic games use the option of punishment for reasons unrelated to the transgression of norms; and second, the illegitimacy of punishment, as a result of the lack of adequate procedures to the allocation of roles and normative ambiguity. I conclude with recommendations about how to tackle these problems.

Key words: punishment, social norms, ecological validity, legitimacy, experimental economics

Resumen

El presente artículo es la primera parte de un trabajo que estudia la conexión entre el castigo y las normas en estudios experimentales en economía. Estos asumen que la transgresión de normas de justicia y equidad es la causa de la aplicación del castigo. A pesar de que hay evidencia empírica que refuerza tal conexión, existen razones para pensar que su validez ecológica puede verse socavada debido a ciertos artefactos experimentales. Estos artefactos experimentales se refieren a dos conjuntos de problemas vinculados con la validez ecológica: primero, la “irrelevancia normativa” del castigo, esto es, que los participantes de los juegos económicos experimentales utilizan la opción de castigo por motivos no relacionados a la transgresión de normas; y segundo, la ilegitimidad del castigo, producto de la falta de procedimientos adecuados para la asignación de roles y de la ambigüedad normativa.

Palabras clave: castigo, normas sociales, validez ecológica, legitimidad, economía experimental.

1. Introducción

Los experimentos económicos en los que los participantes tienen la posibilidad de castigar a otros participantes ("experimentos de castigo" de aquí en adelante) ofrecen dos resultados centrales a las ciencias sociales: en primer lugar, el libro clásico de Ostrom, Walker y Gardner (Rules, games, and common-pool resources) y los pioneros trabajos de Fehr y Gächter (Cooperation and punishment in public goods experiments) muestran que los niveles de cooperación pueden mantenerse estables mediante la imposición de sanciones informales, es decir, el castigo puede contribuir a sostener la pro-socialidad; y en segundo lugar, sugieren que la motivación principal para imponer una sanción proviene de la percepción del castigador de que el castigado transgredió una norma social (típicamente normas de equidad). Este último resultado depende de la inferencia de que la transgresión normativa es *causa* de la aplicación de la sanción. En este trabajo me propongo llamar la atención sobre la dificultad de asumir la validez de esta inferencia sin reservas.

Las ciencias sociales giran en torno al concepto vertebrador de norma social. Sin embargo, éste es un término equívoco. Para Ostrom las normas sociales son "una comprensión compartida sobre las acciones que son obligatorias, permitidas, o prohibidas" (Ostrom 2000:143). En buena medida, el elemento de la sanción está presente en casi todas las definiciones de norma social que se emplean en la literatura experimental en economía. Por ejemplo, la siguiente definición incorpora explícitamente el elemento punitivo: "1) una regularidad comportamental; que está 2) basada en una creencia social común sobre cómo uno debe comportarse; y que motiva 3) la exigencia de cumplimiento (*enforcement*) del comportamiento prescripto por medio de sanciones sociales informales" (Fehr y Gächter 2000:166). La investigación sobre normas sociales ha crecido exponencialmente en la literatura de economía experimental. La situación actual de la literatura no refleja un consenso, sino que más bien pone en evidencia la ausencia de acuerdos básicos en torno a la definición misma de norma social. Se suele asumir, y en esto sí hay reflejado un cierto consenso, que las normas sociales poseen poder motivacional. Esto es, para motivar su cumplimiento no es necesario que las normas estén equipadas con un arsenal de incentivos, ya sea en la forma de recompensas o de castigos. Por el contrario, las normas tienen autonomía y poder motivacional en sí mismas, lo que algunos autores, e. g. Sripada y Stich (A framework for the psychology of norms) han llamado "la independencia normativa de las normas". Menos se sabe, sin embargo, acerca de cuáles son los mecanismos específicos por medio de los cuales se produce dicha motivación. Este estado de cosas es desalentador si se piensa que las ciencias sociales pueden estar cimentadas sobre una noción cenagosa cuya influencia sobre diferentes comportamientos es difícil de determinar. Conocer los mecanismos específicos que subyacen a la observancia de normas es un tema acuciante, ya que uno de los aspectos de relevancia aplicada de este campo es poder modificar o cambiar conductas no deseadas (normas sociales como la infibulación, distintas formas de machismo, el trabajo infantil o la corrupción) y las intervenciones destinadas a ello pueden centrarse en el cambio de normas sociales. Para que dichas intervenciones resulten exitosas, es necesario entender los mecanismos relevantes que intervienen.

