MEMÒRIA CIENTIFICOTÈCNICA DEL PROJECTE */ *MEMORIA CIENTÍFICO-TÉCNICA DEL PROYECTO **

INVESTIGADORS PRINCIPALS DEL PROJECTE / *INVESTIGADORES PRINCIPALES DEL PROYECTO*

Vicent Botti Navarro, Ana García Fornes

TÍTOL DEL PROJECTE / *TÍTULO DEL PROYECTO*

guardIA: GUArdian Personal Virtual en Reali**D**ades Soc**IA**les Hibridas

RESUM DEL PROJECTE / *RESUMEN DEL PROYECTO*

En los próximos 10-15 años, los medios sociales y la tecnología nos permitirán unirnos en entornos virtuales verdaderamente inmersivos, estas realidades sociales híbridas (RSH) permitirán que la interacción social sea realizada en una forma sin precedentes. Los humanos serán asistidos por agentes físicos y digitales, cuyos comportamientos podrían ser indistinguibles. En este contexto, los retos se originan en la interacción entre los bots inteligentes con los humanos, pero también en la creciente necesidad de detectar intenciones o contenidos poco fiables o maliciosos. La posibilidad de fundamentar las opiniones en una realidad innegable y en un entorno verificable de confianza se reducirá, mientras que los humanos y los agentes virtuales podrán fabricar pruebas o información (incluso en tiempo real) en apoyo de una información que podría ser falsa o errónea. Además, al igual que hacen hoy en día los sistemas de recomendación y los sistemas de publicidad dirigida basados en el comportamiento en línea, los agentes que controlan las interacciones y los flujos de información en las futuras RSH tendrán aún más ventaja para explotar las debilidades de los sujetos que interactúan en dichos entornos. De este escenario, en el que no es fácil distinguir en que se puede confiar o no, ya que todo puede ser construido, surge el objetivo general de ***guardIA***: *Garantizar una interacción social más segura y sofisticada en los futuros espacios sociales de realidad híbrida aprovechando las capacidades de la IA.* Para ello se propone desarrollar un Guardián Personal Virtual (GPV) basado en técnicas de Inteligencia Artificial, integrado en la RSH, que interaccione con el usuario final, sensibilizándolo y concienciándolo sobre posibles intentos maliciosos y prácticas manipuladoras en la RSH y mitigando así sus efectos.

MEMÒRIA CIENTIFICOTÈCNICA DEL PROJECTE D'INVESTIGACIÓ
MEMORIA CIENTÍFICO-TÉCNICA DEL PROYECTO DE INVESTIGACIÓN

Objetivo:

En los próximos 10-15 años, los medios sociales y la tecnología nos permitirán unirnos en entornos virtuales verdaderamente inmersivos, estas realidades sociales híbridas[1] (RSH) permitirán que la interacción social sea realizada en una forma sin precedentes. Los humanos serán asistidos por agentes físicos y digitales, cuyos comportamientos podrían ser indistinguibles. En este contexto, los retos se originan en la interacción entre los bots inteligentes híbridos con los humanos, pero también en la creciente necesidad de detectar intenciones o contenidos poco fiables o maliciosos. La posibilidad de fundamentar las opiniones en una realidad innegable y en un entorno verificable de confianza se reducirá, mientras que los humanos y los agentes virtuales podrán fabricar pruebas o información (incluso en tiempo real) en apoyo de una información que podría ser falsa o errónea. Además, al igual que hacen hoy en día los sistemas de recomendación y los sistemas de publicidad dirigida basados en el comportamiento en línea, los agentes que controlan las interacciones y los flujos de información en las futuras RSH tendrán aún más ventaja para explotar las debilidades de los sujetos que interactúan en dichos entornos.

A medida que los mundos online y físico convergen, ¿cómo podemos garantizar una interacción social más segura y satisfactoria en los espacios sociales del futuro? De este escenario surge el objetivo general de ***guardIA***:

> ***Garantizar una interacción social más segura y sofisticada en los futuros espacios sociales de realidad híbrida aprovechando las capacidades de la IA.***

[1] Igor Perko. Hybrid reality development - can social responsibility concepts provide guidance?. https://www.emerald.com/insight/0368-492X.htm

Para alcanzar esta meta, el proyecto ha establecido los siguientes 3 sub-objetivos específicos como pasos necesarios:

1. **Desarrollar un paradigma teórico y tecnológico para anticipar cómo las redes sociales hibridas (RSH) redefinirán las relaciones sociales y qué nuevas formas de interacción social surgirán en los entornos virtuales, caracterizando especialmente las de mayor riesgo como pueden ser las interacciones sociales maliciosas y manipuladoras.**

La otra cara de la moneda de los riesgos de interacción social maliciosa y manipuladora, que se pueden producir en las RSH, es que la inmersión de las RSH también tiene el potencial de ayudar a los individuos que interactúan en estos espacios a navegar por estos flujos de información e interacciones sociales de una manera más segura y productiva, y, por lo tanto, ayudar a mitigar estos riesgos. Para ello, como objetivo específico clave, ***guardIA*** propone: investigar el potencial y desarrollar prototipos de trabajo de un Guardián Personal Virtual (GPV), integrado en la RSH como sustrato tecnológico, asegurando una comunicación veraz y de confianza, específicamente:

2. **Crear un asistente personalizado de código abierto basado en la IA, que participe en una sofisticada interacción con el usuario final, sensibilizándolo y concienciándolo, ayudándole a reconocer posibles intentos maliciosos y patrones manipuladores en la RSH y mitigando así sus efectos.**

3. **Apoyar la comprensión y la percepción pública de las herramientas basadas en la IA que abordan la desinformación en entornos inmersivos, aplicables en diversos campos.**

1. CONCEPTO

En los próximos 10-15 años, los medios sociales y la tecnología nos permitirán unirnos en entornos virtuales verdaderamente inmersivos. El lado positivo es que esto proporcionará oportunidades para interacciones sociales que nunca hemos tenido antes: podremos "estar" juntos mientras seguimos separados físicamente. Ahora la distancia no es un límite para la comunicación - en el futuro tampoco lo será para la presencia gracias a los desarrollos tecnológicos que se están produciendo (las soluciones de estudios virtuales o la realidad aumentada, por ejemplo, que proporciona Brainstorm Multimedia[2] son una prueba de esto).

El uso de estas tecnologías, junto con los avances tecnológicos exponenciales en campos como la realidad virtual (RV), la realidad aumentada (RA), la realidad mixta y extendida (RX), las tecnologías de detección de emociones y la inteligencia artificial (IA), combinados con la creciente comprensión de la psicología humana, también abrirán las puertas a posibles amenazas a la seguridad de la interacción social del futuro, dentro de una perspectiva que favorecerá la aparición de innumerables e imprevisibles nuevas formas de prácticas maliciosas y manipuladoras, o incluso de efectos secundarios no deseados e imprevistos. Una evidencia de estas malas prácticas de la tecnología la podemos encontrar ya en las Deepfake como resultado del mal uso de las redes generativas antagónicas[3] (generative adversarial networks -GANs).

Habrá nuevas herramientas a disposición de quienes quieran manipular a los usuarios en las RSH del futuro, ya sea con fines comerciales, políticos o de cualquier tipo. Ya no se centrará sólo en la esencia de un mensaje, sino también en su forma. Podríamos esperar bots impulsados por la IA con edad, género y voz configurados específicamente para desencadenar nuestras emociones tratando de influir en nuestro estado anímico, nuestra capacidad de tomar decisiones o nuestra capacidad juicio. Del mismo modo se podrían utilizar imágenes de personas en las que confiamos para engañarnos como por ejemplo mediante el uso de modelos de machine learning como "deepfake". Si la participación de los usuarios en el RSH es el motor de un beneficio monetario o político, como ocurre con las redes sociales actuales, por ejemplo, debemos asumir que esto ocurrirá, incluso sin intención maliciosa.

Para evitar este panorama distópico o, al menos, mitigar su impacto, serán necesarias acciones activas y focalizadas. Así pues, ***guardIA*** pretende responder a la siguiente pregunta: ¿cómo hacer que la interacción social del futuro en las RSH sea más segura y gratificante? El proyecto aportará una contribución única a este reto estableciendo un nuevo paradigma teórico y tecnológico para anticipar futuras formas de interacciones sociales maliciosas y manipuladoras en las RSH y proporcionando una solución

[2] https://www.brainstorm3d.com/es/

[3] Ian J. Goodfellow∗, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. Generative Adversarial Nets. Proceedings os Neurips. 2019.
https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf
Alankrita Aggarwal. Mamta Mittal. Gopi Battineni. Generative adversarial network: An overview of theory and applications. International Journal of Information Management Data Insights 1 (2021) 100004.

centrada en el ser humano e impulsada por la IA para mitigarlas: ***guardIA*** GPV (el Guardian Personal Virtual de ***guardIA***). La idea central de ***guardIA*** es que las mismas características de las RSH que subyacen a estos riesgos, también nos proporcionan las herramientas para afrontarlos. Las posibilidades que ofrece este nuevo entorno pueden aprovecharse para facilitar interacciones más seguras y eficaces en las RSH que las que son posibles en los entornos sociales físicos, proporcionando así una experiencia más satisfactoria.

Para abordar un problema, que aún no ha surgido, en un entorno futuro, se hacen algunas suposiciones básicas al diseñar el concepto. A continuación presentamos los supuestos y su validez para ***guardIA***".

- RV, RA y RX: suponemos que en 10-15 años el desarrollo en este campo permitirá la interacción libre en espacios sociales virtuales híbridos. Estos espacios podrían ser, entre otros, plataformas de medios sociales, estudios de noticias, aulas virtuales u oficinas (el equipo de investigación de este proyecto colabora con Brainstorm Multimedia empresa líder en el desarrollo de tecnología para desarrollo de humanos digitales virtuales, realidad mixta y extendida, y realidades inmersivas, y que facilita el acceso a sus productos con fines de investigación);
- La tecnología multisensorial (wearables, detectores de reconocimiento facial, etc.) permitirá un seguimiento fácil, asequible y en tiempo real de las emociones de los usuarios (el equipo de investigación de este proyecto colabora con Noldus[4] empresa líder en el desarrollo de sensores y tecnología para la percepción del comportamiento humano, empresa que facilita el acceso a sus productos con fines de investigación);
- La desinformación y la manipulación son fenómenos crecientes[5] que podrían adoptar nuevas formas con los rápidos avances tecnológicos. Aunque se están dedicando importantes esfuerzos de investigación a contrarrestarlos (por ejemplo, la detección de deepfakes[6]), no sería prudente ignorar la posibilidad de que lleguemos a un punto en el que sea inviable distinguir el audio, las imágenes, los vídeos e incluso las experiencias genuinas de las falsas en las redes sociales y, especialmente, en las RSH que son más sensibles a la información unilateral, o de otras formas que explotan las limitaciones cognitivas humanas.
- Los asistentes personales (AP) ya forman parte de nuestra vida cotidiana, pero todavía se limitan a realizar unas pocas tareas (algunas dependiendo de la dinámica de nuestro hogar) y a entablar conversaciones limitadas en lenguaje hablado. Sin embargo, los dispositivos de interacción persona-ordenador están evolucionando a gran velocidad y se espera que, en 10-15 años, los asistentes personales muestren un comportamiento inteligente para convertirse en verdaderos compañeros humanos[7].

Basándonos en esta imagen del futuro, también identificamos la necesidad de aumentar la concienciación pública sobre las posibilidades de futuras formas de desinformación como un primer paso esencial para apoyar a una ciudadanía informada. Por lo tanto, ***guardIA*** apoya la transformación de lo que podría ser un escepticismo creciente sobre los medios de comunicación y la formación online, así como los futuros flujos de información en los que se pueden producir difusiones de información falsa e intentos de desinformación o manipulación.

Al mismo tiempo, facilitar el proceso de adaptación de los usuarios a la tecnología ***guardIA*** se concibe como una prioridad para garantizar su plena aceptación, atendiendo a la necesidad de preparar un nuevo tipo de experiencia de usuario en este incipiente cambio sociotecnológico[8]. La medida del éxito del ***guardIA*** GPV dependerá en gran medida del grado en que la gente la utilice, se adapte a ella y acabe tolerando sus posibles deficiencias. Por ello, se ha prestado una gran atención a las estrategias de diseño de ***guardIA***, fomentando una IA explicable y la participación activa de los usuarios en el codiseño mediante estudios de validación y actividades de prueba específicas con el objetivo último facilitar el aumento de la confianza percibida hacia ***guardIA*** y su tecnología desarrollada.

