https://doi.org/10.34024/prometeica.2021.23.11319

REVISITANDO CRÍTICAMENTE NOCIONES DE SIMILITUD EN LA TEORÍA DE JUEGOS COMPORTAMENTALES[1]

REVISITING CRITICALLY SIMILITUDE NOTIONS IN BEHAVIORAL GAME THEORY

REVENDO CRITICAMENTE NOÇÕES DE SIMILARIDADE NA TEORIA DOS JOGOS COMPORTAMENTAIS

Carlos Maximiliano Senci
(Instituto de Investigaciones Económicas y Sociales del Sur
Consejo Nacional de Investigaciones Científicas y Técnicas
Universidad Nacional del Sur
Universidad Salesiana - Bahía Blanca)
maxisency@msn.com

Recibido: 22/10/2020
Aprobado: 12/07/2021

RESUMEN

El objetivo de este trabajo consiste en revisar críticamente diferentes concepciones de similitud en la teoría de juegos que se encuentran presupuestas en diversos conceptos de solución. Pasamos revista a cuatro concepciones: 1) similitud empírica, 2) la doctrina Harsanyi-Aumann, 3) Superracionalidad, y 4) pensamiento mágico. Evaluamos críticamente las diferentes nociones en función de la similitud que presuponen, la viabilidad para resolver problemas de coordinación y/o cooperación basados en ellas, y su fundamentación racional.

Palabras clave: teoría de juegos. similitud. superracionalidad. razonamiento en equipo. pensamiento mágico.

ABSTRACT

This work aims to critically revise different conceptions of similarity in game theory that are presupposed in various solution concepts. We review four conceptions: 1) empirical similarity, 2) the Harsanyi-Aumann doctrine, 3) Superrationality, and 4) magical thinking. We critically evaluate the different notions based on the similarity they presuppose, the viability of solving coordination and cooperation problems based on those notions, and their rational foundation.

Keywords: game theory. similarity. superrationality. team reasoning. magical thinking.

RESUMO

O objetivo deste trabalho é revisar criticamente diferentes concepções de similaridade na teoria dos jogos que são pressupostas em vários conceitos de solução. Revisamos quatro concepções: 1) similaridade empírica, 2) a doutrina Harsanyi-Aumann, 3) Superracionalidade e 4) pensamento mágico. Avaliamos criticamente as diferentes noções

[1] Agradezco los comentarios y sugerencias de Fernando Tohmé a una versión previa de este trabajo.

com base na semelhança que pressupõem, na viabilidade de resolução dos problemas de coordenação e cooperação a partir delas e no seu fundamento racional.

Palavras-chave: teoria dos jogos. similaridade. superracionalidade. raciocínio de equipe. pensamento mágico.

Introducción

Un ámbito de aplicación crecientemente extendido de la teoría de juegos (TJ a partir de ahora) consiste en el estudio de las propiedades de la interacción social de agentes intencionales. Pero hay que tener en cuenta que la TJ no es una teoría social, sino matemática. Como dice Ross (2008: 60):

> La teoría de juegos *per se* no es una rama de las ciencias sociales, sino una rama de la matemática: es una teoría de la maximización simultánea de ciertas clases de funciones en objetos abstractos llamados "campos de preferencia" cuando las funciones en cuestión deben tomar recíprocamente los outputs de los demás entre sus inputs. El "agente racional" aquí es un tipo técnico, no un tipo empíricamente determinado.

Como herramienta formal la TJ es útil para modelar la interacción de dos o más agentes, y es muy útil por la capacidad de ofrecer predicciones precisas acerca del comportamiento de dichos agentes, en el sentido de que las decisiones simultáneas de los jugadores unívocamente determina un resultado o un conjunto de resultados.[2] Las predicciones pueden ser precisas justamente porque se basan únicamente en la estructura de pagos de los juegos, especificada por las funciones de utilidad de los agentes. Como consecuencia de que los agentes que modela la TJ son altamente idealizados, las predicciones pueden no ser descriptivamente adecuadas. En este sentido la TJ es normativa, ya que no está interesada en los procesos de toma de decisiones tal como en realidad se llevan a cabo por seres humanos de carne y hueso. Las predicciones son capturadas por distintos conceptos de solución, esto es, reglas que prescriben lo que deben hacer los agentes para que los resultados que pretenden conseguir cumplan ciertos criterios, entre ellos criterios de racionalidad, de eficiencia, entre otros. Los conceptos de solución asignan una combinación de perfiles de estrategia, i. e., las soluciones, a cada juego estratégico finito.

El concepto de solución más extensamente usado en la TJ no-cooperativa es el concepto de equilibrio de Nash (EqN a partir de ahora; Nash, 1950; Myerson 1991: 105). Informalmente, un EqN es una tupla de estrategias, tales que ninguno de los jugadores tiene incentivos en modificar su estrategia, dadas las estrategias de los otros jugadores. Dicho de otra manera, cada jugador tiene conjeturas correctas acerca de las acciones de los otros, y actúa racionalmente, esto es, adopta su mejor estrategia (dadas esas conjeturas) Desde los trabajos pioneros de Aumann sabemos que las condiciones epistémicas que requiere el EqN son sumamente demandantes. En particular, en el caso de dos jugadores la condición epistémica de equilibrio demanda que haya conocimiento mutuo (cada uno sabe lo que sabe el otro) de la racionalidad de los jugadores (i. e., los jugadores son similares en cuanto a su racionalidad), de los pagos, y de las conjeturas (Aumann y Brandenburger, 1995). En el caso de equilibrios en estrategias puras esto significa que hay conocimiento mutuo acerca de qué acciones van a elegir los otros.

En muchas situaciones ubicuas el EqN no provee de soluciones satisfactorias en relación con al menos algún criterio. El caso típico es el Dilema del Prisionero, en el que no pueden satisfacerse al mismo tiempo un criterio de racionalidad individual (entendida como maximización individual de los pagos) y un criterio de eficiencia colectivo. El EqN tampoco parece brindar soluciones aceptables en otra clase de interacciones, e. g., juegos de coordinación como la Caza del Ciervo (*Stag Hunt*) o el Hi-Lo. En estos juegos hay equilibrios que son de recompensa dominante[3] (*payoff-dominant*) y sin embargo el EqN no brinda las herramientas para discriminar entre estos equilibrios y aquellos que no lo son. Estos juegos son instancias de una amplia clase de juegos en los que existen múltiples EqN. En estos casos el EqN no

[2] Para ser más precisos, estoy excluyendo juegos en los que puede introducirse un factor exógeno de incertidumbre en los que los resultados pueden depender de procesos aleatorios (Myerson 1991, capítulo 2).
[3] Un resultado es de recompensa dominante si todos los jugadores obtienen una recompensa mayor que la que obtendrían en cualquier otro resultado.

provee de predicciones precisas, a lo sumo puede ofrecer clases de soluciones. Esta insatisfacción respecto de las soluciones ofrecidas ha generado un programa de refinamientos al EqN (Harsanyi y Selten, 1988). La TJ comportamental también puede verse como una forma de refinamiento del abordaje estándar en la TJ, aunque no necesariamente de sus conceptos de solución (Ross, 2019), cuyo objetivo es proponer conceptos de solución no-ortodoxos, i. e., que modifican o adicionan supuestos y restricciones impuestos por el EqN. En este artículo me enfoco en un conjunto de nociones alternativas que van más allá de la visión clásica apartándose del EqN.

