PRINCIPIA **29**(1): 73–94 (2025) doi: 10.5007/1808-1711.2025.e96263
Published by NEL — Epistemology and Logic Research Group, Federal University of Santa Catarina (UFSC), Brazil.

SESGO IMPLÍCITO, EXTERNALISMO Y SEGUNDA PERSONA

IMPLICIT BIAS, EXTERNALISM AND SECOND PERSON

JUAN R. LOAIZA

Universidad Alberto Hurtado, CHILE
jloaiza@uahurtado.cl
https://orcid.org/0000-0003-0570-3832

Abstract. In this text, I argue that the psychological study of implicit bias can benefit from second-person approaches. Specifically, I show that dominant experimental approaches based on the Implicit Association Test (IAT) presuppose an internalist account of implicit bias, according to which the possession of an implicit bias is identified with the possession of a conceptual association. By equating the possession of implicit bias with the possession of a conceptual association, the traditional view reduces bias to a particular activation in memory and ignores the sociocultural aspects involved in this phenomenon. To avoid this, I suggest adopting an externalist perspective, considering a second-person approach that allows sociocultural and normative aspects to be included in the identification of bias. Finally, I explore some consequences of the externalist proposal for the measurement and intervention on implicit bias.

Keywords: implicit bias • social cognition • representation • stereotypes

RECEIVED: 11/09/2023 REVISED: 30/08/2024 ACCEPTED: 31/01/2025

Introducción

El sesgo implícito es un fenómeno que ha ganado atención en las últimas décadas. En una caracterización general, el sesgo implícito cubre reacciones en las cuales nuestra conducta está guiada de manera a menudo inconsciente e involuntaria por estereotipos. Este fenómeno psicológico puede invocarse para explicar algunas formas de discriminación en las cuales los individuos llevan a cabo acciones discriminatorias sin percatarse plenamente de las injusticias cometidas. Así pues, el estudio de la estructura psicológica del sesgo implícito puede dar pistas sobre cómo podemos intervenir sobre estructuras sociales injustas.

En este artículo, exploraré algunas limitaciones tradicionales de la caracterización del sesgo implícito. En particular, mostraré que la manera tradicional de estudiar el sesgo implícito presupone una imagen internalista del fenómeno que impide ver cómo aspectos socioculturales y normativos hacen parte de la configuración del mismo. En contraste con este internalismo, esbozo una caracterización externalista del sesgo que busca resaltar estos elementos socioculturales como constituyentes del

fenómeno. Esto permite explicar mejor la evidencia empírica disponible a la vez que nos ofrece avenidas nuevas para la medición y la intervención sobre el sesgo.

Este texto se divide en tres partes. En la primera me ocupo de lo que podríamos llamar la visión tradicional del sesgo implícito y cómo esta visión lleva a una operacionalización específica del fenómeno basada en el IAT. Con esto pretendo sentar las bases para la discusión posterior. En la segunda sección expongo algunas limitaciones de la visión tradicional y de la operacionalización mencionada del sesgo implícito. En particular, defiendo que la operacionalización del sesgo implícito en términos de los tiempos de reacción del IAT no es suficiente para explicar algunos aspectos del fenómeno relacionados con factores socioculturales. Finalmente, esbozo una propuesta externalista inspirada en los enfoques de segunda persona sobre los estados mentales y muestro cómo este nos puede abrir puertas a formas más adecuadas de operacionalización.

Antes de comenzar, me gustaría aclarar algunos compromisos y limitaciones del argumento. Primero, no quiero asumir que la perspectiva de segunda persona es superior a otras perspectivas sobre la mente en general. Creo que hay muchos contextos en los que las perspectivas de primera y tercera persona pueden ser de gran valor para comprender toda clase de fenómenos psicológicos. Sin embargo, esto no implica que la perspectiva de segunda persona no sea relevante para el estudio empírico de lo mental. Lo que mostraré es que el estudio empírico del sesgo implícito podría beneficiarse de ver el sesgo desde la perspectiva de segunda persona y de cuestionar algunas formas de internalismo prevalentes en la noción estándar del fenómeno.

1. La tradición del sesgo implícito

Para la psicología experimental y otras ciencias de lo mental, no es misterio que buena parte de los procesos que explican nuestro comportamiento son procesos de los cuales no somos conscientes y que yacen por fuera de nuestro control voluntario. Esto no significa, sin embargo, que estos procesos sean irrelevantes para la explicación de nuestro comportamiento. Explicar, por ejemplo, cómo nuestro sistema visual pasa de irritaciones de superficie (Quine 1969) a un mundo perceptualmente rico y significativo es cuando menos parte de una explicación satisfactoria de varios aspectos de nuestro comportamiento (por ejemplo, cómo navegamos el espacio usando información visual).

En la psicología de la cognición social, el sesgo implícito se entiende como uno de estos procesos a nivel subpersonal que afectan la manera en la que nos relacionamos con otros. En una caracterización inicial, podemos decir que una persona exhibe un sesgo implícito si su comportamiento manifiesta una concepción estereotipada de otros, incluso en casos en los que la persona no se compromete explícitamente con

estos estereotipos. Estos comportamientos pueden incluir pensamientos pasajeros, reacciones emocionales automáticas, y hasta reacciones conductuales abiertas claras.

Un ejemplo sencillo puede aclarar el fenómeno. Pensemos en el caso de Carlos, un hombre joven y consciente de toda clase de luchas feministas e interseccionales, una persona que constantemente revisa sus creencias y manifiesta resistirse a la educación tradicional machista que ha recibido. Un día Carlos se encuentra conduciendo y de repente nota un carro que sale de un parqueadero de manera torpe y descuidada. Sin ver quién conduce el carro, Carlos exclama: "¡Qué mujer tan descuidada!". En este caso, la conducta de Carlos es, primero, una conducta sesgada. Carlos no tiene evidencia alguna de que sea una mujer quien maneja el carro que ha salido del parqueadero, y aunque este hubiese sido el caso (y aunque Carlos hubiese tenido evidencia de esto), el pensamiento de Carlos no ocurrió tras haber examinado detenidamente quién conducía o de haber considerado la evidencia disponible. Por el contrario, la exclamación de Carlos es producto de un estereotipo muy frecuente en su cultura, según el cual las mujeres conducen de manera torpe y descuidada.

El caso ficticio de Carlos permite presentar dos propiedades importantes del sesgo implícito. Primero, el sesgo implícito es precisamente un *sesgo* debido a que se trata de conducta que, en algún sentido, está injustificada. En el caso de Carlos, su exclamación se debe a la presencia de un estereotipo, un estereotipo para el cual Carlos no tiene evidencia alguna y que por todo lo demás sabemos que es falso (e.g., véase Wiberg 2006). Segundo, el sesgo de Carlos es implícito, pues Carlos no es plenamente consciente de tenerlo. Es probable que Carlos, siendo una persona sensible a la discriminación por género, se sorprenda de su propia exclamación y sienta vergüenza. Es incluso posible (o al menos deseable) que Carlos permanezca atento y reevalúe si a futuro encuentra otras conductas que revelan este tipo de sesgos. En cualquier caso, lo que hace que este sesgo sea implícito es que este sesgo actúa de manera involuntaria y a pesar de los posibles compromisos explícitos del agente.[1]

En psicología experimental, la forma más común de detectar y medir el sesgo implícito es el Test de Asociación Implícita (IAT por sus siglas en inglés, *Implicit Association Test*). El IAT originalmente se propuso como una prueba de asociación conceptual general (Greenwald; McGhee; Schwartz 1998). La prueba consiste en un estímulo que se presenta en la parte central de la pantalla y dos categorías a cada lado. El sujeto experimental debe mover una palanca o presionar un botón hacia la dirección de la categoría correspondiente lo más rápido posible. Por ejemplo, si el estímulo consiste en la imagen de un gato, y las categorías son "Felino" y "Canino" a la izquierda y la derecha respectivamente, el sujeto deberá mover la palanca a la izquierda o presionar el botón izquierdo lo más rápido posible. Hasta aquí no hay mucho que nos sorprenda de la prueba. La prueba se torna interesante cuando empezamos a sofisticar las categorías que presentamos a los lados del estímulo.