Al final del proyecto, la tecnología ***guardIA*** GPV se validará en un entorno relevante con el objetivo de proporcionar asistencia individual y personalizada a los usuarios para ayudarlos a navegar y protegerlos en los futuros espacios sociales virtuales,

[4] https://www.noldus.com

[5] Chesney, B., & Citron, D. (2019). Deep fakes: a looming challenge for privacy, democracy, and national security. Calif. L. Rev., 107, 1753/ Floridi, L. (2018). Artificial intelligence, deepfakes and a future of ectypes. Philosophy & Technology, 31(3), 317-321.

[6] Güera, David, and Edward J. Delp. "Deepfake video detection using recurrent neural networks." In 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1-6. IEEE, 2018.

[7] Cohen, P., Cheyer, A., Horvitz, E., El Kaliouby, R., & Whittaker, S. (2016, May). On the future of personal assistants. In Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems (pp. 1032-1037).

[8] Oh, C., Lee, T., Kim, Y., Park, S., & Suh, B. (2017). Us vs. Them: Understanding Artificial Intelligence Technophobia Over the Google DeepMind Challenge Match, In Proceedings of the 2017 CHI Conference on Human factors in computer systems, pp. 2523 - 2534.

garantizando una interacción social más segura y gratificante. Existen diversos dominios donde resultaría adecuado validar la tecnología ***guardIA*** como el e-learning, los medios de comunicación de masas y los medios sociales, además, durante el desarrollo del proyecto, se podrían evaluar y explorar otros dominios de aplicación adecuados.

De entre los diversos candidatos se ha elegido como caso de estudio y validación el modelo de e-learning, que en los dos últimos cursos académicos, debido a la situación pandémica vivida, se ha consolidado especialmente en la Universitat Politècnica de València, en concreto se ha elegido el dominio de 'Aula Virtual Abierta' en el que podemos aportar los conocimientos y la experiencia del equipo investigador, donde se han identificado los problemas subyacentes que confirman la conveniencia de un GPV que asista a las/los alumnas/os, más aún si, como se pretende, adoptamos un modelo de aula virtual abierta, en el que puedan participar, de forma colaborativa, agentes humanos y virtuales no necesariamente benevolentes (algunos de estos agentes humanos o virtuales pueden tener intereses de desinformación o manipulación) y donde se garanticen interacciones más seguras y eficaces.

Los modelos, basados en datos externos e internos tanto al entorno RSH (en el caso de estudio este entorno lo proporcionará la plataforma de e-learning de la UPV donde está previsto poder integrar en un prototipo de desarrollo alguno de los productos facilitados por Brainstorm Multimedia y Noldus, **se adjuntan cartas de apoyo de ambas empresas**) como al usuario (análisis de contenido, fuente de información, claves biométricas, etc.) permitirán al GPV asesorar mejor al usuario para que su experiencia sea más segura y satisfactoria.

La tecnología facilitada por Brainstorm y Noldus permitirá desarrollar un prototipo de demostración (incluido en el caso de estudio) para poder crear los espacios de RSH y permitir la recopilación de datos necesaria mediante la realización de experimentos estilizados y, posteriormente, mediante casos de validación estructurados (T5.2).

La personalización, basada en la percepción individual, la comprensión y la modelización del significado, los valores y las características individuales, reforzada por los datos sociodemográficos, ayudará al usuario reduciendo los efectos adversos de los prejuicios y minimizando los efectos de la desinformación dirigida. La detección de los estados afectivos de un usuario, que influyen en su formación de opiniones y creencias, así como su previsible vulnerabilidad a la desinformación en función de sus antecedentes sociales, advertirá sobre las emociones y los sesgos cognitivos y heurísticos que conducen a un razonamiento subóptimo.

Por último, las capacidades de comunicación de ***guardIA*** GPV incluirán sistemas de corregulación y diálogo afectivo[9]. La ampliación del marco con estos parámetros permitirá al GPV tener en cuenta cómo responder, por ejemplo, a las expresiones negativas o positivas del usuario. Las estructuras genéricas de conocimiento y los modelos de memoria permitirán tener en cuenta las interacciones anteriores y aprovecharlas para las conversaciones posteriores. La tecnología persuasiva abordará el reto de cómo debe comunicarse la GPV de forma que no resulte intrusiva, condescendiente o irritante, sin dejar de cumplir el objetivo de ayudar al usuario.

De este modo, la creciente necesidad de asistencia personalizada para nuestras futuras interacciones sociales y de comunicación de la GPV será abordada plenamente por ***guardIA***. El proyecto también impulsará el desarrollo de una tecnología segura y fiable que sea responsable, ética y verdaderamente útil para los seres humanos, combinando las capacidades alcanzadas para el primer guardián personal virtual diseñado para las Realidades Sociales Híbridas. Para ello se tendrán que proporcionar las siguientes características y funcionalidades de ***guardIA*** GPV:

1. Detección de comportamientos maliciosos en realidades construidas.
   1.1. Detección de patrones de desinformación y comportamiento malicioso.
   1.2. Anticipación, orientación y advertencia adaptadas a las características de cada usuario.
2. Percepción, tratamiento de la información y personalización.
   2.1. Detección de prácticas persuasivas comunes.
   2.2. Mapeo de características personales, heurísticos y sesgos.
   2.3. Personalización para un mayor impacto beneficioso.
   2.4. Modelización del estado de creencias del usuario.
   2.5. Modelización de la confianza en la red social (quién confía en quién).
3. Detección y modelado de emociones.
   3.1. Detección de los estados afectivos del usuario, que influyen en la formación de opiniones y creencias.
   3.2. Detección de las emociones relacionadas con las prácticas persuasivas relacionadas con las emociones por parte de

[9] Claudia Marinetti. Penny Moore. Pablo Lucas. Brian Parkinson. Emotions in Social Interactions: Unfolding Emotional Experience . In Emotion-Oriented Systems. DOI: 10.1007/978-3-642-15184-2_3

los personajes virtuales y otros usuarios.
3.3. Modelización de la empatía y los vínculos afectivos con otras personas o agentes.
3.4. Representación transcultural de las emociones para evitar errores en la percepción, interpretación y expresión.
4. Comunicatividad.
4.1. Un modelo de corregulación y un nuevo marco de diálogo.
4.2. Comunicación afectiva.
4.3. Modelos de memoria para personalizar los diálogos repetidos.
4.4. Habilidades de persuasión.

Como se comentaba previamente encontramos diversos dominios específicos como el e-learning, los medios de comunicación de masas y los medios sociales, donde la tecnología ***guardIA*** puede tener un impacto significativo. A continuación se describe brevemente las aportaciones con las que ***guardIA*** puede contribuir en estos dominios.

E-Learning.

La actual pandemia de COVID ha impuesto la adopción acelerada del e-learning a nivel mundial. Lo que se esperaba para los próximos años en este campo es hoy una realidad que ha llegado para quedarse y para la que no estábamos del todo preparados. Hoy en día es imprescindible proporcionar tanto a los profesores como a los alumnos los medios necesarios para gestionar su aprendizaje en línea de forma eficaz (por ejemplo, adaptándose a sus necesidades, estilos de aprendizaje y objetivos) y respetando criterios de igualdad, transparencia, fiabilidad y equidad. Por ejemplo, debe evitarse la suplantación de identidad. Además, las posibles barreras digitales y el uso masivo de los cursos en línea pueden hacer totalmente imposible que el profesor ofrezca una atención personalizada a cada alumno sin la ayuda de sistemas de aprendizaje reforzados por la tecnología (technology-enhanced learning systems TEL).
Johnson et al (2000)[10] plantea varios aspectos que se han beneficiado de la aplicación de agentes inteligentes en entornos de aprendizaje y en los que la tecnología ***guardIA*** puede producir avances significativos en el estado del arte. Concretamente:

- Nuestro entorno RSH y el GPV permitirán realizar demostraciones interactivas y guiar a los alumnos hacia sus objetivos de aprendizaje ayudando a sus habilidades de aprendizaje paso a paso;
- El GPV será un asistente personal afectivo. La gestión de las emociones es crucial para tener éxito en el proceso de enseñanza-aprendizaje debido a su papel en la motivación, haciendo la experiencia más atractiva, natural y entretenida.
- El entorno RSH permitirá realizar simulaciones sin precedentes de casos prácticos y experiencias de aprendizaje con tareas complejas que requieren coordinación y trabajo en grupo. EL GPV puede desempeñar diferentes papeles, como entrenador en sustitución de los compañeros ausentes.
- El GPV será un tutor inteligente que permitirá interacciones pedagógicas, promoviendo la orientación en un entorno RSH confiable, anticipando y advirtiendo a profesores y alumnos de comportamientos maliciosos, como la manipulación y la suplantación de identidad. Pero, sobre todo, ***guardIA*** GPV puede ser un apoyo eficaz al proceso autónomo de aprendizaje e investigación tanto para profesores como para alumnos en escenarios de RX online y abiertos, más allá de su interacción directa en el aula virtual.

Medios de comunicación

Los medios de comunicación de masas, como fuente principal de información, son un campo lógico en el que ***guardIA*** GPV podría tener un impacto positivo. Las prácticas habituales destinadas a mantener la atención del espectador, la explotación de los sesgos cognitivos, las técnicas de persuasión, los contenidos personalizados y otros riesgos identificados por este proyecto podrían ser una realidad en los medios de comunicación del futuro. En esta situación, los telespectadores podrían beneficiarse mucho de una protección personalizada que identifique estos intentos malintencionados. Además, es probable que en los medios de comunicación del futuro los espectadores puedan realmente interactuar con el presentador, con la posibilidad de ser proyectados en estudios virtuales y tener experiencias de audiencia completamente nuevas. En tal escenario, el GPV marcará la diferencia al ayudar a la audiencia a calibrar la veracidad de lo que percibe.

Medios de comunicación social

Es conocido el giro radical que ha dado la irrupción de los medios de comunicación en la forma en que el público accede,

[10] Rickel, J., Johnson ,L.W. (200)0) Task-Oriented Colaboration with Embodied Agents in Wirtual Worlds. In J.Cassell, J.Sullivan&S.Orevost (Eds). Embodied Conversational Agents (pp. 95-122). Cambridge, MA: MIT Press.

consume y difunde la información. En algunos países, las redes sociales son ya más importantes como fuente de información general que la prensa escrita (Badillo, 2019: 20)[11] y es probable que esta tendencia persista y aumente en los próximos años. En el futuro es plausible que la mayoría de las redes sociales se desarrollen en plataformas RSH. Las RSH permitirán innumerables e inéditas formas de compartir información. Más allá de las preferencias y preocupaciones de los usuarios, las plataformas de medios sociales son las que más interés tienen en mantener a raya la desinformación y la manipulación, sabiendo el precio que podría suponer la pérdida de credibilidad de sus usuarios en el entorno. La tecnología proporcionada por ***guardIA*** ayudará a realizar una transición hacia estas futuras redes sociales de forma segura y confiable.

2 METODOLOGÍA

La metodología adoptada en ***guardIA*** está diseñada, ante todo, para responder a la pregunta planteada en la sección de Concepto: a medida que los mundos online y físico convergen, ¿cómo podemos garantizar una interacción social más segura y satisfactoria en los espacios sociales del futuro? Para ello se promueve un proceso coherente, que se refleja efectivamente tanto en la estructura del programa de trabajo como en la metodología técnica.

Para alcanzar el ambicioso objetivo de ***guardIA*** y, al mismo tiempo, garantizar un plan de trabajo realista orientado a los resultados, se ha estructurado el proyecto en torno a un proceso de 5 pasos.

1. La definición de un nuevo paradigma teórico y tecnológico para anticipar las futuras formas de prácticas maliciosas y manipuladoras en RSH y cómo este entorno cambiará la interacción social. Esto se asegurará, por un lado, mediante modelos matemáticos detallados para el comportamiento individual de los actores humanos en un entorno social virtual (modelando los aspectos relevantes de la recopilación de información, la percepción, el procesamiento y la comprensión, las preferencias individuales, la formación de opinión, la confianza, a nivel microscópico, mesoscópico y macroscópico); por otro lado, a través del potencial de innovación del mapeo capaz de superar significativamente el estado del arte de los asistentes virtuales basados en la IA y el software de protección en línea.
   Esta primera fase de trabajo común permitirá al equipo centrarse en acordar las interdependencias, los parámetros teóricos que servirán como referencia para los modelos matemáticos y los requisitos de desarrollo de ingeniería de software y está marcada por un conjunto de entregables en M6 y M12.

