INTELIGENCIA ARTIFICIAL Y CONFIANZA A TRAVÉS DE UNA COMUNICACIÓN CIENTÍFICA COMPRENSIBLE: UN COMPROMISO ÉTICO CON LA CIUDADANÍA

ARTIFICIAL INTELLIGENCE AND TRUST THROUGH UNDERSTANDABLE SCIENTIFIC COMMUNICATION: AN ETHICAL ENGAGEMENT WITH THE PUBLIC

Antonio luis Terrones Rodríguez
Universitat de València / Instituto de Filosofía-CSIC[*]

Resumen: El enorme potencial del desarrollo de la inteligencia artificial (IA) puede observarse en un aumento de su presencia en diversos entornos para ofrecer soporte en actividades que destacan por un alto nivel de complejidad. En ese sentido, instituciones como la Comisión Europea (CE) advierten la necesidad de promover un cultivo de la confianza en la IA para garantizar una gobernanza que sitúe al ser humano en el centro de sus preocupaciones y beneficios. Un cultivo que implica un compromiso con el fortalecimiento de la habilidad comunicativa de los expertos científicos como una estrategia para favorecer la comprensión ciudadana y su participación. Una comunicación científica preocupada por el principio ético de la explicabilidad implica poner en valor la interacción con la ciudadanía para reconocer sus testimonios y experiencias de vida, incorporándolos en los procesos de investigación e innovación.

---

* Investigador posdoctoral de la Universitat de València. Realizó una estancia de investigación en el Grupo de Ética Aplicada (GEA) del Instituto de Filosofía del CSIC durante los años 2022 y 2023. Email: Antonio.Terrones@uv.es. Esta investigación cuenta con la financiación del Ministerio de Universidades del Gobierno de España y la Unión Europea (Next Generation EU) en el marco de las Ayudas Margarita Salas para la formación de jóvenes doctores del programa de recualificación del sistema universitario español. Además, esta publicación es parte del proyecto TED2021-131295B-C31, Ingeniería de Valores de IA-VAE, financiado por MCIN/AEI/10.13039/501100011033 y la Unión Europea Next Generation EU/PRTR.

PALABRAS CLAVE: Inteligencia artificial confiable, explicabilidad, ciencia, comunicación, ética, ciudadanía.

ABSTRACT: The enormous potential of the development of artificial intelligence (AI) can be seen in its increased presence in various environments to offer support for activities that stand out for a high level of complexity. In this sense, institutions such as the European Commission (EC) warn of the need to cultivate trust in AI to guarantee a governance that places the human being at the center of its concerns and benefits. A culture that implies a commitment to strengthening the communications skills of scientific experts as a strategy to promote citizens' understanding and participation. A scientific communication concerned with the ethical principle of explainability implies valuing interaction with citizens to recognize their testimonies and life experiences, incorporating these into research and innovation processes.

KEYWORDS: Trustworthy artificial intelligence, explainability, science, communication, ethics, citizenship.

1. Introducción

El enorme potencial del desarrollo de la inteligencia artificial (IA)[1] puede observarse en el auge de los sistemas automatizados para la toma de decisiones. Los intelectos sintéticos están cada vez más presentes en entornos complejos para ofrecer soporte en actividades que destacan por un alto nivel de complejidad, como la medicina, el derecho o la economía. En ese sentido, instituciones como la Comisión Europea (CE) advierten la necesidad de promover un cultivo de la

[1] No existe una única definición de IA, pues los investigadores observan esta tecnología desde diferentes perspectivas. La expresión IA fue introducida en la década de los 50 partir de lo acontecido en el Dartmouth Summer Research Project on Artificial Intelligence (McCorduck, 1991). Una forma de describir la IA que puede servir como aproximación, es la definición ofrecida por Margaret A. Boden: "La inteligencia artificial (IA) tiene por objeto que los ordenadores hagan la misma clase de cosas que puede hacer la mente. Algunas (como razonar) se suelen describir como «inteligentes». Otras (como la visión), no. Pero todas entrañan competencias psicológicas (como la percepción, la asociación, la predicción, la planificación, el control motor) que permiten a los seres humanos y demás animales alcanzar sus objetivos. La inteligencia no es una dimensión única, sino un espacio profusamente estructurado de capacidades diversas para procesar la información. Del mismo modo, la IA utiliza muchas técnicas diferentes para resolver una gran variedad de tareas. [...] La IA tiene dos objetivos principales. Uno es tecnológico: usar los ordenadores para hacer cosas útiles (a veces empleando métodos muy distintos a los de la mente). El otro es científico: usar conceptos y modelos de IA que ayuden a resolver cuestiones sobre los seres humanos y demás seres vivos (2017: 11-12).

confianza en la IA para garantizar una gobernanza que sitúe al ser humano en el centro de sus preocupaciones y beneficios. Este cultivo exige un compromiso previo con la sociedad por parte los expertos del ámbito de la IA y sus gestores técnicos y políticos. Pueden mencionarse tres de razones de peso para justificar tal compromiso: en primer lugar, que los afectados por los efectos de la IA deben tener voz en los procesos de investigación e innovación, como muestra de reconocimiento del valor de sus testimonios y mecanismo de inclusión y fortalecimiento democrático para promover la legitimidad de los productos de esta tecnología; en segundo lugar, que gran parte de los fondos que recibe la investigación en IA son de naturaleza pública; y, en tercer lugar, una razón instrumental que tiene que ver con la participación y la anticipación, pues resulta útil contar con varios testimonios para valorar los efectos de esta tecnología, incrementar su apoyo público, promover la formación cívica sobre asuntos que puedan resultar de interés y reducir la probabilidad de errores (Fiorino, 1990).

Un aspecto muy importante de la participación pública y responsable en la IA para el cultivo de la confianza implica un compromiso con el establecimiento de un hilo comunicativo con la ciudadanía como estrategia de comprensión en torno a aspectos que podrían ser relevantes para su bienestar. En ese sentido, tanto el reconocimiento de la necesidad de contar con la ciudadanía para el desarrollo de una IA europea, como la construcción de un correcto hilo comunicativo, son dos aspectos fundamentales para la confianza. Illah Nourbakhsh (2009) señala la relevancia moral de la retórica empleada por los investigadores de robótica, pudiéndose trasladar este aspecto también a la IA. Para Nourbakhsh el discurso de los investigadores y el modo de interactuar con la sociedad, son factores a tener en cuenta para el establecimiento de una correcta comunicación científica y comprensión ciudadana.

