INTERNATIONAL DOCTORAL
SCHOOL OF THE USC

Ilia
Stepin

PhD Thesis

Argumentative Conversational Agents for Explainable Artificial Intelligence

Santiago de Compostela, 2023

Doctoral Programme in Information Technology Research

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

DOCTORAL THESIS

ARGUMENTATIVE CONVERSATIONAL AGENTS FOR EXPLAINABLE ARTIFICIAL INTELLIGENCE

Ilia Stepin

INTERNATIONAL PHD SCHOOL OF THE UNIVERSITY OF SANTIAGO DE COMPOSTELA

DOCTORAL PROGRAMME IN INFORMATION TECHNOLOGY RESEARCH

SANTIAGO DE COMPOSTELA
2023

Declaración del autor de la tesis

D. Ilia Stepin
Título de la tesis: Argumentative Conversational Agents for Explainable Artificial Intelligence

Presento mi tesis, siguiendo el procedimiento adecuado al Reglamento, y declaro que:

1. *La tesis abarca los resultados de la elaboración de mi trabajo.*
2. *De ser el caso, en la tesis se hace referencia a las colaboraciones que tuvo este trabajo.*
3. *Confirmo que la tesis no incurre en ningún tipo de plagio de otros autores ni de trabajos presentados por mí para la obtención de otros títulos.*
4. *La tesis es la versión definitiva presentada para su defensa y coincide la versión impresa con la presentada en formato electrónico.*

Y me comprometo a presentar el Compromiso Documental de Supervisión en caso de que el original no esté en la Escuela.

En Santiago de Compostela, 30 de Junio de 2023

Fdo. Ilia Stepin

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Autorización del Director/Tutor de la Tesis
Argumentative Conversational Agents for Explainable Artificial Intelligence

D. José María Alonso Moral, Profesor Titular de la Universidad de Santiago de Compostela

D. Alejandro Catalá, Profesor Ayudante Doctor de la Universidad de Santiago de Compostela

INFORMAN:

Que la presente tesis, se corresponde con el trabajo realizado por ***D. Ilia Stepin****, bajo nuestra dirección/tutorización, y autorizamos su presentación, considerando que reúne los requisitos exigidos en el Reglamento de Estudios de Doctorado de la USC, y que como directores/tutores de esta no incurre en las causas de abstención establecidas en la Ley 40/2015.*

De acuerdo con lo indicado en el Reglamento de Estudios de Doctorado, declaramos también que la presente tesis doctoral es idónea para ser defendida en base a la modalidad de COMPENDIO DE PUBLICACIONES, en las que la participación del doctorando/a fue decisiva para su elaboración y las publicaciones se ajustan al Plan de Investigación. La utilización de estos artículos en esta memoria, está en conocimiento de que ninguno de los trabajos aquí reunidos podrá ser presentado en ninguna otra tesis doctoral.

En Santiago de Compostela, 30 de Junio de 2023

Fdo. José María Alonso Moral	Fdo. Alejandro Catalá
Director/a tesis	Director/a tesis

Acknowledgements

First and foremost, I am extremely grateful to my supervisors, Dr. José María Alonso Moral and Dr. Alejandro Catalá for their guidance and support throughout my entire doctoral study. I would like to extend my gratitude to Dr. Martín Pereira-Fariña for co-supervising my doctoral project for about two years. The completion of this thesis would have been impossible without their helpful advice over the years.

I would like to extend my sincere thanks to all my fellow (ex-)colleagues from the Singular Research Center for Information Technology (Centro Singular de Investigación en Tecnoloxías Intelixentes) of the University of Santiago de Compostela whose company made me feel greatly integrated to the workplace and far beyond it. I also gratefully acknowledge the assistance of the staff of the University of Santiago de Compostela whose guidance was particularly helpful when resolving various administrative issues that emerged along the way.

Special thanks to the funding agencies that provided financial support for my doctoral project. Namely, I would like to express my gratitude to the Spanish State Research Agency (Agencia Estatal de Investigación) which funded, in its entirety, my predoctoral contract (ref.: FPI2019-090153) as well as three nationwide-level projects (ADHERE - U: RTI2018-099646-B-I00; XAI4SOC: PID2021-123152OB-C21; DeepR3.gal: TED2021-130295B-C33) where I participated as a research assistant. In addition, I would like to extend thanks to the Department of Education and University Management (Consellería de Educación e Ordenación Universitaria) and the European Regional Development Fund (ERDF), which funded the project of excellence "Explica-IA" (ED431F 2018/02) that I had great pleasure of co-working on.

Importantly, the present piece of research is, in part, a product of active international collaboration. I would thus like to recognise the effort of members of the Laboratory of the New Ethos (Warsaw University of Technology, Poland) who played an important role in shaping up the methodological basis of the argumentative framework proposed in this thesis. I would particularly like to thank Prof. Dr. Katarzyna Budzynska, my secondment supervisor, and Dr. Marcin Koszowy for their active cooperation on the final part of my doctoral project as well as numerous outdoor activities.