A pesar de que la observancia de las normas pueda responder a diversos motivos, el más estudiado es, por lejos, el castigo: esto es, el hecho de que evitar ser sancionado puede ser una motivación fuerte para observar una norma. Una de las formas preferidas de los economistas para medir normas sociales es el "experimento de castigo". La idea, que encuentra una justificación teórica en la noción de la "función expresiva del castigo", es sencilla: el castigo, cuando es aplicado, señaliza la presencia de una norma (sobre esta idea ver la amplia literatura de *Law and Economics*, que incluye trabajos clásicos de Kahan, Sunstein, Cooter y otros). La función expresiva del castigo consiste en que este no solo genera un incentivo para

evitar las consecuencias provenientes del castigo, sino que además brinda información acerca de la prescripción o proscripción de cierta acción. La inferencia de los investigadores es la siguiente: si en un contexto de laboratorio los participantes utilizan el castigo (que se instrumenta típicamente como un descuento de dinero) es porque perciben que el castigado ha transgredido algún tipo de norma. Para evitar interpretaciones erróneas, en el artículo no me voy a comprometer con la tesis de que el castigo en los experimentos refleje lo que sucede en situaciones de la vida real. Sigo en esto lo sostenido por Francesco Guala (Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate).

Mi argumento es el siguiente: en el caso de que se aceptara una tesis más débil de acuerdo con la cual el dispositivo experimental del castigo fuera una buena herramienta para estudiar los procesos y mecanismos psicológicos que intervienen en la práctica del castigo en la vida real (de la transgresión de normas sociales), aun así, los experimentos no estarían a la altura para captar dichos mecanismos porque adolecerían de un déficit de validez ecológica.

Para mostrar este déficit, me propongo estudiar algunos desafíos que surgen en la literatura experimental y que dificultan la justificación de la conexión entre castigo y normas sociales, socavando la validez ecológica que los experimentos pretenden alcanzar. Por "validez ecológica" se entiende el grado en que una investigación refleja el fenómeno de la vida real o cotidiana que se pretende estudiar. Siguiendo a Brewer, en un contexto experimental una tarea puede carecer de validez ecológica si, en su totalidad o en parte, difiere sustancialmente de lo que el participante experimentaría en la vida real. Dicho de otra manera, el experimento tiene que ser representativo de la situación real que se pretende estudiar. Brewer ha distinguido dos formas de realismo: el realismo mundano y el psicológico. El primero "refers to the extent to which the research setting and operations resemble events in normal, everyday life" (Brewer 2000:12); mientras que el realismo psicológico "is the extent to which the psychological processes that occur in an experiment are the same as psychological processes that occur in everyday life" (Brewer 2000:12). Los experimentos económicos procuran, en general, preservar este último. Ya que, en buena medida, los experimentos de castigo tienen por objetivo último diseñar intervenciones tendientes a resolver problemas aplicados (como por ejemplo la erradicación de normas ineficientes o destructivas para el conjunto de la sociedad o alguna minoría), la imposibilidad de generalizar sus resultados a situaciones de la vida cotidiana representaría una objeción relevante.

Los desafíos antes mencionados se refieren a dos conjuntos de problemas vinculados con la validez ecológica: primero, la "irrelevancia normativa" del castigo, esto es, que los participantes de los juegos económicos experimentales utilizan la opción de castigo por motivos no relacionados a la transgresión de normas; y segundo, la ilegitimidad del castigo, producto de la falta de procedimientos adecuados y de la ambigüedad normativa.

El argumento que voy a sostener es que la suma de estos problemas atenta contra el realismo psicológico de los experimentos, e impiden, por ende, que sus resultados puedan ser generalizados sin reservas a situaciones de la vida real.

Comenzaré con una breve descripción de las formas en las que se estudia la conformidad normativa en experimentos de castigo. Luego presentaré las objeciones o problemas antes mencionados. Dejaré para la Parte II (que deberá seguir a este artículo) algunas recomendaciones respecto de cómo atacar estos problemas.

2. El castigo en la economía experimental

La literatura sobre normas señala típicamente dos rutas explicativas de la conformidad normativa. Por un lado, lo que podemos llamar la tesis de los incentivos; y por otro, la tesis de la autonomía. De acuerdo con la visión más estrecha propuesta por la primera, las personas no deberían actuar conforme a las normas a menos que estas vengan equipadas con el suficiente arsenal de incentivos, ya sea en forma de recompensas o de castigos, que motiven su cumplimiento. Por su parte, de la segunda tesis se deduce que las normas poseen independencia normativa, esto es, que las normas son autónomas y tienen poder motivacional en sí mismas y, por lo tanto, son irreductibles a la optimización racional.

Entre los mecanismos instrumentales, el más comúnmente empleado como explicación de la conformidad es *el castigo* de los transgresores por parte de aquellos que observan la norma, a pesar de que fuera del laboratorio no es común que las personas tomen en sus propias manos la implementación de castigos materiales. Como ha señalado Francesco Guala: "there are probably good reasons why decentralised, spontaneous material punishment is so rare outside the laboratory. In modern states decentralised sanctioning is explicitly forbidden by law, and anti-social behaviour is curtailed in ways that minimize the risk of feuds. Retaliation is controlled by imposing a monopoly of state violence, and the cost of punishment is recouped by compensating 'professional punishers' (e.g., policemen)" (Guala 2012:9). Es posible que existan ciertos ámbitos específicos, sin embargo, en los cuales el castigo tal como es utilizado en los experimentos tenga una mayor validez ecológica, particularmente en aquellos ámbitos que tienen de por sí un componente artificial, como por ejemplo las interacciones de los wikiworkers en Wikipedia o los sistemas de reputación de ciertos comercios electrónicos (e. g., eBay, Mercado Libre).

