Caracterización dinámica de explicaciones en sistemas autónomos con participación humana*

Antoni Mestre[1], Miriam Gil[2], Manoli Albert[1], Jose Ignacio Panach[2] , Vicente Pelechano[1]

[1] VRAIN Institute, Universitat Politècnica de Valencia, Camino de Vera, s/n, 46022 Valencia, Valencia, Spain
[2] Departament d'Informàtica, Universitat de València, Avenida de la Universidad, s/n, 46100 Burjassot, Valencia, Spain

Abstract. La colaboración humano-sistema es un modelo de trabajo que permite combinar conocimientos y habilidades de humanos y máquinas. El objetivo de esta colaboración es superar situaciones complejas y garantizar una forma de trabajo adecuada y confiable. Para lograr una colaboración humano-sistema efectiva y eficiente, los sistemas deben ser transparentes, comprensibles y confiables para los humanos. Las explicaciones que el sistema ofrece a los humanos son mecanismos clave para lograr este tipo de sistemas. Sin embargo, el diseño de las explicaciones plantea una serie de desafíos en cuanto a las características que deben tener estas. Por ejemplo, ¿cuál es el contenido necesario para la explicación? ¿en qué momento se debe dar? ¿debe ser muy intrusiva para captar la atención del usuario?. En este trabajo, proponemos un modelo conceptual para caracterizar las explicaciones y, en base a este modelo conceptual, se construye un sistema que infiere las características que debe tener la explicación a ofrecer de acuerdo a la acción a explicar, el contexto del usuario y su perfil. En este trabajo, nos centramos en el dominio de la Smart Home, pero el enfoque es extrapolable a otros dominios.

Keywords: Explicaciones, sistemas autónomos, tareas humano-sistema, nivel de atención, diseño centrado en el usuario, aprendizaje automático

1 Introducción

Los sistemas en los que es necesaria la participación de un humano para llevar a cabo tareas colaborativas (esto es, tareas en las que tanto humano como sistema llevan a cabo parte de la funcionalidad), deben ser diseñados haciendo especial hincapié en la comprensibilidad del sistema. Esta comprensibilidad es esencial para lograr el éxito de la participación del humano en el sistema. Una herramienta clave de la que disponen estos sistemas para lograr su comprensibilidad son las explicaciones [16]. En el contexto de las tareas colaborativas, las explicaciones que da el sistema pueden estar orientadas a:

* Trabajo financiado por la Generalitat Valenciana bajo el proyecto GV/2021/072 y cofinanciado por el MINECO bajo el proyecto AVANTIA PID2020-114480RB-I00.

1) requerir la participación del humano cuando le corresponda a este llevar a cabo alguna acción o 2) explicar las acciones que ha llevado a cabo el sistema.

Pero ¿qué tipo de explicación se debe dar? ¿cuál debe ser el contenido de la explicación? ¿se debe proporcionar mucha o poca información? ¿de forma muy intrusiva o poco intrusiva? ¿en el momento actual o puede ser más tarde? ¿la repetimos o solo la damos una vez? Podría pensarse que cuanta más información dé la explicación y cuanto más intrusiva sea, más probable es que el usuario entienda lo que está haciendo el sistema o que el usuario sea consciente de qué acción debe llevar a cabo. Sin embargo, no siempre es así [18]. Proporcionar al usuario información excesiva, ya sea por el contenido o por la forma, puede llevar a un resultado contrario al esperado, ocasionando que el usuario deje de prestar atención a las indicaciones del sistema.

Especificar en tiempo de diseño el tipo de explicación que debe ofrecer el sistema no es una tarea fácil. Se debe realizar un diseño de las explicaciones adecuado a cada usuario para conseguir la atención de éste pero al mismo tiempo sin molestarle, y esto depende de muchos factores. Por ejemplo, en función del perfil de usuario la explicación requiere una complejidad mayor o menor, o en función de si el usuario está más o menos ocupado la explicación requiere un grado de intrusividad diferente. Además, todo esto depende de las preferencias del usuario; el diseñador no siempre sabe qué tipo de explicación es más adecuado para los usuarios.

En este trabajo, se realiza una propuesta en la que de forma dinámica se determinan las características de las explicaciones que se proporcionan al usuario en las tareas colaborativas. Para ello, se tiene en cuenta la acción a explicar, el contexto del usuario que participa en la tarea y su perfil. Además, la propuesta sigue el enfoque de diseño centrado en el usuario (DCU)[21], ya que los propios usuarios participan en el diseño del sistema que determina las características más adecuadas para cada explicación. El diseño de este sistema se estructura en dos componentes: 1) un modelo predictivo que infiere el nivel de atención adecuado para una explicación en función del contexto del usuario, y 2) un método que determina las demás características de la explicación utilizando la inferencia del modelo predictivo, el perfil del usuario y las características de la acción a explicar. La propuesta se realiza en el dominio de la Smart Home. El trabajo utiliza la propuesta de [10] como marco conceptual para la especificación de las tareas colaborativas.