This doctoral project would have been impossible without my family's support. I would therefore like to thank my mother, Liubov Stepina, my father, Nikolai Stepin, my brother, Dmitrii Stepin, and my dog, Irzhick, for always being by my side, even from all the way over there.

Last but not least, I would like to express my gratitude to some of great individuals whose company made this four-year journey particularly joyful. More precisely, I would like to thank my ex-partner Luis Alberto Cambón Rey for very many amazing times that we have shared. Special thanks to Borja Fernando Gómez Castro for his invaluable encouragement prior to and, especially, during the COVID-19 lockdown. I would like to extend thanks to Felipe Arturo Partarrieu Mejías, MD, for his enormous support, especially during the first month of the Russia-Ukraine war. Many thanks to Dra. Yaiza Bugallo Codeseira for her great loyalty to our friendship over the years. Finally, I wish to thank François Philippe Loncke and Ruddymar El Hashim Riera for proving that true love knows no reason, no boundaries, and no distance.

30 June 2023

Contents

Resumo

A Intelixencia Artificial (IA) xoga un papel cada vez máis importante nunha gran cantidade de ámbitos da vida diaria. As aplicacións de IA atópanse en numerosos produtos, dende a banca ata a asistencia sanitaria, pasando pola fabricación ata a educación. Non obstante, o rápido progreso da IA actual suscita preocupacións relacionadas coa súa interpretación e explicabilidade. Por unha banda, os modelos de IA estanse facendo demasiado complexos para que un público xeral comprenda a natureza dos sistemas de toma de decisións dos que forman parte. Isto pode socavar a confianza nas decisións automatizadas producidas por tales sistemas e aumentar a reticencia a usalos. Por outra banda, a transición de algoritmos de IA baseados en regras codificadas a técnicas de aprendizaxe automática (AA) guiadas por datos provocou que a natureza destes algoritmos se ocultase incluso aos seus desenvolvedores.

Para desmitificar tales algoritmos de "caixa negra" tanto para usuarios inexpertos como para especializados do dominio, investigadores de numerosos campos da ciencia solicitaron que a IA actual sexa *explicable*. Isto deu lugar a innumerables proxectos de investigación que forman a base da comunidade de AI eXplicable (XAI) xurdida recentemente. En consonancia coas aspiracións científicas, o uso omnipresente da IA provocou importantes cambios na normativa legal, que se reflicten, por exemplo, no Regulamento Xeral de Protección de Datos (RXPD) da Unión Europea (UE) ou na Proposta da Lei de IA (LIA), que foi votada polo Parlamento da UE en xuño de 2023, estando na fase final antes de converterse en lei e entrar en vigor en cada estado membro.

As escasas capacidades explicativas dos modelos de IA de "caixa negra" motivaron discusións sobre un uso favorable dos chamados modelos interpretables. En diante, o concepto de modelos interpretables refírese á familia de algoritmos cuxos compoñentes internos son comprensibles por humanos. De feito, os modelos de AA máis interpretables pero (posiblemente) menos precisos poden ser mais apropiados para resolver varios problemas desafiantes que os algoritmos máis robustos pero menos transparentes, especialmente en casos de decisións de alto risco. Non obstante, as subtarefas relacionadas coa xeración de explicacións como, por exemplo, a avaliación e a comunicación seguen sendo tarefas esixentes mesmo para modelos interpretables.

Unha gran cantidade de investigacións interdisciplinares sobre a natureza da explicación

afirma que as explicacións deben satisfacer unha serie de propiedades para que sexan eficaces. En primeiro lugar, as explicacións efectivas deben ser *contrastivas*, é dicir, non só explican por que a decisión ou predición automatizada dada é o caso, senón que tamén dan razóns polas que non son aplicables os resultados alternativos. Ademais, deben ser *selectivas*, é dicir, deben incluír só un número suficientemente pequeno das causas ou factores máis relevantes que conducen á decisión ou predición dada. Por último, pero non menos importante, as explicacións considéranse *sociais*, é dicir, son produto da interacción entre o explicador (o axente que explica o fenómeno dado) e o explicado (o destinatario da explicación). Como resultado desta hipótese, as explicacións automatizadas terán a máxima utilidade se se modelan de acordo cos requisitos descritos anteriormente.

De acordo coa normativa vixente da Escola de Doutoramento Internacional da Universidade de Santiago de Compostela, a presente tese preséntase en forma de compendio de publicacións. En xeral, a tese divídese en nove capítulos. O Capítulo 1 introduce o tema da tese. No Capítulo 2 exponse a hipótese probada e os obxectivos xerais e específicos da tese. No Capítulo 3 descríbese a metodoloxía aplicada para acadar os obxectivos da tese. O capítulo 4 ofrece unha discusión xeral sobre as propiedades de explicación modeladas e as ferramentas de xeración de explicacións desenvolvidas nesta tese. O capítulo 5 recolle as principais contribucións que xurdiron neste proxecto. Os capítulos 6-8 están relacionados coa metodoloxía e as conclusións recollidas nos artigos de revistas que forman o núcleo desta tese. O capítulo 9 extrae as principais conclusións da presente tese. A continuación, resumimos capítulos 6-8 con máis detalle.