En los estudios de economía experimental se suele modelar el castigo utilizando las herramientas formales de la teoría de juegos, al menos de tres formas distintas. Una primera forma consiste en la transformación de la matriz de pagos. Por ejemplo, si en el Dilema del Prisionero (DP de aquí en adelante) se introduce un parámetro negativo en el pago que recibe un jugador por no-cooperar (lo que resulta en que la utilidad que recibe un jugador es menor si elige no cooperar, y por lo tanto la no-conformidad es castigada), entonces la estrategia cooperativa se puede volver un equilibrio del juego. Esta es la manera en la que Bicchieri ha propuesto modelar las normas sociales. Para Bicchieri una norma es un equilibrio de Nash, y nunca es una respuesta a problemas de coordinación (i.e., cuando hay equilibrios múltiples), sino que surge como respuesta a situaciones (o juegos) de motivación mixta (*mixed-motive games*), es decir, en las que existen conflictos de interés entre los actores, pero al mismo tiempo hay espacio para lograr una mejora paretiana. Y precisamente, las normas transforman "juegos de motivación mixta en juegos de coordinación" (Bicchieri 2006:3). Según esta caracterización, las normas surgirían típicamente en dilemas sociales. Este primer tipo de modelado del castigo es más apropiado cuando se intenta representar un costo moral o emocional en que incurre un actor por transgredir una norma.

La segunda opción consiste en expandir los juegos para habilitar la posibilidad de sanción. Los experimentos sobre castigo han utilizado típicamente juegos experimentales en los que la norma que suscita el castigo es una norma de equidad. Por ejemplo, en el clásico juego del dictador, uno de los participantes (llamado el dictador) recibe del experimentador una suma $X, que tiene que distribuir (típicamente) con otro participante anónimo. La norma de justicia prescribe que el participante que recibió el dinero debería darle al otro $X/2, de manera tal que cada cual reciba la mitad. En juegos diseñados para estudiar el castigo de terceros se introduce un tercer jugador que puede observar la distribución propuesta y destinar parte de su dinero para castigar (i.e., descontar dinero) al dictador. Algo similar ocurre con el Juego de Bienes Públicos (JBP de aquí en adelante), en el que luego de una primera etapa de

transferencias, se añade una etapa en la que los jugadores pueden castigar a otros. El castigo consiste normalmente en un descuento de puntos.

La tercera opción, el castigo de segundas partes, también puede tomarse como una medida de la presencia de normas sociales, pero la inferencia que permite es más ambigua. Por ejemplo, en el Juego de Ultimátum el rechazo de ofrecimientos bajos suele tomarse como evidencia de que las personas interpretan un ofrecimiento bajo como la transgresión de una norma de justicia. Sin embargo, hay interpretaciones alternativas, según las cuales recibir una oferta baja puede ser interpretada como la imposición de un estatus inferior.

3. Evidencia sobre el castigo en juegos experimentales

La evidencia sobre la eficacia del castigo en juegos experimentales es abundante. Uno de los experimentos más representativos de la eficacia del castigo sobre la cooperación proviene de Fehr y Gächter (2000). Sobre la base de un JBP, los autores realizaron dos tratamientos con grupos de extraños (tratamientos *stranger* según la jerga), con y sin posibilidad de castigo, y dos tratamientos con grupos de compañeros (*partner* según la jerga*)*, también con las mismas variantes, con y sin castigo. Bajo los supuestos estándar de conocimiento común de racionalidad y egoísmo, la predicción es que los participantes no deberían cooperar, dado que de esa manera se obtiene el mayor beneficio individual. No obstante, si todos los miembros del grupo cooperan completamente, se maximiza el pago agregado. Las contribuciones son simultáneas y en todas las condiciones los participantes conocen, al final de la etapa de contribuciones, cuánto contribuyó cada uno de los otros miembros del grupo (esto es, la contribución y las ganancias en cada ronda son conocimiento común). La diferencia entre condiciones con y sin castigo, consiste en que en las últimas se agrega una etapa, posterior a la etapa de contribuciones, en la que los participantes pueden asignar parte de sus recursos a *descontar* puntos de otros participantes. Los autores establecieron los parámetros de manera tal que el castigo, al igual que la cooperación, nunca sea parte del equilibrio perfecto en sub-juegos (bajo los supuestos usuales de conocimiento común de racionalidad y egoísmo). Tres resultados son dignos de mención: 1) en primer lugar, la posibilidad de castigo aumentó considerablemente la contribución promedio tanto en las condiciones *partner* como *stranger* respecto de las condiciones sin castigo. El aumento de las contribuciones representó, en promedio, un 58% de la dotación (de la cantidad de dinero que habían recibido los participantes de los experimentadores); 2) en la condición *partner* sin castigo, las contribuciones convergieron a cero, mientras que en la condición con castigo el patrón fue el opuesto, dado que los participantes contribuyeron casi el total de su dinero; 3) en ambas condiciones, el castigo fue una función del desvío respecto de la contribución promedio del grupo, cuanto más se alejó hacia abajo la contribución de un participante del promedio del grupo, tanto más se lo castigó. Un punto importante consiste en que los participantes observaron una norma de cooperación condicional; esto es, cooperaron si los demás cooperaban, y aumentaron su contribución en función del incremento de la contribución promedio. Este último aspecto, dicen los autores, es evidencia de que el castigo acontece a causa de la presencia de normas sociales.