Una comunicación científica responsable e interesada en la explicabilidad implica poner en valor la interacción con la ciudadanía para reconocer sus testimonios y experiencias de vida, incorporándolos en los procesos de investigación e innovación. A modo de ejemplo, diversos estudios (Palmisciano, P., et al. 2020; Vasiljeva, T., Kreituss, I. y Lulle, I., 2021; Young MD., et al., 2021; Layard Horsfall, et al., 2021) sobre las actitudes del público hacia la IA apuntan al valor que posee la incorporación de los testimonios de los afectados en los procesos de investigación e innovación. Un claro ejemplo del papel que juega la comunicación para influir en la evaluación tecnológica, puedes observarse en la ciencia ficción (Miller y Bennet, 2008; Stephenson, 2011).

El objetivo de este trabajo consiste en justificar la necesidad del fortalecimiento de la comunicación social de la ciencia y su interacción con la ciudadanía como una estrategia de defensa del principio ético de explicabilidad y una garantía para el cultivo de confianza en el ámbito de la IA. Pues a pesar de asistir a un momento histórico con altos niveles de conexión, no se garantiza una comunicación efectiva entre las personas (Domingo Moratalla, 2021). A menudo, los desarrolladores de IA realizan su trabajo de un modo endogámico, destacando exclusivamente aspectos técnicos, en lugar de favorecer la comprensión ciudadana. Este hecho provoca un distanciamiento entre los expertos y los usuarios, y por ello resulta esencial el tratamiento en la esfera pública de aquellos conocimientos problemáticos y controvertidos para responder a la articulación de un discurso público en el que participan diversos grupos de interés en un escrutinio ético (Joss, 2002). Si bien el interés por la explicabilidad es positivo para la adquisición de confianza, en este trabajo se argumentará que para su garantía es fundamental fortalecer las habilidades comunicativas de la ciencia en un contexto democrático.

2. El interés de la Comisión Europea en la confianza y la innovación social

En junio de 2018 la Comisión Europea (CE) constituyó el Grupo de expertos de alto nivel sobre IA, tras haber anunciado el mismo año la voluntad de impulsar un grupo de trabajo para abordar cuestiones éticas de esta tecnología. De ese modo recogían el testigo del Grupo Europeo de Ética en Ciencia y Nuevas Tecnologías (EGE) (2018). Más tarde, en 2019, publicó el documento *Ethics guidelines for trustworthy AI*, donde ofrecían un conjunto de siete requisitos claves que los sistemas de IA deben cumplir para ser considerados confiables: 1) acción y supervisión humanas, 2) solidez técnica y seguridad, 3) gestión de la privacidad y de los datos, 4) transparencia, 5) diversidad, no discriminación y equidad, 6) bienestar ambiental y social, y 7) rendición de cuentas. Además, cuatro principios éticos que se encuentran arraigados en los derechos fundamentales: respeto a la autonomía humana, prevención del daño, equidad y explicabilidad. A estas propuestas hay que sumarles el *Libro Blanco sobre inteligencia artificial* (CE, 2020), donde se muestra un claro interés por un ecosistema de confianza.

Para el Grupo de Expertos de Alto Nivel la explicabilidad es un principio ético crucial para el cultivo de confianza en la sociedad, con respecto a la IA. Se sostiene sobre la necesidad de comprensibilidad y transparencia de los procesos e implica comunicar abiertamente las capacidades y finalidades de los sistemas de IA a las partes afectadas (Comisión Europea, 2019: 16). La CE ha demostrado en los últimos años un importante interés por el impulso de una IA «made in Europe» que exige la construcción de unos cimientos que permitan transitar hacia un ecosistema de confianza. También es importante la referencia explícita a la transparencia que se hace en el Reglamento General de Protección de Datos (GDPR) del Parlamento Europeo (PE) y el Consejo de la Unión Europea (2016: 11).

En sintonía con el interés de la CE en una IA confiable y la importancia que se le concede al principio ético de explicabilidad, se encuentra Horizonte 2020 (H2020) (Comisión Europea, 2015), el Programa Marco de Investigación e Innovación de la Unión Europea. La CE adquirió el compromiso de fortalecer sus sociedades a través del impulso de estrategias de inclusión, innovación y reflexión en un escenario de constantes transformaciones en el marco de la globalización. Europa se enfrenta a importantes retos que exigen un compromiso por parte del conocimiento científico.

El H2020 se despliega en una serie de esferas de interés entre las que se encuentra «Ciencia con y para la Sociedad», donde se plantea una Investigación e Innovación responsables (RRI, siglas en inglés de *Responsible Research and Innovation*)[2]. Estas siglas han adquirido una relevancia cada vez mayor en las políticas de la UE en los últimos años. Un modelo apoyado por la CE que aboga por una nueva relación entre la sociedad, la investigación y la innovación (von Schomberg, 2011). Como reza en su web, «implica que los actores de la sociedad (investigadores, ciudadanos, responsables políticos, empresas, organizaciones del tercer sector, etc.) trabajen juntos durante todo el proceso de investigación e innovación para alinear mejor tanto el proceso como sus resultados con valores, necesidades y expectativas de la sociedad» (CE, 2015). Consiste en un enfoque novedoso para la gobernanza de la ciencia, la investigación y la innovación que persigue ampliar los procesos de toma de decisiones para lograr procesos

[2] Existen cuatro relatos sobre el modelo RRI que han dominado las investigaciones desde su surgimiento, estos son: Máire Geoghegan-Quinn (2012), actual comisaria europea de Investigación, Innovación y Ciencia, Rene Von Schomberg (2013), Stilgoe et al. (2013) y Jeroen van den Hoven (2014).

éticamente aceptables, socialmente deseables y sostenibles. El empeño demostrado en los últimos años por parte de la CE en torno a una RRI destaca una preocupación por la gobernanza de la ciencia, la investigación y la innovación. En ese sentido, la RRI nace de la voluntad de intervenir en las primeras etapas del proceso I+D+i, en lugar de dirigir la mirada hacia los riesgos y beneficios. Se basa en formas deliberativas de gobernanza y busca democratizar la I+ D mediante la participación de los grupos de interés y la sociedad. De esta forma se promueve una responsabilidad compartida en la que diversos actores controlan y dirigen la I+D+i en busca de un equilibrio entre la aceptabilidad y la deseabilidad en aras de la integración social (Timmermans, 2020).