Una intuición empírica que se puede adicionar a la estructura formal de los juegos consiste en la similitud entre los jugadores. A veces la similitud puede referirse a aspectos normativos o formales, tales como la racionalidad. En otras ocasiones los aspectos relevantes pueden ser meramente empíricos y no estar referidos a aspectos normativos. Como señala Morton (1994: 22):

> …los agentes humanos se reconocen claramente entre sí como similares: en la familia, en el entorno social o simplemente como especie. Y este reconocimiento juega un papel en sus interpretaciones de la acción de los demás y sus propensiones a la cooperación. (p. 22; mi traducción; todas las traducciones son propias a menos que se indique lo contrario).

También ciertas nociones de similitud han permeado en la literatura comportamental, en general como reelaboraciones de nociones previas que pueden encontrarse en la psicología social. La intuición básica consiste en que tiene sentido para un jugador reducir la incertidumbre que experimenta cuando enfrenta a otros jugadores considerando la similitud con los demás. Este fenómeno conductual ha sido documentado en una serie de trabajos recientes, que estudiaron el efecto de formas diferentes de *priming* de similitud en juegos de coordinación y/o cooperación (ej., Fischer, 2009; Di Guida y Devetag, 2013; Mussweiler y Ockenfels, 2013; Chierchia y Coricelli, 2015; Rubinstein y Salant, 2016).

Hay diferentes formas por medio de las cuales los teóricos de juegos han intentado capturar diversos sentidos a partir de los cuales pueden pensarse relaciones de similitud entre jugadores. En lo que sigue voy a esbozar una taxonomía, que no pretende ser excluyente, acerca de diferentes formas de pensar la similitud entre jugadores. En primer lugar, lo que llamaré la *concepción inductiva o empírica de la similitud*, en la que discutiré dos abordajes principales: la teoría de puntos focales y el pensamiento en equipo (*team reasoning*).

En segundo lugar, abordaré lo que se ha llamado la *doctrina Harsanyi*. En tercer y cuarto lugar me referiré a soluciones que no han tenido tanta difusión en la literatura económica, pero que especialmente la segunda de ellas ha tenido relevancia en el ámbito filosófico: me refiero a la noción de *superracionalidad* y a la de *pensamiento mágico*.

Similitud empírica en el razonamiento entre los seres humanos

En su reciente libro *Understanding Institutions*, Francesco Guala (2016) profundiza en lo que él llama el problema de la lectura mental (*mind reading*) en los juegos de coordinación. La coordinación depende de que las personas tengan una "teoría de la mente", es decir, la capacidad de atribuir estados mentales a los demás. Aunque su forma de enmarcar el problema es de naturaleza lewisiana, Guala rechaza la conocida solución de Lewis, y aplica otros dos conceptos de solución al problema de la lectura mental: el enfoque del pensamiento de solución (*solution thinking*) de Morton (Morton 1994), así como el enfoque de razonamiento en equipo (RE a partir de ahora) desarrollado por Bacharach, Sugden y Gold (Bacharach 1999, 2006; Gold, 2012; Gold y Sugden 2007; Sugden, 2000, 2003). A pesar de sus similitudes, el RE y el pensamiento de solución poseen un enfoque distinto del de la TJ ortodoxa. Este enfoque se basa en la TJ psicológica (Geanakoplos, Pearce y Stacchetti, 1989), que modifica "la teoría ortodoxa mediante la introducción de principios formales de razonamiento que pueden ayudar a explicar las observaciones empíricas y las intuiciones ampliamente compartidas que la teoría ortodoxa deja sin explicar" (ver Colman, 2003: 149-150), mientras que la solución de Lewis y el enfoque del punto focal se basan en los equilibrios de Nash derivados de creencias mutuamente consistentes. Lo que tienen estos

enfoques en común es que proponen una forma de razonamiento que modifica la forma de razonamiento estándar e introducen principios formales cuyo objetivo es explicar regularidades empíricas que en muchos casos se presentan como anomalías para la TJ estándar. Es importante tener en cuenta que en última instancia la justificación acerca de las soluciones que pueden brindar a diferentes problemas como el de coordinación, o cooperación, es una justificación empírico-inductiva que apela a la experiencia individual y a la psicología.

En las teorías de coordinación de Schelling, Lewis, Bacharach y Sugden se hacen varias suposiciones sobre las características del razonamiento que son comunes a (y conocimiento común dentro de) poblaciones relevantes de potenciales jugadores. Estas son *regularidades empíricas*, basadas en características comunes de la psicología humana y la experiencia compartida de los jugadores en la vida humana. Por lo tanto, puede haber situaciones en las que sea razonable que un jugador asuma que el razonamiento de los demás será similar al suyo.

Schelling abordó el problema de coordinación tratándolo como un problema de predicción. La posibilidad de la coordinación requiere que primero sea posible coordinar las predicciones, esto es, "leer el mismo mensaje en una situación común, identificar el (único) curso de acción en el que las expectativas mutuas puedan converger" (Schelling, 1981: 64). Cualquier inferencia que funcione deberá contener varias creencias de orden superior. Para evitar una regresión al infinito, el proceso de razonamiento debe terminar en alguna etapa intermedia. En muchas ocasiones existen señales que coordinan las expectativas de todas las partes. De acuerdo con Schelling, los elementos que pueden funcionar como señales de ese tipo pueden ser muy diversos, siempre que sean lo suficientemente salientes (*salient*) para los jugadores, es decir, que tengan el tipo adecuado de prominencia o notoriedad. Nótese que las señales son irrelevantes desde el punto de vista de las recompensas y pueden "depender de analogía, precedente, disposición accidental, simetría, configuración estética o geométrica, razonamiento casuístico, o de quiénes son las partes y qué saben unos de otros" (Schelling, 1981, 57). Basado en esta idea, Schelling elaboró su solución de *punto focal* para el problema de coordinación.

Para cumplir su tarea los puntos focales deben satisfacer una propiedad de unicidad. De nada nos sirven puntos focales que no nos permitan discriminar entre diferentes posibles soluciones. Una de las formas de satisfacer este requisito consiste en que los puntos focales se destaquen *para todos los individuos por igual*, es decir, tienen que ser salientes para todos los jugadores que participan de la interacción. Para ello es evidente, a su vez, que los individuos tienen que ser lo *suficientemente similares* como para extraer la misma información de las señales: "los sujetos deberán ser similares en la importancia relativa que otorgan a diferentes aspectos de comparación; pero a menudo son suficientemente similares para resolver el problema" (Lewis, 1969: 35). Sea cual sea la forma de explicar de qué modo los puntos focales permiten la coordinación, lo cierto es que a la base de estas nociones hallamos el supuesto de similitud de los jugadores.

La teoría del RE de Bacharach y Sugden ha sido propuesta como una solución alternativa a ciertos problemas de coordinación. Según Bacharach, el RE es un modo de razonamiento práctico (volveremos más adelante sobre esto) motivado por una clase de juegos que tienen margen para una recompensa común (*scope for common gain*), entendida como la posibilidad de una mejora de Pareto. En esta clase de juegos encontramos puzles de cooperación, como la paradoja de Hi-Lo (Schelling, 1981, p. 291) y el ya mencionado Dilema del Prisionero, para los cuales la teoría de la decisión estándar ofrece soluciones que son cuestionables en términos de su racionalidad.

