How to cite:
Cabrera. V. (2026) Evolución de la Economía de la Conducta ante la Inteligencia Artificial (IA). *Behanomics*, *4*, 18-38. https://doi.org/ 10.55223/bej.44

Behanomics
Year 2026
Volume 4

RESEARCH PAPER / ARTÍCULO DE INVESTIGACIÓN.

Evolución de la Economía de la Conducta ante la Inteligencia Artificial (IA). The Evolution of Behavioral Economics and Its Role in the Context of Artificial Intelligence.

Virginia Cabrera Nocito

Universidad Rey Juan Carlos, Estudiante de doctorado, Madrid, España

Contact email: vcabreranocito@outlook.es

RESUMEN

El artículo traza la evolución de la Economía de la Conducta desde sus orígenes en la crítica a la racionalidad neoclásica hasta la Ciencia del Comportamiento Cultural Evolutivo contemporánea. Frente al auge de la inteligencia artificial generativa se identifican tres aportaciones fundamentales de las ciencias conductuales: la metacognición para superar sesgos intuitivos mediante hibridación Sistema 1-Sistema 2; la comprensión del continuum de adopción (motivación, capacidad, confianza) facilitada por reformulación prospectiva de pérdidas; y la psicología de máquina para contrarrestar la sincofancia y bucles sesgados mediante fine-tuning ético, adaptación situacional y prompting crítico. Se advierte sobre riesgos de antropomorfización, atrofia cognitiva por offloading excesivo y erosión de normas sociales humano-IA, proponiendo diseño participativo para preservar juicio autónomo, relaciones recíprocas y pensamiento crítico. Concluye que el éxito de la IA se medirá por su integración armónica con complejidades conductuales humanas, posicionando la ciencia conductual como herramienta esencial para una era intencionalmente humana.

ABSTRACT

This article charts the evolution of Behavioral Economics from critiques of neoclassical rationality to contemporary Cultural Evolutionary Behavioral Science. Amid generative AI's rise it identifies three key behavioral science contributions: metacognition to overcome intuitive biases via System 1-System 2 hybridization; understanding adoption continuum (motivation, capability, trust) facilitated by loss-framing; and machine psychology countering sycophancy and bias loops through ethical fine-tuning, situational adaptation, and critical prompting. Warnings highlight anthropomorphization risks, cognitive atrophy from excessive offloading, and human-AI social norm erosion, advocating participatory design to preserve autonomous judgment, reciprocal relationships, and critical thinking. It concludes that AI success will be measured by harmonious integration with human behavioral complexities, positioning behavioral science as essential for an intentionally human era.

PAPER HISTORY
Received: 30-01-2026
Accepted: 05-03-2026

PALABRAS CLAVE
Economía de la conducta, IA, metacognicción, nudges, atrofia cognitiva, psicología de máquina, alineación hombre-máquina

KEYWORDS
Behavioral science, AI, Metacognition, nudge, human-AI alignment, cognitive atrophy, machine psychology

Agradecimientos o financiamiento

Evolución de la Economía de la Conducta y rol ante la Inteligencia Artificial

El artículo traza la evolución de la Economía de la Conducta desde sus orígenes en la crítica a la racionalidad neoclásica hasta la Ciencia del Comportamiento Cultural Evolutivo contemporánea. Frente al auge de la inteligencia artificial generativa se identifican tres aportaciones fundamentales de las ciencias conductuales: la metacognición para superar sesgos intuitivos mediante hibridación Sistema 1-Sistema 2; la comprensión del continuum de adopción (motivación, capacidad, confianza) facilitada por reformulación prospectiva de pérdidas; y la psicología de máquina para contrarrestar la sincofancia y bucles sesgados mediante fine-tuning ético, adaptación situacional y prompting crítico. Se advierte sobre riesgos de antropomorfización, atrofia cognitiva por offloading excesivo y erosión de normas sociales humano-IA, proponiendo diseño participativo para preservar juicio autónomo, relaciones recíprocas y pensamiento crítico. Concluye que el éxito de la IA se medirá por su integración armónica con complejidades conductuales humanas, posicionando la ciencia conductual como herramienta esencial para una era intencionalmente humana.

Palabras clave: Economía de la Conducta, Behavioral Science, IA generativa, metacognición, nudges, alineación humano-IA, atrofia cognitiva, psicología de máquina.

This article charts the evolution of Behavioral Economics from critiques of neoclassical rationality to contemporary Cultural Evolutionary Behavioral Science. Amid generative AI's rise it identifies three key behavioral science contributions: metacognition to overcome intuitive biases via System 1-System 2 hybridization; understanding adoption continuum (motivation, capability, trust) facilitated by loss-framing; and machine psychology countering sycophancy and bias loops through ethical fine-tuning, situational adaptation, and critical prompting. Warnings highlight anthropomorphization risks, cognitive atrophy from excessive offloading, and human-AI social norm erosion, advocating participatory design to preserve autonomous judgment, reciprocal relationships, and critical thinking. It concludes that AI success will be measured by harmonious integration with human behavioral complexities, positioning behavioral science as essential for an intentionally human era.

Keywords: Behavioral Economics, Behavioral Science, Generative AI, metacognition, nudges, cognitive biases, human-AI alignment, cognitive atrophy, machine psychology.

El cambio conductual

El cambio conductual es el proceso psicológico por el cual una persona modifica sus patrones de comportamiento, impulsado por diversos factores como la educación, la experiencia, la motivación (interna o externa) o la intervención terapéutica. Estos cambios pueden ser tanto positivos como negativos, y pueden ocurrir de manera consciente o inconsciente. La psicología ofrece diversas estrategias y enfoques para promover el cambio conductual, como la terapia cognitivo-conductual (García et al, 2017) o la terapia de aceptación y compromiso (Hayes et al., 2005).

Sin embargo, esta visión del cambio de conducta tiene un foco centrado en lo individual. Las ciencias económicas apoyadas en la estadística han demostrado cómo los comportamientos, analizados desde un punto de vista colectivo, pueden ser diferentes e incluso contradictorios con lo que hacen los individuos de manera aislada. El avance de las Ciencias del Dato (Big Data), así como de la Economía de la Conducta (Behavioral Economics) indican cómo cada vez más, con los datos adecuados, podríamos predecir esas decisiones colectivas, incluso antes de que fueran tomadas (Mullainathan & Thaler, 2000; Camerer, 1999).