Además de centrar su interés en la participación de diversos actores, en el diseño sensible a los valores democráticos, la ética, el equilibrio entre la aceptabilidad y la deseabilidad o la gobernanza, la RRI destaca por abordar el valor público de la ciencia (Fisher y Rip, 2013). Asumiendo que es necesario fortalecer el vínculo entre la ciencia y la ciudadanía en torno a sus necesidades, pues a veces sus objetivos no se comprenden completamente, incluso cuando son bien intencionados. Los cambios sucedidos en los últimos años y, en ocasiones, las carencias de un hilo comunicativo entre la ciencia y la sociedad, son dos razones que han aumentado la preocupación de la sociedad en los asuntos científicos y a veces han contribuido a una cierta desconfianza sobre los intereses de la ciencia. En este trabajo se sitúa el foco de atención en la relación entre la comunicación social de la ciencia, la implicación ciudadana y la confianza.

Las actitudes comunicativas de la ciencia se construyen sobre las racionalidades políticas que Cecilie Glerup y Maja Horst (2014) han diagnosticado en torno a la responsabilidad de la ciencia hacia la sociedad: Demarcación, Reflexividad, Contribución e Integración. Estas racionalidades responden a los debates sobre la responsabilidad de la ciencia que vienen planteándose desde la segunda mitad del siglo XX. Steven Shapin (2008) describe cómo los científicos trasladaron sus discusiones a la esfera pública sobre las implicaciones morales existentes en torno al desarrollo y uso de bombas nucleares. Unas discusiones que siguen estando presentes en la actualidad, pero contextualizadas en nuevas problemáticas (McLeish y Nightingale, 2007). Además, estos debates favorecieron en la década de los años 70 la inspiración para comenzar a crear nuevas organizaciones gubernamentales dedicadas a la evaluación tecnológica y el asesoramiento político (Bimber, 1996).

En la senda del diagnóstico establecido por Glerup y Horst (2014), en el modelo RRI se encuentran presentes elementos de las racionalidades de Contribución e Integración. Para la primera, la ciencia debe articularse como parte de la sociedad y sirve a determinados fines societarios:

> En esta racionalidad, una visión particular de lo que es bueno para la sociedad es inherente a los fines específicos que persigue la ciencia. De acuerdo con la racionalidad de la Contribución, es primordial que la sociedad tenga un papel decisivo en la configuración de estas visiones y objetivos, y que los científicos se vean a sí mismos trabajando para producir una contribución valiosa a la sociedad (2014: 39).

Esta racionalidad se sostiene sobre dos prescripciones sociales: la primera, que la ciencia debe ser innovadora y aportar conocimientos y tecnologías para mejorar el crecimiento nacional y regional; y la segunda, que las actividades científicas deben estar en sintonía con las preferencias públicas expresadas por la ciudadanía y sus instituciones, debiendo estar sujetas al escrutinio público (Glerup y Horst, 2014: 40). En esta racionalidad los científicos tienen la responsabilidad de establecer un hilo comunicativo con la ciudadanía para que pueda configurar su perspectiva. Para esta racionalidad la ciencia está al servicio de la sociedad.

En cuanto a la racionalidad de la Integración, se construye sobre la premisa del trabajo conjunto de los actores de la ciencia y la sociedad, pues apunta a la falta de integración de otros actores que cuentan con legitimidad epistémica y política. A diferencia de la racionalidad de la Contribución, que promueve un resultado concreto, la Integración no define un objetivo social fijo. Los objetivos de la ciencia y la sociedad serán el resultado de un proceso en el que diversos actores acuerden juntos sus preferencias. En esta racionalidad la deliberación juega un papel muy importante.

La comunicación social de la ciencia es un elemento fundamental de estas dos racionalidades que configuran la gobernanza. En ese sentido, el interés de la CE en un IA confiable y en el modelo RRI como una herramienta para la innovación social, exige un modus operandi de la comunidad científica que incorpore la voluntad de comunicar con mayor interactividad. Para promover un ecosistema confiable de IA la actividad científica debe plantearse a partir de una comunicación interactiva para favorecer la comprensión y participación de la ciudadanía en asuntos de interés público. Por ello son importantes el cultivo de habilidades cívicas y comunicativas, la construcción de puentes de diálogo y la participación entre los grupos de interés para situar los problemas y controversias de la IA en la esfera pública para su deliberación.

3. La importancia de la explicabilidad

La explicabilidad se ha convertido en un principio muy habitual en las investigaciones sobre ética de la IA (Jobin *et al.*, 2019). Cuando la CE señala que el desarrollo de la IA debe estar fundamentado en el principio ético de explicabilidad, está sugiriendo la necesidad de ofrecer posibilidades de comprensión a la ciudadanía, es decir, de una comunicación comprensible sobre su diseño y desarrollo. El interés por la explicabilidad favorece la comprensión, inspección y reproducción de las decisiones tomadas y los datos empleados en los sistemas inteligentes.

Cuando se habla de explicabilidad en el contexto de la IA a menudo se hace referencia a la transparencia y su relación con la comprensión de conocimientos como una condición sine que non para adquirir confianza (Ribeiro, Singh & Guestrin, 2016). Además, la evaluación de la confianza de los usuarios en la aplicación de la IA implica una suposición habitual formulada en la literatura reciente (Miller, 2019), a saber, que la importancia de la transparencia exige tomar en consideración cómo los humanos entienden las explicaciones y cómo evalúan su relación con un determinado servicio o producto. En ese sentido, es evidente que el cultivo de confianza en el ámbito de la IA se encuentra estrechamente vinculado con el principio ético de la explicabilidad. Sin embargo, la exigencia de explicabilidad no está fundamentada en teorías concretas de comunicación social de la ciencia en un contexto democrático, corriendo el peligro de satisfacer únicamente a los desarrolladores de la IA y aumentando la brecha entre quienes crean, regulan y emplean esta tecnología (Miller, Howe y Sonenberg, 2017). Por lo tanto, se debe explorar la posibilidad de fortalecer las habilidades comunicativas y su relación con un modelo comunicativo que favorezca la comprensión ciudadana y persiga el cultivo de la confianza.