Nesta tese, levouse a cabo o deseño, implementación e validación dun novo marco de xeración de explicacións que cumpre todos os requisitos xerais de explicación mencionados anteriormente (é dicir, as explicacións xeradas son contrastivas, selectivas e sociais). Defendemos o uso de modelos interpretables baseados en regras que se poidan utilizar (e cuxas predicións poidan ser explicadas posteriormente) de forma independente ou como representates para explicar as predicións de algoritmos de "caixa negra". De calquera xeito, o marco proposto serve para explicar o resultado dun clasificador interpretable (por exemplo, unha árbore de decisións ou un sistema de clasificación difuso baseado en regras).

No contexto de XAI, a predición dun clasificador pode explicarse (non necesariamente de forma contrastiva) en termos dos trazos máis característicos da instancia que conduciron á predición dada. En diante, referirémonos a tales explicacións como *factuais*. Para que as explicacións resultantes sexan *contrastivas*, buscamos modelar explicacións complementarias ás factuais, é dicir, que opoñen explícitamente o resultado da clasificación realmente predito a resultados alternativos hipotéticos. Noutras palabras, a predición do clasificador dado explícase non só en función das características que son máis relevantes para a predición, senón tamén en termos de clasificacións non preditas. Ademais, tales explicacións poden suxerir cambios mínimos nos valores das características para que o resultado previsto cambie da forma desexada. En XAI, estas explicacións tamén se denominan comúnmente como *contrafactuais* (CF). Nótese que contrafactuais refírense a exemplos xa observados no pasado, mentres que *transfactuais* refírense a exemplos sintéticos aínda non vistos pero que se espera que se observen no futuro. Non obstante, para simplificar a notación, remitirémonos ás explicacións CF no resto desta tese, sen importar se os cambios suxeridos involucran a creación de exemplos sintéticos.

Para fins ilustrativos, consideremos un escenario bancario común. Unha cliente dun banco, unha muller de mediana idade cuxos ingresos son 4.000 euros mensuais, solicita un préstamo a un ano de 30.000 euros. Adestrado para prever se as solicitudes de préstamo deben ser aceptadas ou rexeitadas, o sistema de clasificación bancaria suxire que o funcionario bancario debe rexeitar a solicitude da cliente. Para proporcionarlle ao cliente a recomendación máis relevante sobre como se podería modificar a decisión, o sistema presenta a seguinte suxestión de CF: “A solicitude de préstamo da cliente sería aprobada se os seus ingresos mensuais fosen polo menos 5.000 euros e tivese polo menos un préstamo activo menos”. Como se desprende do exemplo anterior, as explicacións de CF no contexto de problemas de clasificación son inherentemente contrastivas, xa que se opoñen de xeito explícito a diferentes resultados de clasificación.

As explicacións contrastivas e, máis específicamente, CF son estudadas dende hai moito tempo nunha ampla gama de ciencias. Por exemplo, dise que forman parte integrante do razoamento humano. Arguméntase tamén que os contrafactuais representan o nivel máis alto de causalidade. Ademais, pódense xerar para calquera clasificador. Por estes motivos, as explicacións CF chamaron a atención de numerosos investigadores da XAI nos últimos anos. Convencionalmente, os contrafactuais considéranse explicacións agnósticas do modelo, post-hoc e locais. Son locais porque explican o comportamento do sistema a partir das súas predicións individuais. Sábese que os contrafactuais explican as predicións de forma post-hoc, xa que se xeran despois de obter a saída do sistema. Notablemente, esta familia de explicacións é coñecida, en xeral, por ser independente do modelo, xa que os correspondentes métodos de xeración de explicacións están deseñados para operar só na entrada dada e na saída prevista do sistema sen acceder necesariamente aos elementos internos do sistema. Non obstante, a diversidade dos métodos de xeración de explicacións contrastivas e CF recentemente emerxentes mostra que non se limitan necesariamente a esta definición convencional.

No Capítulo 6, revisamos as teorías existentes sobre a explicación contrastiva e CF dunha ampla gama de ciencias. Ademais, analizamos os marcos computacionais de última xeración deseñados para a xeración dos dous tipos de explicación mencionados anteriormente. Ademais, inspeccionamos o grao de sinerxía entre os enfoques teóricos da explicación contrastiva e CF e as súas contrapartes computacionais de última xeración.