También existe evidencia relevante proveniente del castigo de terceros, tanto en el Juego del Dictador, como en el DP (Dilema del Prisionero). En el caso del DP simultáneo, un tercer jugador hace las veces de observador y puede decidir sancionar a los otros jugadores. Los experimentos de castigo de terceros son apropiados para estudiar la existencia de normas sociales, ya que el tercero en este caso no se ve afectado por las decisiones de los otros jugadores, ni tampoco obtiene ningún beneficio al sancionar. Fehr y Fischbacher (Third-party punishment and social norms) encuentran que alrededor del 50% de los sujetos en el rol de observador imparcial están dispuestos a castigar la defección (esto es, las decisiones no

cooperativas), mientras que no sucede lo mismo con las decisiones cooperativas. Además, la defección de un jugador es castigada de manera más severa cuando el otro jugador ha decidido cooperar. De ello, según los autores, se puede extraer la inferencia de que los terceros aplican sanciones solo cuando se transgreden normas sociales.

En suma, los participantes tienden a castigar a otros participantes que fueron injustos con otros (extraños, debido a la norma de anonimato que se sigue en los experimentos). La idea que subyace es que las sociedades humanas han evolucionado de manera tal que el castigo ha resultado ser un instrumento para fomentar la cooperación.

4. La irrelevancia normativa del castigo

Una particularidad de las sanciones en los juegos económicos consiste en que toman la forma de una privación o costo económico para el castigado. En español diríamos que se trata de "multas" o sanciones monetarias. Mi argumento acerca de la irrelevancia normativa de las sanciones monetarias en los juegos económicos consiste en mostrar que los sujetos experimentales pueden interpretar el castigo como un precio. El castigo persigue un efecto disuasorio sobre la conducta reprochada. Idealmente para que el castigo sea legítimo tiene que estar ligado a la transgresión de una norma. El "precio" hace saliente la solución de compromiso entre el cumplimiento de la norma y el castigo: si el castigo excede el beneficio potencial de transgredir la norma, entonces conviene evitarlo; mientras que, si el beneficio supera el castigo, entonces es conveniente incurrir en el costo monetario. En este caso el castigo puede tener un efecto opuesto al perseguido (como muestra el ejemplo discutido a continuación). Esto quiere decir que, en lugar de asociar una determinada conducta con una proscripción, cuando el castigo toma la forma de un costo monetario puede interferir con la motivación normativa intrínseca de los sujetos, es decir, con su disposición a cumplir una norma por la norma misma. Más abajo desarrollo el argumento de que dicha interferencia sobreviene a causa de la implementación del castigo como un descuento monetario, que tiene como consecuencia una modificación en el modo de transacción apropiado.

Evidencia empírica al respecto podemos encontrar en el experimento de campo que realizaron Gneezy y Rustichini en guarderías privadas (*day-care centers*) de la ciudad de Haifa (Israel). Los autores mencionan que antes del estudio no existían multas para los padres que llegaran tarde a buscar a sus hijos, pero que además era poco frecuente que eso suceda. En la mitad de las guarderías implementaron un tratamiento que consistía en una multa por arribar tarde. El resultado paradójico que obtuvieron consistió en que en las guarderías en las que se implementó la multa, el promedio de padres que llegó tarde a buscar a sus hijos se incrementó notablemente (alrededor del doble) en comparación con el tratamiento control (la otra mitad de guarderías en las que no se implementó ninguna multa). Lo que muestra el experimento es que el castigo, en lugar de disuadir la conducta objeto de reproche, puede de manera no intencionada redundar en consecuencias opuestas a las buscadas. Los autores sugieren que la explicación hay que buscarla en la idea de que "the fine changes the agents' perception of the social situation in which they are involved" (Gneezy y Rustichini 2000:10).