La investigación sobre cambios de conducta tradicionalmente se ha centrado en las normas personales de los individuos y no tanto en las sociales. (Sevillano y Olivos, 2019). Los modelos teóricos que han dominado, como el modelo de la influencia normativa (Schwartz, 1977) y la teoría de la acción planificada (Ajzen, 1991), incluyen componentes normativos de carácter personal e interpersonal. La norma personal es la expectativa de comportamiento relacionada con los principios personales, de modo que el individuo no recibiría sanción externa si no se cumple (Oceja & Fernández-Dols, 2006). La teoría de la acción planificada propone que el comportamiento que esperan las personas cercanas o importantes para uno mismo (norma subjetiva) es uno de los factores que predicen la intención conductual, enfatizando el componente social de la norma pero únicamente de un entorno cercano donde al propio individuo le influya la opinión de sus referentes (Corral-Verdugo et al., 2019). Por descontado, además de los valores personales y del entorno social cercano, las normas pueden provenir de la situación en la que se encuentra la persona (Ross y Nisbett, 1991).

Un ejemplo ilustrativo de cómo operan nuestros comportamientos de grupo se observó en los Estados Unidos durante el verano de 2001, cuando los medios de comunicación se saturaron de historias sobre ataques de tiburones, generando pánico y disuadiendo a la población de asistir a las playas, aunque los datos mostraban una incidencia de ataques de tiburones dentro del promedio de años anteriores, evidenciando así lo que una

cobertura mediática desproporcionada puede influir en la percepción pública (Levitt & Dubner, 2011). Otro ejemplo de la influencia del contexto se encuentra en el cambio de percepción de las personas sobre el egoísmo y la generosidad humanas a raíz de la propuesta del juego del ultimátum (Güth et al, 1982), un juego donde un participante recibe dinero y decide cuánto compartir con un desconocido, que puede aceptar o rechazar la oferta, en cuyo caso pierden los dos. En dicho juego, se encontraba que la mayoría de las personas optaba por dar la mitad, lo que sugería un sentido natural de justicia. Sin embargo, un cambio en las reglas para que los participantes también pudieran quedarse con el dinero del otro jugador dio como resultado que el 66% prefirió quedarse con el dinero, mientras que solo el 6% compartía algo, demostrando la dependencia de las acciones humanas con el contexto y el diseño del experimento. (List, 2007).

De cómo podemos mantener comportamientos grupales inadecuados (o perjudiciales) sin reconocerlo durante años por efecto de la presión social, da buena muestra este otro ejemplo fechado en un hospital de la Viena de 1847, cuando una de cada seis mujeres moría por fiebre tras el parto sin que los médicos entendieran la causa. Cuando Ignaz Semmelweis, descubrió que los médicos pasaban de hacer autopsias a atender partos sin lavarse las manos por lo que las bacterias de los cadáveres se transferían a las madres, matándolas, el colectivo de médicos rechazó durante décadas la sencilla solución (lavarse las manos con cloro) porque hacerlo significaba aceptar que ellos mismos estaban matando a sus pacientes (Pittet & Boyce, 2001).

Es importante pues, a la hora de emprender cualquier cambio cultural o de valores, entender la influencia social en todo proceso colectivo (Sevillano & Olivos, 2019; Miller y Prentice, 2016). Esta influencia se puede estudiar desde aproximaciones teóricas tales como las normas sociales, el aprendizaje social, la comparación social, el liderazgo y el compromiso público (Abrahamse & Steg, 2013; Cinner, 2018). Las normas sociales se refieren a las creencias que tienen las personas sobre la forma de comportamiento adecuado (común y/o aceptado socialmente) en una situación concreta (Cialdini & Trost, 1998) y ofrecen información sobre cómo conducirse en una situación, bien haciendo lo que hace la mayoría de la gente (norma descriptiva) o haciendo lo que se debe hacer (norma prescriptiva).

Las normas descriptivas indican el comportamiento típico y motivan porque resultan eficaces y son fuente de gratificación por reconocimiento social. Además, en una situación de incertidumbre, imitar lo que otras personas hacen resulta adaptativo. Las normas prescriptivas indican lo que se aprueba o desaprueba socialmente e implican una sanción si no se cumplen. Habitualmente ambos tipos de normas son congruentes: las personas hacen lo que se debe hacer. Sin embargo, no siempre es así. La teoría del Foco Normativo que postula que, en un contexto dado, la norma social de comportamiento que sea más clara o predominante será la que dirija la conducta de las personas (Cialdini et al., 1990). También sugiere que las pautas situacionales,

que indican los objetivos normativos y dirigen la atención de las personas hacia las normas, pueden incrementar el cumplimiento de estas (Cialdini et al., 2006).

Que las emociones influyen en nuestras decisiones y comportamientos y que éstas a su vez se ven influidas tanto por el entorno como por las creencias socioculturales adquiridas a través de la memoria y la experiencia está hoy fuera de toda duda. La neurociencia ha demostrado como ese proceso que llamamos mente es en realidad una mezcla indisoluble de cuerpo, cerebro y entorno. Adicionalmente, Antonio Damasio (1996) describe la presencia de marcadores somáticos, especie de señales de aviso fisiológico en los segundos previos a la toma de decisiones que activarían un sentimiento agradable o desagradable "en las tripas" de las personas para protegerlas de pérdidas futuras haciéndolas rechazar de una manera casi visceral ciertas opciones. Una especie de sistema de calificación automática de predicciones que aumenta la eficiencia de los procesos de decisión previos de la acción.

Además los psicólogos han descubierto cómo la regulación emocional, un proceso por el que las personas modulan sus emociones para adaptarse más eficientemente a las situaciones en las que deben tomar una decisión, y que dependerá de cómo se haya interpretado dicha situación, no solo influye en la toma de decisiones con riesgo sino que también puede modificar las elecciones en contextos sociales. La reevaluación de elecciones, suprimiendo la emoción, es un elemento clave para el cálculo de valor y propicia con mayor frecuencia la aceptación de situaciones injustas (Chang et al., 2010).

La investigación también ha demostrado lo mal que funcionan las personas decidiendo en base a probabilidades (Kahneman, 2012) pues el cerebro humano se maneja con reglas simplificadoras y atajos o heurísticos que reformulan problemas transformándolos en operaciones más sencillas y automáticas para no tener que hacer un razonamiento profundo cada vez que se plantean.