Cabe destacar que as explicacións CF posúen unha serie de propiedades importantes que se poden utilizar como medidas de utilidade da explicación (validez, proximidade, accionabilidade, diversidade, por citar algunhas). Nesta tese, centrámonos na modelización de CFs que se presume que son suficientes co seguinte subconxunto de propiedades. En primeiro lugar, os CF deben ser *válidos*, é dicir, deben levar a predicións correctas correspondentes ao resultado alternativo desexado. En segundo lugar, espérase que unha explicación CF inclúa só un conxunto de cambios mínimos nos pares característica-valor da instancia para que cambie a clasificación prevista. De feito, o axente explicado está interesado en recibir a explicación CF máis relevante para a instancia que se está a considerar. Tendo en conta o exemplo bancario anterior, mentres que un CF que indica que os ingresos mensuais deberían ser de 6.000 euros ou máis seguirán sendo válidos neste escenario, o usuario final está interesado en manter este valor o máis próximo posible aos seus ingresos reais, polo que a explicación CF sería preferible indicar que os ingresos mensuais deberían ser de 5.500 euros (sempre que os dous CF sexan válidos). En terceiro lugar, espérase que unha explicación CF sexa *accionable*, é dicir, só as características que se poidan modificar de xeito viable forman parte da explicación. De feito, se o sistema suxire que se reduza a idade do cliente, o CF correspondente é inútil, aínda que todos os demais cambios se poidan facer con éxito. Por último, espérase que os contrafactuais sexan *diversos*, é dicir, que abrangan varios CF válidos distintos (un punto único ou agrupados en conxuntos) que teñan un poder explicativo equivalente e, idealmente, que se atopen dispersos nas rexións de datos que cobren. De feito, o usuario final pode atopar explicacións máis satisfactorias que conteñan non só valores específicos das características CF dadas, senón intervalos de tales valores. A presentación de CFs tan diversas pode aumentar a flexibilidade do comportamento do usuario, xa que o destinatario da explicación ten a posibilidade a escoller o escenario a seguir que mellor lle conveña. A diversidade de explicacións de CF baseadas en regras pode manifestarse de varias maneiras. Por unha banda, as características de explicación pódense representar numericamente, en forma de intervalos (por exemplo, "5.500 $\leq$ renda $\leq$ 6.000"). Por outra banda, pódense ofrecer no seu lugar as correspondentes descricións textuais (por exemplo, "a renda é alta"). En ambos casos, os CF xerados automaticamente inclúen un conxunto de datos CF que permiten ao usuario final escoller o valor alternativo máis axeitado para as funcións dadas entre o intervalo de valores suxerido. Non obstante, non está claro se tales etiquetas lingüísticas (é dicir, "alta" do exemplo anterior) fan as características explicativas correspondentes máis comprensibles ou fáciles de usar e, polo tanto, a explicación xeral máis efectiva.

Para facer as explicacións *selectivas*, confiamos no uso de elementos internos de modelos baseados en regras. Algúns destes algoritmos de clasificación interpretables por deseño agregan información sobre as características que son máis relevantes para a predición dada. Por exemplo, as árbores de decisión conteñen os valores de características máis relevantes no camiño de decisión. Polo tanto, a explicación fáctual pódese reconstruír resumindo a información agregada no camiño desde a raíz ata o nodo folla previsto. Non obstante, a xeración de explicacións CF efectivas para familias específicas de algoritmos interpretables, como árbores de decisión ou sistemas de clasificación baseados en regras difusas, seguen sendo pouco estudadas. Ademais, estes clasificadores baseados en lóxica difusa ofrecen ferramentas que, por deseño, permiten aos desenvolvedores constituír explicacións textuais equivalentes utilizando o repertorio de termos lingüísticos. Así, potenciar árbores de decisión (difusas) e clasificadores baseados en regras difusas con novos métodos de xeración de explicacións CF permítenos mellorar aínda máis o seu potencial explicativo.

No Capítulo 7, propoñemos tres algoritmos de xeración de explicacións CF que producen explicacións textuais factuais e CF para clasificadores baseados en regras preseleccionadas. En primeiro lugar, deseñamos un algoritmo (en diante, denomínase XOR) que xera explicacións CF ordenando as representacións vectorizadas das regras CF de acordo co seu grao de relevancia para a instancia de proba. Supoñemos que tales explicacións baseadas en regras levan á xeración de explicacións CF válidas. Posteriormente, introducimos unha variante alternativa do mesmo algoritmo (en diante, denomínase EUC) que relaciona os vectores de funcións de pertenza difusa coas regras CF mediante a medición da distancia euclidiana entre todos os pares de tales vectores. Para ambos algoritmos, propoñemos ademais o mecanismo de *aproximación lingüística*, é dicir, un método para asociar intervalos de características numéricas a termos lingüísticos. Esta extensión permítenos xerar diversas explicacións automáticas equivalentes en formato numérico ou puramente textual.