El principio de explicabilidad debería implicar la exigencia de compromiso a los expertos para que adapten sus conocimientos a las características y necesidades de la ciudadanía, esclareciendo las intenciones científicas que fundamentan el desarrollo de la IA. Esta adaptación se reflejaría en una comprensión y facilitación de conocimientos interactiva con la ciudadanía, donde los deseos y valores están muy presentes. En ese sentido, para que la información sea abierta y transparente se requiere que las personas afectadas sepan evaluar los riesgos de la IA, ofreciéndoles la posibilidad de cuestionar las decisiones de los expertos, promoviendo una crítica razonable en los entornos científicos. Así pues, la atención a los aspectos

específicos de la tecnología, el contexto en el que se despliega y las perspectivas de la ciudadanía afectada son elementos que hay que considerar a la hora de asegurar la explicabilidad.

Un estudio reciente donde la mitad de los participantes desconocían que un algoritmo está implicado en el suministro de noticias de Facebook, planteó una herramienta con la que se contrastaban noticias filtradas y sin filtrar por parte del algoritmo (Eslami *et al*., 2015). Este experimento puso de relieve la existencia de aspectos positivos de la explicabilidad cuando se establece una comunicación interactiva con los usuarios, situando cuestiones científicas en la esfera pública para su deliberación.

AI4People es el primer foro impulsado por el PE que reúne a grupos de interés para valorar el impacto social de las nuevas aplicaciones de IA. Su objetivo principal consiste en la creación de un espacio público común para establecer los principios éticos fundamentales, las políticas y las prácticas que pueden contribuir en la construcción de una buena sociedad de IA. Luciano Floridi *et al*. elaboraron el documento *An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations* donde se presenta una síntesis de cinco principios éticos que deben adoptarse para el desarrollo de la IA y una serie de recomendaciones. Basándose en propuestas ya existentes (*Asilomar AI Principles*, 2017; *Montreal Declaration for a Responsible Development of Artificial Intelligence*, 2017; *IEEE Initiative on Ethics of Autonomous and Intelligent Systems*, 2017; *European Group on Ethics in Science and New Technologies*, 2018; *Partnership on AI*, 2018) ofrecen cinco principios éticos para la IA: beneficencia, no maleficencia, autonomía, justicia y explicabilidad. Los cuatro primeros son ya conocidos en el ámbito de la bioética, mientras que el de explicabilidad representa una nueva incorporación a su propuesta principialista.

Este foro muestra su interés en el principio ético de la explicabilidad de la misma manera que anteriormente lo hizo el Grupo de Expertos de Alto Nivel. AI4People señala que la motivación para el planteamiento de este principio radica en la situación de desigualdad existente entre una minoría de expertos de IA y una mayoría social que desconoce los aspectos relativos a su diseño y desarrollo. Se interesa en este principio en un doble sentido: epistemológico, de «inteligibilidad»; y ético, de «rendición de cuentas» (2018: 700). Así pues, destacan su preocupación por un desarrollo de la IA que siente unas bases fácilmente comprensibles para la ciudadanía.

La IA es extraordinariamente compleja, pues las decisiones que lleva a cabo son específicas del contexto e incorporan miles o millones de factores. Esto supone que gran parte del comportamiento de la IA resulte extraño, inesperado y errático en muchos casos, como puede observarse en predicciones erráticas en el reconocimiento facial. En ese sentido, para garantizar la confianza en los sistemas inteligentes es importante poner en valor la explicabilidad, aunque resulta difícil definir en qué consiste. Daniel S. Weld y Gagan Bansal (2018) han mostrado su interés en la problemática de la explicabilidad, destacando que una posible estrategia puede consistir en hacer que el sistema de explicación interaccione con los usuarios para que puedan profundizar hasta satisfacer la comprensión. A su vez, reconocen que la clave del desafío para diseñar una IA explicable se origina en la complejidad de la comunicación del proceso computacional, señalando como requisito habilidades interdisciplinarias. Si bien estos investigadores muestran su interés por la explicabilidad de la IA, lo hacen desde una perspectiva principalmente técnica, haciendo un especial hincapié en modelos de funcionamiento computacional.

Por su lado, Freddy Lecué y Jiewen Wu (2018) afirman que la mayoría de los enfoques existentes se centran en la explicación basada en datos y carecen de interpretación semántica, lo que supone restar importancia a las explicaciones centradas en el ser humano. Proponen un enfoque que explota la semántica de datos representativos. En primer lugar, sugieren la selección de datos representativos y la elaboración de límites de decisión de los clasificadores; en segundo lugar, la extracción y codificación de la semántica de estos datos utilizando ontologías de dominio; y finalmente computación de las explicaciones informativas basadas en la optimización de ciertos criterios aprendidos de las explicaciones diarias de los humanos (2018: 2). Esta propuesta combina el razonamiento semántico y el *Machine Learning* (ML) mediante la revisión de los límites de las decisiones y sus elementos más representativos.

Reconociendo el valor de las diferentes aportaciones en torno a las posibilidades de explicabilidad en el ámbito de la IA, entiendo la necesidad de llevar a cabo una propuesta concreta desde una perspectiva cívica de la comunicación científica. A continuación, se planteará la exploración de un camino centrado en la comunicación científica y en la importancia que adquiere como un elemento fundamental para articular mecanismos de comprensión ciudadana para favorecer la explicabilidad del desarrollo de la IA como una garantía de cultivo de confianza y humanización de la sociedad digital (Domingo Moratalla, 2021). Las diferentes aportaciones mencionadas en este apartado son muy importantes para

el futuro de la IA, sin embargo, ninguna de ellas centra su interés en aspectos cívicos y políticos que vayan más allá del principialismo, es decir, en dinámicas científicas que podrían favorecer una comprensión ciudadana en el contexto de la democracia.