En palabras de Bacharach: "alguien 'razona en equipo' si encuentra la mejor combinación de acciones posible para todos los miembros de su equipo, y entonces hace su parte en ello" (Bacharach 2006: 121). El RE equivale a pensar en uno mismo como miembro de un grupo, y derivar las implicancias prácticas de ello. Al permitir que se consideren como agentes unidades colectivas, el RE tuerce la sintaxis, por así decirlo, de los juegos estratégicos y permite a los jugadores pensar en sí mismos como miembros de un equipo y adoptar las preferencias del equipo como propias. Por lo tanto, la clave para entender el RE es reemplazar al agente individual por un equipo de razonadores. En este sentido, el RE es una forma de

enmarcar (*frame*) la interacción con otros agentes. Bacharach llama a esto "transformación agencial" (Bacharach, 2006). La idea básica es que al identificarnos con un grupo enmarcamos nuestras propias preferencias de una manera grupal. Como consecuencia de esta transformación, un individuo que se identifica con un grupo "se considera intercambiable con otros miembros del mismo grupo. Percibe esta equivalencia aunque sabe que es 'numéricamente distinto'" (Bacharach 2006: 73). Una traducción directa de esta propiedad es la condición de simetría. Los razonadores de equipo deberían satisfacer una propiedad de invariancia respecto de la simetría del juego: esto es, al rotar la matriz sobre la diagonal los jugadores deberían obtener los mismos resultados en un juego en forma normal.

Gold y Sugden (2007) han propuesto analizar las intenciones del grupo como resultado de un modo específico de razonamiento práctico, instanciado por el RE. El elemento básico es el concepto de *esquema de razonamiento práctico*, inspirado en el silogismo práctico de Aristóteles. Se define la validez de un esquema como una propiedad que asegura la consecución del éxito, es decir, maximiza los pagos (entendidos como índices de utilidad).

Veamos cómo funciona el esquema aplicado al juego Hi-Lo (Tabla 1):

Ejemplo 1: Hi-Lo

Tabla 1. Matriz de pagos del juego Hi-Lo

		Jugador 2	
		H	L
Jugador 1	H	2, 2	0, 0
	L	0, 0	1, 1

El juego Hi-Lo plantea un problema de coordinación ya que los jugadores prefieren coincidir en sus estrategias a no coincidir. Sin embargo, la teoría ortodoxa es estéril para proporcionar un resultado determinado, ya que en realidad hay dos Equilibrios de Nash, en los que ambos jugadores eligen la misma estrategia. Sin embargo, uno de los equilibrios (H, H) es de recompensa dominante, y por lo tanto parece trivial que los jugadores deben coordinarse en ese equilibrio. Esta no es, sin embargo, la manera en que la teoría ortodoxa resuelve este problema. De hecho, dado que el jugador 1 elige L, la mejor respuesta del otro jugador consistiría en elegir la misma estrategia. Y por lo tanto, (L, L) se convierte en un equilibrio de Nash. Parece insatisfactorio desde un punto de vista normativo y tampoco está a la altura de proporcionar un análisis descriptivo adecuado capaz de explicar cómo los individuos se coordinan exitosamente en juegos como este, como lo demuestran los resultados experimentales (ver Colman *et al*. 2008; Bardsley *et al*. 2010).

Consideremos el esquema práctico de racionalidad colectiva de Gold y Sugden:

1. Debemos elegir entre (H, H), (H, L), (L, H), o (L, L)
2. Si elegimos (H, H) el resultado será R1
3. Si elegimos (H, L) el resultado será R2
4. Si elegimos (L, H) el resultado será R3
5. Si elegimos (L, L) el resultado será R4
6. Queremos conseguir R1 en lugar de R2, R3 o R4.
7. Debemos elegir (H, H).

El esquema ofrece una conclusión práctica: teniendo en cuenta las premisas, que incluyen proposiciones sobre los objetivos del grupo, debemos inferir como conclusión lo que debemos hacer para garantizar nuestro objetivo. En la TJ ortodoxa los individuos maximizan una función de utilidad individual. Por el

contrario, en el RE la unidad de agencia es colectiva, es un "nosotros", cuyo objetivo es maximizar una función de utilidad grupal. Un individuo no tiene razones para favorecer una u otra de las posibles estrategias en el juego si razona su camino hacia la conclusión aplicando preferencias estándar. Pero el objetivo práctico del agente colectivo debería consistir en maximizar el bienestar total del grupo.

La teoría del RE, tal como la proponen Bacharach y Sugden, no brinda una respuesta acerca de cuándo la racionalidad requiere de un jugador que adopte el RE (esta crítica ha sido formulada entre otros por Browne, 2018). Eso es correcto, pero la idea de que este es un problema presupone algo así como la doctrina Harsanyi (ver próxima sección), es decir, que cada juego tiene una solución racional única, que es trabajo del teórico encontrar o anunciar. Bacharach y Sugden argumentan explícitamente que esta doctrina está fundamentalmente injustificada. En relación con el RE, lo máximo que puede hacer una teoría de la racionalidad (instrumental) es decirles a los agentes qué hacer, una vez establecidos sus objetivos. No puede decir qué agentes (individuos o equipos) deberían existir racionalmente. Ser racional, desde el punto de vista de la TJ, no brinda un acceso epistémico a conocer o saber en qué situación se debería cambiar la forma de razonamiento y adoptar una preferencia de grupo. Bacharach y Sugden tratan la cuestión acerca del modo de razonamiento de los jugadores como una cuestión empírica, y proponen varias formas de responder esa pregunta en términos de la psicología y la experiencia de los jugadores (sobre esto ver Bacharach, 2006). En este punto las teorías de ambos difieren en un punto sustancial. Para Bacharach es racional decidir en favor del equipo aun cuando no lo sea individualmente. Para Sugden, en cambio, es racional jugar en equipo siempre y cuando se cumpla una condición de garantía (*assurance*) que brinde seguridad respecto de que los demás jugadores también «razonan en equipo». Es elocuente el ejemplo del dilema del prisionero. Mientras que la cooperación mutua es la mejor solución en términos del beneficio colectivo, desde el punto de vista estrictamente individual sigue siendo más conveniente no cooperar, independientemente de la elección del otro jugador. La elección de no-cooperar protege al jugador también en el caso de que el otro jugador elija no-cooperar. No hay ninguna razón instrumental válida que pueda persuadir a un jugador racional de que la estrategia que debe seguir es la cooperación. Por eso el RE se presenta a sí mismo como una teoría alternativa acerca de la racionalidad.

Bacharach introduce un parámetro w ($0 \leq w \leq 1$), que representa la probabilidad de que un individuo se identifique con un grupo, o en otras palabras, razone en equipo. En otras versiones de la teoría de Bacharach (Gold, 2012), w representa la probabilidad de que el RE tenga poder motivacional para un individuo. Lo distintivo de la teoría de Bacharach es que admite situaciones en las que $w < 1$, esto es, en las que un individuo no tiene garantías de que los demás razonen en equipo. Eso lleva a que la teoría de Bacharach admita que los perfiles de estrategia fuera de la diagonal (C, D) y (D, C) sean ponderados mejor que el equilibro de Nash (D, D) (ver ejemplo 2). Por lo tanto, si consideramos que la recompensa de equipo es la suma de las recompensas individuales, un jugador de equipo podría elegir cooperar aun cuando supiera con certeza que los otros jugadores no cooperarían ($w = 0$), ya que (C, D)= (D, C) = 7 > 2 = (D, D).