2. Integración, desarrollo y optimización del GPV. Esta segunda fase aborda plenamente el desarrollo y la integración del GPV, con una hoja de ruta de desarrollo planificada a lo largo de 4 años y cuyos hitos cruciales son las cuatro versiones previstas (a partir de M16 y las dos últimas de código abierto) de la solución. El diseño del agente se basará en los principios de la arquitectura BDI. La semántica y la comunicación del agente se construirán utilizando el lenguaje AgentSpeak[12] Para la implementación del agente revisaremos diferentes plataformas como Jason[13], GenIA[14], GOAL[15], o SPADE[16]. Para las comunicaciones entre agentes utilizaremos protocolos como XMPP (eXtensible Messaging and Presence Protocol), considerado como el protocolo estándar universal de mensajería instantánea por entidades como el IETF o el W3C, y con un uso extendido en la industria, siendo algunas de las empresas que lo utilizan (o alguna variante) Whatsapp, Google Talk, Facebook Messenger o iMessage de Apple, entre otras.
3. Validación y evaluación de la tecnología desarrollada mediante un caso de estudio del dominio de e-learning. En concreto se ha elegido el dominio de 'Aula Virtual Abierta' en la Universitat Politècnica de València. Este caso de estudio permitirá la validación técnica, verificando la alineación de la solución con los requisitos planteados por el paradigma teórico, los escenarios identificados para los experimentos y las pruebas finales. Cada flujo de validación se estudiará cuidadosamente, se simulará en caso de ser necesario y se optimizará durante una fase experimental de 23 meses de duración. Los ajustes y la optimización técnica siguen en una fase paralela de 21 meses en la que los resultados se integrarán en la infraestructura técnica de ***guardIA***. M43 marca el final de los estudios de validación. A

[11] Badillo, A. (2019). La sociedad de la desinformación: propaganda, «fake news» y la nueva geopolítica de la información. Documento de trabajo 8/2019 - May, 14th. 2019 Real Instituto Elcano.

[12] Anand S. Rao. AgentSpeak(L): BDI agents speak out in a logical computable language. MAAMAW '96: Proceedings of the 7th European workshop on Modelling autonomous agents in a multi-agent world : agents breaking away: agents breaking away. Pp. 42-55. Springer. 1996

[13] Rafael H Bordini, Jomi Fred H¨ubner, and Michael Wooldridge.Program-ming multi-agent systems in AgentSpeak using Jason, volume 8. John Wiley& Sons, 2007.

[14] Bexy Alfonso, Emilio Vivancos, and Vicente Botti. Toward formal modeling of affective agents in a BDI architecture.ACM Transactions on InternetTechnology (TOIT), 17(1):5, 2017.

[15] Hindriks, K. V. (2009). Programming rational agents in GOAL. In Multi-agent programming (pp. 119-157). Springer, Boston, MA.

[16] Miguel Escrivá Gregori, Javier Palanca Cámara, and Gustavo ArandaBada. A jabber-based multi-agent system platform. InProceedings of the fifth international joint conference on Autonomous agents and Multiagent systems, pages 1282–1284, 2006.

continuación, los resultados se integrarán en el prototipo final para una última fase de optimización, que también tiene como objetivo facilitar al máximo que otras partes retomen el concepto de ***guardIA*** a partir de ahí.

4. Las actividades de difusión apoyan la investigación participativa basada en la comunidad durante y después del proyecto. Las cuestiones éticas y las prácticas de Investigación e Innovación Responsables (RRI) se integrarán plenamente en todo el proceso de investigación y desarrollo técnico.

Este enfoque se refleja directamente en la estructura de los paquetes de trabajo, tal y como se destaca en la figura 1 de la sección 7 (Plan de Trabajo).

A nivel de gestión, para garantizar la calidad y aplicabilidad de la tecnología ***guardIA***, se utilizará una metodología de desarrollo ágil (SCRUM) combinada con la gestión de proyectos y las medidas de garantía de calidad. Esta metodología de gestión de proyectos se basa en el desarrollo iterativo de varias versiones de los resultados, entregables y versiones del proyecto, utilizando herramientas de asignación de tickets y software de control de versiones (GIT). Así, los riesgos previstos del proyecto pueden limitarse a su mínima probabilidad e impacto. El desarrollo se gestionará a lo largo de iteraciones de duración fija, denominadas Sprints (normalmente de 2/4 semanas de duración). Después de cada iteración se comprobará si los desarrollos se ajustan a las necesidades del proyecto y también se controlará el presupuesto. El proceso SCRUM se apoyará en herramientas de software específicas compartidas por el equipo.

3 ANÁLISIS DE SEXO Y/O GÉNERO

En ***guardIA*** se considera que tanto los elementos relacionados con el sexo (físicos) como los relacionados con el género (sociales) deben ser abordados desde las primeras fases del proyecto, no sólo en el diseño y la implementación de los desarrollos técnicos, sino también en la recogida y publicación de datos desagregados siempre que sea posible.

La bibliografía[17] ha demostrado que el género y la personalidad son factores que influyen en la forma en que los individuos perciben las situaciones, en la manera en que reaccionan afectivamente ante determinados estímulos y en la forma en que se comportan y toman decisiones. Por tanto, ***guardIA*** abordará el análisis de género y sexo como una prioridad transversal. Además, a pesar de que los efectos del género en distintas capacidades sociales, cognitivas y afectivas ha sido ampliamente estudiada y desarrollada en la literatura, hasta ahora este factor apenas se ha abordado en los modelos basados en agentes. Se hará un esfuerzo por destacar y abordar cuestiones y aspectos hasta ahora no considerados en los que el género puede desempeñar un papel discriminatorio, incluso de forma no intencionada.

Se creará un comité de género para garantizar que los aspectos relacionados con el género se tengan en cuenta en todo el proyecto de la siguiente manera (1) las soluciones tecnológicas se diseñan en función de las necesidades de sexo y género, (2) los datos son suficientemente representativos, (3) los algoritmos de IA evitan los sesgos de género, (5) la perspectiva de género está presente en las actividades de difusión y explotación y (6) se registran otras cuestiones relacionadas con el género que puedan surgir a lo largo del proyecto. En el caso de algunas actividades y resultados específicos del proyecto se establecerán grupos de discusión para garantizar que se incluya la dimensión de género.

4 AMBICIÓN

El análisis del estado del arte muestra que ninguna de las soluciones existentes garantiza tanto una dinámica de interacción totalmente receptiva y afectiva (capaz de recibir y procesar las aportaciones de los usuarios a través de una variedad de canales que van desde la voz hasta el seguimiento de las emociones y, al mismo tiempo, capaz de expresar empatía) como un alto grado de protección frente a las prácticas manipuladoras (incluyendo capacidades de personalización, basadas en las características, valores y prejuicios individuales de los usuarios). ***guardIA*** es, por tanto, el primer intento de diseñar un asistente dirigido por la IA que sea totalmente receptivo, ético, responsable, afectivo y efectivamente personalizado, y que ayude a concienciar a los usuarios de RSH tanto sobre la información potencialmente maliciosa o poco fiable como sobre el comportamiento sospechoso de los demás, garantizando una experiencia más satisfactoria.

Potencial de innovación

[17] Zelenski, J. M., Baumeister, R., & Loewenstein, G. (2008). The role of personality in emotion, judgment and decision making. Do Emotions Help or Hurt Decision Making? A Hedgefoxian Perspective, 117-132.
Rueckert, L., & Naybar, N. (2008). Gender differences in empathy: The role of the right hemisphere. Brain and cognition, 67(2), 162-167.
Segerstrom, S. C., & Smith, G. T. (2019). Personality and coping: Individual differences in responses to emotion. Annual review of psychology, 70, 651-671.

El posicionamiento del proyecto demuestra que la solución propuesta es innovadora, ya que hasta ahora no existe ninguna iniciativa de este tipo en el mercado ni se está estudiando actualmente.

Sin embargo, el potencial de innovación de ***guardIA*** reside en varias dimensiones complementarias:

1. Una nueva agenda de investigación para la IA aplicada a la interacción social ***guardIA*** está preparando el camino para una nuevo enfoque de investigación que contrasta con el enfoque clásico de utilizar la IA para abordar de forma general los problemas actuales. En su lugar, el proyecto se centra en la anticipación y la comprobación de las amenazas emergentes, combinadas con la personalización de la solución de IA.
   El enfoque promovido por el proyecto, tal y como se describe en la metodología, tiene como objetivo asegurarse de que nuestra comprensión de las futuras formas de interacción humana en el ámbito de las RSH siga el ritmo de los avances en las interacciones humanas artificiales, de modo que estos dos campos puedan obtener el máximo beneficio el uno del otro. En este sentido, el aprendizaje y la adaptabilidad pueden desarrollarse y asegurarse eficazmente en ***guardIA*** GPV. Nuestro modelo de GPV será el primer agente de software empático capaz de valorar las emociones humanas y ayudar a los humanos a interactuar de forma fiable y segura. La empatía fomenta las relaciones sociales entre los individuos creando vínculos duraderos que permiten el desarrollo de comportamientos prosociales[18] y tiene un efecto positivo en las interacciones. Los agentes con capacidades empáticas han demostrado, mediante estudios empíricos, ser más confiables, amables y creíbles[19]. Nuestro GPV será capaz de interactuar de forma empática con el usuario, mejorando la relación de confianza entre el GPV y el usuario en sus interacciones a largo plazo. Para desarrollar las interacciones afectivas del agente nos basaremos en diferentes teorías de la psicología y la sociología, como las teorías de valoración (también conocidas como teorías de appraisal)[20] y los modelos de representación dimensional[21] y categórica de las emociones[22].

2. Los psicólogos cognitivos han estudiado desde hace mucho tiempo cómo la percepción humana es limitada, y cómo el procesamiento de la información y la toma de decisiones pueden ser irracionales al basarse en heurísticos y estar sesgados de cierta manera[23]. Se sabe que muchos de estos heurísticos y sesgos son beneficiosos desde una perspectiva evolutiva, pero los sistemas de IA diseñados para maximizar el beneficio de los agentes comerciales (o algo peor) son capaces de explotarlos, a veces en detrimento del usuario individual. Las vulnerabilidades que se derivan de estos rasgos humanos están destinadas a amplificarse en las RSH. ***guardIA*** innovará mediante la creación de modelos computacionales (incluyendo modelos de aprendizaje automático para garantizar la personalización de estos modelos) capaces de identificar los sesgos y heurísticos más importantes que los individuos utilizan para navegar por las interacciones sociales y los flujos de información en las RSH, y mediante la validación (y, cuando sea necesario, el ajuste) de estos modelos para el contexto específico de las RSH. Entre los ejemplos de sesgos y heurísticos que son candidatos principales para ser modelados en ***guardIA*** se incluyen (pero no se limitan a): el sesgo de confirmación[24] (dado un modelo para el estado de creencias del usuario, el GPV de ***guardIA*** puede detectarlo); el anclaje[25], el heurístico de disponibilidad y el efecto de mera exposición[26] (dado el conocimiento de las experiencias pasadas o recientes del usuario, el GPV de ***guardIA*** lo señala), el sesgo de autoridad (dado un modelo de confianza y reputación, el GPV de ***guardIA*** puede señalar la confianza indebida de su usuario

[18] Mark H Davis.Empathy: A social psychological approach. Routledge, 2018.

[19] Paiva, A., Leite, I., Boukricha, H., & Wachsmuth, I. (2017). Empathy in virtual agents and robots: A survey. ACM Transactions on Interactive Intelligent Systems (TiiS), 7(3), 1-40.

[20] Andrew Ortony, Gerald L Clore, and Allan Collins.The cognitive structure of emotions. Cambridge University Press, 1990. Richard S Lazarus and Richard S Lazarus.Emotion and adaptation. OxfordUniversity Press on Demand, 1991. Klaus R Scherer, Angela Schorr, and Tom Johnstone. Appraisal processes in emotion: Theory, Methods, Research. Oxford University Press, 2001.

[21] Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament.CurrentPsychology, 14(4):261–292, 1996.
James A Russell and Lisa Feldman Barrett. Core affect, prototypical emo-tional episodes, and other things called emotion: dissecting the elephant.Journal of Personality and Social Psychology, 76(5):805, 1999.

[22] Paul Ekman. An argument for basic emotions.Cognition & Emotion,6(3-4):169–200, 1992.