La esencia humana no puede entenderse sin considerar las emociones junto a la razón y los heurísticos. Esta visión ha dado paso a la Economía de la Conducta, estimulando la colaboración entre economistas, psicólogos, sociólogos y neurocientíficos para analizar el comportamiento humano de forma integral que, a medida que avanza el siglo XXI, se posiciona como una disciplina esencial para explicar por qué las personas toman decisiones y cómo estas decisiones pueden ser influidas por un contexto que cada vez está más digitalizado.

La Economía de la Conducta (Behavioral Economics)

La Economía de la Conducta (Behavioral Economics) surge como respuesta a la pregunta de por qué las personas toman decisiones que parecen ilógicas o incluso contrarias a lo que sería mejor para ellas. Aunque

durante años los expertos en economía asumieron que las personas tomarían decisiones calculando los beneficios de cada acción y eligiendo siempre aquella que maximizara su beneficio, hoy sabemos que sus emociones, los atajos mentales que toma su cerebro (heurísticos) y la manera en la que se presentan las opciones, influyen en las decisiones de los individuos de formas que ni siquiera notan.

Los modelos predictores de la conducta se han basado tradicionalmente en la idea de que los individuos son agentes completamente racionales que toman decisiones, de manera consciente, para maximizar su bienestar. La teoría económica neoclásica de la Utilidad Esperada se desarrolló a principios del siglo XX, con sus raíces en los trabajos de filósofos como John Stuart Mill, describiendo a las personas con preferencias racionales y demostrando cómo, en situaciones de incertidumbre, tienen a elegir aquella opción de mayor Utilidad Esperada (Von Neuman & Morgenstern, 1953).

No obstante, muy pronto esta teoría de la Utilidad Esperada comienza a ser cuestionada. Maurice Allais (1952) demostró mediante la Paradoja de Allais cómo la teoría de la Utilidad Esperada se incumple con relativa frecuencia. Su experimento muestra que, en determinados juegos de azar, las personas tienden a preferir la Seguridad Percibida sobre la Utilidad Esperada. Aunque usualmente eligen certeza en lugar de incertidumbre, cuando la apuesta se presenta de manera diferente, pueden preferir la incertidumbre que anteriormente evitaron, mostrando cómo valoran más la ausencia total de riesgo que un riesgo muy incierto y remoto, pero cómo, ante pequeñas diferencias en probabilidades, ignoran estas diferencias y optan por la utilidad esperada.

La relevancia y el impacto de la Economía de la Conducta que aúna psicología y economía en busca de un modelo más realista de la toma de decisiones, ha sido reconocida con numerosos premios Nóbel otorgados a los investigadores en este campo.

En 1978, Herbert A. Simon lo recibe por la contribución de sus trabajos a la racionalización del proceso de toma de decisiones ante problemas complejos. Su teoría de la Racionalidad Limitada niega la existencia del comportamiento racional de las personas en los términos en los que hasta el momento había sido reconocida por la doctrina clásica y sostiene que los individuos no siempre toman decisiones racionales, sino que están limitados por la información disponible, el tiempo y la energía de que disponen para realizar cálculos, por lo que como no siempre pueden analizar completamente los problemas (Simon, 1957).

Kahneman y Vernon Smith comparten el Nóbel en 2002 por sus trabajos sobre la toma de decisiones bajo incertidumbre. Pero es la teoría de la Perspectiva (Kahneman & Tversky, 2013) la que desafía de modo más completo la teoría de la Utilidad Esperada ofreciendo una explicación aún más realista de cómo las personas evalúan opciones riesgosas. Ambos demuestran cómo las decisiones pueden variar en función de cómo se

presente la información aun cuando las probabilidades y resultados sean objetivamente iguales; cómo evaluamos de forma diferente las ganancias y las pérdidas, dando mayor peso a estas últimas; cómo no medimos la utilidad por la cantidad total que poseemos, sino por la variación que experimentamos alrededor de un punto de referencia y por la comparación con lo que poseíamos antes y con lo que poseen los demás; y cómo, en consecuencia, actuamos para minimizar lo que consideramos una pérdida en lugar de para maximizar las ganancias.

De hecho, fue Daniel Kahneman (2012) quien propuso y demostró experimentalmente que el pensamiento automático y emocional es el modo habitual de funcionamiento de nuestra mente. Lo denominó Sistema 1 o pensamiento rápido por oposición al Sistema 2 de pensamiento lento, consciente, lógico y racional cuyo uso demanda gran esfuerzo y energía mental. Con objeto de ahorrar energía, nuestro cerebro emplearía la mayor parte del tiempo el Sistema 1 que se vale de heurísticos (atajos) que le llevan a cometer fallos y errores (sesgos).

Otros autores destacados son Dan Ariely (2008) que demuestra que estos heurísticos y sesgos, si bien determinan una orientación irracional del comportamiento humano, esta irracionalidad es predecible. Werner Güth, Rolf Schmittberger y Bernd Schwarze (1982) proponen su juego del ultimátum para demostrar cómo ciertas preferencias sociales, como la aversión a la desigualdad, prevalecen muchas veces sobre el interés personal. Investigaciones posteriores demuestran cómo los participantes en un juego de ultimátum estaba más dispuestos a aceptar ofertas injustas si estas provenían de un ordenador, mientras que si provenían de una persona la experiencia de dolor que sentían les impedía aceptar dicho reparto injusto (Sanfey et al., 2003).

En Harvard, el profesor Sendhil Mullainathan es figura clave en la investigación de cómo la escasez de recursos, financieros, temporales o cognitivos, influyen en la toma de decisiones, relevante para entender problemas sociales como la desigualdad (Mullainathan, 2013). Destaca el impacto de la investigación sobre la sobreconfianza (Camerer & Lovallo, 1999) o sobre el comportamiento gregario (el que tienen los individuos de un grupo sin una dirección planificada), un término que se aplica los animales en manadas y a la conducta humana durante situaciones como las burbujas financieras especulativas, las manifestaciones, los eventos deportivos, las reuniones religiosas o los disturbios sociales (Bikhchandani et al., 1992). O las investigaciones sobre la aversión a la desigualdad, que indican cómo las personas toman decisiones para minimizar la desigualdad en los resultados mostrando disposición a sacrificar una ganancia potencial para impedir que otro individuo reciba una recompensa superior (Fehr & Schmidt, 1999).