4. El fortalecimiento de la habilidad comunicativa y la interacción social

La creciente importancia pública de la ciencia y la tecnología, y también la politización como un recurso de poder y legitimación, han ocasionado una preocupación institucional por la comprensión científica de la ciudadanía en diversos formatos de comunicación (López Cerezo, 2005: 353). La ciudadanía se ve cada vez más afectada por los impactos de la tecnología, lo que implica una convivencia cada vez más íntima con cuestiones científicas. La reciente crisis sanitaria derivada de la pandemia ocasionada por virus SARS-CoV-2 ha puesto de relieve el significado de la habilidad comunicativa que debe fortalecer la ciencia y sus gestores para despertar la confianza en la ciudadanía en asuntos tan trascendentales para el conjunto de la sociedad como la campaña de vacunación.

En ocasiones se maneja una visión lineal de la comunicación científica, propia del modelo de déficit científico, para el que la ciencia es maravillosa y el público es idiota al no disponer de suficientes conocimientos, en palabras de Sarah Perrault (2013: 4). En contraste con el modelo de déficit científico, de carácter lineal, se encuentran un conjunto de personalidades que combinan el aprecio por la ciencia con la importancia de su análisis crítico. Las figuras que defienden un modelo que fomenta un compromiso democrático con la ciencia son agrupados por Peter Broks (2006) bajo las siglas CUSP (siglas en inglés de *Critical Understanding of Science in Public*). Para Perrault (2013), el modelo CUSP ofrece un tipo de comunicación científica que es multidimensional y contextual, centrada en un trabajo enfocado en la creación de significados relacionales en interacción con la ciudadanía.

Perrault señala que el modelo CUSP ofrece cuatro ventajas respecto a otros modelos. En primer lugar, presenta un enfoque relacional, asumiendo que las esferas científicas y sociales interactúan y se afectan entre sí. De este modo, pensadores como Alan Irwin y Brian Wynne (1996) apuntan que la ciencia no se encuentra separada de otras instituciones y que participa de diversas conexiones.

En segundo lugar, este modelo de comunicación científica reconoce que la experiencia se lleva a cabo a partir de múltiples construcciones, tomando distancia de enfoques unitarios. Esta consideración pone en valor que la comunicación sobre la ciencia tiene lugar entre diversos grupos sobre la base de que todos pueden contribuir y la existencia de un interés mutuo en el resultado de las deliberaciones (Trench, 2008: 132). En tercer lugar, mantener informado al público implica también promover una actitud crítica. En este caso, la crítica es considerada como un complemento del conocimiento, pues además de comunicar no debe perderse de vista la crítica como una herramienta. Y, en cuarto lugar, el modelo CUSP se adapta a las diferentes perspectivas del público sobre la ciencia, estableciendo un equilibrio con el contexto social, cultural y material susceptible de crítica. De ese modo este modelo se aleja de posiciones dogmáticas que observan las afirmaciones científicas como verdades sagradas e incuestionables.

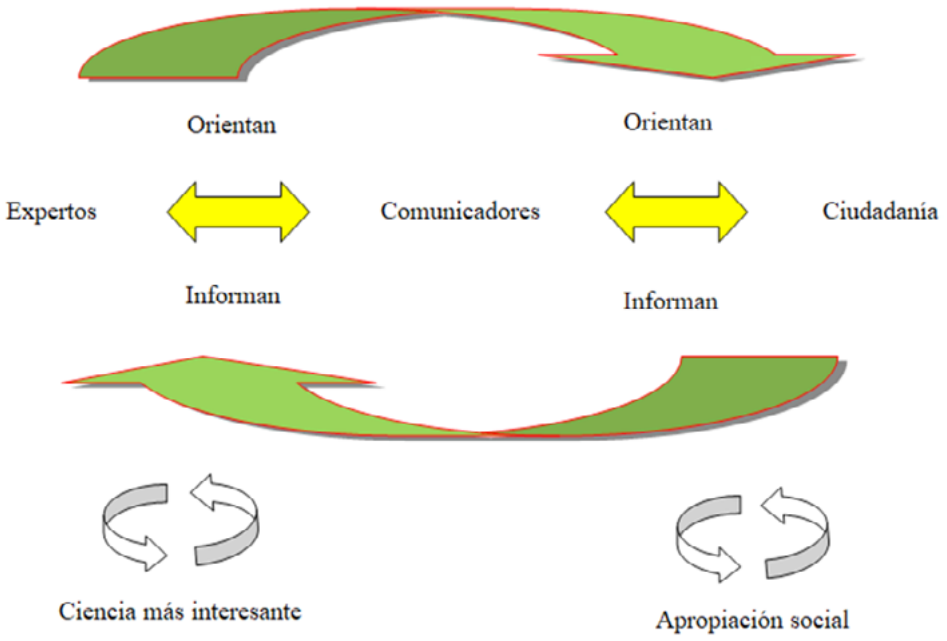
En la estela del CUSP se encuentra el modelo interactivo defendido por José Antonio López Cerezo (2005). Para este investigador, a menudo, se maneja una visión pasiva del proceso de enculturación científica, descuidando aspectos importantes como los comportamientos y las idiosincrasias en el proceso de apropiación (2005: 354). Esta visión es el resultado de un modelo lineal de difusión del conocimiento que empobrece la experiencia científica al no interactuar con la ciudadanía y situarse en la senda de un clásico modelo tecnocrático de gestión pública de los asuntos relacionados con la ciencia y la tecnología (López Cerezo, 2005: 355). Un modelo determinista y piramidal que privilegia los enfoques *top-down*, no incluye a la ciudadanía y se construye sobre una tradición comunicativa unidireccional. Esta manera de comunicar ha resultado de utilidad durante mucho tiempo, sin embargo, la creciente complejidad de los problemas y el progreso moral que ha experimentado la sociedad advierten la necesidad de cuestionar la brecha que existe entre quienes crean, regulan y usan las tecnologías (Hishiyama, 2012). Pues la unidireccionalidad ofrece pocas oportunidades para impulsar un proceso de innovación social más democrático en los entornos de generación de conocimiento científico. En ese sentido, en la senda del interés de la CE por una IA confiable sostenida en la explicabilidad en el contexto de H2020, el desafío se origina en cómo evitar una visión pasiva y lineal del proceso de enculturación y cómo enriquecer la conceptualización de la cultura científica contribuyendo en el cultivo de nuevas posibilidades y habilidades democráticas de gestión y políticas públicas.