Ejemplo 2: Dilema del Prisionero

Tabla 2. Matriz en forma normal del Dilema del Prisionero.

		Jugador 2	
		C	D
Jugador 1	C	4, 4	0, 7
	D	7, 0	1, 1

Para Sugden, esta solución no es aceptable. En efecto, en trabajos recientes Sugden (2011, 2015) se aparta de la versión de Bacharach del RE. En lugar de comparar el resultado del RE con el resultado no-

cooperativo, para Sugden el punto de referencia es el mejor pago que podría obtener el individuo unilateralmente. Por eso Sugden no puede aceptar que los perfiles fuera de la diagonal principal puedan ser válidos para el RE, a menos que los jugadores cuenten con la suficiente garantía de que los demás también jugarán la opción cooperativa. De esta manera Sugden no se aparta drásticamente del concepto de racionalidad estándar.

Lo anterior pareciera implicar que la teoría de RE de Bacharach no presupone una noción de similitud, en la medida en que admite que la identificación con el grupo puede ser unilateral, incluso *contrario sensu* de lo que prescribiría la racionalidad instrumental. Muy por el contrario, las últimas versiones de RE, que debemos a Sugden, parecen reintroducir la idea subyacente de que los jugadores tienen que ser, en un sentido sustancial, similares: tienen que tener la garantía (o la creencia) de que los demás jugarán la misma estrategia con un alto grado de probabilidad (o tener la garantía de que jugarán la misma estrategia). ¿Cuán alta debe ser esa probabilidad para que un jugador coopere en un dilema del prisionero a sabiendas de que cooperar no es su mejor elección *independientemente* de lo que haga el otro jugador?

Este es un tema que excede, desde luego, los límites de este trabajo, aunque de manera provisoria podría decirse que no parece haber una respuesta genuina a este interrogante (para intentos de solución, e interrogantes ver Gold y Colman, 2018; Gold, 2018).

3. La «Doctrina Harsanyi» debe su nombre a Aumann (1976), quien la bautizó en referencia a los supuestos que incluye el desarrollo teórico de Harsanyi acerca de los tipos. También suele aparecer en la literatura como la doctrina Harsanyi-Aumann. Según esta doctrina existen principios normativos universales de racionalidad que pueden identificar una solución única (posiblemente probabilística) para cada problema de decisión descripto en forma completa. Por lo tanto, en un mundo en el que hay conocimiento común de racionalidad, los jugadores racionales obtendrán las mismas soluciones (es decir, serán similares de un único modo racional).

Como es bien conocido, Harsanyi (1968) introdujo su teoría de tipos como una forma de reducir los juegos de información incompleta a juegos de información completa. La idea es que en la noción de tipo está incluida la información de un agente (sus creencias) acerca de su propio tipo y el tipo de los demás. Esto es, la incertidumbre respecto de la función de utilidad de los demás queda condensada en la noción de tipo. Para Harsanyi las únicas diferencias posibles entre jugadores se derivan de diferencias de información, y las diferencias respecto de las creencias individuales solamente pueden ser atribuidas a diferencias de información. Esto se traduce en dos condiciones que deben cumplir los jugadores: primero, los tipos son conocimiento común, y segundo, los jugadores deben tener una distribución de probabilidad *a priori* común acerca del conjunto de los estados del mundo. Esta es una restricción normativa al comportamiento de los jugadores en un juego que se puede traducir fácilmente en la simetría del juego. Dicho de otra manera, el supuesto de distribución *a priori* común de las creencias implica que los tipos de los jugadores no difieren respecto de las expectativas que tienen acerca de los eventos del mundo si poseen la misma información. Desde luego, podría ser el caso de que los jugadores estuvieran en posesión de diferentes conjuntos de información, y entonces las diferencias en cuanto a la forma de jugar el juego podrían atribuirse a diferencias de información. Esto es, los jugadores sólo diferirán en función de nueva información que adquieran.

La extensión de Aumann (1976) de la doctrina Harsanyi elimina las diferencias de información al argumentar que si los individuos tienen la misma distribución común *a priori* respecto del evento, i .e. respecto del juego, entonces sus distribuciones de probabilidad *a posteriori* con respecto a ese evento tienen que estar de acuerdo porque los individuos racionales revisan sus creencias a través de la actualización bayesiana cuando se enfrentan a diferencias informativas. En el mundo real hay especuladores que compran y otros que venden activos financieros, unos creyendo que van a subir de precio y los otros, que van a perder valor. Obviamente alguno está equivocado. Pero la transacción se lleva a cabo porque no tienen conocimiento común de qué creencias tienen. En el momento en que descubren que las creencias que tienen son inconsistentes, si son racionales tienen que revisar sus creencias hasta que converjan. Como lo expresó Aumann, “los jugadores no pueden estar de acuerdo en

estar en desacuerdo" (*agreeing to disagree*). A este resultado le subyace una noción de similitud entre jugadores.

Estos supuestos en conjunto implican lo que Hargreaves-Heap y Varoufakis (2004) llaman *alineación consistente de creencias*, y que está a la base del equilibrio de Nash. Esta noción implica que ningún jugador racional desde el punto de vista instrumental puede esperar de otro jugador similar (i. e., en cuanto a su racionalidad) que juegue de manera diferente[4]. Si ambos poseen la misma información, y son igual de racionales, entonces razonarán de la misma manera, y por ende, terminarán concluyendo de manera independiente lo mismo. Y el resultado tendría que ser necesariamente un EqN. Para que el resultado no fuese un EqN, tendría que ser falso o que los jugadores sean racionales, o que no tuvieran conocimiento común de racionalidad. Como señalan Hargreaves-Heap y Varoufakis (1995: 60), no podemos esperar que los jugadores tengan siempre sus creencias alineadas, a menos que hubiera una única forma racional de jugar el juego para cada jugador. Pero una cosa es suponer que de existir una única solución racional a un juego, esta solución sería la que resultaría de la interacción de jugadores racionales, bajo el supuesto de conocimiento común de racionalidad. Y otra distinta es suponer que siempre existe una única forma racional de jugar un juego. Con buen tino, Hargreaves-Heap y Varoufakis le llaman a este supuesto Principio de determinación racional (p. 60). Sigo en la discusión de este principio a Hargreaves-Heap y Varoufakis. Este principio está sustentado por la doctrina Harsanyi-Aumann en base a dos argumentos: que si los jugadores racionales tienen la misma información entonces sacarán las mismas conclusiones; y que los jugadores tienen la misma información respecto de las reglas del juego. Estos dos elementos conforman lo que podríamos llamar la *similitud racional* de los jugadores.