Supongamos que el estímulo ahora es la cara de un hombre blanco. Bajo la des-

cripción inicial podríamos tener categorías como “Hombre” y “Mujer”, por ejemplo. Sin embargo, la prueba comienza a ser informativa cuando las categorías constituyen categorías disyuntivas como “Hombre o Fuerte” y “Mujer o Débil”. Para este tipo de categorías disyuntivas, la respuesta será exactamente la misma; después de todo, la cara del hombre sí corresponde a la categoría “Hombre o Fuerte” cuando pedimos a los sujetos que categoricen en virtud del género percibido. No obstante, la prueba comienza a generar un efecto importante: en casos en los que tenemos estereotipos asociados a un concepto, nuestro tiempo de reacción cambia. En este caso, dado el estereotipo “Los hombres son fuertes”, las personas tienden a categorizar caras de hombres bajo la categoría “Hombre o Fuerte” (la condición “congruente” en tanto corresponde con el estereotipo) más rápido que a la categoría “Hombre o Débil” (la condición “incongruente” en tanto que contradice el estereotipo) o “Mujer o Fuerte” en caso de que el estímulo sea la cara de una mujer (una segunda condición “incongruente”). Esto sugiere que la persona asocia automáticamente los conceptos “Hombre” con “Fuerte” y “Mujer” con “Débil”, incluso si no es consciente de ello.

El IAT ha sido usado en una variedad de dimensiones. Si bien la mayoría de experimentos han usado categorías raciales (Amodio & Devine 2006), también se han observado efectos sobre el género (Rudman; Greenwald; McGhee 2001) y el peso corporal (Phelan et al. 2015), entre otras propiedades. A su vez, la psicología experimental ha encontrado disociaciones entre los compromisos evaluativos explícitos y los resultados de la prueba, reforzando la idea de que estos sesgos no constituyen realmente formas implícitas que tienen efectos importantes en nuestro comportamiento.

2. El problema representacional del sesgo implícito

A pesar de ser la prueba estándar para medir el nivel de sesgo implícito de una persona, existen varias fuentes de escepticismo sobre lo que esta prueba puede mostrar. A mi parecer, una falla importante con el IAT puede formularse en términos de su interpretación como una operacionalización satisfactoria del sesgo implícito. Recordemos que el IAT pretende detectar una asociación conceptual implícita, asociación que puede ser general y concerniente a la semántica de conceptos no necesariamente relacionados con el mundo social (e.g. “gato” y “felino”). En el caso del sesgo implícito, sin embargo, se usan conceptos relacionados con sesgos implícitos de orden social, sea en términos de raza, género, etc. En cualquier caso, la presuposición de base consiste en asumir que el IAT sirve para detectar sesgos implícitos operacionalizando el sesgo implícito en términos de una asociación conceptual. No obstante, esta operacionalización del sesgo implícito acarrea algunas limitaciones importantes.

Para comenzar, es claro que hay buenas razones para pensar que el sesgo implíci-

to está relacionado en algún sentido con asociaciones conceptuales. En la literatura existen dos caracterizaciones principales sobre el sesgo implícito. Para la primera, el sesgo, pensado en términos de su ontología cognitiva, tiene una estructura proposicional. Para esta perspectiva, podemos intuitivamente pensar en el sesgo implícito en términos de creencias inconscientes o involuntarias (Mandelbaum 2016). Bajo esta lectura, tener un sesgo implícito hacia un grupo social consiste en tener una creencia injusta e injustificada, aunque inconsciente o involuntaria, sobre dicho grupo. Esta lectura del sesgo implícito en términos de creencia lleva entonces a ver el sesgo implícito como una actitud proposicional.[2]

Ahora bien, si el sesgo implícito tiene formato proposicional, entonces es claro por qué el IAT puede ser una prueba satisfactoria. Si el formato del sesgo es proposicional, y las proposiciones son representadas mediante conceptos, entonces el sesgo puede detectarse midiendo cuán asociados están dos conceptos, presuntamente a raíz de que figuran en una representación de una proposición que es constitutiva de la creencia. En otras palabras, si el sesgo implícito es una creencia, es claro que el IAT sirve para detectar esta creencia mediante la detección de asociaciones conceptuales que en principio constituyen la misma.

Para la segunda caracterización, el sesgo no tiene una estructura proposicional, sino asociativa. En lugar de consistir en un conjunto de creencias sobre grupos sociales, el sesgo bajo esta visión consiste simplemente en la asociación—quizás al modo del aprendizaje hebbiano—entre distintas representaciones mentales.[3] Para ver con claridad el contraste, consideremos el siguiente argumento por parte del asociacionismo en contra de la tesis proposicional.

El argumento central en favor de la tesis proposicional sobre el sesgo implícito apela a que el sesgo parece ser sensible a consideraciones lógicas y racionales. Toribio (2018), en respuesta a Mandelbaum (2016) resume este argumento de la siguiente manera:

(P1) Si las actitudes implícitas responsables por la conducta implícitamente sesgada son estados mentales con una estructura asociativa, entonces la conducta implícitamente sesgada puede ser modificada o eliminada *únicamente* alterando algunas contingencias ambientales, i.e., sea mediante extinción o contracondicionamiento.

(P2) Factores lógicos y evidenciales modifican o eliminan la conducta implícitamente sesgada.

(C1) Las actitudes implícitas responsables por la conducta implícitamente sesgada no son estados mentales con una estructura asociativa.

(C2) Las actitudes implícitamente sesgadas responsables por la conducta implícitamente sesgada son sólo creencias. (Toribio 2018, p.47; traducción propia)[4]

De la reconstrucción de Toribio es claro que el argumento es problemático. En primer lugar, el paso de (C1) a (C2) no es claro, es decir, que las actitudes responsables por la conducta sesgada no sean estados mentales con estructura asociativa no implica que entonces sean creencias. Toribio menciona este problema, citando a Levy (2015), aunque no sea el foco de su crítica. Para la filósofa, el problema yace en (P1); para ella, es falso que los sesgos sólo sean modificables o eliminables mediante extinción o contracondicionamiento si su estructura es asociativa. Así pues, Toribio sostiene que incluso bajo estructuras asociativas el sesgo puede ser modificable o eliminable bajo consideraciones lógicas o evidenciales.

En cualquiera de los dos casos, sea que el sesgo tenga una estructura proposicional o asociativa, podemos todavía justificar el IAT como prueba para la presencia del sesgo implícito. Como puede notarse, a pesar de sus diferencias, las principales caracterizaciones de la ontología cognitiva del sesgo implícito asumen que la asociación conceptual es, de un modo u otro, una forma plausible de detectar el sesgo implícito. Sea porque lo que constituye el sesgo es la asociación misma, o porque lo que constituye el sesgo es una actitud proposicional compuesta de dichas asociaciones, en cualquier caso es la asociación la que nos sirve como criterio para decir si alguien está sesgado o no.