El modelo interactivo ofrece una alternativa al modelo lineal y unidireccional de difusión, pues contempla a la ciudadanía como un elemento proactivo

en el proceso comunicativo. Este modelo promueve la apropiación social de los contenidos, ya que es capaz de destacar la presencia de incertidumbre y valores subyacentes en el despliegue de la ciencia y de políticas públicas (López Cerezo, 2005: 357). Además, ofrece una oportunidad para la generación de un espacio para el encuentro público y la deliberación sobre asuntos científicos y tecnológicos que resultan problemáticos o controvertidos. En el contexto particular de la IA, los expertos, gestores y la ciudadanía pueden participar en el establecimiento de un flujo de información desde diversas miradas para la puesta en común y la deliberación sobre condicionantes axiológicos de diversa índole que deben ser considerados para enriquecer el proceso de generación de conocimiento y promover un cultivo de confianza.

López Cerezo refleja de la siguiente manera un modelo interactivo de comunicación que conecta la cultura científica y la participación ciudadana.

Modelo interactivo



López Cerezo, 2005: 355.

En este modelo la cultura se encuentra entrelazada con la participación, configurando una sinergia para el aprendizaje social que pone en valor un concepto de cultura científica abierto al enriquecimiento a través de nuevas experiencias. Como López Cerezo señala, «una cultura científica de calidad es una cultura crítica y responsable, es conocimiento no sólo de las potencialidades de la ciencia, sino también de sus incertidumbres, de sus riesgos, y de los interrogantes éticos que plantea» (2005: 357). Esta visión de la ciencia pone de relieve la importancia de la interacción social, pues ofrece numerosos beneficios derivados de una retroalimentación que conlleva nuevos y valiosos conocimientos que son el resultado de saberes y experiencias de vida ciudadana que impulsan la construcción de nuevas infraestructuras para la producción de conocimiento e inteligencia colectiva, pues como Néstor-Hernando Parra y Francisco Arenas Dolz señalan:

> El conocimiento absoluto no es posible. Es por esta razón que resulta casi vital la colaboración para «el reconocimiento y el enriquecimiento mutuo de las personas, y no el culto de comunidades fetichizadas o hipostasiadas». El intercambio de conocimiento y experiencias, donde la diferencia es una manera de enriquecer el saber, nos aleja de una uniformada de pensamiento […] De esta manera, si juntamos todos esos microsaberes, crearemos una inteligencia colectiva, que parte del principio de que cada persona sabe sobre algo. Por tanto, nadie tiene el conocimiento absoluto. De lo que resulta fundamental la inclusión y participación de los conocimientos de todos (Parra y Arenas-Dolz, 2015: 123).

Una cultura de calidad implica a diversos agentes, interactúa con aquellos que no son expertos y demanda también su interés e implicación. Por ello hay que asumir que una cultura científica enriquecedora es el resultado de un aprendizaje social caracterizado por la participación. Según López Cerezo, la participación no debe concebirse de un modo restrictivo, con un marco procedimental y metodológico concreto y definido, limitándose a acciones como las encuestas de opinión o la audiencia pública (2005: 358). En ese sentido, es importante la ampliación de oportunidades de implicación ciudadana en aquellos problemas y controversias en torno al diseño, la gestión, la regulación y la motivación de la ciencia y la tecnología.

Espacios como los cafés científicos, laboratorios abiertos, talleres científicos, *science shop* y conferencias de consenso, ofrecen métodos comunicativos desarrollados en los últimos años sobre la base de la interacción social. Estos métodos son facilitadores de espacios de encuentro entre los expertos y la ciudadanía que

se reúnen para compartir y discutir aspectos problemáticos y controvertidos. El propósito de estos espacios es promover una comprensión compartida de los problemas y favorecer el tratamiento de las diferencias axiológicas que influyen en la toma de decisiones sobre el conocimiento. Por un lado, estas diferencias son el resultado de los valores proyectados sobre la tecnología en el diseño, reconociendo que sus productos constituyen una mediación empleada por la acción humana que incorpora discursos de los diferentes responsables (Toboso y Aparicio, 2019); y, por otro lado, estas diferencias también son provocadas por el proceso a través del cual la tecnología se integra y asume en la sociedad.

Experiencias como el proyecto para la innovación social *Ethos Living Lab* son un claro ejemplo de las posibilidades que reúne la generación de espacios para el encuentro ciudadano y científico con el objetivo de conectar explicabilidad y participación. El concepto *Living Lab* ha sido descrito en numerosas ocasiones desde diversas perspectivas, principalmente como una metodología, organización, sistema, terreno o enfoque de innovación sistémica (Ballon, Pierson, y Delaere, 2005; Eriksson, Niitamo y Kulkki, 2005; Feurstein *et al*., 2008). Este proyecto fue impulsado en el año 2018 por el Departamento de Filosofía de la Universitat de Valencia como un proyecto de innovación docente que perseguía el tejido de hilos comunicativos entre la comunidad universitaria y varios colegios profesionales de la Comunitat Valenciana. El establecimiento de este vínculo comunicativo procuraba el cultivo de buenas prácticas para fortalecer y mejorar las competencias del estudiantado y los profesionales involucrados desde una perspectiva ética. *Ethos Living Lab* ha servido para desarrollar habilidades comunicativas y profesionales, impulsar la creatividad y estimular un ethos de compromiso, colaboración e implicación (Arenas Dolz, 2021: 3).

Estos espacios deliberativos, como resultado de la aceptación del valor que posee la habilidad comunicativa de la ciencia para construir un vínculo entre los expertos y la ciudadanía, abren nuevos caminos hacia un futuro situado en la senda de la innovación social y brindan la oportunidad de poner en práctica un ejercicio de confluencia de perspectivas para compartir soluciones mutuamente deseables y aceptables (Gregory y Miller, 2000). Además, el fortalecimiento de habilidades comunicativas con interés cívico en el ámbito de la ciencia puede contribuir en la construcción de un ecosistema de IA confiable, en la senda de la hoja de ruta de la CE.