La contribución de este trabajo se enmarca en el ámbito de la Ingeniería del Software Dirigida por Modelos (ISDM). La propuesta para el diseño de explicaciones usa: (1) modelos para la especificación de las tareas colaborativas y las explicaciones que se deben ofrecer, y (2) modelos en tiempo de ejecución para la selección dinámica de las características de las explicaciones. Además, la propuesta combina la ISDM con técnicas de aprendizaje automático para abordar la complejidad de especificar la lógica que determina las características adecuadas para una explicación a partir del contexto del usuario y su perfil.

El resto del trabajo se estructura de la siguiente manera. En la Sección 2 se identifican y analizan trabajos relacionados con la propuesta. En la Sección 3 se presenta el caso de estudio. En la Sección 4 se define un marco conceptual para las explicaciones en el contexto de las tareas colaborativas. La Sección 5 introduce un método que infiere el nivel de atención adecuado para las explicaciones en función del contexto del usuario

y en la Sección 6 se utiliza un modelo de características para definir la variabilidad en la selección de características de las explicaciones de acuerdo a la acción a explicar, el perfil del humano y su contexto. En la Sección 7 se presenta la propuesta para caracterizar las explicaciones de las tareas colaborativas en tiempo de ejecución. Finalmente, la Sección 8 presenta las conclusiones del trabajo y el trabajo futuro.

2 Estado del Arte

Durante más de cien años, muchos trabajos en los campos de la filosofía, la psicología social y la cognitiva han estudiado qué constituye una explicación, cómo se estructura y cómo las personas generan y evalúan la explicación [20]. Hace más de tres décadas, hubo una extensa investigación sobre las explicaciones en el contexto de los sistemas expertos [4]. Recientemente, el papel de la explicación ha resurgido en el campo de la Inteligencia Artificial con la noción de Inteligencia Artificial Explicable (XAI) y en agentes y robots autónomos como una capacidad importante de éstos [20]. Hellström y Bensch [14] describen cómo se captura el estado del sistema en la mente de un ser humano. Cuando no se explica el comportamiento del sistema, es posible que el estado en la mente no sea coherente con el estado real, lo que podría conducir a situaciones peligrosas. Además, la falta de un modelo mental para el ser humano para que pueda estimar las acciones del sistema puede desencadenar riesgos para la seguridad [3].

En el ámbito de los sistemas autónomos, Drechsler et al. [7] describen los primeros pasos hacia un marco conceptual para un sistema ciberfísico auto-explicativo. En su aproximación proponen agregar una capa de auto-explicación que incluye un modelo abstracto del sistema, y proponen ajustar la granularidad de las explicaciones para diferentes grupos de personas usuarias. Proponen construir cadenas de causa-efecto para acciones observables utilizando el modelo abstracto. Las personas usuarias pueden acceder a estas cadenas para comprender la causa de las acciones. Wüest et al. [25] utilizan un enfoque de ciclo de retroalimentación para identificar situaciones en las que es valioso pedirle a una persona usuaria retroalimentación sobre el comportamiento del sistema. Los autores comparan el comportamiento de la persona usuaria con un modelo de objetivos y solicitan comentarios cuando las personas logran subobjetivos o cuando se desvían de un subobjetivo esperado. Blumreiter et al. [2] proponen un framework de referencia para construir sistemas ciberfísicos auto-explicativos introduciento capacidades de auto-reflexión en el bucle de control MAPE-k. Li et al. [16], han investigado el papel de la explicación como un mecanismo para mejorar la comprensión de las personas cuando están involucrados en un papel de supervisión, ya sea para aprobar o rechazar la propuesta de acción que ha hecho el sistema autónomo. Sin embargo, en ninguno de estos trabajos se tienen en cuenta las necesidades de las personas ni el contexto para adaptar el tipo de explicación ofrecida.

Otros trabajos en el ámbito de los sistemas inteligentes se han centrado en racionalizar y verbalizar el comportamiento de los agentes autónomos. Las racionalizaciones no necesitan reflejar con precisión el verdadero proceso de toma de decisiones, pero dan algunas explicaciones como las que darían las personas en situaciones similares. Harrison et al. [13] racionalizan las acciones de un agente mediante el uso de una red

neuronal codificadora-decodificadora para traducir la información del estado de la acción a lenguaje natural. En [22], las experiencias del agente en una ruta se verbalizan convirtiendo los datos del sensor en lenguaje natural como respuesta a las consultas de los usuarios con diferentes niveles de abstracción, especificidad y localidad. Otro enfoque para generar explicaciones en tiempo de ejecución es utilizar un agente multimodal que se pueda consultar 'a petición' [23]. En este trabajo, los comportamientos del sistema se mapean en una versión modificada de árboles de fallos, que los autores llaman 'modelo de autonomía', el cual captura los posibles estados del sistema [6].