A mi parecer, la discusión entre las tesis proposicional y asociacionista revela una limitación general de la operacionalización tradicional del sesgo implícito, a saber, que se piensa en el sesgo implícito como dependiente únicamente de los estados mentales internos de un agente, y no se tienen en cuenta factores externos al agente que, en mi opinión, son también *constitutivos* del fenómeno del sesgo implícito. Puesto de otra manera, el problema con la operacionalización y discusión filosófica hasta ahora sobre el sesgo implícito ha sido asumir una forma de internalismo que deja por fuera factores externos que hacen parte del fenómeno. Si tengo razón, debemos entonces pensar en el sesgo desde una perspectiva externalista.

3. Internalismo sobre el sesgo implícito

En esta sección, mostraré por qué el internalismo ha dominado la interpretación de los resultados del IAT como evidencia de sesgo implícito. Para ello, distinguiré varias tesis internalistas y sostendré que incluso en su versión más débil, algunas de estas tesis continúan asumiéndose en la literatura filosófica y psicológica en torno al sesgo implícito. Esto sentará las bases para que la propuesta externalista que esbozaré más adelante tenga lugar como alternativa a la interpretación dominante.

Siguiendo la discusión anterior, podemos caracterizar la tesis internalista como aquella según la cual el fenómeno del sesgo implícito depende de los estados mentales internos (creencias, asociaciones, conceptos, etc.) del agente. De esta tesis general

podemos distinguir tres interpretaciones. En una interpretación fuerte, la tesis sostiene que el sesgo depende *exclusivamente* de estados mentales internos. Esto significa que estos estados internos son necesarios y suficientes para decir que alguien está sesgado o no. A su vez, podemos distinguir dos interpretaciones débiles: una en la cual los estados mentales internos son necesarios, mas no suficientes, para el sesgo, y otra según la cual los estados mentales son suficientes, mas no necesarios, para el mismo. Así pues, tenemos las siguientes tres tesis internalistas:

(I1) La posesión de ciertos estados mentales internos (creencias, asociaciones, conceptos, etc.) del agente son *necesarios y suficientes* para la posesión de un sesgo implícito.

(I2) La posesión de ciertos estados mentales internos (creencias, asociaciones, conceptos, etc.) del agente son *necesarios* para la posesión de un sesgo implícito.

(I3) La posesión de ciertos estados mentales internos (creencias, asociaciones, conceptos, etc.) del agente son *suficientes* para la posesión de un sesgo implícito.

A mi parecer, hay buenas razones para aceptar (I2). Es difícil pensar que una persona que incurre en conducta sesgada no tenga algunos estados mentales internos como creencias o asociaciones que expliquen su conducta y que, en ese sentido, sean al menos parte necesaria de la identificación del sesgo. Volviendo al ejemplo de Carlos, es difícil pensar que Carlos no tenga ninguna creencia o asociación que sea cuando menos sintomática de su sesgo. Así pues, mis objeciones no estarán dirigidas a (I2) ni a posiciones que asuman esta tesis.

¿Qué posiciones en el debate asumirían (I3) (y por consiguiente (I1), que es la conjunción de (I2) e (I3))? Por una parte, la discusión entre una las tesis asociativa y proposicional mencionadas anteriormente en la discusión entre Toribio y Mandelbaum parecen presuponer (I3), toda vez que tanto asociaciones como creencias se entienden en el debate como estados mentales, y la discusión gira en torno a la estructura de tales estados. Por otra parte, la reconstrucción del debate Holroyd et al. (2017) también parece apoyar la lectura internalista de la discusión. Según Holroyd y colegas, buena parte de la discusión en torno a la realidad psicológica del sesgo implícito puede dividirse entre posturas que sostienen que el sesgo implícito debe ser modelado sobre el modelo de las creencias o debe adoptarse un modelo de estado mental sui generis. A estos modelos se contraponen modelos que asumen que el sesgo implícito debe entenderse en términos de rasgos (e.g., Machery 2016) o el eliminativismo (que no atribuyen todavía a ninguna posición, pero véase Schimmack (2021)). En el caso de los dos primeros modelos, tenemos nuevamente una discusión sobre qué tipo de estado mental constituye el sesgo implícito y cuál es su estructura, asumiendo que basta con la posesión de alguno de estos estados para que exista un

sesgo. Finalmente, esta presuposición también puede notarse en discusiones más recientes en torno a si el sesgo implícito consiste en la posesión de una disposición o de una representación específica (Nanay 2021; Kalis & Ghijsen 2022).

Además de que buena parte de la discusión filosófica en torno al sesgo implícito parece asumir compromisos internalistas, es importante notar que la historia del IAT también sugiere compromisos de esta naturaleza. Como recuenta Machery (2022), el IAT surgió en el campo del estudio de las actitudes implícitas en psicología social. Los autores y autoras del test, Greenwald et al. (1998) sostienen que el test "busca medir las actitudes implícitas midiendo sus evaluaciones automáticas subyacentes" (p.1464, traducción propia). Por "actitudes", los autores y autoras entienden "trazas de experiencias pasadas introspectivamente no identificadas (o identificadas de manera imprecisa) que median favorable o desfavorablemente el sentimiento, pensamiento o la acción hacia objetos sociales" (Greenwald & Banaji 1995, p.8, citado en Greenwald et al. 1998, pie de página 1). Bajo la interpretación original del IAT, entonces, el test proveería evidencia de que hay un sesgo al proveer evidencia de evaluaciones automáticas y actitudes implícitas, entendidas como estados mentales ("trazas de experiencias pasadas"), y su presencia sería interpretada como evidencia suficiente de la posesión de un sesgo. Esta interpretación ha sido mantenida por Greenwald en otras ocasiones (Mitchell & Tetlock 2017).

En este punto, podría surgir la objeción de que se está confundiendo la ontología del sesgo implícito (i.e., qué tipo de estado o constructo es) con su medición. Según esta objeción, podemos usar el IAT como medición del sesgo implícito aun cuando aceptemos que el sesgo implícito como fenómeno no agota aquello que el IAT mide. Podríamos pensar entonces que el IAT sirve para detectar la presencia del sesgo sin comprometerse con una ontología determinada del mismo, lo que permitiría que el sesgo implícito se conceptualice como algo distinto a los estados mentales que detecta este test. La objeción cobra fuerza si notamos que podemos detectar un fenómeno sin que la señal que nos informa de su presencia nos informe de todas las propiedades del mismo. En un caso simple, podemos detectar por ejemplo la presencia de un intruso en un lugar mediante una luz o sonido de alarma sin con ello concluir que el intruso sea una luz o un sonido. Para el caso del sesgo, el IAT funge como señal de la presencia del sesgo, pero, según la objeción, no nos serviría para determinar las propiedades ni la ontología del mismo.