Relacionado con esta interpretación, los autores también mencionan explícitamente una explicación del fenómeno basada en normas sociales, ya que la introducción de la sanción monetaria transformaría una acción que no se considera una mercancía (puesto que el maestro cuida del niño después de hora porque es una buena persona o tiene buena voluntad o porque es su deber, y eso no forma parte de una relación mercantil), en una mercancía, esto es, la acción de cuidar al niño después de hora tiene un precio, y por ende puedo comprar tanto de esa mercancía como quiera. Es evidente que un aumento en la conducta

objeto de reproche a partir de la introducción de la sanción iría en contra de la predicción de la hipótesis de la función expresiva de las sanciones. Sin embargo, la posibilidad de pagar por el cuidado fuera de hora transforma la acción en una mercancía, y ofrece una excusa a quien paga para transgredir la norma sin sentir culpa o vergüenza. En este caso puntual, la imposición de un "precio" impide que el castigo funcione expresando una norma, y más bien obstruye esa función al deslocalizar la acción de un modo de transacción a otro transformándola en mercancía. Los modos de transacción se refieren al conjunto de normas y reglas que forman el entramado normativo de una esfera de acción. En palabras de Andvig: "A transactional mode (or mode of micro-coordination) specifies a set of rules for the engagement between at least two persons, a decision-making, information, and motivational structure guiding the actions of the agents operating in that mode" (Andvig 2006:329).

El argumento que defiendo es que las sanciones deben tener un significado normativo para ser efectivamente consideradas "normativas". Por esta razón, no pueden ser reducidas a un precio. Las sanciones no solo cambian la estructura de incentivos materiales, esto es, monetarios, sino que cambian los significados asociados a ciertas acciones. Esto quiere decir que el castigo tiene que cambiar el estatus normativo del sancionado, asignándole un significado distinto. Es común asociar el castigo a la privación de un derecho (de la libertad o de los propios bienes). En ausencia de una reversión en el estatus del castigado, es difícil que la acción que motivó el castigo adquiera un nuevo significado a la luz de la sanción. Ya no se trata de una acción "no permitida" o "proscripta", sino meramente una acción que puede realizarse a un cierto costo. En el caso de que la sanción efectivamente se interprete como un precio, entonces el rol que el experimentador pretende otorgarle al castigo se vería tergiversado por la estructura de incentivos. Mientras que el castigo, siguiendo la tesis de la función expresiva de las sanciones, señala la presencia de una norma, en los casos en los que carece de significado normativo puede obrar de manera contraria, esto es, modificando la lógica o el modo de transacción apropiado para tal interacción, como en el caso comentado anteriormente de las guarderías. Es evidente que hay ciertas transgresiones para las cuales no puede convenirse que el modo apropiado para castigar a quien las haya cometido sea el monetario (e. g., el homicidio).

El razonamiento anterior recuerda los argumentos contra los mercados repugnantes de Sandel (*What money can't buy: The moral limits of markets*), de acuerdo con los cuales ciertos bienes no pueden transformarse en bienes transaccionales porque repugnan a nuestra moral (por ejemplo, no pueden existir mercados de personas, de órganos, etc.). En el mismo sentido, hay transgresiones de normas para las cuales el castigo apropiado no puede ser el monetario. Si es posible "expiar" una transgresión normativa en un juego económico experimental por medio de una sanción monetaria, lo que ocurre es una modificación del modo de transacción apropiado, es decir, del entramado normativo que debería servir de telón de fondo para comprender "normativamente" las acciones de los jugadores. La expresión del castigo como un descuento de dinero es un artefacto experimental que no guarda relación con lo que sucede en la vida real y, por lo tanto, no es realista en el sentido del realismo mundano; pero más importante aún, he argumentado que la expresión del castigo a través de una multa monetaria tiene consecuencias para la manera en que los participantes encuadran la situación desde un punto de vista normativo y, por lo tanto, el castigo monetario no es psicológicamente realista como expresión del modo en que las comunidades enfrentan las transgresiones normativas.

Otro aspecto que contribuye a que las sanciones modifiquen los significados de las acciones consiste en que el castigo tiene que seguir reglas públicamente estipuladas (en lo posible con el acuerdo de los participantes en la práctica social relevante): tienen que ser realizados por quienes legítimamente están autorizados para ejercer el castigo y tiene que estar regulado por reglas de proporcionalidad entre la

ofensa y el castigo. Estas son condiciones que todos los sistemas de sanciones deberían estar en condiciones de cumplir. A diferencia de lo que sucede en la vida real (al menos lo que debería suceder), no es difícil observar, sin embargo, que estas condiciones raramente se cumplen en los experimentos de castigo. Se me podrá objetar que en los experimentos económicos las reglas son información pública y que, en ciertas circunstancias, *en promedio*, se ha encontrado proporcionalidad entre la ofensa y el castigo, ya que la diferencia entre la contribución del castigado y la del promedio del grupo suele ser la variable explicativa más importante (ver Henrich *et al. Costly punishment across human societies*).