Para categorizar el tipo de explicaciones se han definido diferentes taxonomías. Lim y Dey [17] definieron una taxonomía con diferentes tipos de explicación y bajo qué circunstancias usar cada tipo en el ámbito de los sistemas sensibles al contexto. Además, demostraron empíricamente la efectividad de algunos tipos de explicaciones [19]. Sin embargo, en estudios posteriores, encontraron que los usuarios también podían razonar de manera diferente a lo previsto y tener diferentes preferencias para los tipos de explicación incluso para las mismas tareas [18]. Por tanto, es importante que el tipo de explicación se adapte a cada persona. Chari et al. [5] construyeron una ontología para modelar diferentes primitivas de las explicaciones con el objetivo de soportar el diseño de sistemas IA centrados en el usuario. Nuestra propuesta incorpora ciertas primitivas propuestas por estas taxonomías pero además incorpora otras características de las explicaciones adecuadas para las tareas colaborativas y la interacción no intrusiva.

3 Caso de estudio

El caso de estudio propuesto es un escenario del hogar inteligente. En el escenario ejemplo participan los dos residentes que se describen en la Tabla 1. Para cada residente se describe su perfil de usuario utilizando el modelo OWC el cual categoriza las características de los humanos en tres factores: *opportunity*, *willingness* y *capability* [8].

Tabla 1. Descripción de los residentes utilizados en el escenario ejemplo

Residente	**Características**	***Opportunity***	***Willingness***	***Capability***
Jane	74 años, mujer, jubilada, vive con su hijo Paul.	Ubicación = habitación	NivelEstres = bajo; EstadoCognitivo = medio	Experiencia-Digital = baja
Paul	40 años, informático, trabaja en casa, vive con su madre Jane	Ubicación = salón	NivelEstres= medio; EstadoCognitivo = alto	Experiencia-Digital = alta

El escenario propuesto implica a la tarea de la compra automática de alimentos cuando la nevera se está quedando sin existencias. En esta tarea, el sistema: 1) detecta la falta de existencias de algunos alimentos, 2) realiza la lista de la compra, 3) informa al humano de que se necesita realizar la compra, 4) le solicita una confirmación para llevar a cabo la compra, 5) el usuario confirma y 6) se realiza la compra. Las acciones que requieren de una explicación son la 3 y la 4. Estas explicaciones deberían ser diferentes según la situación. Por ejemplo, tres situaciones diferentes son:

1. Las explicaciones son para Paul y en el momento en que se ejecuta la tarea de la compra automática está trabajando.
2. Las explicaciones son para Paul y en el momento en que se ejecuta la tarea de la compra automática está viendo la tele.
3. Las explicaciones son para Jane y en el momento en que se ejecuta la tarea de la compra automática está haciendo cosas personales.

A lo largo del trabajo, se mostrará cómo la solución propuesta caracteriza de forma dinámica las explicaciones a ofrecer por el sistema en cada una de estas situaciones.

4 Explicaciones en tareas colaborativas ¿cómo deben ser?

En trabajos anteriores [10] se ha propuesto un marco conceptual para la especificación de tareas colaborativas en tiempo de diseño. Esta especificación describe:

- *¿Quién puede participar en la tarea?*: Una tarea colaborativa requiere de un **perfil humano** necesario para llevar a cabo la tarea. El perfil se especifica utilizando el modelo OWC.
- *¿Cuándo se puede ejecutar la tarea colaborativa?*: Se requiere una situación de contexto apropiada para ejecutar una tarea. Esta situación de contexto puede considerarse como la **precondición** de la tarea. La precondición representa el contexto en el que el sistema, su entorno y el ser humano están preparados para ejecutar la tarea en las condiciones adecuadas para lograr el desempeño adecuado de la tarea.
- *¿Qué secuencia de acciones seguirá la interacción humano-sistema?*: Una tarea está formada por un conjunto de acciones que debe llevar a cabo el humano y que debe llevar a cabo el sistema. Se distinguen diferentes tipos de acciones:

— *Acciones propias de la tarea*, acciones necesarias para realizar la tarea. Deben permitir una adecuada simbiosis entre el humano y el sistema, ofreciendo al humano los mecanismos de interacción y la información adecuada para poder ejecutar la funcionalidad requerida. Estas acciones se corresponderían con las acciones 1, 2, 5, 6 del ejemplo del caso de estudio.
— *Acciones de retroalimentación*, se corresponden con acciones donde el sistema informa al humano. Ofrecen información para que el humano comprenda el funcionamiento del sistema y deposite su confianza en él. Estas acciones se corresponderían con las acciones 3 y 4 del ejemplo del caso de estudio.
— *Acciones de preparación*, correspondientes a acciones para alcanzar las condiciones adecuadas que permitan al humano hacer su tarea, es decir, satisfacer la precondición.

Cada acción viene caracterizada por:

— Condiciones de contexto que se deben cumplir para ejecutar la acción.
— Nivel de atención requerido por la acción. Este nivel de atención indica el grado de demanda cognitiva que requiere la explicación del usuario y va a determinar el mecanismo de interacción a usar.

En el presente trabajo nos centramos en las acciones de retroalimentación. Estas acciones van ligadas a una explicación. El objetivo del trabajo es adaptar el tipo de explicación a ofrecer en estas acciones dependiendo de la acción en particular, el contexto del usuario y su perfil. Para especificar el tipo de explicación a ofrecer, proponemos un marco conceptual con primitivas que caracterizan la explicación en el ámbito de las tareas colaborativas (ver Figura 1).