A pesar de que os dous algoritmos introducidos anteriormente ofrecen explicacións facilmente interpretables, son específicos do modelo, é dicir, requiren acceso aos elementos internos do modelo e non se poden aplicar directamente a calquera clasificador. Non obstante, os algoritmos de xeración de explicacións CF independentes do modelo son capaces de explicar universalmente a saída de calquera clasificador tanto de xeito factual como contrafactual. Así, tamén propoñemos un algoritmo de xeración de explicacións CF xenética independente do modelo (en diante, denomínase GEN) que produce explicacións CF optimizando a poboación inicial de forma iterativa ata que se identifique o único punto de datos máis próximo á instancia de proba. En conxunto, ambos grupos de algoritmos de xeración de explicacións CF (é dicir, aqueles específicos do modelo e os agnósticos do modelo) poden usarse de forma complementaria entre si, especialmente se as explicacións resultantes veñen en diferentes formatos.

Co fin de comparar a utilidade das explicacións específicas do modelo baseadas en coñecementos imprecisos fronte a outras que apuntan a puntos de datos específicos, realizamos dous estudos de avaliación humana (é dicir, *Survey GM* e *Survey TS*) onde comparamos a eficacia das explicacións resultantes para as instancias de proba preseleccionadas. En ambos estudos, adestramos sistemas de inferencia difusa de Mamdani que fan predicións utilizando o algoritmo FURIA e xeran as explicacións correspondentes utilizando todos os algoritmos propostos (é dicir, XOR, EUC e GEN). Adestramos aos clasificadores nun conxunto de datos de clasificación de tipos de cervexa para xerar posteriormente explicacións lingüísticas para os estímulos da enquisa preditos correctamente. Cabe sinalar que todas as explicacións xeradas supoñense accionables debido á estrutura do conxunto de datos utilizado nos experimentos: en calquera caso, sempre é posible modificar os valores das características dentro dos intervalos de valores suxeridos.

Survey GM está deseñado para que o usuario poida avaliar a calidade da explicación automatizada en base ás catro máximas de Grice (cantidade, calidade, relevancia e forma), que transformamos en cinco aspectos explicativos (informatividade, fiabilidade, precisión, relevancia e lexibilidade) para que os participantes do estudo comprendan máis facilmente a súa tarefa de avaliación. Durante o estudo, os participantes avaliaron cada aspecto das tres explicacións (unha por cada método de explicación proposto) empregando unha escala Likert de 7 puntos. Pola súa banda, *Survey TS* é unha variante simplificada de 5 puntos baseada na escala Likert empregada en *Survey GM* onde se avalía unha única explicación en termos de fiabilidade e satisfacción xeral. Ambas enquisas leváronse a cabo para un público obxectivo que tiña suficiente coñecemento do dominio e un alto nivel de experiencia. A investigación e os protocolos experimentais foron aprobados polo comité ético da Universidade de Santiago de Compostela.

Mentres que se propuxeron un gran número de métricas computables automaticamente para estimar a calidade das explicacións automatizadas, estas adoitan servir para avaliar a calidade desde o punto de vista algorítmico (por exemplo, a distancia xeométrica á instancia de proba). Non obstante, as métricas automáticas que estiman aspectos da percepción do usuario seguen sendo escasas. Para abordar este problema, propoñemos a métrica da complexidade da explicación percibida, é dicir, unha estimación do complexa que parece ser unha explicación desde o punto de vista do usuario ao lela. A nova métrica proposta está inspirada no Gunning Fog Index (un indicador da facilidade de entender o texto por parte do público destinatario). En particular, baséase en dous factores que están presentes nas explicacións textuais baseadas en regras xeradas mediante os métodos de xeración de explicacións propostos: a lonxitude da explicación e a relación agregada do número de termos lingüísticos de todas as características utilizadas na explicación. Os resultados dos estudos de avaliación humana realizados mostran que a complexidade da explicación percibida ten unha correlación positiva moderada coa informatividade estimada polo usuario e unha forte correlación negativa coa relevancia e a lexibilidade estimadas polo usuario, mentres que as puntuacións métricas propostas non se correlacionan coa fiabilidade ou a precisión. Polo tanto, pódese concluír que a métrica proposta comprende tres dos aspectos de explicación mencionados anteriormente para usuarios que teñan coñecementos e experiencia suficientes no dominio. Ademais, usar a puntuación de complexidade da explicación percibida pode ser útil para diminuír os custos de avaliación humana (para o público obxectivo), xa que se pode calcular para substituír (en parte) os estudos de usuarios correspondentes.

Compre sinalar que os algoritmos propostos no Capítulo 7 limítanse á xeración de explicacións contrastivas selectivas non sociais, é dicir, carecen de calquera interacción directa co usuario final e só representan os datos CF máis relevantes desde o punto de vista algorítmico. Nestas configuracións, o usuario final ten que tomar unha decisión sobre a fiabilidade das explicacións automatizadas ofrecidas sobre a base dunha única información. Se a explicación non se considera o suficientemente fiable ou satisfactoria, o usuario final pode querer descartala aínda que sexa válida. Polo tanto, é indispensable que o usuario final explore o espazo de explicación se o considera necesario.