5. Conclusiones

La reflexión desplegada a lo largo de las páginas ha girado en torno a la necesidad de fortalecer la habilidad comunicativa de los expertos en IA en un contexto democrático como un elemento fundamental para promover la explicabilidad y de este modo garantizar una IA confiable. En el primer apartado se destacó el interés de la CE en una IA confiable a través de la formulación de una serie de principios éticos entre los que se encontraba la explicabilidad y su relación con el modelo RRI en el marco del programa H2020. Posteriormente se mencionó la relevancia que adquiere la explicabilidad en el contexto de la IA y se señalaron algunas propuestas para su articulación, tanto principialistas como técnicas. Y en el último apartado se puso de relieve la importancia de la comunicación social de la ciencia como una estrategia de explicabilidad para favorecer la comprensión de la ciudadanía en aquellos asuntos que resultan de interés público.

Los retos éticos y políticos de la IA abren un camino de oportunidades que deben articularse a través de la implicación ciudadana. Por ello resultará fundamental la exploración de nuevas habilidades comunicativas en el entorno de los expertos y los gestores, con la vista puesta en un equilibrio entre la deseabilidad y la aceptabilidad. Sin duda, esta nueva praxis comunicativa de la ciencia estimulará la interacción ciudadana en aras de un ecosistema confiable de IA.

El recorrido para fomentar la explicabilidad puede resultar complejo, puesto que existen importantes desafíos relacionados con un cambio en la cultura científica. El aprovechamiento de universidades y centros de investigación como entornos proclives para el encuentro entre diversos saberes, supondría una importante oportunidad para dar los primeros pasos en el cultivo de habilidades comunicativas y significaría un llamado a la ciudadanía para su implicación. Además, situaría a universidades y centros de investigación en un nuevo marco de responsabilidad social para la asunción de un compromiso con el potencial comunicativo de la ciencia.

Bibliografía

ARENAS DOLZ, Francisco (2021). "Competencias para la innovación y el emprendimiento social. El caso de Ethos Living Lab". *Revista CIDUI: Más allá de las competencias: nuevos retos en la sociedad digital, XI Congreso CIDUI 2020+1*.

Ballon, Pieter; pierson, Jo & delaereerlin, Simon (2005). "Open Innovation Platforms for Broadband Services: Benchmarking European Practices". 16th European Regional Conference. Oporto, Portugal.

Bimber, Brucer A. (1996). *The politics of expertise in congress: The rise and fall of the office of technology assessment*. Albany: State University of New York Press.

Boden, Margaret A. (2017). *Inteligencia Artificial*. Madrid: Turner Noema.

Brocks, Peter (2006). *Understanding Popular Science*. Milton Keynes. United Kingdom: Open University Press.

Comisión Europea (2015). *Portal Horizonte 2020*. Extraído el 24 de enero de 2022 desde https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation#Article

Comisión Europea (2018). *Artificial intelligence: Commission kicks off work on marrying cutting-edge technology and ethical standards*. Extraído el 24 de enero de 2022 desde https://ec.europa.eu/commission/presscorner/detail/en/IP_18_1381

Comisión Europea (2019). *Ethics guidelines for trustworty AI*. Extraído el 24 de enero de 2022 desde https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Comisión Europea (2020). *Libro Blanco sobre la inteligencia artificial: un enfoque europeo orientado a la excelencia y la confianza.* Extraído el 24 de enero de 2022 desde https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

Domingo moratalla, Agustín (2021). *Del hombre carnal al hombre digital: vitaminas para una ciudadanía digital*. España: TEELL Editorial.

Ege, European Group on Ethics in Science and New Technologies (2018). Statement on Artificial Intelligence, Robotics and "Autonomous" Systems, Comisión Europea, Dirección General de Investigación e Innovación. Extraído el 24 de enero de 2022 desde https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf.

Eriksson, Mats; niitamo Veli-Pekka; kulkki Seija (2005). State-of-the-Art in Utilizing Living Labs Approach to User-centric ICT innovation - a European approach. State-of-the-art in utilizing Living Labs approach to user-centric ICT innovation. Extraído el 27 de enero de 2022 desde https://www.academia.edu/2548057/State-of-the art_in_utilizing_Living_Labs_approach_to_user-centric_ICT_innovation_Automotive_Rural_eEngineering_and_Renewable_Energy_LLs_in_Hungary

Eslami, Motahhare; rickman, Aimee; vaccaro, Kristen; aleyasen, Aleyasen; vuong, Andy; karahalios, Karrie; hamilton, Kevin y sandvig, Christian (2015). ""I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in news feeds". *CHI 2015 - Proceedings of the 33rd Annual CHI Conference on Human Factors in Computing Systems,*New York: ACM, pp.153–162.

FEURSTEIN, K; HESMER, A.; HRIBERNIK, KARL A.; THOBEN, Klaus-Dieter; SCHUMACHER, Jens (2008). "Living Labs: A New Development Strategy". *European Living Labs - A New Approach for Human Centric Regional Innovation.* Eds. Schumacher, J. & V. P. Niitamo, V. P. Berlin: Wissenschaftlicher Verlag, pp. 1-14.

FIORINO, Daniel J. (1990). "Citizen Participation and Environmental Risk: A Survey of Institutional Mechanism". *Science, Technology, & Human Values* 15/2, pp. 226-243.

FISHER, Erik & RIP, Arie (2013). "Responsible Innovation: Multi-Level Dynamics and Soft Intervention Practices". *Responsible Innovation Managing the Responsible Emergence of Science and Innovation in Society*. Eds. Owen, R; Bessant, J. & Heintz, M. New York: John Wiley & Sons, pp. 165-183.

FLORIDI, Luciano; COWLS, Josh; BELTRAMETTI, Monica; CHATILA, Raja; CHAZERAND, Patrice; DIGNUM, Virginia; LUETGE, Christoph; MADELIN, Robert; PAGALLO, Ugo; ROSSI, Francesca; SCHAFER, Burkhard; VALCKE, Peggy & VAYENA, Effy (2018). "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations". *Minds and Machines* 28, pp. 689-707.

GEOTHEAN-QUINN, Máire (2012). Commissioner Geoghegan-Quinn Keynote Speech at the "Science in Dialogue" Conference, Odense, April 23–25. Extraído el 24 de enero de 2022 desde https://ec.europa.eu/archives/commission_2010-2014/geoghegan-quinn/headlines/speeches/2012/documents/20120423-dialogue-conference-speech_en.pdf

GLERUP, Cecilie & HORST, Maja (2014). "Mapping 'social responsibility' in science". *Journal of Responsible Innovation* 1/1, pp. 31-50.