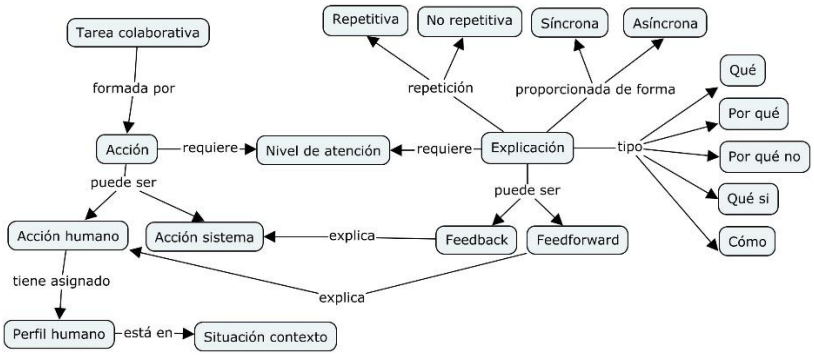

Fig. 1. Conceptos propuestos para caracterizar las explicaciones.

Una explicación en el ámbito de una tarea colaborativa asociada a una acción de retroalimentación puede ser de dos tipos: 1) de "**feedback**" si sirve para explicar una acción que hace el sistema y el objetivo de la explicación es informar y/o justificar al humano la acción que el sistema ha llevado a cabo, por ejemplo siguiendo con el caso de estudio, *informar al usuario de la compra autónoma (acción 3)*, o 2) de "**feedforward**" si sirve para reclamar la participación del humano en la tarea colaborativa y explicarle la acción que debe realizar, por ejemplo, *solicitar al humano confirmación de compra (acción 4)*. De acuerdo al trabajo de Lim y Dey [17], la explicación puede ser de diferentes **tipos** según a qué pregunta responden:

1. "**Qué**": ¿Qué ha hecho el sistema?
2. "**Por qué**": ¿Por qué el sistema hizo X?
3. "**Por qué no**": ¿Por qué el sistema no hizo Y?
4. "**Qué si**": ¿Qué haría el sistema si sucediera W?
5. "**Cómo**": ¿Cómo puedo hacer que el sistema haga Z, dado el contexto actual?

Al igual que las acciones propias de una tarea colaborativa, una explicación requiere de un **nivel de atención**. El nivel de atención nos indica el grado de demanda cognitiva que requiere la explicación por parte del usuario. En trabajos anteriores [9, 15] se ha investigado cómo la gestión de la atención afecta a la capacidad de un humano para cooperar con el sistema. Por tanto, es necesario controlar el nivel de atención requerido por una explicación para que el usuario entienda la explicación pero evitando abrumarlo con un exceso de información. Por último, si tenemos en cuenta restricciones temporales de una explicación, una explicación podría ser **síncrona** o **asíncrona** (si se da en el

momento de la colaboración humano-sistema o se puede dar en otro momento) [11] o **repetible** o **no repetible** (si se da varias veces o solo una vez).

Todas estas primitivas van a determinar el tipo de explicación a utilizar en función del perfil del humano y su situación de contexto.

5 Inferencia del nivel de atención

Como hemos visto en la Sección 4, el nivel de atención es una característica de las explicaciones relacionada con el grado de la demanda cognitiva que la explicación requiere del usuario. Especificar el nivel de atención adecuado para una explicación en tiempo de diseño es difícil, ya que esto depende de muchos factores. Entre estos factores se encuentran: (1) las preferencias de los usuarios, habrá usuarios que se sentirán más molestos ante notificaciones intrusivas que otros, y (2) variables de contexto del usuario, por ejemplo, la actividad actual del usuario, la hora actual, etc.

En esta sección, se propone un modelo predictivo que infiere en tiempo de ejecución el nivel de atención adecuado para una explicación. El modelo predictivo se ha construido utilizando aprendizaje automático supervisado. El uso del aprendizaje automático nos permite gestionar la gran cantidad de factores que influyen en las preferencias del usuario y su contexto. En concreto, la inferencia se realiza a partir de los datos que se tiene del contexto del usuario (determinado por la situación actual del hogar inteligente). Esta información viene dada por los datos obtenidos a través de dispositivos del hogar. En este trabajo se han considerado los siguientes (ver Tabla 2):

Tabla 2. Datos de entrada utilizados para la inferencia del nivel de atención

Componentes	Nombre de las variables	Descripción
0 - 6	Armario, tv, horno, portátil, frigorífico, sofá, cama	Si se está haciendo uso [1] o no [0] de los dispositivos/muebles integrados en la casa.
7 - 13	Luz de oficina, de salón, de cocina, de pasillo, de habitación, sobre lacama, de baño	Si las luces de las estancias están activas [1] o no [0].
14-18	Puerta principal, de oficina, de cocina, de habitación, de baño	El estado de la puerta, abierta [1] o cerrada [0].
19-23	Pestillo de puerta principal, de puerta oficina, de puerta cocina, de puerta habitación, de puerta baño	El estado del pestillo de las puertas, activo [1] o no [0].
24-28	Alfombra de oficina, de salón, de cocina, de habitación, de baño	Si se detecta presencia en las estancias [1] o [0].
29	Actividad[1]	Actividad del usuario [dormir|trabajar|personal|ocio|etc.]