Para facer *sociais* as explicacións resultantes, modelamos a interacción entre o sistema e o usuario en forma de diálogo explicativo argumentativo onde o usuario é capaz de discutir sobre pezas de explicación específicas e, deste xeito, explorar o espazo de explicación ata que poida tomar unha decisión informada sobre a predición do sistema. Para iso, ampliamos os nosos algoritmos de xeración de explicacións cun módulo de xeración de diálogos explicativos. Deseñado como un axente conversacional, o modelo de diálogo resultante garante unha comunicación dialóxica interactiva entre o sistema e o usuario onde este está habilitado para solicitar e procesar as explicacións necesarias ofrecidas de forma comprensible e humana. Ademais, esta extensión mellora o marco de xeración de explicacións proposto coa opción de formar explicacións interactivas dinámicas en contraste coas xenéricas estáticas.

No Capítulo 8 propoñemos o denominado “xogo de diálogo explicativo”, un modelo formal de diálogo explicativo baseado no enfoque correspondente á modelización do diálogo a partir da teoría da argumentación. O modelo de diálogo está deseñado de forma descendente, é dicir, baséase nun protocolo de diálogo predefinido que contén catro posibles solicitudes de usuarios (as de explicación factual ou CF, detalle, aclaración e explicación alternativa), con varias posibles respostas do sistema asociadas a distintas solicitudes do usuario. Ademais, xeneralizamos o modelo de diálogo explicativo na forma dunha gramática de diálogo sen contexto para facelo universalmente aplicable á saída de calquera sistema de clasificación baseado en regras mellorado cun explicador que é capaz de producir explicacións textuais baseadas en regras.

Validamos o modelo de diálogo resultante realizando un estudo de avaliación humana mediante tres casos de uso: clasificación da posición do xogador de baloncesto, clasificación do tipo da cervexa e clasificación da enfermidade da tiroide. Nestes escenarios de diálogo de *busca de información*, unha das partes do diálogo (neste caso, o usuario) é inicialmente informada sobre os datos que se están procesando e despois pretende dar sentido á información dada pola outra parte (neste caso, a predición do sistema). Nos tres casos de uso, os diálogos pretenden explicar a predición dun único sistema para unha instancia de datos previamente seleccionada e clasificada correctamente. Como neste experimento se aborda o aspecto comunicativo da xeración de explicacións, utilízanse árbores de decisión nítidas como clasificadores, xunto co método XOR empregado para xerar as explicacións correspondentes, co fin de garantir a transparencia dos resultados experimentais.

Para analizar as transcricións de diálogos recollidas, aplicamos técnicas de minería de procesos que tratan as instancias de diálogo explicativo como fíos de proceso. En particular, realizamos a denominada comprobación de conformidade para relacionar o protocolo de diálogo e o corpus de diálogos explicativos realmente rexistrados. Concluimos que os participantes no estudo fan un uso activo de todos os tipos de solicitudes ofertadas. Polo tanto, o procedemento de verificación da conformidade confirma a utilidade do modelo de diálogo proposto inicialmente na súa totalidade. Ademais, rexistramos un gran número de solicitudes de explicacións de CF alternativas (segunda e terceira mellor clasificadas polo sistema) para case todas as clases de CF en todos os casos de uso. Esta observación apunta ademais á necesidade de presentar diversos CFs múltiples en diálogos de busca de información.

Cabe destacar que todos os traballos de revistas e software que constitúen a base da presente tese están a disposición do público. Finalmente, realizamos observacións finais e esbozamos direccións para o traballo futuro no Capítulo 9.

Summary

Artificial Intelligence (AI) plays an increasingly important role in a large number of daily life activities. AI applications are found in numerous products, from banking to health care to manufacturing to education. However, the fast progress of the present-day AI raises concerns related to its interpretability and explainability. On the one hand, AI models are rapidly becoming overly complex for a general audience to understand the nature of the decision-making systems that they make part of. This may undermine trust in automated decisions produced by such systems and increases reluctance to use them. On the other hand, shifting from hard-coded rule-based to data-driven machine learning (ML)-based AI algorithms has resulted in the nature of such algorithms being concealed even from their developers.

In order to demystify such "black-box" algorithms to both lay users and domain experts, researchers from numerous fields of science called for making the present-day AI *explainable*. This resulted in countless research projects forming the basis of the recently emerged eXplainable AI (XAI) community. In line with scientific aspirations, the ubiquitous use of AI has led to major changes in legal regulation, which are reflected in, for example, the European Union's (EU) General Data Protection Regulation (GDPR) or the recently proposed Artificial Intelligence Act (AIA) which was voted for by the EU Parliament in June 2023, being in the final stage before becoming law and coming into force in each member state.

Poor explanatory capacities of "black-box" AI models have motivated discussions on a favourable use of so-called interpretable models, instead. Hereinafter, the concept of interpretable models refers to the family of algorithms that grant access to their human-comprehensive internals. Indeed, more interpretable but (possibly) less accurate ML models may appear to be more efficiently applicable to solving various challenging problems than more robust but less transparent algorithms, especially in cases of high-stakes decisions. Nevertheless, such explanation generation-related sub-tasks as, for example, evaluation and communication remain being demanding tasks even for interpretable models.