[23] Kahneman, Daniel, Amos Tversky, and Paul Slovic, eds. 1982. Judgment Under Uncertainty: Heuristics & Biases. Cambridge, UK: Cambridge University Press.

[24] Oswald ME, Grosjean S (2004). "Confirmation Bias". In Pohl RF (ed.). Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory. Hove, UK: Psychology Press. pp. 79–96.

[25] Zhang Y, Lewis M, Pellon M, Coleman P (2007). "A Preliminary Research on Modeling Cognitive Agents for Social Environments in Multi-Agent Systems"

[26] Bornstein RF, Crave-Lemley C (2004). "Mere exposure effect". In Pohl RF (ed.). Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory. Hove, UK: Psychology Press. pp. 215–234.

en la autoridad). Además, el mismo nombre/etiqueta de emoción puede interpretarse de forma diferente en distintas culturas[27]. Por tanto, para evitar errores culturales es necesario personalizar los modelos emocionales al entorno cultural del usuario.

Además, aparte del impacto directo de ***guardIA*** GPV a nivel individual (microscópico) (el beneficio directo para un usuario de VPG), también investigaremos y modelaremos los efectos de ***guardIA*** GPVa nivel mesoscópico de los círculos sociales, y a nivel macroscópico de la sociedad. Si es adoptada por un gran número de individuos que interactúan en una gran RSH (de escala similar a las redes sociales actuales), los efectos mesoscópicos y macroscópicos sobre la polarización de la opinión, el riesgo de conflicto, la propagación de la confianza y la influencia, etc., pueden ser sustanciales. Los modelos actuales para estos fenómenos siguen limitándose a una comprensión muy superficial y poco dimensional de los sujetos que interactúan entre sí, y a modelos simples de propagación de la información o la confianza entre individuos. Los modelos basados en ***guardIA*** tienen el potencial de ir mucho más allá.

3. El enfoque radicalmente nuevo hacia la interacción multimodal. La mayoría de los marcos de diálogo se centran en los movimientos de diálogo que se realizan y menos en la estructura general de una conversación. Aunque el enfoque de la Actualización del Estado de la Información (Information State Update ISU[28]) incluye una agenda como parte del estado de un agente, los marcos existentes dicen muy poco sobre cómo modelar y mantener una agenda para una conversación, dejando la tarea de construir dicha agenda al diseñador de la conversación o al usuario. En cambio, el Marco de Conversación Natural (NCF[29]) se centra más en los aspectos estructurales de la conversación y propone una larga lista de patrones que se dan de forma natural en la conversación. El NCF también tiene más que decir sobre cómo se gestiona una conversación que la mayoría de las teorías del diálogo. Sin embargo, el NCF está pensado principalmente para guiar la práctica del diseño de experiencia de usuario (UX) conversacional, y no explica cómo los patrones conversacionales podrían informar a un módulo de gestión del diálogo, de forma parecida a como se integra el ISU con los movimientos conversacionales. ***guardIA*** definirá un marco computacional basado en agentes que integre ambos, lo que permitirá una forma radicalmente más flexible de gestión del diálogo de una conversación.

4. Situar el proyecto en un entorno de RSH: hacer frente a las futuras formas de desinformación y manipulación en RSH.

   guardIA es el primer asistente personal que trabaja para frenar los efectos negativos de fenómenos como las burbujas de filtro (un efecto debido a la omnipresencia de los sistemas de recomendación y personalización, que explotan los puntos débiles individuales) y las cámaras de eco (causadas en gran parte por el sesgo de confirmación), estando diseñado para apoyar el grado de conciencia de los usuarios hacia las formas e intenciones de la información que reciben en RSH.

   guardIA es una solución concreta a los retos que plantea la sobrecarga de información (incluida la desinformación y los equivalentes de RSH al cebo sensacionalista), que permite a los usuarios comprender por qué han sido objeto de un determinado conjunto de información, y tener una experiencia interactiva personalizada en línea.

5. IMPACTO

Se espera que los resultados de la investigación desarrollada den lugar a avances en contribuciones científicas y tecnológicas, que mejoren la aceptación de las tecnologías de IA por los usuarios y para proporcionar una interacción social más segura. En concreto:

1. Desarrollo de una herramienta de protección de la interacción social basada en la IA, primera en su género, en el ámbito de la RSH.
2. Desarrollo de un marco de gestión del diálogo innovador y más flexible con capacidades de corregulación y afectivas que garanticen una interacción sofisticada.
3. Estudio de experiencias multisensoriales para una interacción empática y de confianza.
4. Seguimiento multisensorial de las emociones abstraído a un nivel conceptualmente alto.
5. Diseñar estrategias para promover la IA explicable apoyando todo el proceso de definición de la tecnología.
6. Asegurar la implicación y la participación de los grupos objetivo en el diseño, el uso y las pruebas de la tecnología, ayudará a generar confianza y a mejorar la creación de valor.

[27] Joshua Conrad Jackson, Joseph Watts, Teague R. Henry, Johann-MattisList, Robert Forkel, Peter J. Mucha, Simon J. Greenhill, Russell D. Gray,and Kristen A. Lindquist. Emotion semantics show both cultural variationand universal structure.Science, 366(6472):1517–1522, 2019.

[28] Traum, D. R., & Larsson, S. (2003). The information state approach to dialogue management. In Current and new directions in discourse and dialogue (pp. 325-353). Springer, Dordrecht.

[29] Moore, R. J., & Arar, R. (2019). Conversational UX Design: A Practitioner's Guide to the Natural Conversation Framework. ACM.

7. Creación de un código de conducta que tenga en cuenta las necesidades y percepciones de los usuarios, como se indica en la T7.3.
8. Evaluación del nivel de confianza de los usuarios en las interacciones sociales virtuales a lo largo del proyecto para alimentar el diseño y la implementación de la solución ***guardIA.***
9. Las actividades de difusión demostrarán las capacidades del sistema a un público más amplio.
10. El paradigma tecnológico propuesto por ***guardIA*** podría ser utilizado por diversas entidades: usuarios finales, motores de búsqueda, plataformas de medios sociales y medios de comunicación. El proyecto también investigará otros posibles objetivos.
11. Si la solución propuesta por ***guardIA*** se adopta de forma generalizada y logra sus objetivos, podría tener un efecto disuasorio: la producción de prácticas de interacción social engañosas y manipuladoras disminuiría.

6. DISEMINACIÓN Y EXPLOTACIÓN DE RESULTADOS.

La difusión de los resultados y las actividades de comunicación son cruciales dentro del proyecto ***guardIA***, como demuestra un paquete de trabajo completo (WP7). Se realizarán grandes esfuerzos para garantizar una comunicación eficaz sobre los resultados del proyecto y sobre el propio proyecto. Uno de los medios previstos para difundir los resultados del proyecto será la publicación científica. Se elegirá preferentemente editoriales que respeten los intereses de los autores y acepten la publicación en acceso abierto (con un periodo de embargo). Se utilizará un repositorio de acceso abierto, conectado a las herramientas propuestas por la Comisión Europea (open AIRE), para dar acceso a las publicaciones y a unos metadatos bibliográficos en formato estándar.

Objetivos de difusión
Se realizarán grandes esfuerzos para garantizar una comunicación bidireccional continua con los actores interesados y los grupos objetivo identificados Para ello se realizarán actividades participativas capaces de involucrar de forma efectiva a las partes interesadas, los usuarios finales y el público en general en la contribución al proceso de investigación de ***guardIA***, la definición de escenarios, las actividades de prueba (tanto para la recopilación de datos iniciales como para los estudios de validación) y el diseño de GPV. El proyecto ***guardIA*** se dirige a un amplio grupo de partes interesadas para su difusión. A continuación se presenta una primera muestra:

- Comunidad investigadora: ***guardIA*** aportará avances innovadores en su campo y, por lo tanto, los resultados del proyecto llegarán a toda la comunidad investigadora y se dirigirán específicamente a ella. Esto puede conducir a nuevas mejoras de la solución desarrollada.
- Partes interesadas privadas: Los profesionales de los medios de comunicación, los medios sociales y el sector educativo son cruciales para la aceptación de la solución. Se les presentará el potencial de rendimiento futuro en términos de innovación social o económica para que puedan integrarlo en su estrategia de desarrollo.
- Otras partes interesadas son: Responsables políticos, inversores, usuarios finales, otros proyectos relevantes.

Medios de difusión

Se llegará a estos destinatarios utilizando medios de difusión específicos, y en particular

- Base de datos técnica: Asegurar el acceso a la información relativa al desarrollo del conocimiento y la innovación que se producirá durante el proyecto. Estos datos estarán disponibles cuando sea posible en el sitio web de ***guardIA*** o se protegerán cuando sea necesario.
- Presentaciones en eventos internacionales: se realizarán presentaciones en ferias/conferencias profesionales para presentar los resultados del proyecto ante industriales, inversores y autoridades públicas.
- Artículos de divulgación: se publicarán al menos 4 artículos de divulgación en revistas no científicas. Se dirigirán a inversores, proveedores tecnológicos, responsables políticos, plataformas sociales, pero también a potenciales usuarios finales y al público en general, presentando los beneficios sociales de la solución ***guardIA***.
- Asistencia a eventos relevantes como IJCAI, AAMAS, IVA, ECAI, AAAI, AIES, Persuasive Technologies, etc.
- Publicaciones científicas: se publicarán al menos 15 artículos en revistas científicas para presentar los resultados destacados que se espera obtener en el proyecto ***guardIA***. Se considerarán las revistas científicas de alto factor de impacto revisadas por pares de los editores científicos más relevantes (por ejemplo, JAI, IEEE Transactions on Affective Computing, Nature Machine Intelligence, IEEE Computational Intelligent Magazine, Engineering Applications of Artificial Intelligence, JAAMAS, JAIR, AIJ).

Las actividades adicionales de apoyo a los objetivos de difusión incluirán:

- Diseño y gestión de un sitio web dedicado al proyecto
- Gestión de los canales sociales dedicados al proyecto: Los canales de medios sociales de ***guardIA*** se utilizarán para salvar las fronteras disciplinarias, construir y comprometer a la comunidad de investigación ampliada de ***guardIA***

(T7.5) y asegurar el intercambio de conocimientos con los actores relevantes de la industria.
● Otros: Se enviarán comunicaciones o boletines periódicos a los medios de comunicación, a las entidades privadas y públicas de e-learning y a los representantes públicos. También ser enviarán boletines informativos a las principales partes interesadas identificadas. Se publicarán varios artículos de difusión abierta en periódicos nacionales y conferencias de prensa.

HOJA DE RUTA PARA EL DESARROLLO

Las estrategias de explotación detalladas se desarrollarán como parte del paquete de trabajo 7. Estas estrategias prepararán el camino para las actividades de explotación que se llevarán a cabo a lo largo del proyecto y deberán facilitar la transición desde el entorno de la investigación y el desarrollo tecnológico hasta la asimilación por parte del mercado. Durante estas actividades, se establecerá una distribución de los resultados esperados del proyecto ***guardIA*** en función de las capacidades de explotación y las expectativas de cada beneficiario.
La UPV garantizará la explotación científica de los resultados de ***guardIA*** a través de TAILOR, una red de centros de investigación de excelencia sobre los Fundamentos de la Inteligencia Artificial Confiable, y utilizará los hallazgos y soluciones más relevantes de ***guardIA*** para la mejora y el desarrollo posterior de productos conectados como GRSK[30], Pesedia[31], U-tool[32] y su plataforma de e-learning.

PLAN DE GESTIÓN DE DATOS

Al inicio del proyecto, se creará un Plan de Gestión de Datos (DMP) detallado (D8.4).

¿Qué tipos de datos generará/recopilará el proyecto?

guardIA recogerá datos de voluntarios adultos. Los puntos de recogida de datos en los que pueden surgir cuestiones éticas son: el análisis de los miembros del estudio de casos (WP2, T2.6, WP5); las actividades de difusión (WP7); los casos de validación y los datos resultantes (WP5). Se recogerán los siguientes tipos de datos: biométricos (reconocimiento facial, etc.), datos personales estándar (edad, sexo, etc.). Se crearán dos categorías de datos, la primera (biométricos) para la creación de conjuntos de datos para los experimentos y el entrenamiento de la IA serán anonimizados de forma completa e irreversible, la segunda (direcciones de correo) para la difusión del proyecto cumplirán con las normas LGPD.

¿Qué normas se utilizarán?