[1] Esta variable se obtiene directamente del simulador. El usuario debe indicar en el simulador a través de botones qué acción se está llevando a cabo de forma simulada.

Además, puesto que se deben tener en cuenta las preferencias de los usuarios para determinar el nivel de atención adecuado, el modelo predictivo se construye utilizando aprendizaje automático con participación del humano. De esta forma, los datos de entrenamiento son etiquetados por los propios usuarios, ya que son estos los que deben decidir qué nivel de atención es adecuado ante cada situación (lo que un diseñador puede pensar que es adecuado podría no serlo para los usuarios). En la siguientes secciones se detalla las características del modelo predictivo construido.

5.1 Conjunto de datos de entrenamiento

Para obtener un conjunto de datos de entrenamiento para el algoritmo de aprendizaje supervisado, se ha utilizado una extensión del simulador OpenSHS. OpenSHS es un simulador 3D multiplataforma para generar datos en el campo del Internet de las Cosas (IoT) [1]. El simulador permite que un participante interactúe con un hogar inteligente en un contexto específico, que ha sido diseñado por investigadores. Durante la interacción, se pueden simular eventos de la vida diaria (como dormir, moverse por la casa, cocinar, etc.). El simulador incluye dispositivos y sensores como sensores de presión (p. ej., alfombra activada, cama, sofá, etc.), sensores de puertas, interruptores de electrodomésticos (TV, horno, nevera, etc.) y controladores de luz. El simulador se ha extendido de forma que permite etiquetar los datos de acuerdo a los 3 niveles de atención que se quiere inferir. La Figura 2 muestra una imagen del simulador con los botones para etiquetar la situación que el simulador está recreando.

Fig. 2. Simulador con los botones *high*, *slight* y *low* para etiquetar la situación recreada

Este simulador ha sido utilizado por 30 estudiantes del Grado de Ingeniería Informática de la Universitat Politècnica de València. que han recreado diferentes situaciones del hogar inteligente y han etiquetado estas situaciones con el nivel de atención que han considerado adecuado en cada una. Estas simulaciones permitieron obtener un conjunto de datos formado por 594 registros. Cada registro es un vector con los datos de la

Tabla 2 más la variable etiqueta que se muestra en la Tabla 3. Esta variable se utiliza para la construcción de los datos de entrenamiento etiquetados.

Tabla 3. Conjunto de datos de entrenamiento

Componente	Nombre de las variables	Descripción
Etiqueta	Nivel de obtrusividad	Nivel de intrusividad de las notificaciones: alto [2], medio [1] o bajo [0].

5.2 Algoritmo de Aprendizaje Automático

Los modelos predictivos construidos clasifican los datos de entrada según 3 clases: (0) bajo, (1) intermedio o (2) alto. Los datos de entrada son un vector con los datos de la Tabla 2. Para la construcción del modelo predictivo se ha hecho uso de algoritmos de clasificación de aprendizaje automático supervisado, los cuales mapean una entrada en una salida basándose en pares de ejemplo entrada-salida. Los algoritmos de clasificación identifican la relación que hay entre las entradas y salidas de los pares de ejemplo y crea un modelo predictivo que puede usarse para mapear nuevos pares. Los algoritmos han sido entrenados con el conjunto de datos de entrenamiento etiquetado de pares entrada-salida presentado en la Sección 5.1, el cual contiene 594 muestras. Entre los diferentes algoritmos de clasificación que existen, en este trabajo se han explorado los siguientes: Red Neuronal (RN), Random Forest (RF), y K-Nearest Neighbors (KNN). De entre estos tres, el algoritmo con el que se ha conseguido mayor precisión es RF con un 93,2% (frente al 84,6 que se conseguía con RN y 82,6 con KNN). El algoritmo RF se basa en la construcción de múltiples árboles de decisión que son entrenados con un subconjunto aleatorio, y diferente para cada árbol, de valores de las muestras de entrenamiento. La selección aleatoria de valores de las muestras reduce la correlación entre los árboles, generando un bosque de decisión robusto. El modelo RF implementado en este trabajo consta de 50 árboles de decisión, ha sido entrenado con el 90% de los datos y el 10% restante ha sido utilizado para test.

6 Selección de características de la explicación

De acuerdo al marco conceptual propuesto en la Sección 4, para caracterizar una explicación se debe definir, además del nivel de atención, ¿qué partes de la explicación se deben dar? ¿se debe explicar el por qué? ¿el por qué no? ¿se debe repetir la explicación? ¿hay que ofrecerla en tiempo real o se puede dar más tarde de que ocurra la tarea? Estas características van a depender de:

1. El tipo de acción a explicar. En concreto el nivel de criticidad de la acción. En este trabajo este nivel se define en tres grados: crítico, intermedio, no-crítico. El nivel de criticidad afectará a si la explicación se debe repetir u ofrecer en tiempo real.
2. El nivel de atención inferido. Este nivel de atención afecta a las partes que debe contener la explicación y a si se debe repetir.