A large body of interdisciplinary research on the nature of explanation claims that explanations should satisfy a number of properties for them to be effective. First, effective explanations are claimed to be *contrastive*, i.e. they do not only explain why the given automated decision or prediction is the case but also give reasons why alternative outcomes are not applicable. In

addition, explanations should be *selected*, i.e. they should include only an adequately small number of the most relevant causes or factors that lead to the given decision or prediction. Last but not least, explanations are deemed *social*, i.e. they are a product of interaction between the explainer (the agent that explains the given phenomenon) and the explainee (the recipient of the explanation). As a result, automated explanations are hypothesised to have maximal utility if modelled in accordance with the requirements outlined above.

In accordance with the actual regulations of the international doctoral school of the University of Santiago de Compostela, this thesis is presented in form of a compendium of publications. Overall, it is divided into nine chapters. Chapter 1 introduces the topic of the thesis. Chapter 2 states the hypothesis tested and the general and specific objectives of the thesis. Chapter 3 describes the methodology applied to reach the thesis objectives. Chapter 4 provides the reader with a general discussion on the explanation properties modelled and the explanation generation tools developed in this thesis. Chapter 5 lists main contributions that emerged as part of the doctoral project. Chapters 6-8 relate to the methodology and findings reported in the journal papers that form the core of this thesis. Chapter 9 draws main conclusions from the present thesis. Let us now summarise Chapters 6-8 in more detail.

We design, implement, and validate a novel explanation generation framework whose output explanations are claimed to meet all the aforementioned general requirements to explanation (i.e. being contrastive, selected, and social). We advocate the use of interpretable rule-based models that can be used (and whose predictions can be subsequently explained) independently or as proxies to explain predictions of "black-box" algorithms. Either way, the proposed framework serves the purpose of explaining the outcome of a given rule-based interpretable classifier (e.g., a decision tree or a fuzzy rule-based classification system).

In the context of XAI, a classifier's prediction can (not necessarily contrastively) be explained in terms of the most characteristic features of the test instance that led to the given prediction. Hereinafter, we refer to such explanations as *factual*. To make the output explanations *contrastive*, we pay particular attention to modelling explanations that are complementary to factual ones, i.e. they explicitly oppose the actually predicted classification outcome to hypothetical alternative outcomes. In other words, the given classifier's prediction is explained not only in terms of the features that are the most relevant to the prediction but also in terms of non-predicted classifications. Further, such explanations can suggest minimal changes in feature values so that the predicted outcome changes in a desired way. In XAI, these are also commonly referred to as the so-called *counterfactual* (CF) explanations. Notice that, counterfactuals refer to examples already observed in the past while *transfactuals* refer to synthetic examples not seen yet but expected to be observed in the future. Anyway, for simplicity of notation, we will refer to CF explanations in the rest of this thesis, no matter if suggested changes involve the creation of synthetic examples.

For illustrative purposes, let us consider a common banking scenario. A client of a bank, a middle-aged woman whose income equals €4.000 per month, is applying for a one-year loan of €30.000. Trained to predict whether loan applications should be accepted or rejected, the banking classification system suggests that the bank officer should decline the client's request. To provide the client with the most relevant recommendation of how the decision can be changed, the system outputs the following CF suggestion: "The client's loan application would be approved if her monthly income were at least €5.000 and if she had at least one active loan less." As follows from the example above, CF explanations in the context of classification problems are inherently contrastive, as they explicitly oppose different classification outcomes.

Contrastive and, more narrowly, CF explanations have long been studied in a wide range of sciences. For instance, they are claimed to make an integrative part of human reasoning. Further, counterfactuals are argued to represent the topmost level of causation. In addition, they can be generated for any classifier under consideration. For these reasons, they have attracted attention of numerous XAI researchers in recent years. Conventionally, counterfactuals are considered local post-hoc model-agnostic explanations. They are local because they explain the system's behaviour on the basis of its individual predictions. Counterfactuals are known to explain predictions in a post-hoc manner, as they are generated after the system's output has been obtained. Remarkably, this family of explanations is, in general, known to be model-agnostic, since the corresponding explanation generation methods are designed to operate only on the given input and predicted output of the system without necessarily accessing the system's internals. However, the diversity of the newly emerging contrastive and CF explanation generation methods shows that they are not necessarily limited to this conventional definition.

In Chapter 6, we review existing theories of contrastive and CF explanation from a wide range of sciences. Further, we analyse the state-of-the-art computational frameworks designed for generation of the two aforementioned kinds of explanation. In addition, we therein inspect the degree of synergy between the theoretical approaches to contrastive and CF explanation and their state-of-the-art computational counterparts.