La objeción nos recuerda la importancia de mantener separadas categorías epistémicas de ontológicas, y por la misma vía, de distinguir las herramientas de detección y medición de los hechos que queremos estudiar. Sin embargo, para el caso del sesgo implícito y el IAT, la objeción puede desarmarse de la siguiente manera aunque la señal de detección o medición no deba parecerse al objeto que representa, el instrumento de detección o medición sí debe guardar cierta relación con el fenómeno que pretende detectar o medir. En la analogía anterior, una alarma que nos informe de la

presencia de un intruso en un lugar detectará cosas que sean cuerpos extensos con cierto tamaño únicamente, y no detectará, por ejemplo, la presencia de algún gas. Así, nuestra pretensión de usar cierto instrumento para detectar o medir un fenómeno sí nos compromete con algunas suposiciones sobre qué tipo de fenómeno estamos estudiando. En este caso, el uso del IAT como instrumento de medición o detección del sesgo implícito nos compromete con que el sesgo implícito sea algo detectable o medible mediante asociaciones conceptuales(o en la formulación original de Greenwald et al., evaluaciones automáticas subyacentes) que son las que directamente mide el IAT. Más aún, el uso del IAT como herramienta de detección nos compromete con la idea de que no es necesaria más información para la detección del sesgo que aquella contenida en el tiempo de reacción en una tarea de asociación conceptual, esto es, basta con detectar el cambio en el tiempo de reacción de un sujeto experimental para detectar el sesgo y no es necesario evaluar ninguna otra fuente de evidencia para concluir que existe el sesgo. En este orden de ideas, el IAT nos compromete, al menos en su aplicación estándar y en su formulación original, con un modelo internalista del sesgo implícito según el cual basta con la presencia de cierto tipo de estado mental (proposicional, asociativo, similar a la creencia o sui generis) para que tenga sentido decir que una persona posee un sesgo de esta naturaleza. Esto, por supuesto, no implica que no pueda haber interpretaciones alternativas de IAT, pero como hemos visto, estas no son las estándar en la práctica científica actual ni en la conceptualización original del test.[5]

Con esto, espero haber ofrecido argumentos de que la discusión filosófica en torno al sesgo implícito y el uso e interpretación del IAT como herramienta principal de medición asumen compromisos internalistas frente al sesgo implícito. Entiendo por compromisos internalistas comprometerse con la idea según la cual basta con la presencia de algún tipo de estado mental para la posesión del sesgo. Si este compromiso fuese correcto, esto explicaría el presunto éxito de las herramientas actuales de detección, medición y estudio del sesgo. Sin embargo, como intentaré mostrar en la próxima sección, existen en la literatura fuertes críticas al uso de estas herramientas que podemos explicar sobre la base de los compromisos internalistas que acarrean. , como ya anotamos,

4. Objeciones al internalismo sobre el sesgo implícito

En esta sección, sostendré que hay buenas razones para sospechar de (I3) y con ello de (I1) (que, como ya anotamos, es la conjunción de (I2) e (I3)). Puesto de manera más clara, creo que los estados mentales internos como creencias o asociaciones, incluso si son necesarios, no son suficientes para establecer la presencia de un sesgo. Esto debido a que (1) los resultados experimentales sugieren que podemos identificar

el sesgo de manera más robusta considerando factores contextuales externos a los individuos, (2) la identificación del sesgo involucra aspectos normativos, y (3) el sesgo implícito involucra elementos interactivos y no meramente disposicionales.

4.1. Resultados experimentales a nivel individual vs. poblacional

Desde hace unos años, ha habido fuertes críticas metodológicas al constructo del sesgo implícito desde la psicología experimental. Payne, Vuletich y Lundberg (2017) resumen estas críticas en tres enigmas. Primero, pese a que los estudios sobre sesgo implícito muestran efectos promedio robustos, estos efectos parecen ser demasiado variantes a lo largo del tiempo. La misma persona evaluada con un mes de diferencia puede mostrar niveles de sesgo completamente distintos. Así pues, la estabilidad temporal del sesgo, de acuerdo a su medición con el IAT, parece ser muy baja, pese a que en principio los sesgos implícitos constituyen sesgos temporalmente estables. Segundo, y en relación con el primero, los niños parecen exhibir niveles de sesgo similares a los de los adultos, pero no es claro cómo esto es posible siendo los sesgos tan variantes en lapsos de tiempo tan cortos según el primer enigma.

Tercero, y más relevante para el argumento, los efectos individuales observados tienden a ser relativamente débiles, es decir, no hay efectos grandes al medir individuos. Sin embargo, al agregar estos resultados y evaluar efectos a nivel grupal (por ejemplo a nivel de ciudades o grupos nacionales), los efectos son mucho más robustos. Esto sugiere que el sesgo parece detectarse con mayor claridad a nivel grupal que a nivel individual. Además, los efectos grupales parecen estar fuertemente relacionados con efectos contextuales. Por ejemplo, en Estados Unidos, estados con mayor proporción de población afrodescendiente muestran efectos diferenciales mayores en términos de preferencias raciales e incluso búsquedas de insultos raciales en línea, sugiriendo que en estos estados existe mayor conflicto interracial (Rae; Newheiser; Olson 2015). A su vez, países con mayor número de población con sobrepeso muestran mayor sesgo en favor de personas delgadas que países donde estos grupos están más balanceados (Marini et al. 2013).

En mi opinión, estos resultados nos pueden enseñar dos lecciones importantes. Primero, que no es claro cómo identificar sesgos a nivel individual. Si bien tenemos una idea de cómo detectar estos sesgos, todavía hay varios artefactos que hacen difícil la interpretación de resultados sobre individuos concretos. Segundo, estos resultados sugieren que el sesgo implícito es un fenómeno contextual. En otras palabras, estos resultados muestran que el contexto es un ingrediente vital en la presencia de sesgos, y que sin considerar el contexto en el que hacemos la medición, perderemos de vista efectos importantes y relevantes.

¿Cómo se conecta esta evidencia con el problema del internalismo sobre el sesgo implícito? Los resultados empíricos resumidos por Payne et al. parecen apoyar la

idea según la cual el estudio de los estados mentales de individuos no son un criterio suficiente para la detección del sesgo implícito. Más bien, es con ayuda de elementos contextuales—tanto temporales como espaciales—que podemos detectar cuándo existen sesgos en los miembros de una población.

Payne et al. proponen un modelo según el cual el sesgo implícito debe concebirse como un constructo a nivel poblacional y no individual. En palabras de los investigadores:

> [...] aunque el sesgo implícito puede en principio existir como un atributo de personas o un atributo de situaciones, la evidencia empírica es más consistente con la visión situacional. Al cambiar el énfasis de un análisis basado en personas a un análisis basado en situaciones, llegamos a una reinterpretación de los datos empíricos. Esta nueva interpretación sugiere que las medidas de sesgo implícito son significativas, válidas y confiables. Contrario a muchas presuposiciones, sin embargo, éstas son medidas significativas, válidas y confiables de situaciones en lugar de personas. (Payne et al. 2017, p.236)

Si bien comparto el espíritu de la propuesta de Payne y colegas, creo que podemos resistir la reconstrucción del sesgo implícito como una propiedad de situaciones en lugar de una propiedad de personas. Puesto de manera concreta, son las personas quienes incurren en conductas sesgadas, no las situaciones las que son sesgadas. Así pues, pasar completamente del nivel individual al nivel situacional parece dejar de lado que el sesgo se aplica sobre individuos.

No obstante, como lo reconocen los investigadores, podemos de todas maneras aprender de la evidencia empírica que reseñan para decir algo sobre la presencia de sesgos a nivel individual. En mi opinión, la tesis externalista ofrece precisamente este marco conceptual. Para el externalismo que propongo, son los individuos quienes son sesgados, pero los criterios mediante los cuales decimos que están sesgados apelan a factores contextuales externos y no únicamente a sus estados mentales individuales. Es por cómo interactúan estos individuos en un contexto determinado y en presencia de otros individuos que decimos que su conducta es sesgada, de modo que podemos integrar elementos contextuales y elementos individuales en la explicación e individuación del fenómeno del sesgo implícito.

Antes de continuar, podemos considerar una objeción sobre la interpretación de los resultados de Payne et al. que resulta pertinente. Si bien es verdad que los resultados de Payne et al. muestran que los efectos individuales son poco robustos, podría objetarse que se está asumiendo un estándar epistémico demasiado alto para el IAT. Según esta objeción, los resultados del IAT no deben ser infalibles para ser útiles, y por lo tanto no debería criticarse al IAT por su falibilidad. Estos problemas de falibilidad, sostiene la objeción, son solucionables usando otros métodos y triangulando la información que ofrece el IAT con información adicional que permita resultados más

fiables, de modo que basar el argumento sobre las limitaciones epistémicas del IAT pone el argumento sobre bases débiles.