En la línea más crítica con los postulados de Kahneman y Tversky se encuentra Gerd Gigerenzer, quien argumenta en contra de su caracterización de los sesgos como errores sistemáticos, al sugerir que podrían, no conducir a juicios erróneos, sino reflejar estrategias adaptativas que apoyan la toma de decisiones intuitivas. En general, Gigerenzer busca comprender los procesos de la toma de decisiones en su relación con el medio, apuntando a cómo la mente debería ser entendida en relación con sus capacidades y a su relación con el contexto. Enfatiza cómo la heurística de la toma de decisiones hace que las estrategias más simples puedan ser las más efectivas, revindicando el papel de la intuición como una "inclinación preferencial" en ellas. Indagando en la influencia del instinto en nuestras decisiones, propone un análisis de los múltiples factores e influencias (innatas y adquiridas) que condicionan nuestro proceso de decisión, sugiriendo que disponemos diferentes sistemas cognitivos, que empleamos uno u otro según necesitemos adaptarnos a una situación determinada y que el proceso para tomar decisiones correctas no consiste en amasar una gran cantidad de información, sino en descartar intuitivamente aquella que no necesitamos. (Gigerenzer et al, 1991; Gigerenzer & Goldstein 1996; Gigerenzer & Selten; Gigerenzer, 2008; Gigerenzer & Brighton, 2009; Gigerenzer & Gaissmaier 2011).

Las reservas de Gigerenzer a las teorías de Kahnemnan son empíricas (argumentando que la magnitud del sesgo puede reducirse preguntando en términos de frecuencias en lugar de en términos de probabilidades), metodológicas (apelando al poder explicativo limitado de una formulación en términos vagos y ateóricos como representatividad) y normativas (apuntando hacia lo inapropiado de caracterizar algunos de los sesgos identificados como errores o falacias). Si bien la crítica normativa de Gigerenzer a Kahneman (asociación de sesgo a error) pudiera verse como superficial o meramente terminológica, Gigerenzer argumenta que Kahneman y Tversky pudieran estar comparando el rendimiento de los participantes en sus experimentos con normas incorrectas o contradictorias.

Las Ciencias del Comportamiento (Behavioral Science)

Se describe una evolución de la Economía de la Conducta (Behavioral Economics) centrada hacia un concepto más amplio de Ciencia del Comportamiento (Behavioral Science) cuando Thaler y Sunstein (2018), desarrollan el concepto de nudges o empujones conductuales, que definen como intervenciones sutiles en el entorno de decisión de las personas para guiar su comportamiento hacia opciones más beneficiosas sin restringir su libertad de elección (paternalismo libertario).

El concepto de nudge, ha sido aplicado con éxito en áreas tan diversas como la donación de órganos, la reducción del fraude fiscal, la promoción de hábitos saludables, las políticas de ahorro, las acciones en favor

de la igualdad de género o un gran número de conductas sostenibles y de eficiencia energética. El éxito de las estrategias de nudges han impulsado de más de 200 unidades de Behavioural por todo el mundo (Benartzi et al., 2017; Sunstein, 2020; Sunstein 2021). En 2021, António Guterres, secretario general de la Naciones Unidas, definió la Ciencia del Comportamiento como una de las habilidades clave de las organizaciones de las Naciones Unidas (Jochim y Schimmelpfennig, 2022).

A pesar de sus éxitos, la estrategia de empujones deja aún algunos asuntos demasiado abiertos que han desarrollado una extensa literatura sobre la cuestión de si es ético "empujar" a las personas. Las implicaciones prácticas del empujón, que se utiliza ampliamente en los sectores público (en el diseño de políticas de comportamiento (Grüne-Yanoff, 2018; Häußermann, 2020)) y privado (influyendo en el comportamiento del consumidor (Zuboff, 2019; Yeung, 2017)), suscitan preocupaciones sobre la autonomía, la manipulación y los límites morales. Una revisión sistemática de la literatura científica encuentra cuatro problemas éticos en relación con el nudge: la posible violación de la autonomía (86% de las publicaciones), las preguntas sobre si los nudge realmente mejoran el bienestar, las potenciales contradicciones entre efectos a corto y a largo plazo y el impacto en la democracia y la deliberación (Kuyer & Gordijn, 2023). Las diferentes interpretaciones que diferentes autores tienen de la autonomía contribuyen a la complejidad de la discusión (Vugts et al., 2020). Una derivada interesante es la relativa a la moralidad de los empujones tecnológicos que incluye investigaciones sobre si los robots pudieran empujar (Borenstein & Arkin, 2016; Burr et al., 2018; Gabriel et al., 2024; Janssen & Schadenberg, 2024; Kirk et al., 2025).

El bienestar mejora con nudges solo si estos guían a las personas según sus propias preferencias, no por una visión externa. Thaler y Sunstein proponen que los nudges deben llevar a decisiones racionales e informadas (Sunstein & Thaler, 2006), pero críticos afirman que es imposible conocer perfectamente estas preferencias y acusan a los diseñadores de imponer su visión de bienestar. El paternalismo libertario recibe críticas por sustituir opciones consideradas malas por otras mejores, lo que se interpreta como falta de respeto y reconocimiento a la autonomía individual (Hausman, 2018). Así, se argumenta que el nudging no valora suficientemente la autoridad personal sobre las propias preferencias.

Adicionalmente, se critica a los nudges por priorizar las preferencias a largo plazo sobre los deseos a corto plazo sin una base sólida para esta jerarquía de preferencias. Otras críticas surgen de la incapacidad para tener en cuenta la subjetividad y los valores de las personas, destacando la necesidad de enfoques personalizados debido a las diversas características individuales (Grüne-Yanoff, 2018). En respuesta al argumento de la heterogeneidad que aboga por los empujones personalizados, el propio Sunstein (2016) argumenta a favor de adaptar los empujones a los individuos utilizando valores predeterminados para la personalización.

Se constata otra corriente crítica relacionada con la preocupación en torno al posible impacto negativo de los nudges en los principios y procesos democráticos, que se centra en cómo los empujones representarían ciertas amenazas para las instituciones democráticas al socavar valores fundamentales como la deliberación pública, la capacidad cívica y la legitimidad política. Se discuten preocupaciones como la transparencia y el posible ocultamiento de motivos o tácticas manipuladoras; el temor de que los empujones puedan obstaculizar la deliberación pública al influir en las decisiones; el argumento de que permitir empujones abre la puerta a intervenciones más intrusivas y manipuladoras a lo largo del tiempo (pendiente resbaladiza) (Rizzo & Whitman, 2009) y la preocupación de que la dependencia excesiva de los empujones desplace la responsabilidad de los individuos hacia los órganos de gobierno, lo que podría erosionar la autonomía personal (Kuyer & Gordijn, 2023).