Los datos recogidos y creados durante el proyecto serán utilizados para entrenar y mejorar los algoritmos, con el fin de aumentar la capacidad del GPV y se pondrán a disposición del público cuando sea posible. Los datos se tratarán de acuerdo con la legislación aplicable en cada estado (incluido la LGPD) y se someterán a una revisión ética nacional, regional o institucional adecuada.

¿Cómo se explotarán y/o compartirán/se harán accesibles estos datos para su verificación y reutilización?

Los datos se almacenarán en una base de datos única en un servidor europeo para garantizar que se apliquen las leyes europeas e irán acompañados de metadatos que ayuden a encontrarlos y utilizarlos. Todos los datos se almacenarán también en un servidor de copia de seguridad para evitar una posible pérdida. Se utilizarán conjuntos de datos abiertos para el proyecto, con el fin de alimentar el proceso de sugerencia de prototipos. Los resultados del proyecto se pondrán a disposición del público, incluyendo tanto el código fuente del análisis de datos como los conjuntos de datos, con el fin de facilitar su reproducibilidad. Se llevará a cabo un análisis de datos sensibles para evitar problemas inesperados de fuga de datos y garantizar el pleno anonimato de los datos registrados. El consorcio también aplicará el requisito de adherirse al Open Research Data Pilot.

7. PLAN DE TRABAJO.

En la figura 1 se muestra el plan general de trabajo donde se incluye la relación entre los distintos WPs del proyecto. En la figura 2 se muestra la planificación temporal de WPs y las tareas que los constituyen.

[30] http://www.gti-ia.upv.es/sma/tools/GRSK/overview.php

[31] https://riunet.upv.es/handle/10251/69308

[32] http://gti-ia.webs.upv.es/results/utool.php

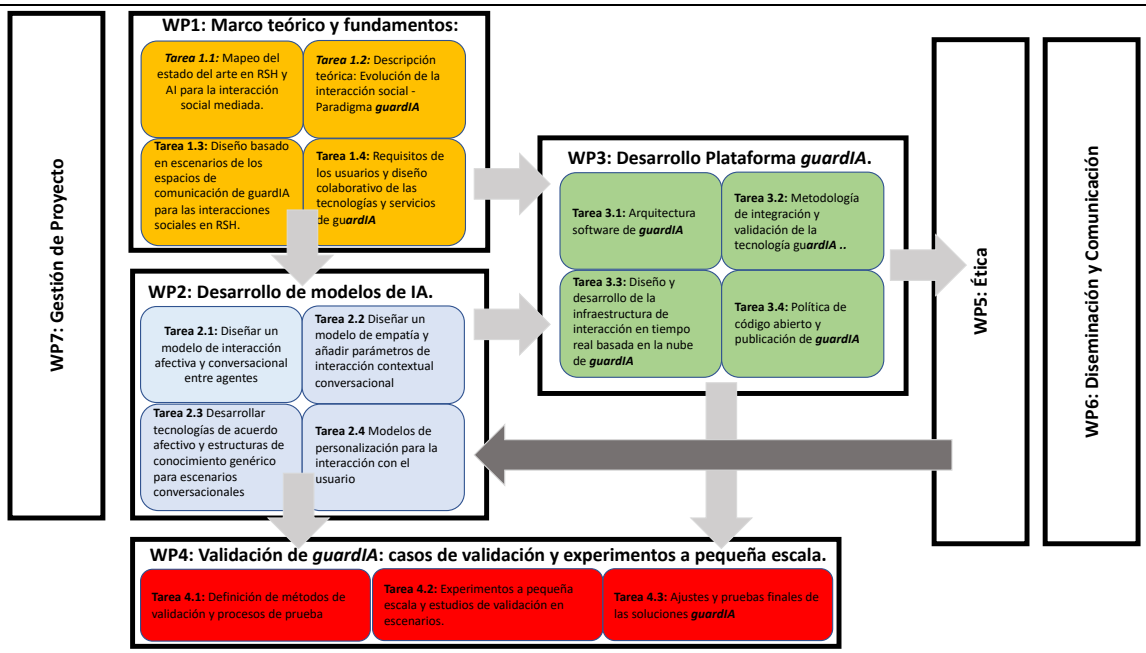

Figura 1: Plan General de Trabajo

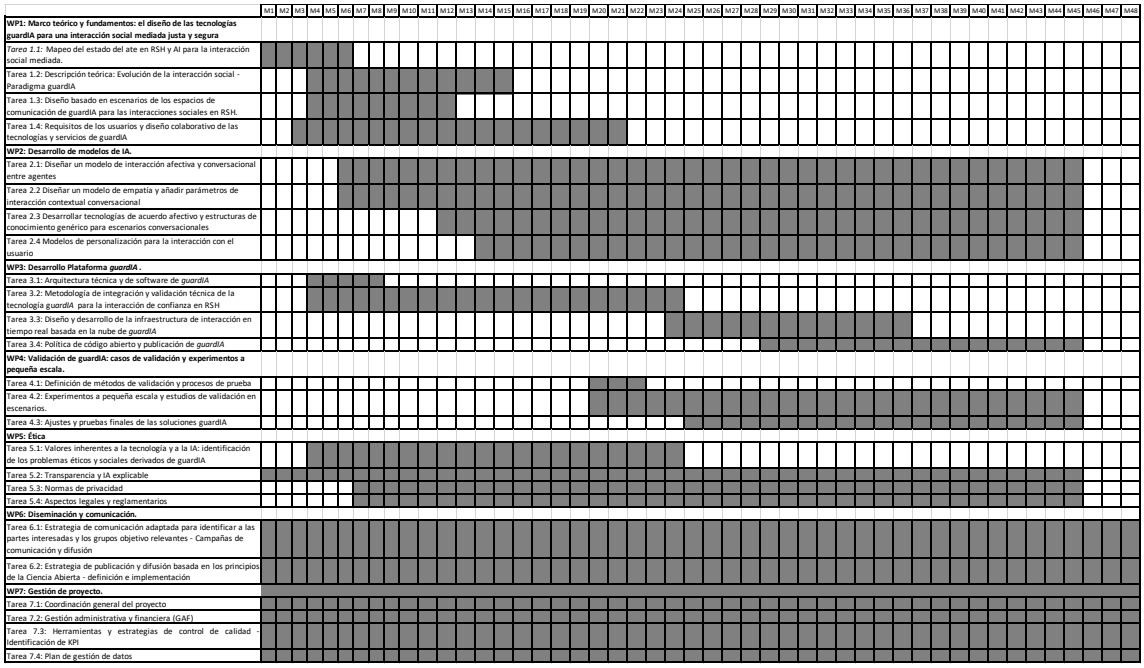

Figura 2. Planificación Temporal de Tareas

WP1: Marco teórico y fundamentos: el diseño de las tecnologías guardIA para una interacción social mediada justa y segura

Objetivos: Es objetivo de este WP el desarrollo del paradigma ***guardIA***, un modelo integral de interacciones sociales en RSH dotado de una arquitectura de sistemas para mejorar su seguridad y eficacia.
Además, en este WP se identificarán los campos tecnológicos que formarán parte de la plataforma tecnológica ***guardIA*** para una interacción social segura y mediada, y se especificará cómo se diseñarán, desarrollarán y ampliarán. Se investigarán las siguientes áreas principales 1) Espacios comunicativos basados en escenarios en RSH; 2) Requisitos de los usuarios.
Para ello, en primer lugar, se realizará un análisis del estado del arte en cada una de estas áreas y, a continuación, se identificarán los retos y las oportunidades relacionados con el desarrollo de la plataforma tecnológica ***guardIA***.

Tarea 1.1: Mapeo del estado del arte en RSH y AI para la interacción social mediada
Calendario: M1 a M6
Objetivo de la tarea: Análisis de la investigación relevante en ciencias sociales que pueda ser útil para mediar la interacción social en entornos virtuales para mejorar la seguridad y la eficacia.
Plan de acción: La investigación y las tecnologías relevantes que deben estudiarse, así como sus interdependencias, se han trazado en D1.1, para incluir:

a. Tecnología de medición de la percepción y el comportamiento;
b. Asistentes digitales personales basados en la IA, incluida la investigación y las tecnologías de agentes empáticos y conversacionales;
c. Psicología cognitiva (incluidos los sesgos cognitivos y la heurística, los modelos de valoración emocional y cognitiva de la información y las teorías de la persuasión);
d. Tecnologías de IA y enfoques de modelado computacional para modelar la interacción social multilateral y los comportamientos emergentes en un contexto social (ciencia de las redes; modelos de formación y propagación de opiniones; modelos de maximización de la influencia; modelos de confianza y teoría del equilibrio social; causas de la polarización, burbujas de filtro y cámaras de eco);
e. Tecnologías de IA para el modelado de usuarios (medidas teóricas de la información sobre el interés subjetivo de la información; modelado de máxima entropía de los estados de creencia; modelos de preferencia; modelado de sesgos cognitivos y heurísticos).

Resultado: D2.1.

Tarea 1.2: Descripción teórica: Evolución de la interacción social - Paradigma ***guardIA***
Calendario: M4 a M15
Objetivo de la tarea: Desarrollar el paradigma ***guardIA***, un modelo completo de interacciones sociales en entornos virtuales, y una arquitectura de sistemas para mejorar su seguridad y eficacia.
Plan de acción: D1.1 y D2.1 proporcionarán una base sólida y las herramientas necesarias para desarrollar el paradigma ***guardIA***. En un primer paso, se desarrollarán modelos matemáticos/causales formales para el comportamiento individual de los actores humanos en un entorno social virtual (nivel microscópico), modelando los aspectos relevantes de la recogida, percepción, procesamiento y comprensión de la información, las preferencias individuales, la formación de opiniones, la confianza, las emociones, etc. En un segundo paso, estos modelos se utilizarán para estudiar y simular matemáticamente/causalmente el comportamiento emergente mesoscópico (a nivel de subred) y macroscópico (a nivel de red) de dichos modelos individuales.
En comparación con el estado del arte, los modelos desarrollados tendrán en cuenta explícitamente el hecho de que los actores humanos serán conscientes de que su espacio social puede estar poblado tanto por actores humanos como por bots, pero que sus comportamientos son indistinguibles. Las repercusiones de esta conciencia y de la falta de información pueden repercutir en la confianza mutua. Además, los modelos podrán incorporar más información procedente de múltiples sensores, lo que permite que los modelos sean más precisos y eficaces. Esto nos lleva al resultado D1.2. Estos modelos revelarán las dependencias de la eficacia y la seguridad de estas interacciones en una serie de parámetros controlables del modelo. Dichos parámetros incluirán rasgos y variables de estatus a nivel humano (por ejemplo, las creencias y opiniones previas, la tendencia personal a la confianza frente a la sospecha, la emocionalidad frente a la racionalidad y la dependencia de la heurística cognitiva, los sesgos cognitivos y la susceptibilidad a diferentes tipos de información, como las opiniones extremas y las conspiraciones frente a la opinión moderada y mayoritaria), y para el nivel mesoscópico y macroscópico también propiedades a nivel de red (grado medio, coeficiente de agrupación, asertividad de grado, varios tipos de heterogeneidad, etc.). Al menos en principio, estos parámetros podrían ser mediados por ***guardIA*** GPV. Un tercer paso en esta tarea identificará estos parámetros controlables y evaluará su potencial para mejorar la seguridad y la eficacia. Esto nos lleva a la tarea D1.3.
Resultado: D1.2, D1.3

Tarea 1.3: Diseño basado en escenarios de los espacios de comunicación de ***guardIA*** para las interacciones sociales en RSH
Calendario: M4 a M12
Objetivo de la tarea: Identificar los escenarios de interacción social en entornos virtuales que tienen el mayor potencial para ser impactados positivamente por las tecnologías ***guardIA***, basándose en el caso de estudio seleccionado.
Plan de acción: La tarea comienza con el análisis del caso de estudio seleccionado. Los escenarios definidos servirán de referencia tanto para el punto 2.6 (actividades de recopilación de datos) como para el 5.2 (casos de validación). Actividades:

a. Análisis en profundidad de los campos de aplicación y establecimiento de criterios para fijar escenarios específicos dentro de cada uno de ellos.
b. Análisis de los requisitos específicos dentro de cada escenario elegido (volumen potencial de información, número esperado de usuarios, tipo de interacción social, entre otros).
c. Identificación de los riesgos actuales de manipulación/desinformación dentro de cada escenario.
d. Identificación de KPI (o indicadores clave de desempeño) dentro de cada escenario: qué resultados debemos obtener dentro de cada uno para considerar que nuestro sistema es eficiente.

e. Definición de los escenarios óptimos para probar la tecnología ***guardIA*** y transferirla al equipo de diseño.