A mi parecer, es posible superar esta objeción con dos anotaciones importantes al argumento. En primer lugar, es claro que el IAT, como todo instrumento de medición, es falible en cierto grado y debe triangularse con información adicional. El problema no yace en su mera falibilidad, sino en que cierto tipo de resultados son poco robustos, a saber, los resultados a nivel individual. Quien defienda el IAT deberá, no solamente apelar a la falibilidad estándar de cualquier instrumento de medición, sino que además debe explicar por qué una herramienta diseñada para medir asociaciones conceptuales (como lo es el IAT) falla en detectar sesgos a nivel individual pero es presuntamente exitosa a nivel poblacional. Es este fenómeno específico el que sustenta el argumento según el cual factores contextuales y sociales son relevantes en la detección y medición del sesgo, y no únicamente factores relativos a la asociación conceptual a nivel individual. En segundo lugar, a propósito de la triangulación, es posible conceder que el IAT debe triangularse con otros métodos. No obstante, aun cuando aceptemos que el IAT debe triangularse con otros instrumentos de medición, debemos escoger cuáles son los métodos de detección relevantes. En este punto, conviene entonces revisar los compromisos ontológicos que acarrea el uso de uno u otro instrumento de medición, y es aquí donde notar que los factores contextuales y la sensibilidad de cualquier instrumento de medición a los mismos puede ser un factor a considerar.

Con esto, el argumento que defiendo sostiene que, siguiendo la línea de razonamiento de Payne et al., debemos considerar factores contextuales y sociales en cómo caracterizamos el fenómeno del sesgo implícito (lo que afectaría la elección de métodos de detección y medición apropiados. Sin embargo, en contraste con Payne et al., sostengo que la caracterización del sesgo no recaiga sobre situaciones sino sobre personas aunque considerando factores sociales y contextuales que las acompañen, como quedará más claro más adelante. Por lo pronto, consideraré otros dos argumentos contra la concepción internalista del sesgo implícito.

4.2. Sesgo implícito y normatividad

Un elemento importante en la identificación del sesgo implícito es el de la detección de una conducta como sesgada, esto es, como producida por un estereotipo y, por lo tanto, injustificada en algún sentido relevante. En el ejemplo de Carlos, es importante que la conducta de Carlos no sea una conducta basada en un proceso reflexivo basado en evidencia (que además en este caso no existe). Es precisamente la falta de una justificación adecuada para la conducta de Carlos lo que la hace una conducta sesgada.

Esta característica de los sesgos en general introduce elementos normativos en la

identificación de este tipo de conductas. La conducta sesgada no es, pues, una simple desviación de una conducta frecuente. Al contrario, muchos sesgos implícitos ocurren por ser precisamente conductas y discursos muy frecuentes en un grupo particular. Lo que constituye el sesgo no es la desviación de una práctica común, sino que la conducta sesgada es una conducta que falla en cumplir algún tipo de norma. En el caso de Carlos, es claro que Carlos no *debería* creer que las mujeres son malas conductoras, y es la falta de justificación para su conducta lo que nos lleva a identificar un sesgo.

El carácter normativo del sesgo implícito se hace más patente cuando consideramos algunas variantes del IAT. Cuando el IAT se usa para medir sesgos, usamos categorías escogidas basándonos ya en conocimiento previo de normas sociales. Así es como usamos la categoría "Hombre o Fuerte" y no cualquier otra categoría como "Hombre o Pelo corto". Este conocimiento de las normas ya está incrustado en el diseño experimental mismo usado en los estudios sobre sesgo implícito. Si usáramos otro tipo de asociaciones semánticas, el IAT podría igualmente detectar dichas asociaciones pues podríamos todavía esperar tiempos de reacción reducidos para parejas de conceptos asociados, pero estas asociaciones ya no constituirían sesgo alguno. Encontrar asociaciones entre conceptos como "Mesa" y "Silla" no constituyen sesgo alguno, incluso si son categorías que podríamos perfectamente usar como parte del IAT. Así las cosas, para que el IAT detecte sesgos, debemos usar categorías teniendo en cuenta potenciales fallas en normatividad.

Es por supuesto una pregunta importante qué tipo de normatividad está involucrada en esta conducta. ¿Son todos los sesgos fallas en cumplir normas morales? ¿Existen sesgos implícitos sin carga moral? Por el momento no entraré a discutir estas preguntas, preguntas que requieren más cuidado que el que ofreceré aquí. Lo que me interesa argumentar es que el aspecto normativo del sesgo implícito es, a mi parecer, *constitutivo* del sesgo. No es simplemente que la conducta sesgada correlacione con fallas normativas, sino que la conducta sesgada es sesgada porque incurre en estas fallas.

Si esta idea es correcta, de aquí se sigue que la identificación y consecuente operacionalización del sesgo implícito debe necesariamente considerar factores normativos, y estos factores normativos no requieren de ningún estado mental particular de los agentes. En otras palabras, un agente puede fallar en cumplir una norma y puede incurrir en conductas sesgadas incluso si no tiene ninguna creencia o concepto que le permita considerar la norma. Podríamos decir por ejemplo que el agente no cumple con alguna condición epistémica para ser moralmente responsable de su conducta y esto quizás pueda servir de atenuante o eximirle de responsabilidad moral. Algunos ejemplos en esta dirección son los casos de mala suerte epistémica y moral que Fricker (2007) menciona, en donde podemos atribuir responsabilidad, mas no culpabilidad, a una persona que no tiene los recursos para superar los prejuicios que

llevan a injusticias testimoniales. En cualquiera de estos casos estamos aceptando que pese a haber atenuantes de la responsabilidad moral, la conducta es una conducta sesgada, en tanto sesgada ha incumplido alguna norma. En este sentido, la norma constituye un factor externo que es necesario para la identificación del sesgo, sugiriendo nuevamente una tesis externalista sobre la identificación del fenómeno en cuestión.

4.3. Sesgo implícito e interacción

Según la concepción tradicional del sesgo implícito, la conducta sesgada es producto de estados mentales como creencias o actitudes que, sea por su formato proposicional o asociativo, pueden identificarse midiendo el nivel de asociación conceptual que existe entre determinadas categorías. Así pues, el sesgo se concibe como un estado mental presuntamente disposicional. Alguien se dice sesgado en caso de que según las asociaciones que estén presentes en su mente, existan asociaciones basadas en estereotipos sobre distintos grupos sociales.

Ahora bien, el hecho de que el sesgo implícito sea un estado disposicional es, en mi opinión, correcta. De lo contrario, el sesgo implícito tendría que ser un estado mental ocurrente, esto es, un estado que sólo se da mientras la persona está en dicho estado, como lo sería el dolor o algunos estados emocionales como el temor. Esto implicaría que Carlos sólo está sesgado cuando exclama "¡Qué mujer tan descuidada!" y una vez su atención pasa nuevamente a la carretera, su sesgo desaparece. Esta consecuencia parece no capturar adecuadamente el sesgo, pues no pensamos en el sesgo como algo que desaparece cuando nuestra atención cambia de foco. Por esta razón, pensar en el sesgo implícito como un estado disposicional resulta más adecuado, pues los estados disposicionales no desaparecen pese a cambios de atención.