GREGORY, Jane & MILLER, Steven (2000). *Science in Public-Communication, Culture, and Credibility*. New York: Basic Books.

HISHIYAMA, Reiko (2012). "Outreach Communication". *Field Informatics: Kyoto University Field Informatics Research Group*. Ed. Ishida, T. Berlin: Springer, pp. 157-169.

IRWIN, Alan & WYNNE, Brian (1996). *Misunderstanding Science?: The Public Reconstruction of Science and Technology.* New York: Cambridge University Press.

JOBIN, Anna; IENCA, Marcello & VAYENA, Effy (2019). "The global landscape of AI ethics guidlines". *Nature Machine Intelligence* 1, pp. 389-399.

JOSS, Simon (2002). "Toward the Public Sphere—Reflections on the Development of Participatory Technology Assessment". *Bulletin of Science, Technology & Society,* 22/3, pp. 220-231.

LARSSON, Sefan & HEINTZ, Fredrik (2020). "Transparency in artificial intelligence". *Internet Policy Review*, 9/2.

LAYARD HORSFALL, Hugo; PALMISCIANO, Paolo; KHAN, Danyal Z.; MUIRHEAD, William; KOH, Chan Hee; STOYANOV, Danail & MARCUS, Hani J. (2021). "Attitudes of the

Surgical Team Toward Artificial Intelligence in Neurosurgery: International 2-Stage Cross-Sectional Survey". *World Neurosurgery* 146, pp. 724-730.

Lecue, Freddy & Wu, Jiewen (2018). "Semantic Explanation of Predictions". Extraído el 24 de enero de 2022 desde https://arxiv.org/abs/1805.10587

López Cerezo, José Antonio (2005). "Participación ciudadana y cultura científica". *Arbor: Ciencia, pensamiento y cultura* 715, pp. 351-362.

Mccorduck, Pamela (1991). *Máquinas que piensan*. Barcelona: Tecnos

Mcleish, Caitríona & Nightingale, Paul (2007). "Biosecurity, bioterrorism and the gobernance of science: The increasing convergence of science and security policy". *Research Policy* 36/10, pp. 1635-1654.

Miller, Tim; Howe, Piers & Sonenberg, Liz (2017). "Explainable AI: Beware of Inmates Running the Asylum. Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences". *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*. Extraído el 24 de enero de 2022 desde https://arxiv.org/abs/1712.00547

Miller, Timm (2019). "Explanation in artificial intelligente: Insights from the social sciences". *Artificial Intelligence* 267, pp. 1-38.

Nourbakhsh, Illah (2009). *Ethics in Robotics*. Extraído el 24 de enero de 2022 desde https://www.youtube.com/watch?v=giKT8PkCCv4

Palmisciano, Paolo; Jamjoom, Aimun. A. B.; Taylor, Daniel; Stoyanov, Danail & Marcus, Hani J. (2020). "Attitudes of Patients and Their Relatives Toward Artificial Intelligence in Neurosurgery". *World Neurosurgery* 138, pp. 627-633.

Parlamento Europeo y Consejo de la Unión Europea (2016). *Reglamento General de Protección de Datos*. Extraído el 24 de enero de 2022 desde https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679

Parra, Néstor-Hernando & Arenas-Dolz, Francisco (2015). *Revolución tecnológica y democracia del conocimiento. Por una universidad innovadora*. Valencia: Laboratorio de la Sociedad del Conocimiento.

Perrault, Sarah (2013). *Communicating Popular Science. From Deficit to Democracy*. England: Palgrave Macmillan.

Ribeiro, Marco Tulio; Singh, Sameer & Guestrin, Carlos (2016). "'Why should I trust you?': Explaining the predictions of any classifier". *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135-1144.

Shaping, Steven (2008). *The Scientific Life. A moral History of a Late Modern Vocation*. Chicago: University of Chicago Press.

Stilgoe, Jack Owen, Richard & Macnaghten, Phil (2013). "Developing a framework for responsible innovation". *Res Policy* 42/9, pp. 1568–1580.

TIMMERMANS, Job (2021). "Responsible Research and Innovation". *Decent Work and Economic Growth. Encyclopedia of the UN Sustainable Development Goals*. Eds. Filho, W. L., Azul, A. M., Brandli, L., Lange Salvia, A. y Wall, T. London: Springer, pp. 847-858.

TOBOSO, Mario & APARICIO, Manuel (2019). "Entornos de funcionamientos robotizados. ¿Es posible una robótica inclusiva?". *Dilemata* 30, pp. 171-185.

TRENCH, Brian (2008). "Towards an Analytical Framework os Science Communication Moderls". *Communicating Science in Social Contexts: New Models. New Practiques.* Eds. Cheng, D., M. Claessens, M., Ggascoigne, T., Metcalfe, J., Schiele, B. & Shi, S. New York: Springer, pp. 119-135.

VAN DEN HOVEN, Jeroen (2014). "Responsible innovation: a new look at technology and ethics". *Responsible innovation 1.* Eds. van den Hoven, J., Doorn, N., Swierstra, T., Koops, B.-J., Romijn, H. Netherlands: Springer, pp. 3-13.

VASILJEVA, Tatjana; KREITUSS, Ilmars & LULLE, Ilze (2021). "Artificial Intelligence: The Attitude of the Public and Representatives of Various Industries". *Journal of Risk and Financial Management* 14/8, 339.

VON SCHOMBERG, René (2011). *Towards responsible research and innovation in the information and communication technologies and security technologies fields*, Luxemburgo: European Commission-DG Research and Innovation

VON SCHOMBERG, René (2013). "A vision of responsible research and innovation". *Responsible innovation: managing the responsible emergence of science and innovation in society.* Eds. Owen, R., Bessant, J. R. & Heintz, M. United States: Wiley-Blackwell, pp. 51-74.

WELD, Daniel S. & BANSAL, Gagan (2019). "The challenge of crafting intelligible intelligence". *Communications of the ACM 62*/6, pp. 70-79.

YOUNG, Albert T.; AMARA, Dominic; BHATTACHARYA, Abhishek, WEY, Maria L. (2021). "Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review". *The Lancet Digital Health* 3/9, pp. 599-611.

Recibido: 24/01/2022

Aceptado: 11/02/2022