Metodología principal: a. Investigación documental, b. Entrevistas con expertos, c. Entrevistas con expertos en eLearning, d. Informe final (D1.4) describiendo las situaciones virtuales que se recrearán en el proyecto para probar la solución ***guardIA.***
Resultado: D1.4

Tarea 1.4: Requisitos de los usuarios y diseño colaborativo de las tecnologías y servicios de ***guardIA***
Calendario: M3 a M21
Objetivo de la tarea: El objetivo de esta tarea es desarrollar requisitos para la tecnología ***guardIA*** basados en los resultados de las tareas 1.1 y 1.2 con un enfoque multidisciplinar. Los requisitos dependerán de las tecnologías subyacentes que soportan la arquitectura de ***guardIA***.
Plan de acción: La tarea comienza con la colaboración con expertos procedentes de diferentes disciplinas (ciencias sociales, ciencia de los datos, expertos en el dominio) y el examen de los escenarios definidos en la Tarea 1.3 y los avances iniciales del WP1, en particular los principios derivados de la teoría en torno a la confianza, la valoración de las emociones, etc. Se realizará un análisis de la tarea sobre los escenarios derivando los requisitos a la tecnología y las suposiciones sobre los usuarios finales. La tarea incluye también el desarrollo de visualizaciones, prototipos y maquetas para ejemplificar los requisitos. Los requisitos se determinan mediante la gestión de los perfiles de los usuarios en los gráficos sociales, tanto desde el punto de vista conceptual como funcional. Los gráficos sociales, además de los niveles de interacción, tratarán los rasgos de personalidad y los aspectos culturales. D1.5 incluirá características básicas, requisitos e indicadores de evaluación.
Resultados: D1.5

Entregables WP1:
D1.1 Análisis detallado de las tecnologías existentes y una valoración de su potencial futuro. (M6)
D1.2 Modelos matemáticos/causales para la interacción social a nivel microscópico, mesoscópico y macroscópico, parametrizados con parámetros relacionados con el estado y los rasgos humanos cognitivos, emocionales y sociales. (M12)
D1.3 Descripción detallada de los parámetros en los modelos de D1.2 que pueden ser instrumentales para el GPV (M15)
D1.4 Escenarios virtuales para la interacción futura y fiable de la RSH(M12)
D1.5 Lista de requisitos del usuario para las tecnologías ***guardIA*** (M6, M14, M21)

WP2: Desarrollo de modelos de IA.

Objetivos: El objetivo de este WP es definir y desarrollar los elementos innovadores basados en la IA necesarios para la Plataforma Tecnológica ***guardIA*** en varias áreas principales: organizaciones complejas de agentes emocionales similares a los humanos; computación afectiva y modelado de la personalidad; corregulación y gestión del diálogo; experiencia del usuario/diseño de la interacción. Siguiendo el marco teórico definido en el WP1, tanto los agentes software como las estructuras basadas en sistemas multiagente (SMA) se diseñarán mediante el uso de modelos computacionales basados en las teorías de valoración de las emociones procedentes de los modelos de D1.2 y los parámetros descritos en D1.3. Estos modelos permitirán representar computacionalmente el comportamiento humano racional y afectivo, lo que implica la representación de creencias y la toma de decisiones, teniendo en cuenta los rasgos de personalidad, el estado de ánimo, la empatía y los vínculos afectivos. Además, deberán adaptarse al entorno cultural en el que sean utilizados para evitar sesgos o malinterpretaciones culturales. Nuestros modelos darán lugar a la creación de las habilidades de razonamiento y conversación de la GPV, que se integrarán en el paquete de trabajo 4 con la plataforma de e-learning adoptada. El plan y un calendario de este WP estará coordinado con los WPs 2, 4 y 5.

Tarea 2.1: Diseñar un modelo de interacción afectiva y conversacional entre agentes
Calendario: M6 a M45
Objetivo de la tarea: Diseñar un modelo computacional de agente afectivo conversacional basado en la teoría de valoración.
Plan de acción: Desarrollar el modelo basado en agentes: Esta tarea se llevará a cabo centrándose en el proceso de toma de decisiones de la VPG, y se guiará por las prácticas de RRI del WP6. Un enfoque candidato para guiar el desarrollo de estos modelos es la arquitectura de agentes *GenIA3*. En nuestro modelo, los agentes serán capaces de exhibir un comportamiento racional y afectivo similar al comportamiento humano tal y como se estudia en las teorías desarrolladas en la Tarea1.2 e incluirán los parámetros obtenidos en la Tarea1.3. Esto implica establecer un sistema de percepción (análisis de sentimiento en texto, en audio, reconocimiento facial de sentimientos) que permita al agente conocer el estado de ánimo del usuario basado en los desarrollos de la Tarea 2.5 y crear un modelo de representación de emociones que establezca un marco de representación común para las emociones percibidas en usuarios humanos y agentes software.
Diseñar un marco de conversación basado en ISU y NCF: La corregulación es un aspecto clave de la interacción social, donde todas las partes de la interacción son capaces de alterar la dirección de la misma. Los marcos de diálogo actuales no admiten la corregulación, sino que se basan en diálogos con guión o en estructuras de agenda fija. Al fusionar el enfoque de actualización del estado de la información con el marco de conversación natural (de Moore), desarrollaremos un marco de gestión del diálogo

innovador y más flexible que apoye una interacción más natural.
Resultado: D2.1 Modelos de interacción de agentes conversacionales y afectivos (M16, M24, M34, M45).

Tarea 2.2 Diseñar un modelo de empatía y añadir parámetros de interacción contextual conversacional
Calendario: M6 a M45
Objetivo de la tarea: Diseñar un modelo computacional de un agente empático conversacional.
Plan de acción: Vinculación afectiva y empatía entre el usuario y el agente: Integrar un modelo computacional empático en el GPV afectivo para establecer un vínculo afectivo entre el usuario y el agente. Este agente utilizará un modelo de representación de emociones al usuario que permitirá que el agente represente y exprese las emociones usando la misma terminología que el usuario teniendo en cuenta las variaciones producidas en el entorno cultural. Partiremos de los modelos empáticos analizados en la Tarea 1.1 para adaptarlos e incorporarlos al agente afectivo desarrollado en la Tarea
2.1. Este modelo se utilizará para establecer un vínculo afectivo agente-usuario con el fin de mejorar la confianza del usuario en el agente.
Añadir una lógica para afinar los parámetros contextuales y de interacción (incluyendo, por ejemplo, el afecto)
Para poder afinar las respuestas de los agentes conversacionales, el marco se ampliará con parámetros contextuales y de interacción como el afecto, tal y como se identificó en la tarea 1.3. Al ampliar el marco con estos parámetros podemos tener en cuenta cómo responder, por ejemplo, a las expresiones negativas o positivas de un usuario.
Resultado: D2.2 Modelo Computacional de Empatía (M15, M24, M34, M45) y D3.4 Parámetros de Interacción Contextual Conversacional (M16, M24, M34, M45)

Tarea 2.3 Desarrollar tecnologías de acuerdo afectivo y estructuras de conocimiento genérico para escenarios conversacionales
Calendario: M12 a M45
Objetivo de la tarea: Desarrollar modelos computacionales que permitan a los agentes alcanzar acuerdos considerando sus rasgos afectivos (negociación automatizada, confianza, reputación, organizaciones virtuales, normas, argumentación) y estructuras de conocimiento genéricas para escenarios conversacionales específicos.
Plan de acción: Crear tecnologías de acuerdo afectivo: partiendo de los modelos y estructuras de los posibles flujos de información que tienen lugar en las interacciones entre agentes y entre humanos y agentes, tal y como se definen en D1.2 y D1.3, se crearán tecnologías de acuerdo (por ejemplo, de argumentación, negociación, confianza y reputación) para incluir la influencia que los factores humanos, como el estado de ánimo y la personalidad, ejercen en las interacciones y dinámicas sociales. Se analizarán los componentes normativos de la cognición y la emoción para crear un marco que represente un contexto normativo (reglas legales, éticas y morales). Añadir estructuras de conocimiento genéricas para escenarios conversacionales: Diferentes escenarios requieren diferentes tipos de conocimientos previos. Diseñaremos estructuras de conocimiento genéricas que el agente pueda utilizar en las conversaciones y las relacionaremos con los casos de validación para poder instanciar el conocimiento conversacional necesario para cada caso.
Resultado: D2.4 Integración de Tecnologías del Acuerdo Afectivas en Estructuras de Conocimiento Genéricas para Escenarios Conversacionales (M24, M34, M45).

Tarea 2.4 Modelos de personalización para la interacción con el usuario
Calendario: M14 a M45
Objetivo de la tarea: Adaptar el modelo de agente a las especificidades de cada usuario. *Plan de acción:* Personalizar agentes software capaces de representar e interactuar con humanos mostrando un comportamiento racional y afectivo: las tecnologías desarrolladas en la Tarea 2.3 se conectarán con las desarrolladas en la T2.2 para permitir al GPV mantener una conversación personalizada con humanos y agentes interactuando en los escenarios definidos en el caso de validación de la T4.2. Esto incluirá la adaptación de la representación de las emociones al entorno cultural para evitar errores en la percepción, interpretación y expresión de las emociones o el estado de ánimo. También diseñaremos estructuras basadas en SMA que representen características sociales (por ejemplo, características estructurales), tipos de interacciones sociales (por ejemplo, negociación), roles y dinámicas.
Diseñar tecnologías de persuasión para la formación de opinión y el cambio de comportamiento: El GPV será capaz de entablar diálogos persuasivos con el usuario humano, adaptando el formato y el contenido del mensaje a los rasgos de personalidad, el estado de ánimo, la cultura, las preferencias, etc. del usuario para mejorar los resultados de la interacción entre el GPV y el humano. Añadir modelos de memoria para personalizar los diálogos repetidos con los usuarios: Para tener en cuenta el historial de interacciones personales, añadiremos modelos de memoria que hagan un seguimiento de las interacciones anteriores con los usuarios. También desarrollaremos la lógica para utilizar estos modelos de memoria en las conversaciones posteriores (por ejemplo, para evitar las típicas repeticiones que introducen los agentes conversacionales actuales y para basarse en la experiencia de los usuarios aprovechando intercambios conversacionales anteriores).
Resultado: D2.5 Modelos de Personalización para Interacción de Usuarios (M34, M45)

Entregables:

D2.1 Modelos de interacción afectiva y conversacional del agente ***guardIA*** (M16, M24, M34, M45)
D2.2 Modelo computacional de empatía ***guardIA*** (M16, M24, M34, M45)
D2.3 Parámetros de interacción contextual conversacional ***guardIA*** (M16, M24, M34, M45)
D2.4 Integración de tecnologías de acuerdo afectivo en estructuras de conocimiento genéricas para escenarios conversacionales (M24, M34, M45)
D2.5 Modelos de personalización para la interacción con el usuario (M34, M45)

WP3: Desarrollo Plataforma *guardIA*.

Objetivos: El WP3 se centra en el diseño y desarrollo de la arquitectura software de ***guardIA***. Propondrá una nueva arquitectura que soporte la interacción social en la plataforma de e-learning en la nube y que sea capaz de procesar los datos de detección e interacción en tiempo real. Como primer paso hacia este objetivo, se diseñará una arquitectura que defina los requisitos, la tecnología utilizada y la interacción entre los principales componentes funcionales de ***guardIA***. Sobre la base del diseño de la arquitectura, se definirá una metodología de integración y validación técnica para la integración continua del desarrollo de software realizado en el proyecto con el fin de garantizar una infraestructura de software integrada que pueda utilizarse con fines de prueba y validación (WP4). Esta metodología definirá los procesos de desarrollo que se utilizarán para integrar los distintos componentes desarrollados en el WP1, WP2. Además, también investigaremos y diseñaremos los requisitos en tiempo real para la interacción social en la nube. Para ello, se desarrollará una infraestructura de interacción social basada en la nube. Por último, este paquete de trabajo es responsable de definir la política de código abierto y la liberación real del software desarrollado en el proyecto.