Sin embargo, lo que creo que la tradición ha perdido de vista es que el sesgo implícito es un fenómeno interactivo, no meramente disposicional como la creencia. Con "interactivo" quiero decir que lo relevante del sesgo implícito no es simplemente la existencia de una asociación conceptual, sino la existencia de ciertas disposiciones a formas de interacción particulares. Estas interacciones incluyen, por ejemplo, conductas como cruzar la calle o alejar la mirada automáticamente al ver una persona de un grupo social hacia el cual tenemos un sesgo y pensamos que son peligrosas. Es en nuestra interacción con otros que el estudio del sesgo implícito es importante. De lo contrario, el fenómeno del sesgo implícito se reduce a la simple categorización por estereotipos, un fenómeno relacionado, pero que no creo que agote el fenómeno del sesgo implícito.

Si tengo razón, una caracterización y operacionalización más robusta del sesgo implícito depende entonces de considerar la interacción como parte relevante del fenómeno. Esta interacción, por supuesto, constituye otro elemento externo a los

estados mentales de los agentes. Así pues, la importancia de elementos interactivos, sumada a la evidencia empírica disponible y los elementos normativos involucrados en la identificación del sesgo, sugieren que debemos aceptar que los estados mentales internos de los agentes no son suficientes para una teoría adecuada del sesgo, y, por lo tanto, que debemos optar por un externalismo sobre el sesgo. En lo que sigue, esbozaré una caracterización externalista del sesgo, aunque su desarrollo completo dependerá de investigación y aclaración conceptual futura. Presentaré no obstante algunas ideas sobre cómo pensar el sesgo de modo externalista y exploraré algunas consecuencias importantes para la medición del sesgo, así como para las posibilidades de intervención sobre el mismo.

5. Una caracterización externalista del sesgo implícito

Según la tesis que he defendido, la medición tradicional del sesgo implícito ha aceptado una caracterización internalista del fenómeno que es limitada por las razones ya expuestas. En esta sección, presentaré algunas ideas sobre cómo pensar en el sesgo implícito desde una perspectiva externalista. Mi propósito no es el de proponer una teoría detallada sobre el sesgo, sino más bien marcar la que considero sería la mejor estrategia para elaborar esta teoría. Ofreceré pues una historia sobre el desarrollo ontogenético del sesgo, aunque con propósito especulativo en general.

Para comenzar, pensemos en cómo pensar sobre los estados mentales en modo externalista. Bajo la definición de externalismo que he adoptado, esto significa pensar en los estados mentales de manera que un estado mental no dependa únicamente de estados internos. A mi parecer, una pista importante para comenzar a elaborar una teoría externalista del sesgo implícito es reconocer que el sesgo implícito es un fenómeno que se da primordialmente en nuestra transacción con otros. En otras palabras, el sesgo implícito es, en algún sentido, una forma de interacción social.

¿De qué tipo de interacciones estamos hablando? Es claro que no se trata de interacciones en las cuales atribuimos estados mentales a otros a nuestro alrededor. Más aún, en muchos casos de conducta sesgada los estados mentales de otros son irrelevantes para la producción de dicha conducta. En el caso de Carlos, los estados mentales de quien conduce el otro vehículo no interesan para decir que Carlos ha incurrido en una conducta sesgada. Así pues, se trata de interacciones en donde nuestro comportamiento es producido teniendo en cuenta qué hacen otros a nuestro alrededor, mas no es un comportamiento que requiera de que tengamos representaciones de los estados del otro.[6]

¿Cómo podemos pensar en formas de interacción que no involucran necesariamente la atribución explícita de estados a otros? A mi parecer, podemos pensar en el sesgo implícito sobre la base de la interacción en segunda persona. Las interacciones

de segunda persona son precisamente el tipo de interacciones en donde no necesitamos una atribución explícita de estados mentales a otros, sino en donde guiamos nuestra conducta basándonos en un proceso continuo de reacción frente a los actos de otros. Sobre la base de este tipo de interacciones podemos ver cómo surge el sesgo implícito como una forma en la que, dado un contexto social determinado, aprendemos a interactuar con los demás.

Para caracterizar estas formas de interacción en segunda persona, seguiré la caracterización de Pérez y Gomila (2017), según la cual algunos rasgos comunes de los casos paradigmáticos de la perspectiva de segunda persona son:

1. Hay interacción dinámica cara a cara.
2. Los aspectos expresivos del cuerpo son vistos como directamente significativos, no sólo significativos bajo interpretación.
3. La interacción es recíproca.
4. Dichas interacciones no suponen una actividad metarepresentacional acerca de los estados mentales del otro.
5. No es indispensable la existencia de un mundo compartido.
6. No requiere del lenguaje. (Adaptado de Pérez & Gomila 2017, p.277)

Bajo la caracterización de Pérez y Gomila, la interacción de segunda persona es entonces un modo de interacción en el cual la interacción no depende atribuciones explícitas, sino que implica la coordinación de la conducta en presencia de otros sin actividades metarrepresentacionales o incluso lingüísticas. Por esta razón, Pérez y Gomila piensan que la perspectiva de segunda persona es, dados estos rasgos, ontogenética, filogenética y lógicamente anterior a las atribuciones de primera y de tercera persona (Pérez & Gomila 2017, p.278). Teniendo ya más claro el panorama de la interacción de segunda persona, pensemos cómo podríamos caracterizar el sesgo implícito sobre la base de este tipo de interacción. Para ello, me remitiré a cómo la perspectiva de segunda persona explica fenómenos sobre la cognición social en general.

Las interacciones de segunda persona ofrecen una base para ver cómo evolucionaron y cómo pudimos aprender formas de interacción más complejas que involucran actitudes hacia mis propios estados mentales (primera persona) y los de los otros (tercera persona). Para ellos, es en el hecho de que somos animales que coordinamos nuestro comportamiento mediante el uso de expresiones presuntamente innatas que comenzamos a relacionarnos unos con otros. Así pues, las interacciones infante-cuidador son la primera instancia en la que comenzamos a construir un mundo social, pues es en estos casos en los que comenzamos a aprender a coordinar nuestra conducta.

Durante el desarrollo a lo largo de la infancia y presuntamente la adolescencia, es claro que nuestro comportamiento social se vuelve cada vez más complejo. Esto

se debe a que, sobre la base de las interacciones de segunda persona, comenzamos a hacer caer conceptos y otras habilidades cognitivas que nos facilitan la navegación del mundo social. Una de estas habilidades cognitivas es la habilidad de categorizar personas, esto es, de almacenar información sobre el mundo social haciendo uso de conceptos[7] que nos sirven para dividir a las personas en distintas agrupaciones. Es en este punto donde aparece la forma en la que representamos categorías sociales, esto es, mediante el uso de *estereotipos*.

Los estereotipos son, pues, categorías preestablecidas en el mundo social que guardan información sobre los grupos en los que una determinada estructura social divide a las personas. En esta perspectiva, los estereotipos existen en el nivel del mundo social. ¿Cómo pasamos entonces de estereotipos en el mundo social al sesgo implícito? A mi parecer, una historia plausible es que durante el desarrollo, a medida que vamos insertándonos en el mundo social en el que vivimos y aprendemos distintas clases de normas sociales, internalizamos también los modos de categorización de las estructuras sociales prevalentes. Así pues, aprendemos a dividir el mundo social siguiendo los lineamientos de la estructura prevalente y, en ese sentido, adquirimos modos de categorización que reproducen estructuras a menudo injustas e injustificadas. Es aquí donde aparece el sesgo implícito.

La historia que he desarrollado hasta ahora permite entonces explicar la aparición del sesgo implícito a partir de formas de interacción en las cuales internalizamos categorías preexistentes en nuestro mundo social. Esto explica algunos aspectos externos del sesgo implícito, como su carácter grupal y su sensibilidad al contexto. En esta historia, en tanto que el sesgo implícito se aprende dentro de otros modos de navegación del mundo social, el sesgo se cuela en nuestra psicología junto con otras herramientas que nos sirven para relacionarnos con otros en distintos contextos.