Hargreaves y Varoufakis presentan dos argumentos contra esta versión de la *similitud racional*: por un lado, la doctrina Harsanyi-Aumann parece requerir de los jugadores que estos sean similares en cuanto a su capacidad de procesar información. Si lo pensamos como un programa de computadora, frente a los mismos inputs los jugadores tiene que obtener los mismos outputs. Pero esto presupone que cada situación tiene una resolución algorítmica, esto es, que para cada problema existe un procedimiento mecánico para resolverlo respecto del cual los jugadores coincidirán. Esta visión algorítmica de la razón implicaría que sería siempre posible identificar un conjunto exhaustivo de reglas, cuya aplicación cada jugador estaría en condiciones de determinar. Sin embargo, como señalan Hargreaves y Varoufakis en su crítica a este aspecto de la doctrina Harsanyi-Aumann:

> ningún conjunto de reglas puede ser exhaustivo: Ningún conjunto de reglas puede contener reglas para su propia aplicación. Cualquier configuración nueva donde las reglas puedan aplicarse, siempre puede ser individualizada de tal manera que no esté cubierta por las reglas existentes. En estas circunstancias, o bien es necesario crear nuevas reglas para cubrir cada nueva aplicación (y este proceso amenaza con una regresión al infinito), o bien las personas deben interpretar creativamente de qué forma las reglas existentes se aplican en las nuevas configuraciones. Pero una vez que se reconoce una interpretación de este tipo, no hay razón para suponer que todos los individuos serán creativos, en este sentido, de la misma manera (p. 64).[5]

Por otro lado, la segunda forma de similitud es mucho más fuerte, ya que supone que además de tener las mismas capacidades predictivas, los jugadores deben tener la misma información. El teorema de Aumann consiste en lo siguiente: los desacuerdos respecto de cómo jugar el juego provocan revisiones (bayesianas) de las creencias o probabilidades asociadas a las alternativas del juego, y en el largo plazo, esto es, después de las revisiones necesarias los jugadores terminarán con idéntica información. Pero aplicado a juegos en forma normal el teorema de Aumann parece introducir un proceso de convergencia que no estaría incluido en la descripción del juego. ¿Cómo y en qué momento tendría lugar ese proceso de revisión de creencias?

[4] Por supuesto que Harsanyi asume que cada jugador conoce su propio tipo, es decir, un jugador no puede no saber que es racional.

[5] Aquí Hargreaves y Varoufakis parecen presuponer que el conjunto de principios a aplicar para resolver un juego cualquiera es «abierto», es decir que siempre van a existir casos no cubiertos. Pero la teoría de juegos ha tratado este tema extensamente (ver por ejemplo Lipman, 1991) y no necesariamente la respuesta es negativa. La autorreferencia *per se* no es un problema.

Según Hargreaves y Varoufakis esta forma de similitud introduce un elemento difícil de justificar en juegos *one-shot* que se refiere al proceso de convergencia en una distribución *a posteriori* común:

> …cuando las creencias se refieren a cómo jugar el juego, y las creencias divergentes solo se revelan en el juego, es más que un poco difícil ver cómo se aplica el argumento a las creencias que los agentes sostienen antes de jugar el juego. Naturalmente, cuando se repite el juego, la idea tiene mucho más sentido (p. 66).

Superracionalidad

Un concepto de solución que hace de la similitud entre jugadores su fundamento es la noción de *Superracionalidad*. Esta noción fue propuesta por Hofstadter (1983, 1985) para brindar un fundamento a la solución cooperativa en el Dilema del Prisionero (DP) (Flood y Dresher, 1952). La noción de superracionalidad de Hofstadter comparte un parecido de familia con una serie de nociones que toman la simetría de los jugadores como argumento central para justificar la cooperación en el DP (Campbell, 1989). Estas nociones incluyen un supuesto adicional a la descripción del DP según el cual los jugadores son *similares* en cierto sentido, similitud que se evidenciaría en una facultad epistémica de los jugadores para reconocer que se enfrentan a una misma situación y que poseen las mismas estrategias (Hofstadter, 1985).

La noción de Superracionalidad describe una propiedad de los perfiles en la diagonal de un juego simétrico. Hofstadter parece suponer que los judadores superracionales tienen conocimiento mutuo de que son racionales. Eso parece deducirse de la idea de que los "pensadores superracionales …incluyen en sus cálculos el hecho de que están en un grupo de pensadores superracionales" (Hofstadter, 1985: 748). La noción de superracionalidad se aplica naturalmente a juegos simétricos. De la aceptación del principio de determinación racional, se sigue para los pensadores superracionales que hay una única solución racional al juego. Dado que el juego es simétrico, la estrategia racional debe ser la misma para todos los jugadores, y debe encontrarse en la diagonal principal del juego. Y dado que estos no pueden diferir en cuanto a sus poderes de razonamiento (i. e., son similares) en un meta-nivel los jugadores pueden darse cuenta que de las estrategias disponibles deben elegir aquella que sea óptima.

La única solución racional estaría determinada conjuntamente por dos meta-principios: el Principio de Razón Insuficiente (PRI) (Sinn, 1980) y el principio de coordinación (Bacharach, 1999). Según el primero, ninguna estrategia debería tener mayor probabilidad si son *a priori* indistinguibles. Es evidente que si los jugadores son racionales, y si sus preferencias son indistinguibles en términos de su racionalidad (i. e., como son racionales deben razonar de la misma manera), entonces sus estrategias deberían ser indistinguibles. De acuerdo con este principio, en el ejemplo 2, si los jugadores son superracionales, la aplicación del PRI nos llevaría a considerar como posibles sólo los perfiles de estrategias sobre la diagonal principal del juego (ver tabla 2), ninguno de los cuales debería tener primacía sobre otro. Es decir, si sólo aplicáramos el PRI tanto (C, C) como (D, D) serían igual de probables. Para brindar una solución precisa, superracionalidad requiere de otro principio que pueda arbitrar una solución. La noción de superracionalidad adiciona a la doctrina Harsanyi-Aumann la idea de que la solución racional a cualquier juego debe ser una solución Pareto eficiente o de recompensa dominante (Harsanyi y Selten, 1988). Esta idea está capturada por un principio que vemos asumido en las distintas formulaciones de la teoría del RE, tal como en el principio de coordinación de Bacharach (Bacharach, 1999), o en el principio individual de racionalidad en equipo de Janssen (2001). En ambos casos lo que requiere el principio es que cada jugador haga su parte en una combinación de estrategias Pareto eficiente.

Otra vez, en el ejemplo 2, es evidente que, de los perfiles de estrategias (C, C) y (D, D), la solución identificada por el principio de coordinación es la de recompensa dominante (C, C).

Ahora bien ¿Cuál es la justificación para dar ese paso? Es evidente que puede haber condiciones particulares de la situación o de los jugadores que hagan uno u otro supuesto más plausible. Pueden existir situaciones en las cuales la no-cooperación, o la elección de la opción no-paretiana (pensemos en

cuestiones como la deserción durante una guerra), estuviera penada con la pena capital (y esta fuera prácticamente inexorable), quizás la opción alternativa ni siquiera sería pensable para los involucrados. Sin embargo, este supuesto parece injustificado, incluso en términos racionales.

Pensamiento mágico

El pensamiento mágico es otra noción que ha aparecido bajo diferentes nombres, y que esencialmente se reduce a la idea de que existe alguna manera de saber *a priori* la elección de otro jugador. Pensemos en el contexto de un juego *one-shot* en forma normal (i.e. una interacción que tiene lugar una sola vez sin posibilidad de comunicación), en el que la única información que poseen los jugadores consiste en la estructura estratégica del juego (las estrategias posibles, las recompensas, etc.). Los jugadores no tienen ningún acceso epistémico privilegiado a los estados de conciencia de los demás, por lo tanto la posibilidad de «conocer» la disposición a elegir una u otra estrategia no es en absoluto trivial. El mecanismo preciso que presentan las distintas soluciones que agrupo bajo el concepto de pensamiento mágico pueden variar.