Tarea 3.1: Arquitectura software de ***guardIA***
Calendario: M4 a M8
Plan de acción: Esta tarea identificará los requisitos (no) funcionales y los componentes funcionales, y creará el diseño de alto nivel de la arquitectura funcional de ***guardIA***. Esta arquitectura proporcionará un modelo de las principales funciones del sistema y sus interacciones, definiendo el modo en que estas funciones operarán conjuntamente para implementar los requisitos. Esta tarea se refiere a la elaboración de un documento de diseño en el que se definen detalladamente todas las tecnologías utilizadas en la arquitectura de ***guardIA***. Cada componente se desarrollará utilizando su propio lenguaje de autor específico (por ejemplo, Python, Java, Jason, etc.) y, en la medida de lo posible, se utilizarán las bibliotecas existentes para aumentar el rendimiento funcional general. Estos componentes incluyen componentes de IA desarrollados en el proyecto basados en la tecnología de agentes desarrollada previamente por el equipo. Además, la arquitectura incluirá una base de datos de interacción con el usuario y una tecnología que facilite el procesamiento de grandes flujos de datos.
Resultado: D3.1 Informe: Arquitectura software de ***guardIA*** (M8)

Tarea 3.2: Metodología de integración y validación de la tecnología ***guardIA*** para la interacción de confianza en RSH
Calendario: M4 a M24
Plan de acción: Esta tarea definirá el proceso de ingeniería de software y la gestión de control de versiones utilizando repositorios basados en Git para integrar los diversos subsistemas de componentes desarrollados en los otros paquetes de trabajo (WP2, WP5) en un sistema y definir un enfoque de integración continua para mantener la arquitectura ***guardIA*** más actualizada para las pruebas y la validación (en WP4). Además, se especificará un enfoque de validación mediante la definición y ampliación de un conjunto de pruebas a lo largo del tiempo. El objetivo de este conjunto de pruebas es evaluar si la agregación de componentes ofrece la funcionalidad global y garantizar que los subsistemas funcionan juntos como el sistema previsto. Para que el proceso de integración sea manejable, se especificarán y mantendrán unas API bien definidas para la interacción con otros componentes y la integración en la arquitectura global. Esto también facilitará la reutilización de los componentes de ***guardIA*** en otros sistemas.
Resultado: D3.2 Metodología de integración y validación (M15, M24)

Tarea 3.3: Diseño y desarrollo de la infraestructura de interacción en tiempo real basada en la nube de ***guardIA***
Calendario: M24 a M36
Objetivos de la tarea: Añadir las características en tiempo real y definir la infraestructura basada en la nube, esta arquitectura se basará en arquitecturas previas desarrolladas por el equipo para sistemas de Fabricación 4.0.
Plan de acción: La ejecución en tiempo real es un requisito clave de la arquitectura de ***guardIA***. Aunque no todo el procesamiento tendrá que ofrecer resultados en tiempo real, los componentes de interacción sí necesitan garantizar unos plazos de respuesta (no críticos) para poder facilitar la interacción natural de los usuarios con el sistema ***guardIA***. Para lograrlo, la infraestructura clave se basará en la nube para poder poner a disposición del sistema los niveles necesarios de recursos informáticos. Se necesitarán recursos en la nube para poder procesar los múltiples y variados flujos de datos relacionados con la interacción social con los usuarios, para transmitir contenidos como vídeos, etc. Se utilizarán arquitecturas y tecnología

existentes relacionada con el IoT.
Resultado: D3.3 Infraestructura de interacción en tiempo real en la nube de ***guardIA*** (M36)

Tarea 3.4: Política de código abierto y publicación de ***guardIA***
Calendario: M29 a M45
Objetivos de la tarea: Definir la política de liberación del código base como código abierto al público.
Plan de acción: La tecnología ***guardIA*** se liberará anualmente y como software de código abierto en sus dos últimas versiones (M36, M48). Definiremos una política y un enfoque de código abierto que especifique qué y cómo publicaremos los distintos componentes y sus especificaciones como código abierto para uso público (software, herramientas y otros productos relacionados con el software, bibliotecas) siguiendo los acuerdos de derechos de propiedad intelectual.
Resultado: D3.4 Política de código abierto y publicación de ***guardIA*** (M33, M45)

Entregables:

D3.1 Informe: Arquitectura software de ***guardIA*** (M8)
D3.2 Metodología de integración y validación(M15, M24)
D3.3 Infraestructura de interacción en tiempo real en la nube de ***guardIA*** (M36)
D3.4 Política de código abierto y publicación de ***guardIA*** (M33, M45)

WP4: Validación de *guardIA*: casos de validación y experimentos a pequeña escala.

Objetivos: El WP4 se basa en los resultados de todos los WPs anteriores y contempla la validación tanto del marco teórico como de las soluciones tecnológicas para la experimentación de sistemas reales previstas en T4.2. Basándose en los escenarios definidos en T1.3, el WP4 implementará un conjunto de estudios de validación para experimentos a pequeña escala.

Tarea 4.1: Definición de métodos de validación y procesos de prueba
Calendario: M20 a M22
Plan de acción: La tarea define el proceso de validación, incluidas las funciones de los socios y el flujo de trabajo. La tarea también incluye configuraciones, procedimientos y tareas para T4.2 y T4.3. También mantiene una lista de componentes y soluciones que deben validarse. Estos se traducen en medidas y se seleccionan los métodos de recogida de datos y los enfoques empíricos adecuados.
Resultado: D4.1: Metodología de validación de ***guardIA*** (M22)

Tarea 4.2: Experimentos a pequeña escala y estudios de validación en escenarios.
Calendario: M20 a M45
Objetivo de la tarea: Validar la capacidad de los escenarios de ***guardIA*** para probar el GPV y garantizar la recopilación de datos de retroalimentación para el ajuste y el desarrollo posterior de la GPV de IA.
Plan de acción: Esta tarea permitirá realizar dos tipos de simulaciones o experimentos:

a. Enfrentarse a usuarios reales en un entorno RSH con situaciones estilizadas, y evaluar sus reacciones, decisiones y emociones. Los resultados deberían permitir la mejora del conjunto de datos para seguir entrenando la GPV de ***guardIA***.
b. Una vez creados y entrenados los modelos de IA para el GPV, los experimentos evaluarán la capacidad de las soluciones ***guardIA*** para apoyar una interacción fiable y empática en HSR.

Resultado: D4.2: Evaluación y validación de ***guardIA*** (M43)

Tarea 4.3: Ajustes y pruebas finales de las soluciones ***guardIA***
Duración: M25 a M45
Objetivo de la tarea: El objetivo de la tarea es realizar una prueba sumativa de la solución ***guardIA*** evaluando los indicadores y medidas de éxito definidos en la tarea 5.1.
Plan de acción: La tarea lleva a cabo una prueba incremental de las soluciones ***guardIA*** basada en la metodología T5.1. Además, la Tarea 5.2 informa de la selección de los enfoques empíricos indicando las tareas, los temas y las escenas RSH adecuadas. El trabajo abordará en particular los objetivos de los indicadores de éxito medibles del proyecto ***guardIA.***
Resultado: D4.3: GPV validado en el entorno correspondiente (M45)

Entregables:

D4.1 Metodología de validación de ***guardIA*** (M22)
D4.2 Evaluación y validación de ***guardIA*** (M43)

D4.3 GPV validado en el entorno correspondiente (M45)

WP5: Ética

Objetivos: El objetivo general del WP5 es identificar los principales problemas éticos y sociales potenciales relacionados con el desarrollo de las tecnologías ***guardIA***, siguiendo las directrices y principios éticos exigidos por la legislación española, de la UE e internacional, así como desarrollar estrategias estructuradas para abordarlos. Dichos mecanismos se elaborarán en forma de recomendaciones para la elaboración de reglamentos y códigos de conducta profesionales que permitan abordar eficazmente las cuestiones éticas, legales y sociales (ELSI). En este WP se incluirán los aspectos de genero y se creará el Comité de Genero encargado de garantizar que los aspectos relacionados con el género se tengan en cuenta en todo el proyecto, este Comité tendrá una estrecha relación con la tarea T7.4.

Tarea 5.1: Valores inherentes a la tecnología y a la IA: identificación de los problemas éticos y sociales derivados de ***guardIA***
Calendario: M4 a M24
Objetivo de la tarea: Esta tarea tiene como objetivo cumplir con el requisito de H2020 de integrar la dimensión social en el diseño, desarrollo e implementación tanto de la investigación como de las tecnologías que pueden ayudar a encontrar soluciones a los problemas de la sociedad.
Plan de acción: Para ayudar a que los resultados del proyecto tengan un impacto en los usuarios finales y en la sociedad, T6.1 identificará los KPI para evaluar los resultados sociales de ***guardIA***, identificará e informará sobre los problemas sociales y las recomendaciones que surjan de ***guardIA***, a través de grupos de discusión y Delphi.
Resultado: D5.1. Estudio del impacto ético, legal y social de g***uardIA*** GPV (, M7, M24)

Tarea 5.2: Transparencia y IA explicable
Calendario: M1 a M45
Objetivo de la tarea: Asegurar que las acciones y decisiones del GPV de ***guardIA*** sean confiables y explicables, para garantizar tanto la auditabilidad como la confianza en el GPV.
Plan de acción: Los modelos en los que se basa la GPV de ***guardIA*** se inspirarán a menudo en teorías establecidas y, por tanto, serán modelos de caja blanca, que son transparentes por diseño. En algunos casos, sin embargo, los modelos serán menos transparentes y se utilizarán técnicas recientes desarrolladas en la comunidad del aprendizaje automático y la IA para explicar los modelos de caja negra y sus acciones. A su vez, esto permitirá a al GPV explicar sus propias recomendaciones y consejos a su usuario. Además, en esta tarea se definirá una lista de métricas pertinentes para evaluar el grado de satisfacción, comprensión, fiabilidad y fidelidad de la explicación proporcionada. Adaptaremos las métricas existentes que se han propuesto recientemente en la literatura de la Inteligencia Artificial Explicable (XAI).
Resultado: D5.2 Desideratas éticas para el GPV (M9, M24, M45)

Tarea 5.3: Normas de privacidad
Calendario: M7 a M45
Objetivo de la tarea: Aunque los GPV de ***guardIA*** serán personales, es decir, accesibles únicamente por el "propietario" individual, algunos aspectos del GPV pueden ser entrenados también con datos procedentes de otros usuarios (para evitar el "problema del arranque en frío"). En estos casos, deben existir garantías de privacidad.
Plan de acción: Todo el modelado se llevará a cabo garantizando la privacidad diferencial o las nociones de privacidad relacionadas. Aunque este es un requisito que se tendrá en cuenta a lo largo del proyecto, esta tarea auditará todos los enfoques de modelado utilizados y los adaptará cuando sea necesario para cumplir con los requisitos de privacidad específicos del proyecto.
Resultado: D5.3 Arquitectura para un GPV explicable y confiable (M24, M45)

Tarea 5.4: Aspectos legales y reglamentarios
Calendario: M7 a M45
Objetivo de la tarea: Esta tarea pretende identificar e integrar los aspectos normativos, autorreguladores (por ejemplo, éticos) y legales del sistema GPV ***guardIA*** e identificar cómo esta tecnología emergente y los patrones de interacción afectarán a los códigos éticos y legales existentes.
Plan de acción: El código ético más reciente sobre publicidad y marketing publicado por la Cámara de Comercio Internacional (CCI) constituye un punto de partida para identificar los principios éticos y jurídicos que deben regir el GPV ***guardIA***. El código también promueve la responsabilidad social, la transparencia, la competencia leal, la protección de datos, la privacidad y la confianza. Si bien este código de la CCI puede servir de piedra angular para los aspectos legales y reglamentarios, todavía debe adaptarse en el contexto de GPV ***guardIA*** para tener en cuenta las nuevas formas de formas de interacción provocadas por la aparición de esta nueva configuración sociotécnica.
Resultado: D5.4 Libro blanco sobre los aspectos legales y reglamentarios de la GPV (M45)

Entregables:

D5.1 Estudio del impacto ético, legal y social del GPV g***uardIA*** (, M7, M24)
D5.2 Desideratas éticas para el GPV (M9, M24, M45)
D5.3 Arquitectura para un GPV explicable y confiable (M24, M45)
D5.4 Libro blanco sobre los aspectos legales y reglamentarios del GPV (M45)

WP6: Diseminación y comunicación.