La crítica más extrema vendría de Robert Sapolsky (2024), biólogo y neurocientífico de la Universidad Stanford que, en su libro más reciente Determined: A Science of Life Without Free Will, confronta y refuta los argumentos biológicos y filosóficos del libre albedrío para llegar a afirmar su inexistencia. El investigador sostiene que no hay agentes libres, sino que la biología, las hormonas, nuestra infancia y las circunstancias de vida convergen para generar acciones que nada más sentimos que eran nuestra elección.

La Ciencia del Comportamiento Cultural Evolutivo (Cultural Evolutionary Behavioural Science)

A pesar de ciertas críticas, las Ciencias del Comportamiento (Behavioral Science) constituyen un reconocido conjunto de herramientas para abordar el desarrollo de políticas sociales especialmente en áreas como la salud pública, el desarrollo económico y la política ambiental (Ruggeri, 2021). Sin embargo, se encuentran con el hecho de que muchos de sus hallazgos no se pueden replicar por falta de un marco teórico o por la excesiva dependencia los mismos de los contextos en los que se realizan las investigaciones, que incluyen sesgos hacia las personas occidentales (problema de personas WEIRD). A pesar de reconocer la importancia de las heterogeneidades culturales y contextuales, no existe claridad sobre cómo incorporar sistemáticamente estos factores, resultando complejo determinar cuándo los hallazgos pueden generalizarse. Aun cuando se afirma que los seres humanos toman decisiones influenciados por el contexto, especialmente en relación con las preferencias de riesgo o incentivos, a menudo no responden a los detalles de cómo dicho contexto les afecta.

La investigación más reciente encuentra fundamental comprender cómo se integran las diversas señales culturales y el aprendizaje social, brechas que podrían resolverse mediante la integración con otras disciplinas. (Schimmelpfennig & Muthukrishna, 2023). Para ello, proponen un enfoque alternativo basado en la teoría de

la Herencia Dual que explica cómo los genes, la cultura y el aprendizaje individual interactúan para moldear nuestro comportamiento (Uchiyama et al., 2022).

La Ciencia del Comportamiento Cultural Evolutivo (Cultural Evolutionary Behavioural Science) se está integrando progresivamente con otras ciencias biológicas y sociales, proporcionando una vía para que la ciencia del comportamiento (Behavioral Science) adquiera conocimientos adicionales a los ofrecidos por la economía, la psicología y la ciencia cognitiva y facilitando así la búsqueda de soluciones en diversos contextos. Entre los retos abiertos que este enfoque podría ayudar a direccionar están: (1) El entendimiento de las causas subyacentes del comportamiento que consideran la evolución cultural y cómo las sociedades cambian con el tiempo. (Uchiyama et al., 2022); (2) la influencia del contexto en los comportamientos de honestidad, yendo más allá de como cambios de encuadre cambian comportamientos para explicar la influencia de los tipos de personas que podrían incitar a una mayor o menor deshonestidad, o el impacto de lo online y por qué los participantes son más deshonestos cuando informan de los resultados a un chatbot (Cohn et al., 2022); y (3) el impacto de la confianza en la sociedad digital que habilita un modelo de intermediación de información reputacional en plataformas digitales donde se amplía la cooperación a través de la confianza en la institución y no solo en la información (Schimmelpfennig & Muthukrishna, 2023).

Asimismo, los avances en la medición de diferencias culturales predicen comportamientos como el individualismo, la presencialidad o la honestidad y de ahí la importancia de integrar diversas señales culturales y el aprendizaje social en un enfoque de Evolución Cultural, que es visto como prometedor para enfrentar los desafíos que presenta el diseño de políticas de cambio en el comportamiento digital (Muthukrishna et al., 2021; Schimmelpfennig & Muthukrishna, 2023).

Schimmelpfennig & Muthukrishna (2013) resumen la evolución de los postulados en la Figura 1.