Sobre el primer punto se puede decir que, si bien las reglas son públicas, eso no implica que los roles de castigador y castigado se asignen de acuerdo con reglas que todos consideren legítimas. De hecho, la investigación muestra que la percepción de legitimidad asociada a las diferencias en los roles es importante a la hora de juzgar quién está justificado legítimamente para ejercer el castigo. El castigo está regulado por reglas restrictivas, en la medida en que las personas tienden a considerar apropiado el castigo cuando quien lo ejerce tiene algún tipo de autoridad. De acuerdo con Chaurand y Brauer algunas personas "feel that being citizens who pay local taxes, who regularly use the public property that is being degraded by an uncivil behavior, and who personally suffer the consequences from the behavior gives them the legitimacy to express their disapproval verbally to the perpetrator of the uncivil behavior. Other people may disapprove of an uncivil behavior but, at the same time, not feel that they have legitimacy to exert social control. They may think that only figures of authority (e.g., police officers, guards) should sanction the perpetrators of uncivil behaviors. As such, it is not surprising that the feeling of legitimacy is a strong predictor of people's tendency to exert social control" (Chaurand y Brauer 2008:1709). En los experimentos habitualmente los roles (como el de castigador) no se confieren de acuerdo con procedimientos o reglas que les confieran legitimidad. Puede conjeturarse que este déficit normativo podría afectar negativamente la disposición a castigar o la percepción de legitimidad de este. Algunos trabajos han empezado a tener conciencia de este problema y, en consecuencia, han empezado a evaluar instituciones de castigo endógenamente determinadas (ver próxima sección).

Respecto de la proporcionalidad entre ofensa y castigo, algunos estudios sugieren que sí la hay, pero esa proporcionalidad solo puede evaluarse *en promedio*. Lo que resulta ecológicamente válido es que el castigo sea proporcional respecto de cada ofensa y de cada uno de los transgresores. De lo contrario, puede darse el absurdo de que un participante sea castigado muy duramente, mientras que otro lo sea muy levemente, por la misma ofensa y, por lo tanto, que *en promedio* el castigo sea proporcional a la ofensa. Claramente no es esto lo que tenemos en mente cuando pensamos en la proporcionalidad del castigo. Esto puede crear confusión entre los participantes. Desde el punto filosófico el castigo incluye un supuesto epistémico, que consiste en que quien castiga conoce la culpabilidad del castigado y además conoce la cantidad de castigo que merece. No poder ajustarse a estos requisitos hace que el castigo en juegos experimentales pueda ser percibido por los participantes como arbitrario. Este aspecto se conecta con la forma en que los participantes evalúan la legitimidad del acto de sancionar.

5. Percepción de legitimidad del castigo

Las personas perciben el castigo como legítimo si ayuda a proteger los valores y normas que creen válido resguardar. Por legitimidad se entiende la conformidad o consentimiento que exhiben las personas hacia quienes ejercen la autoridad. Desde el punto de vista psicológico, la legitimidad es una "property of an authority, institution, or social arrangement that leads those connected to it to believe that it is appropriate, proper, and just" (Tyler 2006:375). Por ende, la mejor manera de que se considere un sistema de sanciones como legítimo es que se gobierne de acuerdo con la moralidad predominante de la mayoría.

Y si bien esto puede ser posible en muchas ocasiones, no lo es claramente en todas, ya que las sanciones habitualmente cumplen un rol de reforma social. Por ejemplo, si la moralidad prevalente en una determinada sociedad estuviera caracterizada por niveles de corrupción rampantes, entonces las sanciones buscarían modificar el *statu quo*, no garantizar su continuidad. Las distintas formas de sanción surgen, precisamente, con el objetivo de modificar estados de cosas, o para evitar su deterioro.

La legitimación del castigo es sumamente importante, ya que si los castigados lo perciben como injusto o ilegítimo, podrían actuar en represalia. De hecho, la evidencia experimental sugiere que quienes son castigados tienden a tomar represalias, castigando a su vez a quienes los castigaron, lo que tiene como resultado menores niveles de cooperación. A su vez, los niveles de castigo anti-social correlacionan negativamente con medidas de confianza en las instituciones, lo que sugiere que a medida que la percepción de legitimidad del castigo disminuye, aumentan los niveles de castigo anti-social.