La idea de pensamiento mágico consiste en la creencia de que la propia acción de alguna manera tiene una influencia causal sobre la acción de otro jugador (Shafir y Tversky, 1992; Daley y Sadowsky, 2016). En consecuencia, las creencias respecto de las acciones de los demás varían de acuerdo a las propias acciones. Esto es, si voy a cooperar en el DP, es más probable que crea que el otro jugador también cooperará. Y además, si voy a cooperar en el DP le atribuyo una mayor probabilidad a que el otro jugador también coopere.

Una noción familiar es la de translucidez (*translucency*), que fuera propuesta por Gauthier como fundamento de su teoría del contrato social. La noción de translucidez captura la idea de que los jugadores pueden, a través de algún medio (cognitivo), establecer si otro jugador puede estar dispuesto a cooperar (Gauthier, 1986; Frank, Gilovich, y Regan, 1993; Spiekermann, 2007; Capraro y Halpern, 2015). En forma general, los modelos de pensamiento mágico capturan la creencia de que, debido a las similitudes existentes entre los seres humanos, un jugador puede elegir un modo de razonamiento a sabiendas de que *cualquiera que elija* será también elegido por el otro jugador (si suponemos un entorno de dos jugadores). En contraste, la doctrina de Harsanyi y la superracionalidad implican solo que si el jugador usa el modo de razonamiento que realmente se requiere normativamente, este será el mismo modo de razonamiento que usaría otro jugador completamente racional.

Es claro que el pensamiento mágico requiere alguna justificación filosófica para ser considerado un concepto de solución válido. En general, se ha justificado el pensamiento mágico por medio de teorías psicológicas. Posiblemente la más conocida sea la teoría de la proyección social (Alicke, Dunning y Krueger, 2005; Acevedo y Krueger 2005). Según esta teoría la mayoría de las personas espera que los otros se comporten como ellos lo hacen. Desde luego, la solución que promueven de los dilemas sociales, o los problemas de coordinación es sencilla: si elijo cooperar, y mis co-jugadores se me parecen, existe una probabilidad mayor de que elijan cooperar a que no lo hagan. Desde luego, la proyección social provee una solución psicológica, esto es, empírica, a los dilemas sociales y los problemas de coordinación, pero no una solución que pueda considerarse racionalmente fundamentada.

Hay otras versiones de la teoría de la proyección social que tienen un parecido de familia entre sí. La percepción de la similitud es particularmente importante en la investigación en psicología social, y ha sido propuesta para explicar la cooperación en juegos *one-shot* (Aksoy y Weesie, 2012). El comportamiento de los actores generalmente depende de las creencias sobre los motivos sociales y las estrategias de los demás (e incluso creencias de orden superior, como las creencias sobre las creencias de los demás acerca de las propias motivaciones sociales). Este punto ha sido enfatizado en la literatura de psicología social como se ejemplifica en la hipótesis del triángulo (Kelley y Stahelski, 1970), el sesgo de similitud estructural (SSE) (Kuhlman, Brown y Teta, 1992) y el modelo de cono de Iedema (1993). Estas tres teorías plantean diferentes relaciones entre las preferencias y las creencias acerca de las preferencias de los demás. La hipótesis del triángulo clasifica a los individuos en competidores y

cooperadores, y establece que los competidores esperan menos variación en la población que los cooperadores. Según SSE, los actores clasificados en una categoría de orientación social esperan que otros pertenezcan de manera predominante a su propia categoría. En coincidencia con el SSE, el modelo de cono de Iedema también predice que los actores de todas las categorías sociales esperan una mayor frecuencia de actores de su propia categoría. Esto está en línea con el efecto de falso consenso (Ross, Greene y House, 1977), que consiste en que en ausencia de mayor información, las personas esperan que las demás personas sean similares a ellos en términos de gustos, preferencias y otros atributos socialmente relevantes.

Uno de los problemas de estas nociones consiste en que el supuesto de independencia causal parece ser violado (sobre esta noción ver Bicchieri y Green, 1997), lo que descarta la elección de resultados fuera de la diagonal. Basta con que un jugador se forme la intención de elegir cierta estrategia para asegurarse que el otro jugador hará lo mismo. Mientras que la dependencia causal de las decisiones podría considerarse producto de una forma de razonamiento inválida, la noción de dependencia causal puede volver a entrar en escena en la forma de la correlación intrínseca de creencias (o considerando jerarquías de creencias). Entonces, en cierto modo, incluso si se respeta el principio de independencia causal de que las elecciones de un jugador no dependen causalmente de las de otro, las elecciones podrían estar correlacionados al nivel de las creencias (de unos sobre otros). Como señalaron Brandenburger y Friedenberg (2008), esta es una adaptación a la TJ de la idea del principio de causa común o correlación (Reichenbach, 1956).

Hay otras nociones que parecen seguir la misma línea, en el sentido de que no respetan el principio de independencia causal. Una de esas nociones es la propuesta por Morton (1994). La idea de Morton consiste en que cuando dos individuos tienen un objetivo en común pueden predecir sus acciones recíprocamente e interpretar las creencias de cada uno a través de un procedimiento que llama pensamiento de solución (*solution thinking*) ¿En qué consiste la noción propuesta por Morton? Es un intento de solución para el problema de la lectura mental en los juegos de coordinación. Al igual que el RE y la teoría de puntos focales, la solución propuesta por Morton pretende racionalizar la posibilidad de la coordinación para obtener beneficios mutuos en una amplia variedad de juegos, para los cuales el EqN no ofrece soluciones satisfactorias. El problema, según el autor, parece provenir de los supuestos que incorpora la teoría de la acción racional. Por eso es necesario ir más allá del alcance de la teoría pura, e incorporar intuiciones empíricas (Morton: 22). Para ello Morton recurre explícitamente a un concepto de similitud:

> …el simple hecho de la similitud tiene muchas de las consecuencias que tendría el conocimiento mutuo perfecto. La similitud, utilizada correctamente, logra lo que parece necesitar creencias sobre creencias, creencias sobre creencias sobre creencias, etc. (p. 22)

El método de simulación defendido por Morton es una forma de simulación desde el punto de vista de la primera persona. Consiste en que un agente puede imaginar lo que haría si estuviera en el lugar del otro: lo que una persona haría en la situación de interés, le sirve para predecir lo que harían los demás. Es decir, para entender lo que harían los demás necesitamos ponernos a nosotros mismos en la situación en la que están ellos, y el resultado de la simulación en nosotros mismos podemos atribuirla a los otros.

A mi modo de ver la «*solution thinking*» de Morton es un esquema (necesariamente) inválido de razonamiento, una forma de razonamiento mágico. Para entender por qué sería instructivo ver su representación en el esquema de Gold y Sugden. Seguimos para ello la reconstrucción de Guala de la solución de Morton que tiene la virtud de enfatizar los aspectos de similitud entre jugadores (Guala 2016: 97). Guala discute la solución de Morton en relación con la teoría de puntos focales, es decir, como una noción que se encuadraría dentro de las nociones empíricas de similitud en nuestra taxonomía. Sin embargo, como quedará claro en la reconstrucción que realizo, el lugar adecuado para la solución de Morton es como una forma de pensamiento mágico.