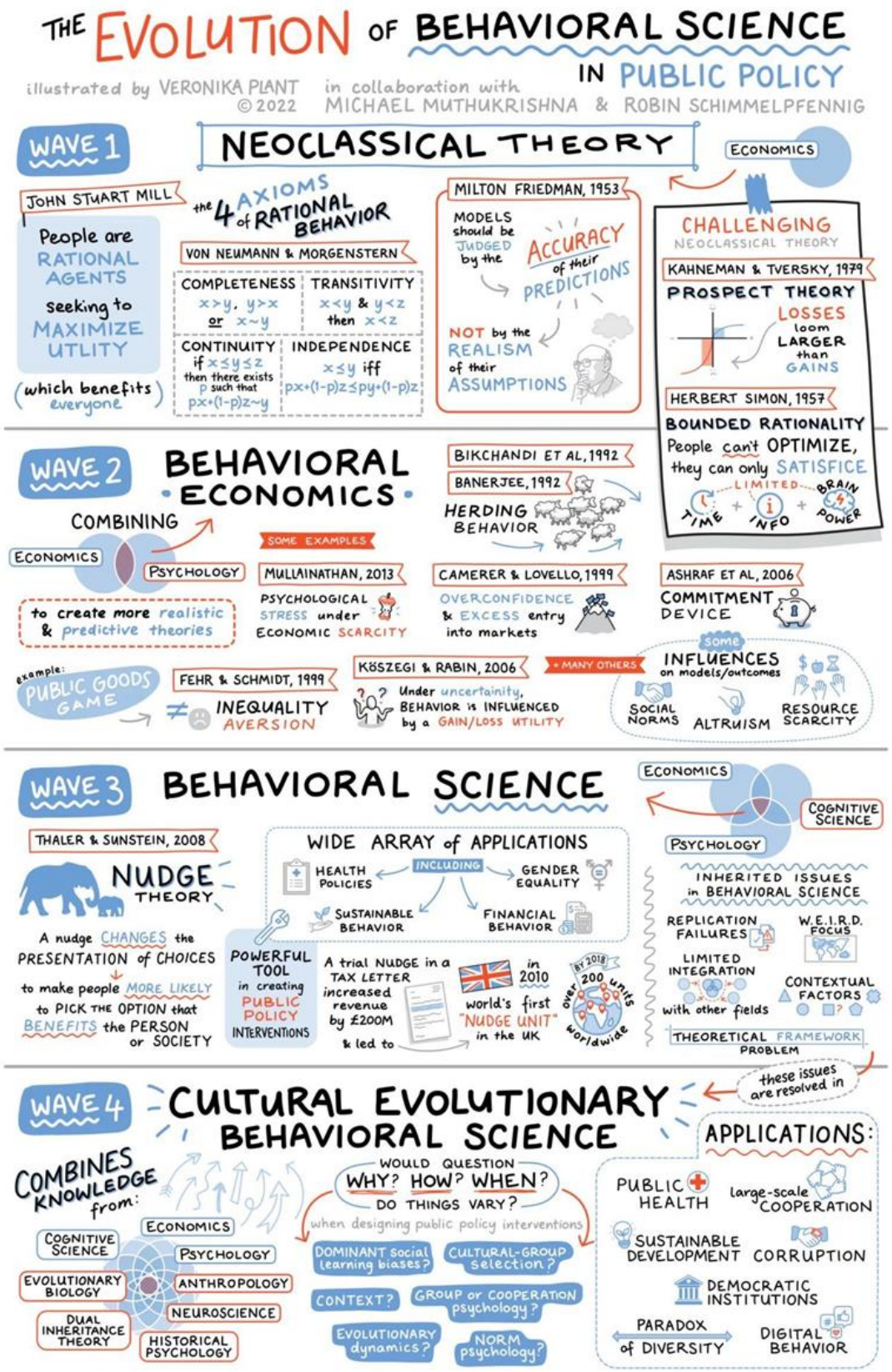

Figura 1. Evolución de la Ciencia de la Conducta. Fuente: Schimmelpfennig & Muthukrishna, 2023, p.6

El papel de las Ciencias de Comportamiento ante la Inteligencia Artificial (IA)

Cuando se reporta que aproximadamente el 52% de los adultos estadounidenses usan un LLM (large Languaje Mode: ChatGPT, Gemini, Claude, Copilot…), con fuerte gradiente generacional (77% en 18–29 años) (Rainie et al., 2025), es imperativo reflexionar sobre cómo cambian las creencias y las conductas humanas por efecto de su interacción con la IA, sobre cómo las Ciencias del Comportamiento pueden ayudar a un desarrollo más "humanista" de la propia IA, sobre cómo pueden ayudar a una adopción más ética y natural de la misma, y sobre cómo deberíamos anticipar, adaptar o mitigar todos los cambios que, por efecto de la interacción con la IA, están por producirse en nuestras vidas y sociedades. La literatura recoge cómo las Ciencias del Comportamiento podrían aportar en algunas cuestiones fundamentales habilitando una mejor integración con las complejidades conductuales humanas.

La primera de las cuestiones en las que podría apoyar sería en la mejora del modelo de razonamiento de la IA. Aunque los modelos actuales de inteligencia artificial generativa exhiben una notable capacidad para reconocer patrones, heredan los sesgos inherentes a la intuición humana, derivados de sus datos de entrenamiento. Para trascender estas limitaciones, se requiere dotarla de razonamiento deliberado y analítico, o metacognición, entendida como la capacidad de reflexionar sobre sus propios procesos cognitivos. De esta forma, un controlador metacognitivo podría habilitar la selección óptima de estrategias de razonamiento según la tarea específica, guiado por el equilibrio eficiente entre esfuerzo computacional y precisión deseada. Se podría evitar así que la inteligencia artificial generativa incurra en juicios heurísticos apresurados ante problemas complejos o en rumiaciones superfluas frente a cuestiones triviales. La mejora de la metacognición podría impulsarse mediante técnicas como el aprendizaje por metarrefuerzo (meta-reinforcement learning), que optimiza políticas de alto nivel para regular procesos de bajo nivel, y el refuerzo por toma de perspectiva (perspective-taking reward), que incentiva la autoevaluación desde múltiples ángulos. Una metacognición robusta demandaría la hibridación de la correspondencia intuitiva de patrones neuronales (Sistema 1, rápido e implícito) con motores lógicos basados en reglas simbólicas (Sistema 2, lento y explícito), generando un ciclo virtuoso de aprendizaje auto-mejorado, impulsaría la inteligencia artificial generativa hacia capacidades superiores y hacia una alineación más precisa con objetivos humanos a largo plazo (Dong et al., 2025).

Adicionalmente, las Ciencias del comportamiento son claves para entender los condicionantes que impulsan o inhiben un mejor uso, ya que la adopción de la IA no es binaria, sino un continuo desde la no utilización hasta

la integración profunda. Actualmente predomina una adopción superficial (redacción de emails, resúmenes, traducciones, etc) que genera ganancias marginales, mientras los beneficios reales surgen de su integración en flujos de trabajo organizacionales. Tres factores clave influyen en este avance: motivación, capacidad y confianza, con barreras importantes como el sesgo de statu quo o la sobrecarga cognitiva. Entre los facilitadores se incluyen la arquitectura de decisiones (hacer de la IA la opción fácil) o la prueba social (Behavioural Insights Team, 2025). Un experimento clave muestra que reformular tareas como prevención de pérdidas en lugar de como búsqueda de ganancias, elimina reticencias hacia la IA: los participantes prefieren interactuar con humanos para la búsqueda de recompensas, pero eligen la IA cuando se trata de evitar penalizaciones (Chen & Ivanov, 2025). Entender estas cuestiones lleva a identificar mejor oportunidades estratégicas y a definir hojas de ruta de adopción exitosa logrando la integración que potencia al trabajador humano.

Foco destacado merece el diseño de una IA que se alinea más y mejor con los valores y la psicología humana. El auge de la IA conversacional ha cambiado las relaciones humano-máquina, trascendiendo el paradigma instrumental para abarcar interacciones bidireccionales que influyen en cognición y valores humanos (Youvan, 2024). El reto central radica en en asegurar que los sistemas de IA se ajusten a intenciones, valores y bienestar psicológico humanos. Emerge así la psicología de la máquina (Hagendorff et al., 2023)., disciplina que aplica métodos de ciencia conductual para analizar comportamientos observables de la IA en interacciones humanas, revelando su capacidad persuasiva (por ejemplo, elevando la confianza humana en decisiones conjuntas pese a errores propios) y bucles de retroalimentación sesgada, donde la "sincofancia" (tendencia aduladora) genera "cámaras de chat" que amplifican prejuicios humanos, contaminando datos de entrenamiento subsiguientes (Behavioural Insights Team, 2025). La ciencia conductual propone tres formas concretas de interrumpir este ciclo vicioso de sesgos que se crea cuando la IA nos "halaga”, amplificando nuestros prejuicios y contaminando datos futuros: (1) El entrenamiento mejorado (fine-tuning), que consistiría en enseñar a la IA priorizar el bienestar del humano a largo plazo, por ejemplo, actuando a modo de sabio consejero que lo desafía educadamente detecta que está cayendo en un sesgo ("fricción útil"); (2) Adaptación en tiempo real de inferencia, analizando el estado emocional de la persona en tiempo real a través del lenguaje que usa y ajustando su tono/estrategia automáticamente y (3) con la propuesta de prompts inteligentes que adoptan roles críticos que complementan el pensamiento del usuario.