3. El perfil del usuario, basado en el modelo OWC. El perfil de usuario afecta a si la explicación se debe repetir, si se debe ofrecer en tiempo real, y a las partes que debe contener la explicación.

La definición de las características de la explicación en función de estas tres variables se puede realizar de diversas formas. Una herramienta apropiada para modelar esta variabilidad son los modelos de características. Los modelos de características nos permiten representar la información de todos los posibles tipos de explicación en términos de características, sus relaciones, y sus restricciones. En este trabajo se opta por utilizar una aproximación basada en modelos de características.

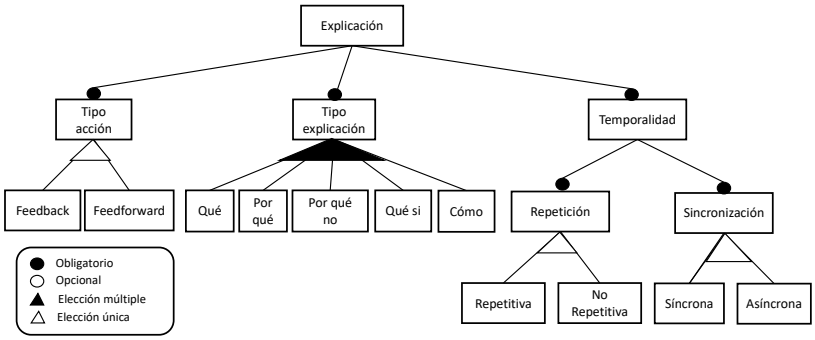

Fig. 3. Modelo de características para las explicaciones

El modelo de características en este trabajo nos permite (1) la gestión de la variabilidad de las características de la explicación para seleccionar el tipo de explicación, (2) la representación de las características de la explicación de manera taxonómica, (3) el uso de modelos en tiempo de ejecución para seleccionar dinámicamente las características de la explicación. La Figura 3 muestra el modelo de características construido para representar las distintas características de las explicaciones definidas en el marco conceptual propuesto en la Sección 4.

El modelo de características se utiliza para especificar qué características de la explicación son necesarias para:

- un nivel de atención, de acuerdo a 3 niveles: alto, medio y bajo,
- un perfil de usuario, basado en el modelo OWC,
- y un nivel de criticidad de acción, de acuerdo a 3 niveles: crítico, intermedio y no crítico.

Para ello, se definen unas características activas que determinan una configuración del modelo de características. Cada configuración se define mediante un conjunto de estados activos e inactivos. Los diseñadores deben definir las posibles configuraciones de explicaciones y asignarlas a un nivel de atención y un posible perfil de humano. Siguiendo con el caso de estudio presentado en la Sección 3, la configuración para la acción *informar al humano de que se necesita realizar la compra* con un nivel de atención alto y un usuario con experiencia digital baja podría ser:

TipoExplicación (AtenciónAlta, ExperienciaDigitalBaja, informar al humano de que se necesita realizar la compra) = {Feedback, Qué, Repetitiva, Síncrona}

7 Caracterización dinámica de las explicaciones

La caracterización de una explicación se lleva a cabo cuando el sistema identifica que hay una acción de una tarea colaborativa que tiene asociada una explicación a ofrecer (esto es, una acción de retroalimentación). Cuando se va a ejecutar una acción de retroalimentación, los pasos a seguir son los siguientes:

1. El modelo predictivo infiere el nivel de atención adecuado a partir del contexto del usuario.
2. Se selecciona la configuración del modelo de características en base al nivel de atención inferido, el nivel de criticidad de la acción y el perfil del usuario.
3. La configuración seleccionada determina las características de la explicación.

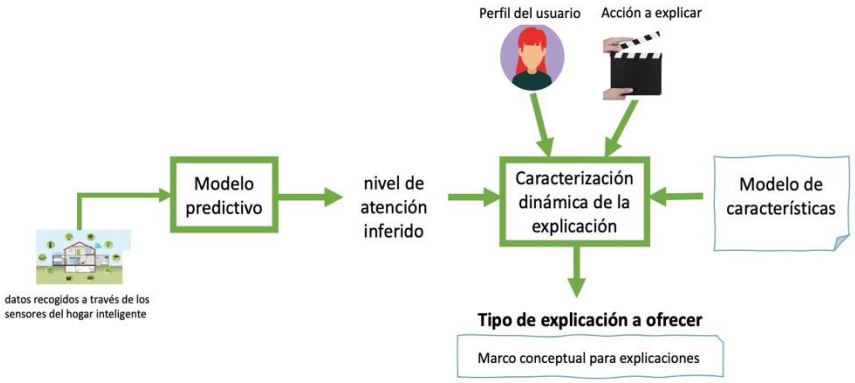

Fig. 4. Arquitectura de la solución propuesta

De esta forma, se determina en tiempo de ejecución el tipo de explicación más adecuado para cada acción a explicar en función de la acción, el perfil del usuario y el nivel de atención. La Figura 4 muestra gráficamente los componentes de la solución propuesta para la caracterización dinámica de explicaciones.