Ahora bien, en esta historia falta todavía ofrecer una historia de un rasgo importante de estos sesgos: su carácter implícito. Hasta el momento, la historia que he ofrecido no dice nada sobre por qué el sesgo implícito se inserta en nuestra psicología de manera a menudo involuntaria e inconsciente. Es en este punto donde la perspectiva de segunda persona cobra mayor fuerza. En la historia que he contado, el sesgo implícito aparece cuando comenzamos a internalizar formas de interacción social que dependen de estructuras sociales preestablecidas. En otras palabras, el sesgo aparece al aprender a movernos en el mundo social dada su estructura. Sin embargo, es importante recordar que este aprendizaje no es un aprendizaje explícito. No se nos enseña a navegar el mundo social mediante manuales de instrucciones que debamos memorizar en nuestra infancia. Al contrario, aprendemos a navegar el mundo social mediante pautas sutiles y a menudo *implícitas* sobre cómo debemos comportarnos en distintos contextos y con distintas personas. Es en esta dinámica en la que muchos de los estereotipos se internalizan sin ser nunca objeto explícito de mención. Es en prácticas implícitas en las que se nos enseña a alejarnos de algunos

grupos o acercarnos a otros, a comportarnos de una u otra manera frente a ciertos grupos que internalizamos estereotipos y aparecen entonces los sesgos como reflejo de la estructura social.

Así pues, podemos caracterizar el sesgo implícito como una forma de internalización de estructuras sociales estereotipadas que aparece durante el desarrollo de nuestra cognición social sobre la base de interacciones donde los estereotipos y las normas sociales pueden operar de manera implícita. En otras palabras, el sesgo implícito es, según esta historia, un fenómeno psicológico en el cual un agente dirige su conducta de forma inconsciente e involuntaria, manifestando ciertas formas de aprendizaje implícito de categorías preexistentes en su mundo social. En este sentido, la identificación del sesgo implícito requiere de conocer no solo la historia psicológica del agente, i.e. sus estados mentales internos, sino también el contexto estructural social en el cual esa persona se desenvuelve.

Esta caracterización externalista del sesgo explica bien los fenómenos que problematizaban la caracterización internalista. En primer lugar, en esta caracterización es claro por qué observamos el sesgo a nivel grupal, pues el sesgo es un fenómeno que requiere de la interacción social a gran escala. A su vez, podemos explicar cómo se insertan aspectos normativos en la identificación del sesgo, pues es de la existencia de normatividad en el mundo social que aparece el sesgo como fenómeno psicológico, y podemos ver por qué el aspecto interactivo del sesgo es vital para una explicación del mismo. En este orden de ideas, la caracterización externalista nos permite ver más facetas del fenómeno que aquellas ofrecidas por la caracterización tradicional.

6. Consecuencias del externalismo sobre el sesgo implícito

La discusión anterior sobre internalismo y externalismo sobre el sesgo implícito, así como la sugerencia de caracterizar el sesgo implícito como un fenómeno basado en interacciones de segunda persona en las que internalizamos formas de categorización social, tienen dos consecuencias importantes. En primer lugar, la caracterización propuesta tiene consecuencias importantes para la operacionalización y medición del sesgo. Segundo, veremos también cómo esta propuesta puede tener impacto sobre cómo podemos intervenir en el sesgo y qué estrategias podríamos implementar para reducir algunas formas de discriminación e injusticia social.

Comencemos considerando las consecuencias de la propuesta para la medición del sesgo implícito. En la primera sección presenté el IAT como la medición principal usada para detectar el sesgo implícito. Bajo la interpretación tradicional del IAT, el sesgo implícito de operacionalizaba como la presencia de una asociación conceptual determinada que tenía como efecto la reducción de los tiempos de reacción en una tarea de categorización. En la sección siguiente, presenté la discusión sobre qué for-

mato tenía esta asociación, si era un formato proposicional o asociativo. En cualquier caso, la interpretación canónica de los resultados del IAT se limitaban a considerar como relevante únicamente la presencia de asociaciones conceptuales.

Bajo la propuesta externalista, podemos ver con claridad las ventajas y limitaciones de la medición del IAT. Por un lado, vemos por qué la medición del IAT ha sido la medición principal del sesgo, toda vez que las asociaciones conceptuales pueden formar parte del sesgo. Sin embargo, el argumento que propongo nos lleva a ver que esta medición no es suficiente para un estudio completo del sesgo, toda vez que los estados internos que el IAT estudia no son suficientes para la identificación de este fenómeno. A mi parecer, esto invita entonces a dos revisiones importantes de la medición del sesgo.

Primero, los diseños experimentales que hacen uso del IAT deben ser explícitos sobre los aspectos externos involucrados en la selección de estímulos y condiciones experimentales. No es gratuito que la mayoría de estudios que hacen uso del IAT usen estímulos de personas racializadas o marcadores (estereotipados) de género y sexo. Al hacer explícitos los factores sociales involucrados en el diseño y medición del IAT, podemos explicar mejor en qué sentido es que el IAT está detectando un sesgo, no solo detectando asociaciones conceptuales, sino también midiendo cómo ciertas formas de categorización social afectan los procesos psicológicos de los individuos.

Segundo, el externalismo sobre el sesgo implícito invita a una reconsideración de los métodos principales para medir el sesgo. En particular, invita a usar métodos interactivos para la detección de estos comportamientos. Métodos como el seguimiento de miradas de manera sincrónica (Wilms et al. 2010) permiten detectar fenómenos sociales en medio de la coordinación de la conducta con otros, de modo que usando este tipo de diseños podemos identificar otros aspectos del sesgo que permanecerían ocultos si nos concentramos únicamente en la medición del IAT.

Por último, me gustaría considerar algunas consecuencias del externalismo para motivos de intervención sobre el sesgo implícito. Desde el enfoque externalista, el sesgo implícito está constituido en parte por factores sociales como estereotipos y normas externas a los individuos. En este sentido, una intervención exitosa sobre distintas formas de sesgo implícito debe pasar necesariamente por una intervención en las estructuras sociales que llevan a este fenómeno psicológico. Este énfasis en la intervención externa y no solo interna abre la puerta para algunas estrategias. Una es intervenir en las formas en las que enseñamos sesgos a lo largo de la infancia. Al hacer explícitas las formas en las que los infantes internalizan formas de sesgo gracias a nuestras formas de interacción como adultos, podemos prevenir el surgimiento de conductas sesgadas. Esto incluye, por ejemplo, atender a cómo nos referimos a otros grupos y el tipo de interacciones que esperamos de los infantes.

Además de esto, la perspectiva externalista también invita a formas de intervención en el mundo social mismo. En otras palabras, para el externalismo, intervenir en

el mundo social mediante formas de activismo o campañas de concientización se ven como llevando a efectos psicológicos importantes. Así, estas intervenciones no tienen el único efecto de buscar aliviar formas tradicionales de injusticia, sino también acarrean formas de reconfigurar nuestra propia psicología.

Referencias

Amodio, D. M.; Devine, P. G. 2006. Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology* 91(4): 652–661. https://doi.org/10/d5k6tj

Frankish, K. 2016. Playing Double: Implicit Bias, Dual Levels, and Self-Control. In: M. Brownstein; J. Saul (Eds.), *Implicit bias and philosophy: Vol. 1: Metaphysics and epistemology*, p.23–46. Oxford: Oxford University Press.