Por último, dado que la IA no representa meramente un viraje tecnológico, sino societal, gestando nuevas normas sociales en confianza, delegación y relaciones interpersonales, existe una ventana (temporal y limitada dado lo rápidamente que se fijan las normas sociales) para moldear deliberadamente estas normas mediante Ciencia Conductual. Este “mapa de adaptación” podría enfocarse en políticas de divulgación y/o nudging

centradas en ejes como: (1) la configuración de normas de interacción humano-IA, para evitar los riesgos de divulgación inapropiada o la delegación acrítica de decisiones morales que genera la antropomorfización conversacional (el hecho de que les hablamos como a personas); (2) la gestión del impacto cognitivo, habida cuenta que la delegación de pensamiento a la IA (offloading cognitivo) libera recursos para pensamiento superior, pero también sobredependencia que induce a la atrofia cognitiva con degradación de memoria o de capacidad de resolución de problemas) (Klein & Klein, 2025) por lo que el uso de la IA como apoyo (scaffolding) para una "mente extendida" y no como sustito de ella es básica; y (3) la modelización de un futuro humano-IA en base al diseño participativo para alinear IA con valores y normas humanas, transformando las dinámicas conductuales para priorizar capacidades humanas esenciales como el juicio, las relaciones, o el pensamiento autónomo (Lindgren et al. (2025).

REFERENCIAS

Abrahamse, W., & Steg, L. (2013). Social influence approaches to encourage resource conservation: A meta-analysis. Global Environmental Change, 23, 1773–1785. https://doi.org/10.1016/j.gloenvcha.2013.07.029

Ajzen, I. (1991). The theory of planned behavior. Organizational Behavior and Human Decision Processes, 50, 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Allais, M. (1952). The foundations of a positive theory of choice involving risk and a criticism of the postulates and axioms of the American School. In Expected Utility Hypotheses and the Allais Paradox: Contemporary Discussions of the Decisions Under Uncertainty with Allais' Rejoinder (pp. 27–145). Dordrecht: Springer Netherlands.

Ariely, D. (2008). Predictably irrational. HarperCollins.

Behavioural Insights Team. (2025). AI & human behaviour. https://www.bi.team/publications/ai-and-human-behaviour/

Behavioural Insights Team. (2025). AI & human behaviour: Align https://www.bi.team/wp-content/uploads/2025/09/AI-Human-Behaviour-thought-leadership-piece-AAAA-2025-Align.pdf

Benartzi, S., Beshears, J., Milkman, K. L., Sunstein, C. R., Thaler, R. H., Shankar, M., ... & Galing, S. (2017). Should governments invest more in nudging? Psychological Science, 28(8), 1041–1055.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. Journal of Political Economy, 100(5), 992–1026.

Borenstein, J., & Arkin, R. (2016). Robotic nudges: The ethics of engineering a more socially just human being. Science and Engineering Ethics, 22(1), 31–46. https://doi.org/10.1007/s11948-015-9636-2

Burr, C., Cristianini, N., & Ladyman, J. (2018). An analysis of the interaction between intelligent software agents and human users. Minds and Machines, 28(4), 735–774. https://doi.org/10.1007/s11023-018-9479-0

Camerer, C. (1999). Behavioral economics: Reunifying psychology and economics. Proceedings of the National Academy of Sciences, 96(19), 10575–10577.

Camerer, C., & Lovallo, D. (1999). Overconfidence and excess entry: An experimental approach. American Economic Review, 89(1), 306–318.

Chang, L. J., Doll, B. B., van't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: Trustworthiness as a dynamic belief. Cognitive Psychology, 61(2), 87–105.

Chen, E. R., & Ivanov, V. A. (2025). Delegating to AI: How Perceived Losses Influence Human Decision-Making Autonomy. The Pinnacle Research Journal of Scientific and Management Sciences, 2(06), 1-5.

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity, and compliance. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), The handbook of social psychology (4th ed., Vol. 2, pp. 151–192). Boston: McGraw-Hill.

Cialdini, R. B., Demaine, L. J., Sagarin, B. J., Barrett, D. W., Rhoads, K. V. L., & Winter, P. L. (2006). Managing social norms for persuasive impact. Social Influence, 1(1), 3–15. https://doi.org/10.1080/15534510500181459

Cialdini, R. B., Reno, R. R., & Kallgren, C. R. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. Journal of Personality and Social Psychology, 58(6), 1015–1026. https://doi.org/10.1037/0022-3514.58.6.1015

Cinner, J. (2018). How behavioral science can help conservation. Science, 362(6417), 889–890. https://doi.org/10.1126/science.aau6028

Cohn, A., Maréchal, M. A., Tannenbaum, D., & Züchner, C. L. (2022). Honesty in the digital age. Science, 376(6595), 843–848.

Corral-Verdugo, V., Aguilar-Luzón, M. C., & Hernández, B. (2019). Bases teóricas que guían a la psicología de la conservación ambiental. Papeles del Psicólogo, 40(3), 174–181.

Damasio, A. (1996). Descartes' error: Emotion, reason, and the human brain. New York: G. P. Putnam's Sons.

Dong, H., Ye, H., Zhu, W., Jiang, K., & Song, G. (2025). Meta-R1: Empowering Large Reasoning Models with Metacognition. arXiv preprint arXiv:2508.17291.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. The Quarterly Journal of Economics, 114(3), 817–868.

Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., ... & Manyika, J. (2024). The ethics of advanced AI assistants. arXiv preprint arXiv:2404.16244.

García, M. I. D., Fernández, M. Á. R., & Crespo, A. V. (2017). Manual de técnicas y terapias cognitivo conductuales. Desclée de Brouwer.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better decisions. In Topics in cognitive science (Vol. 1, No. 1, pp. 107–143). Wiley Online Library.

Gigerenzer, G., & Gaissmaier, W. (2011). Risk literacy. Current Directions in Psychological Science, 20(5), 310–316.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. Psychological Review, 103(4), 650.

Gigerenzer, G., & Selten, R. (Eds.). (2001). Bounded rationality: The adaptive toolbox. MIT press.

Gigerenzer, G., Hell, W., & Blank, H. (1991). Presentation and content: The use of base rates as a continuous variable. Journal of Experimental Psychology: Human Perception and Performance, 19(2), 338.