Quisiera señalar dos aspectos que dificultan que los participantes puedan percibir el castigo como legítimo: por un lado 1) la ausencia de corrección formal o procedimental del castigo; por otro, 2) la ambigüedad normativa, producto de la dificultad de establecer normas claras. Respecto del primer punto seré breve, ya que la investigación experimental ha sido consciente de este problema y ha estudiado formas "restrictivas" de castigo. Preferentemente en JBP hay una línea de investigación que se ha ocupado de diferentes formas de implementar el castigo. En estos estudios los participantes pueden escoger entre diferentes instituciones de castigo. No es extraño que dichos experimentos muestren una diferencia respecto de si la institución punitiva es electa (esto es, endógenamente determinada, lo que otorgaría a la institución del castigo mayor legitimidad) o, si por el contrario, es impuesta exógenamente por los experimentadores (el lector interesado puede encontrar un ejemplo, así como una revisión de la literatura en Markussen, Putterman y Tyran. *Self-organization for collective action: an experimental study of voting on sanction regimes*). Estos experimentos suelen reportar que la eficiencia del castigo suele aumentar cuando los miembros del grupo tienen la posibilidad de votar para elegir la institución de castigo. Si la legitimidad puede provenir del procedimiento mismo o, si por el contrario, el procedimiento necesita, a su vez, una instancia de legitimación ulterior, es una pregunta empírica para la que aún no hay una respuesta clara. Según algunas concepciones positivistas del derecho, basta con que la norma sea promulgada por una autoridad competente para que sea legítima. En esta visión, de tipo decisionista, la legitimidad se resuelve en hacerla colapsar con la legalidad formal, con los consabidos problemas respecto de la noción de justicia que ello puede acarrear. Basta recordar para ello la famosa fórmula de Radbruch, según la cual el derecho injusto no puede ser legal. Dejando de lado esta cuestión, que atañe exclusivamente a la institución del castigo formal (a través de una autoridad previamente establecida), el castigo informal corre el riesgo de ser percibido como ilegítimo por otras razones además de su promulgación.

Quisiera a continuación llamar la atención sobre el segundo punto, el de la ambigüedad normativa, al que se le ha prestado escasa atención. Hay tres formas básicas en las que puede darse ambigüedad normativa y que pueden referirse a: 1) la norma subyacente que motiva el castigo; 2) el conflicto entre la frecuencia estadística del castigo y la prescripción normativa y, por último, 3) la distinción entre comportamiento desviado e incorrecto.

Veamos el primer tipo de ambigüedad, concerniente a la norma que puede motivar la sanción. En el caso de que la norma subyacente sea la equidad contributiva, una heurística plausible indicaría que debiera castigarse a quien contribuye por debajo del promedio grupal (en JBP). Nótese que en el clásico trabajo de Fehr y Gächter (2000), el castigo logra estabilizar la norma de cooperación, pero lo hace a costa de la

eficiencia. El resultado 8 de Fehr y Gächter evidencia este hecho, ya que, a pesar de generar una estabilización de la cooperación, los tratamientos con castigo (considerando solo los tratamientos *stranger*) tienen como efecto no deseado una reducción del agregado de beneficios del grupo, una tendencia que solo en las últimas dos rondas del experimento logra revertirse. Con razón, los sujetos experimentales podrían no tener en mente una norma de equidad, sino de eficiencia, lo que ayudaría a explicar el hecho de que algunos sujetos decidan no castigar, o estén incómodos con esa institución. En efecto, el castigo parece contravenir una norma de eficiencia.

Respecto del segundo tipo de ambigüedad, esto es, el conflicto entre la frecuencia estadística del castigo y la prescripción normativa, en muchos estudios empíricos suelen estar confundidas la legitimidad de una norma con su frecuencia, es decir, el nivel de aceptación del que goza la norma prescriptiva y su frecuencia estadística. Esto responde a una diferencia entre la información descriptiva y la prescriptiva. Mientras que la primera nos informa acerca del nivel de ocurrencia de una conducta, es decir, de su distribución empírica; la segunda nos informa acerca de las conductas que los demás aprueban o consideran legítimas. Muchas veces una conducta puede ser común (pensemos en la corrupción), aun cuando las personas piensen que no es apropiada, lo que indica que la información empírica no está en condiciones, por sí misma, de comunicar la presencia de una norma. Para ello es necesario introducir información prescriptiva. Estas diferencias pueden reproducirse en un juego experimental.

Por ejemplo, en un JBP puede darse el siguiente caso: uno de los participantes del grupo puede iniciar el castigo unilateralmente, y que los otros participantes no lo sigan. Ello puede dar lugar a la percepción de él mismo y de parte de los otros miembros del grupo de que si solo uno de ellos ha utilizado el castigo (mientras que no lo ha hecho así el resto), es porque no existe un consenso respecto de que el comportamiento objeto de consideración deba ser castigado. Nótese que este razonamiento se apoya sobre información descriptiva, que *lejos de facilitar la legitimación del castigo, puede ser fuente de ambigüedad, y servir de excusa para justificar una conducta egoísta o injusta.* También puede darse el caso contrario, y que al castigo iniciado por un participante se sumen otros participantes, pero no necesariamente porque consideren legítimo el castigo, sino simplemente en función de una imitación: el tipo de razonamiento que podría sustentar dicho comportamiento sería el siguiente: “si la mayoría decide castigar, entonces es correcto hacerlo”, y funcionaría a modo de una heurística de la prueba social.