Fricker, M. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Gallistel, C. R.; Matzel, L. D. 2013. The Neuroscience of Learning: Beyond the Hebbian Synapse. *Annual Review of Psychology* 64(7.1–7.32): 169–200. https://doi.org/10.1146/annurev-psych-113011-143807

Greenwald, A. G.; Banaji, M. R. 1995. Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review* 102(1): 4–27. https://doi.org/10.1037/0033-295X.102.1.4

Greenwald, A. G.; McGhee, D. E.; Schwartz, J. L. K. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology* 74(6): 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Hebb, D. O. 1949. *The Organization of Behavior: A Neuropsychological Theory*. New York: John Wiley & Sons, Inc.

Kalis, A.; Ghijsen, H. 2022. Understanding implicit bias: A case for regulative dispositionalism. *Philosophical Psychology* 35(8): 1212–1233. https://doi.org/10.1080/09515089.2022.2046261

Levy, N. 2015. Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements. *Noûs* 49(4): 800–823. https://doi.org/10.1111/nous.12074

Machery, E. 2016. De-Freuding Implicit Attitudes. In M. Brownstein; J. M. Saul (Eds.), *Implicit bias and philosophy: Vol. 1: Metaphysics and Epistemology*, p.104–129. Oxford: Oxford University Press.

Machery, E. 2022. Anomalies in implicit attitudes research. *WIREs Cognitive Science* 13(1): 1–15. https://doi.org/10.1002/wcs.1569

Mandelbaum, E. 2016. Attitude, Inference, Association: On the Propositional Structure of Implicit Bias. *Noûs* 50(3): 629–658. https://doi.org/10/gf4fx4

Marini, M.; Sriram, N.; Schnabel, K.; Maliszewski, N.; Devos, T.; Ekehammar, B.; Wiers, R.; HuaJian, C.; Somogyi, M.; Shiomura, K.; Schnall, S.; Neto, F.; Bar-Anan, Y.; Vianello, M.; Ayala, A.; Dorantes, G.; Park, J.; Kesebir, S.; Pereira, A.; Tulbure, B.; Ortner, T.; Stepanikova, I.; Greenwald, A. G.; Nosek, B. A. 2013. Overweight People Have Low Levels of Implicit Weight Bias, but Overweight Nations Have High Levels of Implicit Weight Bias. *PLOS ONE* 8 (12): 1–9. https://doi.org/10.1371/journal.pone.0083543.

Mitchell, G.; Tetlock, P. E. 2017. Popularity as a Poor Proxy for Utility: The Case of Implicit Prejudice. In S. O. Lilienfeld; I. D. Waldman (Eds.), *Psychological science under scrutiny*, p.164–195. Oxford: Wiley Blackwell.

Nanay, B. 2021. Implicit Bias as Mental Imagery. *Journal of the American Philosophical Association* 7(3): 329–347. https://doi.org/10.1017/apa.2020.29

Payne, B. K.; Vuletich, H. A.; Lundberg, K. B. 2017. The Bias of Crowds: How Implicit Bias Bridges Personal and Systemic Prejudice. *Psychological Inquiry* 28(4): 233–248. https://doi.org/10.1080/1047840X.2017.1335568

Pérez, D. I.; Gomila, A. (2017). Lo que la segunda persona no es. In: D. I. Pérez; D. Lawler (Eds.), *La segunda persona y las emociones*, p.275–293. Buenos Aires: SADAF.

Phelan, S. M.; Puhl, R. M.; Burke, S. E.; Hardeman, R.; Dovidio, J. F.; Nelson, D. B.; Przedworski, J.; Burgess, D. J.; Perry, S.; Yeazel, M. W.; van Ryn, M. 2015. The mixed impact of medical school on medical students' implicit and explicit weight bias. *Medical Education* 49(10): 983–992. https://doi.org/10.1111/medu.12770

Quine, W. V. O. 1969. Epistemology naturalized. In: *Ontological Relativity and Other Essays*, p.69–90. New York: Columbia University Press.

Schimmack, U. 2021. The Implicit Association Test: A Method in Search of a Construct. *Perspectives on Psychological Science* 16(2): 396–414. https://doi.org/10.1177/1745691619863798

Rae, J. R.; Newheiser, A.-K.; Olson, K. R. 2015. Exposure to Racial Out-Groups and Implicit Race Bias in the United States. *Social Psychological and Personality Science* 6(5): 535–543. https://doi.org/10.1177/1948550614567357

Rudman, L. A.; Greenwald, A. G.; McGhee, D. E. 2001. Implicit Self-Concept and Evaluative Implicit Gender Stereotypes: Self and Ingroup Share Desirable Traits. *Personality and Social Psychology Bulletin* 27(9): 1164–1178. https://doi.org/10.1177/0146167201279009

Toribio, J. 2018. Implicit Bias: From social structure to representational format. *THEORIA. An International Journal for Theory, History and Foundations of Science* 33(1): 41–60. https://doi.org/10.1387/theoria.17751

Wiberg, M. 2006. Gender differences in the Swedish driving-license test. *Journal of Safety Research* 37(3): 285–291. https://doi.org/10.1016/j.jsr.2006.02.005

Wilms, M.; Schilbach, L.; Pfeiffer, U.; Bente, G.; Fink, G. R.; Vogeley, K. 2010. It's in your eyes—using gaze-contingent stimuli to create truly interactive paradigms for social cognitive and affective neuroscience. *Social Cognitive and Affective Neuroscience* 5(1): 98–107.

Notes

[1]Es importante aclarar que el hecho de que el sesgo sea implícito no significa necesariamente que sea inconsciente. Como explica Frankish (2016), podemos ganar consciencia del sesgo implícito sin que este deje de ser implícito. Este tipo de sesgos se dicen implícitos dada su disparidad con nuestros compromisos explícitos y la manera automática en la que ocurre. En el caso de Carlos, Carlos puede ser ahora consciente de que posee tal sesgo y, sin embargo, actuar de manera sesgada a futuro en casos en los que su conducta se produzca automáticamente.

[2]Entender el sesgo implícito con un formato proposicional es entenderle en una estructura conceptual que permite inferencias lógicas. Según Mandelbaum (2016), el sesgo no

consiste simplemente en la asociación entre «Hombre afrodescendiente» y «Peligroso», sino en tener una actitud hacia la proposición «Los hombres afrodescendientes son peligrosos» que compromete doxásticamente al agente con la verdad de esta proposicíon. Esto implicaría la posibilidad de atribuirle al agente otras creencias que sean consecuencias lógicas de la proposición mencionada, tales como «Algunos hombres son peligrosos». Esto contrasta con modelos asociacionistas que no permitirían atribuir tales creencias sobre la base de consideraciones lógicas.

[3]El principio del aprendizaje hebbiano consiste en el principio según el cual la conexión entre las neuronas depende de su activación sincrónica. Esto explicaría algunos fenómenos de aprendizaje asociativo. Véase Hebb (1949) y Gallistel y Matzel (2013) para una discusión crítica.

[4]La separación entre premisas y conclusiones se encuentra en el artículo original.

[5]Para la propuesta original del test, véase Greenwald, McGhee, y Schwartz (1998). Para discusiones sobre distintas maneras de modelar los resultados del test, véase Holroyd et al. (2017), Machery (2016) y Schimmack 2021.

[6]Esto nos permite excluir consideraciones basadas en teorías clásicas de la cognición social como la Teoría de la Teoría o la Teoría de la Simulación.

[7]Es importante aclarar que uso la palabra “concepto” aquí como refiriendo a una representación mental individual de una categoría y no a la representación lingüística de una categoría en una comunidad de hablantes.