7.1 Aplicación al caso de estudio

Volvamos al caso de estudio. En la Sección 3 se han planteado tres situaciones diferentes en las que se debe ofrecer una explicación. La Tabla 4 muestra la descripción de cada uno de estos escenarios. Se está ejecutando la tarea colaborativa compra autónoma y el sistema debe ofrecer la explicación: *informar al humano de que se necesita realizar la compra*. Esta acción tiene definido un nivel de criticidad intermedio (definido por el diseñador cuando se define la acción). En este momento el sistema inicia el proceso que determina las características que debe tener la explicación. Los pasos del proceso propuesto son los siguientes:

Tabla 4. Descripción de los tres escenarios en los que se debe dar la explicación

	Escenario	Contexto de usuario[2]	Perfil de usuario[1]
1	Las explicaciones son para Paul y en el momento en que se ejecuta la tarea de la compra automática está trabajando.	portátil=1, luz de oficina=1, pestillo puerta oficina=1, alfombra oficina = 1	ExperienciaDigital = alta
2	Las explicaciones son para Paul y en el momento en que se ejecuta la tarea de la compra automática está viendo la tele.	TV=1, luz de salón=1, puerta salón=1, alfombra salón = 1	ExperienciaDigital = alta
3	Las explicaciones son para Jane y en el momento en que se ejecuta la tarea de la compra automática está haciendo tareas personales.	Armario=1, luz de habitación=1, puerta habitación=1, alfombra habitación = 1	ExperienciaDigital = baja

1. **Determinar el nivel de atención**. El modelo de inferencia a partir del contexto de usuario infiere los niveles de atención que se muestran en la Tabla 5.

Tabla 5. Resultado del modelo predictivo para cada escenario

Escenario	Nivel de atención
1	Bajo
2	Medio
3	Alto

2. **Seleccionar una configuración del modelo de características**. Con el nivel de atención inferido, el perfil de usuario y el nivel de criticidad de la acción, se selecciona una configuración. La Tabla 6 muestra las configuraciones para cada escenario.

Tabla 6. Configuración del modelo de características seleccionada en cada escenario

Escenario	Configuración
1	*TipoExplicación (AtenciónBaja, ExperienciaDigitalAlta, NivelCriticidadIntermedio) = {Feedback, Qué y Por qué, NoRepetitiva, Asíncrona}*
2	*TipoExplicación (AtenciónMedia, ExperienciaDigitalAlta, NivelCriticidadIntermedio) = {Feedback, Qué y Por qué, NoRepetitiva, Asíncrona}*
3	*TipoExplicación (AtenciónAlta, ExperienciaDigitalBaja, NivelCriticidadIntermedio) = {Feedback, Qué, Repetitiva, Síncrona}*

3. **Caracterizar la explicación a ofrecer.** Cada configuración especifica unas características para la explicación. A partir de estas características, se construirá la explicación a ofrecer. Por ejemplo, para el escenario 1 y 2, la explicación construida sería *"Se necesita realizar la compra autónoma porqué faltan productos básicos"*, mientras que para el escenario 3 sería "*Se necesita realizar la compra autónoma".* La construcción de la explicación se podría realizar mediante el uso

[2] Se muestran únicamente las variables relevantes para la selección de características

de una ontología basada en el modelo conceptual. Este paso queda fuera del alcance de la propuesta que se realiza en este artículo; incluyéndose como trabajo futuro en la siguiente sección.

8 Conclusiones y trabajo futuro

Una comunicación humano-máquina comprensible es clave para el éxito de los sistemas autónomos con participación del humano. La necesidad de maximizar esta comprensibilidad lleva a nuevas formas de abordar el diseño de la colaboración humano-máquina. El diseño de esta colaboración no puede hacerse de forma personalizada y adaptable si se aborda en tiempo de diseño, cuando no se conoce el comportamiento de los usuarios ni se sabe qué posibles situaciones se va a encontrar el sistema. Ante este reto, se propone el uso de técnicas de IA para la caracterización dinámica de explicaciones en base a la acción a explicar, el contexto del usuario y su perfil. Este trabajo abre el camino hacia futuros trabajos donde se pretende aprender de la reacción del usuario ante las explicaciones del sistema para continuar aprendiendo y retroalimentando el modelo predictivo.

Como trabajo futuro se pretende validar la propuesta con usuarios. Para ello planteamos reclutar a sujetos con distintos perfiles y ubicarlos en contextos con distintos niveles de atención. Para cada contexto, definiremos tareas experimentales donde el usuario debe indicar dado un prototipo de sistema, cuál es su preferencia de tipo de explicaciones a recibir. Compararemos si la preferencia indicada coincide o no con el resultado proporcionado con el conjunto de datos de entrenamiento y los algoritmos propuestos en el presente artículo. Además, se pretende ir un paso más allá y construir las explicaciones de forma automática a partir de sus características mediante el uso de ontologías.