Veamos el siguiente esquema de racionalidad individual para la solución de Morton:

1. ‘Yo’ identifico el punto focal obvio
2. Dado que es exactamente como yo, el otro jugador también identifica el mismo punto focal
3. Yo derivo mis acciones y las del otro jugador por medio de un simple razonamiento instrumental
4. Predigo lo que hará y lo que el otro jugador cree que haré
5. Quiero lograr (H, H)

Debo elegir H

Nótese que el pensamiento de solución se realiza en el modo de agencia individual, contrario al RE. De acuerdo con Guala, la simulación ocurre entre los pasos 2 y 4. Una vez que se identifica el punto focal cabe hacerse la pregunta acerca de qué fundamentos tenemos para pensar que el otro jugador podrá reconocer el mismo punto focal que yo, y además si lo elegirá. De acuerdo con los pasos (1) y (2) «Yo» identifico el punto focal, y como el otro jugador es exactamente como yo, entonces él/ella también lo identificará. Hasta aquí, todo lo que se requiere es un principio de simetría (que desde luego habría que justificar también). El problema surge en los pasos (3) y (4), que permanecen sin explicación. A partir de la premisa «simple razonamiento instrumental» todo lo que puedo concluir es que sería bueno para mí, y también para el otro jugador, que él/ella y yo eligiéramos H. Si eso fuera suficiente para inferir «Debo elegir H», entonces el paso en el razonamiento según el cual el otro jugador razona de manera *similar* a como razono yo, me permitiría inferir “ÉL/Ella elegirá H”. Pero ¿de dónde pueden provenir estas inferencias si no de alguna forma de pensamiento mágico? Guala no parece proveer un argumento genuino que permita justificar esta inferencia.

Adviértase que si hubiera un argumento que justificara la solución de Morton, ese mismo argumento sería atendible para sostener un argumento en favor de la noción de superracionalidad.

Conclusión

En este artículo se ha pasado revista a una serie de enfoques sobre nociones de solución en teoría comportamental de juegos que presuponen diferentes nociones de similitud. Además se ha ofrecido una revisión crítica de dichas soluciones. Bajo la noción de similitud empírica se han agrupado las teorías de punto focal de Schelling y Lewis y el RE. Se han desarrollado diferentes maneras en que el RE se aparta de la noción de racionalidad estándar en la teoría de juegos ortodoxa, y hemos analizado las diferentes soluciones que ofrecen al dilema del prisionero.

También se ha discutido la llamada doctrina Harsanyi-Aumann, para subrayar las dos formas de similitud entre jugadores que sustenta el EqN: por un lado, la idea de que si los jugadores tienen la misma información, entonces obtendrán las mismas inferencias, por lo que los jugadores deberían ser similares en cuanto a su capacidad cognitiva; por otro lado, la similitud planteada por Aumann respecto de que los jugadores no pueden tener diferencias de información, porque frente a diferencias de información revisarían sus creencias hasta converger. Un problema que se plantea a esta visión de la similitud consiste en que este proceso de convergencia no puede ocurrir en juegos *one-shot*, por lo que parece más apropiado en el caso de juegos repetidos.

Por último, se ha ejemplificado el *pensamiento mágico* a partir de una noción de solución propuesta por Morton y fundamentada en formas de simulación, para mostrar que incurre en una forma de razonamiento inválido.

Diferentes nociones de similitud también han comenzado recientemente a permear en las ciencias del comportamiento. La similitud es un mecanismo cognitivo que tienen las personas para reducir la complejidad de las situaciones. Creemos que proporcionar una representación formal clara es una de las tareas que los filósofos analíticos pueden emprender, y que puede resultar en un trabajo propedéutico para la elaboración de modelos formales.

Referencias

AKSOY, O.; WEESIE, J. (2012). "Beliefs about the social orientations of others: A parametric test of the triangle, false consensus, and cone hypotheses". *Journal of Experimental Social Psychology*. 48. 45-54.

ALICKE, M. D.; DUNNING, D. A.; KRUEGER, J. (2005). *The self in social judgment*. Psychology Press.

AUMANN, R. (1976). "Agreeing to disagree". *The Annals of Statistics*. 4 (6). 1236-1239.

AUMANN, R.; BRANDENBURGER, A. (1995). "Epistemic conditions for Nash equilibrium". *Econometrica*. 63(5). 1161-1180.

BACHARACH, M. (1999). "Interactive team reasoning: A contribution to the theory of cooperation". *Research in Economics*. 53: 117-157.

BACHARACH, M. (2006). *Beyond Individual Choice: Teams and Frames in Game Theory*. Editado por N. Gold y R. Sugden. Princeton. Princeton University Press. NJ.

BARDSLEY, N.; MEHTA, J.; STARMER, C.; SUGDEN, R. (2010). "Explaining focal points: cognitive hierarchy theory versus team reasoning". *The Economic Journal*. 120 (543). 40-79. https://doi.org/10.1111/j.1468-0297.2009.02304.x

BICCHIERI, C.; GREEN, M.S. (1997). "Symmetry arguments for cooperation in the prisoner's dilemma". En Bicchieri, C., Jeffrey, R., y Skyrms, B. (Editores). *The Logic of Strategy*. Kluwer Academic Publisher/Oxford University Press. New York. 229–249.

BRANDENBURGER, A.; FRIEDENBERG, A. (2008). "Intrinsic correlation in games". *Journal of Economic Theory*. 141(1). 28-67. https://doi.org/10.1016/j.jet.2007.09.012

BROWNE, K. (2018). "Why should we team reason". *Economics and Philosophy* 34. 185-198. https://doi.org/10.1017/s0266267117000347

CAMPBELL, R. K. (1989). "The Prisoner's Dilemma and the Symmetry Argument for Cooperation". *Analysis*. 49 (2). 60-65.

CAPRARO, V.; HALPERN, J. (2019). "Translucent players: Explaining cooperative behavior in social dilemmas". *Rationality and Society.* 31(4). 371-408.

CHIERCHIA, G.; CORICELLI G. (2015). "The impact of perceived similarity on tacit coordination: propensity for matching and aversion to decoupling choices". *Frontiers in Behavioral Neuroscience*. 9. 202. https://doi.org/10.3389/fnbeh.2015.00202

COLMAN, A. M. (2003). "Cooperation, psychological game theory, and limitations of rationality in social interaction". *Behavioral and Brain Sciences*. 26. 139-198. https://doi.org/10.1017/s0140525x03000050

COLMAN, A. M.; PULFORD, B. D.; ROSE, J. (2008). "Collective rationality in interactive decisions: Evidence for team reasoning". *Acta Psychologica*. 128(2). 387–397. https://doi.org/10.1016/j.actpsy.2007.08.003

DALEY, B.; SADOWSKY, P. (2017). "Magical Thinking: A Representation Result". *Theoretical Economics*. 12. 909-956. https://doi.org/10.3982/te2099

GAUTHIER, D. (1987). *Morals by Agreement*. Oxford University Press.

GEANAKOPLOS, J.; PEARCE, D.; STACHETTI, E. (1989). "Psychological games and sequential rationality". *Games and Economic Behavior*. 1(1). 60-79.