De acuerdo con Kuran, la esencia de la heurística de la prueba social radica en que “we believe an explanation, assertion, prediction, or evaluation because most others do. (…) the fact that our perceptions are shared assures us of their correctness” (Kuran 1998:163). En efecto, existe evidencia experimental que sugiere que la percepción de legitimidad de una norma puede estar afectada por la frecuencia del comportamiento. De hecho, como afirman Lindström *et al.* tenemos una “tendencia a inferir el valor moral de un comportamiento social a partir de su frecuencia relativa” (2018:3), lo que para los autores constituye una heurística que asigna valor prescriptivo a las acciones frecuentes. Este último fenómeno podría explicar que en muchas ocasiones el castigo pueda emerger en casos en los que no es beneficioso para el grupo, sino incluso cuando va en detrimento de este, esto es, para hacer cumplir normas ineficientes, o sencillamente destructivas, como han mostrado recientemente Abbink *et al*. (*Peer punishment promotes enforcement of bad social norms*).

La tercera fuente de ambigüedad normativa que quisiera mencionar proviene de la dificultad para distinguir entre un comportamiento incorrecto (esto es, socialmente inapropiado) y un comportamiento desviado. Diremos que un comportamiento puede ser desviado sin que sea percibido, necesariamente, como incorrecto. Por ejemplo, es claro que en un JBP no existe *a priori* una conducta que pueda ser

considerada incorrecta. El conjunto de acciones disponibles a los individuos consiste en contribuir dinero a una cuenta grupal. Típicamente se considera la norma como el monto promedio de contribución en el grupo. De hecho, hay evidencia experimental que sugiere que cuanto mayor es la desviación del monto promedio de contribuciones, mayor es la sanción que recibe el jugador. Pero los comportamientos que se desvían de la norma lo hacen no de manera absoluta, sino por grados. Por ende, la norma (si existe) dependerá de las características idiosincráticas de cada grupo particular.

La percepción de ilegitimidad del castigo puede impedir su función expresiva. ¿Cómo afectaría ello la validez ecológica? Si los experimentos pretenden recrear en el laboratorio los mismos procesos psicológicos que pretendidamente se dan en la realidad, entonces, deben procurar recrear la conexión entre castigo y normas. No hace falta abundar en esto, pero la vida cotidiana, a diferencia del contexto aséptico del laboratorio, está repleta de claves contextuales que nos brindan información normativa acerca de las conductas apropiadas o inapropiadas.

Primera conclusión

En esta parte del trabajo me propuse evaluar la hipótesis acerca de que la transgresión normativa es causa de la aplicación de sanciones. Luego de pasar revista a distintas formas de implementación del castigo en experimentos económicos, sostuve que hay una serie de desafíos que dificultan la robustez de dicha hipótesis y que, por ende, socavan la validez ecológica de los experimentos. Esos desafíos son la irrelevancia normativa y la ilegitimidad del castigo, acerca de los cuales desarrollé de qué forma se hacen presentes en los experimentos, y cuáles son sus consecuencias para su validez ecológica.

En la parte II de este trabajo intento brindar una solución a los obstáculos que se le plantean a la utilización del castigo como dispositivo experimental, a partir del cual aprender acerca del funcionamiento de los mecanismos psicológicos que intervienen en el cumplimiento de normas, en su transgresión y, eventualmente, en su sanción.

Bibliografía

Andvig, J. (2006). Corruption and fast change. *World Development 34*(2): 328-340. https://doi.org/10.1016/j.worlddev.2005.03.007

Bicchieri, C. (2006). *The grammar of society*. Cambridge University Press. https://doi.org/10.1017/cbo9780511616037

Brewer, M.B. (2000). Research design and issues of validity. H.T. Reis, C.M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 3-16). Cambridge University Press.

Chaurand, N., Brauer, M. (2008). What determines social control? People's reactions to counternormative behaviors in urban environments. *Journal of Applied Social Psychology 38*(7): 1689-1715. https://doi.org/10.1111/j.1559-1816.2008.00365.x

Fehr, E., Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review 90*(4): 980-994. https://doi.org/10.1257/aer.90.4.980

Gneezy, U., Rustichini, A. (2000). A fine is a price. *The Journal of Legal Studies 29*(1): 1-17. https://doi.org/10.1086/468061

Guala, F. (2012). Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences 35*(1): 1-15. https://doi.org/10.1017/s0140525x11000069

Kuran, T. 1998. Social mechanisms of dissonance reduction. P. Hedström, R. Swedberg (Eds.), *Social mechanisms: an analytical approach to social theory* (pp. 147-171). Cambridge University Press. https://doi.org/10.1017/cbo9780511663901.007

Lindström, B., et. al. (2018). The role of a "common is moral" heuristic in the stability and change of moral norms. *Journal of Experimental Psychology: General 147*(2): 228-242. https://doi.org/10.1037/xge0000365

Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of Economic Perspectives 14* (3): 137-158. https://doi.org/10.1257/jep.14.3.137

Tyler, T.R. (2006). Psychological perspectives on legitimacy and legitimation. *Annual Review of Psychology 57*(1): 375-400. https://doi.org/10.1146/annurev.psych.57.102904.190038

Recibido el 1 May 2019
Aceptado el 24 Ago 2019