Referencias

1. Alshammari, Nasser & Alshammari, Talal & Sedky, Mohamed & Champion, Justin & Bauer, Carolin. (2017). OpenSHS: Open smart home simulator. Sensors. 17. 1003.
2. Blumreiter, M. et al., "Towards Self-Explainable Cyber-Physical Systems," 2019 ACM/IEEE 22nd International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C), 2019, pp. 543-548.
3. Broekens, J., Harbers, M., Hindriks, K., Van Den Bosch, K., Jonker, C. and Meyer, J.-J. "Do you get it? user-evaluated explainable bdi agents," in German Conference on Multiagent System Technologies, 2010, pp. 28–39.
4. Chandrasekaran, B., Tanner, M. C. and Josephson, J. R. "Explaining control strategies in problem solving," IEEE Expert, vol. 4, no. 1, pp. 9–24, 1989.
5. Chari, S., Seneviratne, O., Gruen, D. M., Foreman, M. A., Das, A. K., & McGuinness, D. L. (2020). *Explanation Ontology: A Model of Explanations for User-Centered AI*.
6. Chiyah Garcia, F. J., Robb, D. A., Liu, X., Laskov, A., Patron, P., and Hastie, H. "Explain yourself: A natural language interface for scrutable autonomous robots," in Explainable Robotic Systems Workshop (HRI), 2018.

7. Drechsler, R., Lüth, C., Fey, G., and Güneysu, T. “Towards self-explaining digital systems: A design methodology for the next generation,” in 2018 IEEE 3rd International Verification and Security Workshop (IVSW). IEEE, 2018, pp. 1–6.
8. Eskins, D. And Sanders, W. H.: The Multiple-Asymmetric-Utility System Model: A Framework for Modeling Cyber-Human Systems. QEST '11, 233-242 (2011).
9. Gil, M, Giner, P., and Pelechano, V. “Personalization for unobtrusive service interaction,”Pers. Ubiquitous Comput., vol. 16, no. 5, pp. 543–561, 2012.
10. Gil, M., Albert, M., Fons, J. et al. Modeling and “smart” prototyping human-in-the-loop interactions for AmI environments. Pers Ubiquit Comput (2021).
11. Glomsrud, J. A., Ødegårdstuen, A., St. Clair, A. L., and Smogeli, Ø. Trustworthy versus Explainable AI in Autonomous Vessels, International Seminar on Safety and Security of Autonomous Vessels, 17 - 18 September 2019, Helsinki
12. Goodfellow, I., Bengio, Y. and Courville, A. (2016). Deep Learning. MIT Press.
13. Harrison, B., Ehsan, U., and Riedl, M. O. “Rationalization: A neural machine translation approach to generating natural language explanations,” 2017.
14. Hellström, T. and Bensch, S. “Understandable robots-what, why, and how,” Paladyn, Journal of Behavioral Robotics, vol. 9, no. 1, pp. 110– 123, 2018.
15. Horvitz, E., Kadie, C. M. , Paek, T., and Hovel, D. “Models of attentionin computing and communication: from principles to applications,”Commun. ACM, vol. 46, no. 3, pp. 52–59, 2003.
16. Li, N., Adepu, S., Kang, E., and Garlan, D. “Explanations for human-on-the-loop: A probabilistic model checking approach,” in Proceedings of the 15th International Symposium on Software Engineering for Adaptive and Self-managing Systems (SEAMS), 2020.
17. Lim, B. Y., & Dey, A. K. (2009, September). Assessing demand for intelligibility in context-aware applications. InProceedings of the 11th international conference on Ubiquitous computing (pp. 195-204). ACM.
18. Lim, B. Y., & Dey, A. K. (2013). Evaluating Intelligibility Usage and Usefulness in a Context-Aware Application. In International Conference on Human-Computer Interaction (pp. 92-101). Springer, Berlin, Heidelberg.
19. Lim, B.Y., Dey, A.K, Avrahami, D. “Why and why not explanations improve the intelligibility of context-aware intelligent systems,” in Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, 2009, pp. 2119–2128.
20. Miller, T. “Explanation in artificial intelligence: Insights from the social sciences,” Artif. Intell., vol. 267, pp. 1–38, 2019.
21. Norman, D. A. & Draper, S. W. (Editors) (1986) “User-Centered System Design: New Perspectives on Human-Computer Interaction”. Lawrence Earlbaum Associates, Hillsdale, NJ.
22. Perera, V., Selveraj, S. P., Rosenthal, S., and Veloso, M. “Dynamic generation and refinement of robot verbalization,” in 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). IEEE, Aug. 2016, pp. 212–218.
23. Robb, D. A., Chiyah Garcia, F. J., Laskov, A., Liu, X., Patron, P., and Hastie, H. “Keep me in the loop: Increasing operator situation awareness through a conversational multimodal interface,” in 20th ACM International Conference on Multimodal Interaction (ICMI). ACM, 2018, pp. 384–392.
24. Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S., Vertesi, J., 2018. Fairness and abstraction in sociotechnical systems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 59–68.
25. Wüest, D., Fotrousi, F. and Fricker, S. “Combining monitoring and autonomous feedback requests to elicit actionable knowledge of system use,” in Requirements Engineering: Foundation for Software Quality: Springer International Publishing, 2019, pp. 209–225.