GOLD, N., Ed. (2005). *Teamwork: Multi-Disciplinary Perspectives*. Palgrave Macmillan. New York.

GOLD, N. (2012). "Team reasoning, framing and cooperation". En Okasha, S. y Binmore, K. (Eds.) Publisher. Cambridge University Press. 185-212. https://doi.org/10.1017/CBO9780511792601.010

GOLD, N. (2018). "Team reasoning: controversies and open research questions". En Jankovic, M. y Ludwig, K. (Eds.) Handbook of Collective Intentionality. Routledge. New York. 221-232.

GOLD, N.; COLMAN, A. (2018). "Team Reasoning and the Rational Choice of Payoff-Dominant Outcomes in Games". Topoi. 39(2). 305-316. https://doi.org/10.1007/s11245-018-9575-z

GOLD, N.; SUGDEN, R. (2007). "Collective intentions and team agency". *The Journal of Philosophy*. 104 (3). 109-137.

DI GUIDA, S.; DEVETAG G. (2013). "Feature-Based Choice and Similarity Perception in Normal-Form Games: An Experimental Study". *Games.* 4(4). 776-794.

FLOOD, M.; DRESHER, M. (1952). "Some experimental games". *Research memorandum RM-789*. Rand. Santa Monica, CA.

FRANK, R. H.; GILOVICH, T.; REGAN, D. T. (1993). "The evolution of one-shot cooperation: An experiment". *Ethology and Sociobiology.* 14: 247-256.

GAUTHIER, D. (1986). *Morals by Agreement*. Oxford University Press. Oxford.

GUALA, F. (2016). *Understanding Institutions: The Philosophy and Science of Living Together*. Princeton University Press. Princeton.

HARGREAVES, S.; VAROUFAKIS, Y. (1995). *Game Theory. A critical introduction*. Routledge. London.

HARSANYI, J. C. (1968). "Games with incomplete information played by Bayesian players". Parts I, II, III. *Management Science*. 14. 159-182, 320-334, 486-502.

HARSANYI, J. C.; R. SELTEN, R. (1988). *A General Theory of Equilibrium Selection in Games*. MIT Press. Cambridge. MA and London.

HOFSTADTER, D. R. (1983). "Dilemmas for superrational thinkers, leading up to a luring lottery". *Scientic American*. 248(6).

HOFSTADTER, D. R. (1985). *Metamagical themas: Questing for the essence of mind and pattern.* Basic Books. New York.

HURLEY, S. (2005). "Rational agency, cooperation and mind-reading". En Gold, N. (ed). *Teamwork: multi-disciplinary perspectives*. 9. Palgrave MacMillan. Basingstoke. 200–215. https://doi.org/10.1057/9780230523203_10

IEDEMA, J.; POPPE, M. (1994). "Causal attribution and self-justification as explanations for the consensus expectations of one's social value orientation". *European Journal of Personality*. 8. 395-408.

JANSSEN, T. (2001). "Rationalizing focal points". *Theory and Decision.* 50. 119-148.

KELLEY, H. H.; STAHELSKI, A. J. (1970). "Social interaction basis of 'cooperators and competitors' beliefs about others". *Journal of Personality and Social Psychology*. 16(1). 66-91. https://doi.org/10.1037/h0029849

KRUEGER, J. I.; ACEVEDO, M. (2005). "Social Projection and the Psychology of Choice". En M. D. Alicke, D. A. Dunning, y J. I. Krueger (Eds.). *Studies in self and identity. The Self in Social Judgment*. Psychology Press. 17-41.

KRUEGER, J.I.; DIDONATO, T.E.; FREESTONE, D. (2012). "Social projection can solve social dilemmas". *Psychol Inq*. 23. 1-27. https://doi.org/10.1080/10478 40X.2012.641167

KUHLMAN, D. M., BROWN, C.; TETA, P. (1992). "Judgements of cooperation and defection in social dilemmas: The moderating role of judge's social orientation". En W. LEWIS, D. K. 1969. *Convention: A Philosophical Theory*. Harvard University Press. Cambridge.

LIPMAN, B. (1991). "How to Decide How to Decide How to...: Modeling Limited Rationality". *Econometrica*. 59(4). 1105-1125.

MORTON, A. (1994). "Game Theory and Knowledge by Simulation". *Ratio*. 7(1). 14-25. https://doi.org/10.1111/j.1467-9329.1994.tb00150.x

MUSSWEILER, T.; OCKENFELS, A. (2013). "Similarity increases altruistic punishment in humans". *PNAS*. 110(48). 19318-19323.

MYERSON, R. B. (1991). *Game Theory: Analysis of Conflict*. Harvard University Press. Cambridge.

NASH, J. F. (1950). "Equilibrium points in n-person games". *PNAS*. 36 (1). 48-49. https://doi.org/10.1073/pnas.36.1.48

QUATTRONE G.A.; TVERSKY, A. (1984). "Causal versus diagnostic contingencies: on self-deception and on the voter's illusion". *Journal of Personality and Social Psychology*. 46(2). 237-248. https://doi.org/10.1037/0022-3514.46.2.237

RAPOPORT, A. (1960). *Fights, games, and debates*. University of Michigan. Ann Arbor.

REICHENBACH, H. (1956) *The Direction of time*. University of Los Angeles Press. Berkeley.

ROSS, D. (2016). *Philosophy of Economics*. Palgrave. New York: https://doi.org/10.1057/9781137318756

ROSS, L.; GREENE, D.; HOUSE, P. (1977). "The false consensus effect: An egocentric bias in social perception and attribution processes". *Journal of Experimental Social Psychology*. 13(3). 279-301. https://doi.org/10.1016/0022-1031(77)90049-x

RUBINSTEIN, A.; SALANT, Y. (2016). "'Isn't everyone like me?': On the presence of self similarity in strategic interactions". *Judgment and Decision Making*. 11 (2). 168-173.

SCHELLING, T. (1981). *The Strategy of Conflict*. Harvard University Press. Cambridge.

SHAFIR, E.; TVERSKY, A. (1992). "Thinking Through Uncertainty: Nonconsequential Reasoning and Choice". *Cognitive Psychology*. 24(4). 449-474. https://doi.org/10.1016/0010-0285(92)90015-t

SINN, H. W. (1980). "A rehabilitation of the principle of insufficient reason". *Quarterly Journal of Economics*. 94(3). 493-504. https://doi.org/10.2307/1884581

SPIEKERMANN, K. (2007). “Translucency, assortation, and information pooling: how groups solve social dilemmas”. *Politics, Philosophy and Economics.* 6. 285-306.

SUGDEN, R. (2000). “Team preferences”. *Economics and Philosophy*. 16(2). 175-204

SUGDEN, R. (2003). “The logic of team reasoning”. *Philosophical Explorations*. 6(3). 165-181. https://doi.org/10.1080/10002003098538748

SUGDEN, R. (2011). “Mutual advantage, conventions and team reasoning”. *Int Rev Econ*. 58(1). 9–20. https://doi.org/10.1007/s12232-011-0114-0

SUGDEN, R. (2015). “Team reasoning and intentional cooperation for mutual benefit”. *Journal of Social Ontology*. 1. 143–166. https://doi.org/10.1515/jso-2014-0006