Objetivos: El WP6 tiene como objetivo desarrollar un marco sólido y estratégico para la comunicación efectiva, la difusión y las actividades de compromiso que se desarrollen a lo largo de toda la duración de ***guardIA*** y que garanticen un impacto a largo plazo mucho más allá de su conclusión;

Tarea 6.1: Estrategia de comunicación adaptada para identificar a las partes interesadas y los grupos objetivo relevantes - Campañas de comunicación y difusión
Calendario: M1 a M48
Objetivo de la tarea: El objetivo de esta tarea es proporcionar una estrategia común y un kit de comunicación y difusión durante y después del proyecto.
Plan de acción: El plan de comunicación y difusión (CDP) será al inicio del proyecto (M4) y actualizado regularmente. La difusión de los resultados del proyecto se iniciará en una fase suficientemente temprana del proyecto para permitir una difusión continua y significativa. El CDP detallará además los grupos objetivo, los canales de comunicación previstos, los métodos y las herramientas que pueden maximizar la exposición del proyecto al mayor público posible y los medios para medir los esfuerzos de comunicación y el impacto. Se diseñará un conjunto de herramientas de comunicación y marketing que se pondrá a disposición de todos los socios. Incluirá:

a. La identidad visual del proyecto: logotipo y materiales digitales,
b. Plantillas del proyecto - que se utilizarán para personalizar las herramientas de difusión para eventos específicos,
c. Sitio web del proyecto - para asegurar el acceso a toda la información y los productos principales del proyecto,
d. Canales de medios sociales - utilizados estratégicamente para aumentar el alcance y apoyar la creación y el compromiso de la comunidad de investigación ampliada del proyecto, como se define en T6.5.

Se desarrollarán e implementarán campañas de comunicación para asegurar una amplia concienciación de los resultados del proyecto a las partes interesadas (como se describe en la sección 2 y se indica en D6.2).
Las acciones de divulgación tendrán como objetivo difundir los resultados del proyecto a una amplia audiencia, promoviendo la adopción de los resultados del proyecto para su asimilación por el mercado. La difusión se basará principalmente en acciones de difusión orientadas técnicamente, como se describe en T6.2.
Resultado: D6.1: Plan de difusión y comunicación (M4, M24, M36, M48), D6.2: Kit de herramientas de difusión (folletos, carteles, plantillas, directrices) (M4) D6.3: Sitio web público de ***guardIA*** (M4)

Tarea 6.2: Estrategia de publicación y difusión basada en los principios de la Ciencia Abierta - definición e implementación
Calendario: M1 a M48
Objetivo de la tarea: Asegurar que las actividades de publicación y difusión se basan en los principios de la ciencia abierta a lo largo del proyecto.
Plan de acción: Las acciones de difusión con orientación técnica serán una parte importante de los esfuerzos de difusión de ***guardIA***:

a. Publicación de artículos en revistas (JAI, IEEE Transactions on Affective Computing, Nature Machine Intelligence, IEEE Computational Intelligent Magazine, Engineering Applications of Artificial Intelligence, ..) revisadas por pares y participación en conferencias (IJCAI; AAMAS, ECCAI, PAAMS, …): se producirán al menos 20 publicaciones revisadas por pares.
b. Participación activa de redes locales, nacionales o internacionales como ELLIS, CLAIRE, TAILOR. Todas ellas constituyen un foro válido y un vehículo para promover el desarrollo de g***uardIA*** y sus futuros avances.

Resultado: D6.4: Estrategia de publicación de ***guardIA*** (M4, 12, 24, 36)

Entregables:

D6.1: Plan de difusión y comunicación (M4, M24, M36, M48)
D6.2: Kit de herramientas de difusión (folletos, carteles, plantillas, directrices) (M4)
D6.3: Sitio web público de ***guardIA*** (M4)
D6.4: Estrategia de publicación de ***guardIA*** (M4, 12, 24, 36)

WP7: Gestión de proyecto.

Objetivos: El WP7 garantizará el desarrollo específico y eficiente de los objetivos del proyecto. Se utilizarán herramientas de gestión, un portal de extranet y reuniones del equipo de investigación.

Tarea 7.1: Coordinación general del proyecto
Calendario: M1 a M48
Objetivo de la tarea: El IP, prof. Vicent Botti, será responsable de la gestión y coordinación general del proyecto.
Plan de acción: contemplará las siguientes actividades: a. Control del progreso; b. Comprobación de calendarios, entregables e hitos; c. Análisis de riesgos, preparación y gestión de planes de contingencia.
Resultado: D7.1 Herramientas de gestión (M3)

Tarea 7.2: Gestión administrativa y financiera (GAF)
Calendario: M1 a M48
Objetivo de la tarea: Esta tarea tiene como objetivo garantizar un informe adecuado hacia la GVA, crear herramientas de gestión (D8.1) para apoyar la gestión del proyecto, seguir el progreso y garantizar el éxito de las actividades.
Plan de acción: La UPV garantizará una gestión jurídica y financiera eficaz del proyecto. Abarca la creación y el mantenimiento de registros financieros, la planificación y el seguimiento de los gastos, la preparación de las declaraciones de gastos consolidadas siguiendo las normas y el formato de Política Financiera de la GVA.
Resultado: D7.2: Procedimientos de GAF (M6, M24, M32, M48)

Tarea 7.3: Herramientas y estrategias de control de calidad - Identificación de KPI
Calendario: M1 a M48
Plan de acción: Se elaborarán los procedimientos de evaluación de la calidad: esta información se recopilará en el Plan de Calidad (D8.2). Este Plan de Calidad medirá y documentará la eficacia de las acciones del proyecto en comparación con la propuesta y los objetivos iniciales.
Resultado: D7.3: Plan de calidad de ***guardIA*** (M4)

Tarea 7.4: Plan de gestión de datos
Calendario: M1 a M48
Plan de acción: Se redactará un Plan de Gestión de Datos (PGD) específico en los primeros 6 meses del proyecto, en el que se detallará con precisión el procedimiento de recogida de datos, el procedimiento de consentimiento, el almacenamiento, la protección, la conservación y la destrucción de los datos, y la confirmación de que cumplen la legislación autonómica, nacional y de la UE.
El PGD del proyecto servirá como un documento vivo que abordará todos los aspectos del ciclo de vida de los datos dentro de ***guardIA***, tal y como se describe en la parte 2.2.1 Plan de gestión de datos de la propuesta.
Resultado: D7.4: Plan de gestión de datos (M6)

Entregables

D7.1 Herramientas de gestión (M3)
D7.2: Procedimientos de GAF (M3)
D7.3: Plan de calidad de ***guardIA*** (M4)
D7.4: Plan de gestión de datos (M6)

Tabla de entregables:

Entregable	Nombre	Nº WP	Fecha de entrega
D1.1	Análisis detallado de las tecnologías existentes y una valoración de su potencial futuro.	WP1	M6
D1.2	Modelos matemáticos/causales para la interacción social a nivel microscópico, mesoscópico y macroscópico, parametrizados con parámetros relacionados con el estado y los rasgos humanos cognitivos, emocionales y sociales	WP1	M12
D1.3	Descripción detallada de los parámetros en los modelos de D1.2 que pueden ser instrumentales para el GPV	WP1	M15
D1.4	Escenarios virtuales para la interacción futura y fiable de la RSH	WP1	M12
D1.5	Lista de requisitos del usuario para las tecnologías	WP1	M6, M14,M21
D2.1	Modelos de interacción afectiva y conversacional del agente guardIA	WP2	M16, M24, M34, M45
D2.2	Modelo computacional de empatía guardIA	WP2	M16, M24,M34,M45
D2.3	Parámetros de interacción contextual conversacional guardIA	WP2	M16, M24,M34,M45
D2.4	Integración de tecnologías de acuerdo afectivo en estructuras de conocimiento genéricas para escenarios conversacionales	WP2	M24, M34,M45
D2.5	Personalization models for user interaction	WP2	M34, M45
D3.1	Informe: Arquitectura software de guardIA	WP3	M8
D3.2	Metodología de integración y validación	WP3	M15, M24
D3.3	Infraestructura de interacción en tiempo real en la nube de	WP3	M36
D3.4	Política de código abierto y publicación de guardIA	WP3	M33, M45
D4.1	Metodología de validación de guardIA	WP4	M22
D4.2	Evaluación y validación de guardIA	WP4	M43
D4.3	GPV validado en el entorno correspondiente	WP4	M45
D5.1	Estudio del impacto ético, legal y social del GPV guardIA	WP5	M6, M24
D5.2	Desideratas éticas para el GPV	WP5	M9, M24,M45
D5.3	Arquitectura para un GPV explicable y confiable	WP5	M24, M45
D5.4	Libro blanco sobre los aspectos legales y reglamentarios del GPV	WP5	M45
D6.1	Plan de difusión y comunicación	WP6	M4, M24,M36, M48
D6.2	Kit de herramientas de difusión	WP6	M4
D6.3	Sitio web público de guardIA	WP6	M4
D6.4	Estrategia de publicación de guardIA	WP6	M4,M12,M24,M36
D7.1	Herramientas de gestión	WP7	M3
D7.2	Procedimientos de GAF	WP7	M3
D7.3	Plan de calidad de guardIA	WP7	M4
D7.4	Plan de gestión de datos	WP7	M6

Gestión de riesgos.

Los principales riesgos del proyecto identificados, analizados y discutidos durante la preparación del mismo, así como las medidas previstas para mitigarlos, en su caso, se detallan en la siguiente tabla. Durante la ejecución del proyecto se dispondrá de una matriz de gestión de riesgos, en la que registrará los riesgos abiertos para el proyecto. La combinación de probabilidad de ocurrencia y gravedad generará una clasificación de los riesgos.

Descripción del Riesgo	WP	Medidas de mitigación de riesgo propuestas
El sistema de pruebas multiusuario en tiempo real no permite un número elevado de usuarios en línea. Probabilidad: Alta Gravedad: Baja	WP3	Esta situación ya se da en todos los sistemas multiusuario. El sistema se diseñará y optimizará para proporcionar una experiencia totalmente fluida a un mínimo de tres usuarios, y en base a ello, también se proporcionarán otras configuraciones con menos calidad.
Los escenarios RSH y los diseños de interfaces de GPV finales no cumplen con los requisitos de los experimentos, o se descubre que requieren mejoras durante la validación. Probabilidad: Baja Gravedad: Alta	WP3	Todos los componentes que participen en los escenarios virtuales de RSH y también todos los modelos de GPV se diseñarán y modelarán para que sean intercambiables y con un estándar bien definido, de modo que los cambios en las piezas pequeñas no requieran redefinir escenarios enteros, o estructuras de GPV.
Tecnologías guardIA utilizadas por terceros para crear contenidos/tecnologías manipuladoras basadas en los resultados del proyecto. Probabilidad: Media Gravedad: Media	WP2	Se han hecho grandes esfuerzos para garantizar que las prácticas de Investigación e Innovación Responsables se apliquen en todo el proyecto y son proporcionadas por el WP6
La arquitectura basada en agentes no es lo suficientemente flexible y/o preciso como para representar a un humano real. Probabilidad: Baja Gravedad: Alta	WP2	Reconstruirla arquitectura revisando el paradigma teórico de RAISE. Nuestra metodología de desarrollo iterativo detectará y solucionará estos problemas con antelación para que sean resueltos antes del siguiente paso.
El diseño/la tecnología de la experimentación no es capaz de obtener datos de todas las variables racionales/afectivas necesarias Probabilidad: Media Gravedad: Baja	WP2, WP4	Volver a analizar los casos prácticos para redefinir las variables. Si es necesario, reducir el alcance de la simulación. Investigar posibles sustituciones tecnológicas .
Retraso en los hitos clave o en los resultados críticos. Probabilidad: Baja Gravedad: Media	Todos los WPs	Los resultados de las tareas se entregarán de forma escalonada en cada fase, de manera que las demás tareas que se basen en ellos puedan aprovechar los

_____ , __ d' _____ de 20 __

Investigadors principals del projecte / Investigadores principales del proyecto

BOTTI NAVARRO VICENTE JUAN - 19880706N

Firmado digitalmente por BOTTI NAVARRO VICENTE JUAN - 19880706N
Fecha: 2021.10.27 09:58:33 +02'00'

GARCIA FORNES ANA MARIA - 19893848K
2021.10.27 11:00:47 +02'00'

Firma: Vicent Botti Navarro

Firma: Ana María García Fornes

* L'Historial del grup i la Memòria cientificotècnica del projecte es consideren part integrant i contingut mínim de la sol·licitud, per la qual cosa l'absència o falta de contingut d'aquests documents determinarà la inadmissió d'aquesta. (Art. 14.3 de les bases reguladores)

* *El Historial del grupo y la Memoria científico-técnica del proyecto se consideran parte integrante y contenido mínimo de la solicitud, por lo que la ausencia o falta de contenido de estos documentos determinará la inadmisión de la misma. (Art. 14.3 de las bases reguladoras)*