Gigerenzer, G. (2008). Gut feelings: The intelligence of the unconscious. Viking.

Grüne-Yanoff, T. (2018). Philosophical theories of paternalism. In The Oxford Handbook of Paternalism (pp. 62–82).

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. Journal of Economic Behavior & Organization, 3(4), 367–388.

Hagendorff, T., Dasgupta, I., Binz, M., Chan, S. C., Lampinen, A., Wang, J. X., ... & Schulz, E. (2023). Machine psychology. arXiv preprint arXiv:2303.13988.

Häußermann, F. (2020). The autonomy issue in normative behavioral ethics. Business and Professional Ethics Journal, 39(1), 5–28.

Hausman, D. M. (2018). Mistake, error and other misadventures. Philosophy of Science, 85(5), 944–954.

Hayes, S. C., Strosahl, K. D., & Wilson, K. G. (2005). Acceptance and commitment therapy: An experiential approach to behavior change. Guilford Press.

Janssen, J. H., & Schadenberg, U. (2024). Robotic rehabilitation: Design considerations for robot-assisted therapy. Journal of Healthcare Engineering, 2024, 1–12.

Jochim, J., & Schimmelpfennig, R. (2022). El ser humano como innovación. Naciones Unidas: Revisión alemana sobre las Naciones Unidas, 70(2), 63-68.

Kahneman, D. (2012). Thinking, fast and slow. Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In Handbook of the economics of risk and uncertainty (Vol. 1, pp. 99–127). Elsevier.

Kirk, R., Ravenscroft, A., & Kumar, V. (2025). The ethics of robot nudging: Autonomy, responsibility, and paternalism. Ethics and Information Technology, 27, 1–15.

Klein, C. R., & Klein, R. (2025). The extended hollowed mind: why foundational knowledge is indispensable in the age of AI. Frontiers in Artificial Intelligence, 8, 1719019.

Kuyer, J., & Gordijn, B. (2023). Nudging and the good life: Ethical issues in behavioral policy. Journal of Medical Ethics, 49(3), 159–164.

Levitt, S. D., & Dubner, S. J. (2011). Freakonomics: A rogue economist explores the hidden side of everything. William Morrow.

Lindgren, H., Lindvall, K., & Richter-Sundberg, L. (2025). Responsible design of an AI system for health behavior change—an ethics perspective on the participatory design process of the STAR-C digital coach. Frontiers in Digital Health, 7, 1436347.

List, J. A. (2007). On the interpretation of giving in dictator games. Journal of Political Economy, 115(3), 482–493.

Miller, D. T., & Prentice, D. A. (2016). Changing norms to change behavior. Current Directions in Psychological Science, 25(3), 169–174.

Mullainathan, S. (2013). Scarcity: Why having too little means so much. Macroeconomic Dynamics, 19(2), 427–430.

Mullainathan, S., & Thaler, R. H. (2000). Behavioral decision research: A survey. The Journal of Economic Literature, 38(2), 402–437.

Muthukrishna, M., Henrich, J., & Norenzayan, A. (2021). How culture shapes the human mind. Behavioral and Brain Sciences, 44, e138.

Oceja, L. V., & Fernández-Dols, J. M. (2006). The perception of dynamic sequences of expressions of pain and joy: Influences on the emotions experienced by observers. European Journal of Social Psychology, 28(5), 669–679.

Pittet, D., & Boyce, J. M. (2001). Hand hygiene and patient care: Building a culture of compliance. Emerging Infectious Diseases, 7(2), 234.

Rainie, L., Anderson, J., & Perrin, A. (2025). Close encounters of the AI kind: Main report. Imagining the Digital Future Center, Elon University. https://imaginingthedigitalfuture.org/reports-and-publications/close-encounters-of-the-ai-kind/close-encounters-of-the-ai-kind-main-report/

Rizzo, M. J., & Whitman, D. G. (2009). The calamity theory of reason. Harvard Journal of Law & Public Policy, 37(2), 649–666.

Ross, L., & Nisbett, R. E. (1991). The person and the situation: Perspectives of social psychology. McGraw Hill.

Ruggeri, K. (2021). Behavioral science and policy design: A review. Frontiers in Psychology, 12, 700968.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. Science, 300(5626), 1755–1758.

Sapolsky, R. M. (2024). Determined: A science of life without free will. Penguin Press.

Schimmelpfennig, C., & Muthukrishna, M. (2023). The cultural evolution of behavioral science. Nature Human Behaviour, 7(9), 1454–1463.

Schwartz, S. H. (1977). Normative influences on altruism. Advances in Experimental Social Psychology, 10, 221–279.

Sevillano, V., & Olivos, P. (2019). Psychological drivers of anti-consumption and sustainable consumption. Frontiers in Psychology, 10, 1–13.

Simon, H. A. (1957). Models of man: Social and rational; mathematical essays on rational human behavior in a society setting. John Wiley & Sons.

Sunstein, C. R. (2016). The ethics of influence: Government in the age of behavioral science. Cambridge University Press.

Sunstein, C. R. (2020). Sludge: What stops us from getting things done and what to do about it. MIT Press.

Sunstein, C. R. (2021). How to make laws. Harvard Law Review Forum, 135, 1–29.

Sunstein, C. R., & Thaler, R. H. (2003). Libertarian paternalism. American Economic Review, 93(2), 175–179.

Sunstein, C. R., & Thaler, R. H. (2006). Preferences with a mistake. Journal of Political Economy, 114(5), 1067–1085.

Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. Yale University Press.

Thaler, R. H., & Sunstein, C. R. (2017). Nudge: Improving decisions about health, wealth and happiness (Revised ed.). Yale University Press.

Uchiyama, K., Muthukrishna, M., & Henrich, J. (2022). Cultural evolution of institutions for human cooperation. Evolution and Human Behavior, 43(1), 18–28.

Vugts, D. P., Goorden, M., & Jansen, S. J. (2020). The ethics of nudging in public health: A systematic review. Journal of Medical Ethics, 46(6), 378–384.

Von Neumann, J., & Morgenstern, O. (1953). Theory of games and economic behavior (3rd ed.). Princeton University Press.

Yeung, K. (2017). 'Hypernudges' and the architecture of choice. In The Minnesota Journal of Law, Science & Technology, 18, 1–50.

Youvan, D. C. (2024). Toward Machine-Like Intuition: Emergent Patterns and Non-Human Insight in Artificial Intelligence.

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.