Noteworthy, CF explanations possess a number of important properties that can be used as measures of explanation utility (validity, proximity, actionability, diversity, to name a few). In this thesis, we focus on modelling CFs that are hypothesised to suffice the following subset of such properties. First, CFs must be *valid*, i.e. they must lead to correct predictions corresponding to the desired alternative outcome. Second, a CF explanation is expected to include only a set of minimal changes to the test instance feature-value pairs for the predicted classification to change. Indeed, the explainee is interested in receiving the piece of CF explanation that is the most relevant to the test instance under consideration. Considering the banking example above, whereas the CF stating that the monthly income should be €6.000 or more will still be valid in this scenario, the end user is interested in keeping this value as close as possible to her actual income, so the CF explanation stating that the monthly income should be €5.500 would be preferred (provided that both CFs are valid). Third, a CF explanation is expected to be *actionable*, i.e. only the features that can be feasibly changed make part of the explanation. Indeed, if the system suggests that the client's age be decreased, the corresponding CF is useless, even if all other changes can be made successfully. Last but not least, counterfactuals are expected to be *diverse*, i.e. they should cover multiple distinct (either single-point or grouped in sets) valid CFs that have equivalent explanatory power and, ideally, well dispersed in the data regions that they cover. Indeed, the end user may find more satisfactory explanations that contain not only specific values of the given CF features but ranges of such values. Presenting such diverse CFs may increase the flexibility of user behaviour, as the recipient of the explanation then becomes entitled to choose the scenario to follow that suits him or her best. Diversity of rule-based CF explanations can be manifested in several ways. On the one hand, explanation features can be represented numerically, in form of intervals (e.g., "$5.500 \leq$ income ≤ 6.000"). On the other hand, the corresponding textual descriptions can be offered instead (e.g., "income is high"). In both cases, automatically generated CFs embrace a set of CF data points letting the end user to choose the most suitable alternative value for the given features from the suggested range of values. However, it remains unclear whether such linguistic labels (i.e., "high" from the example above) make the corresponding explanatory features more comprehensible or user-friendly and therefore make the overall explanation more effective.

To make automated explanations *selected*, we rely on the use of internals of rule-based models. Some of such interpretable-by-design classification algorithms aggregate information on the features that are most relevant to the given prediction. For example, decision trees contain the most relevant feature values to the test instance in the decision path from the root to the predicted leaf node. The output factual explanation can therefore be reconstructed by summarising the information aggregated in the decision path. However, effective CF explanation generation for specific families of interpretable algorithms, e.g. (fuzzy) decision trees, fuzzy rule-based classification systems, remains understudied. Further, such fuzzy logic-based classifiers offer means that, by design, enable developers to constitute equivalent textual explanations using the repertoire of linguistic terms. Thus, empowering (fuzzy) decision trees and fuzzy rule-based classifiers with novel CF explanation generation methods allows us to further enhance their explanatory potential.

In Chapter 7, we propose three CF explanation generation algorithms that output textual factual and CF explanations for preselected rule-based classifiers. First of all, we design an algorithm (hereinafter, it is referred to as XOR) that generates CF explanations by ranking vectorised representations of CF rules in accordance with their degree of relevance to the test instance. We assume that such rule-based explanations lead to generation of valid CF explanations. Subsequently, we introduce an alternative variant of the same algorithm (hereinafter, it is referred to as EUC) that relates fuzzy membership function vectors to CF rules by measuring Euclidean distance between all pairs of such vectors. For both algorithms, we additionally propose the mechanism of *linguistic approximation*, i.e. a method of mapping intervals of numerical feature values to linguistic terms. This extension allows us to generate diverse equivalent automated explanations in either numerical or purely textual format.

Despite the fact that both of the algorithms introduced above offer easily interpretable explanations, they are model-specific, i.e. they require access to the internals of the model and cannot be directly applied to any classifier. Notwithstanding, model-agnostic CF explanation generation algorithms are able to universally explain the output of any classifier both factually and counterfactually. Hence, we additionally propose a model-agnostic genetic CF explanation generation algorithm (hereinafter, it is referred to as GEN) which produces CF explanations by optimising the initial population iteratively until the single closest-to-the-test-instance data point is identified. Altogether, both groups of CF explanation generation algorithms (i.e., those model-specific and model-agnostic ones) can be used complementarily to each other, especially if the output explanations come in different formats.

In order to compare the utility of the fuzzy set-based textual model-specific explanations using imprecise knowledge against model-agnostic ones pointing to specific data points, we perform two human evaluation studies (i.e., *Survey GM* and *Survey TS*) where we compare effectiveness of the output explanations for the pre-selected test instances. In both studies, we train Mamdani fuzzy inference systems that make predictions using the FURIA algorithm and generate the corresponding explanations using all the proposed algorithms (i.e., XOR, EUC, and GEN). We train the classifiers on a beer style classification dataset to subsequently generate linguistic explanations for the correctly predicted survey stimuli. It is worth noting that all the generated explanations are assumed to be actionable due to the structure of the dataset used in the experiments: in any case, it is always possible to modify feature values within the suggested ranges of values.

Survey GM is designed to enable the user to assess the quality of the automated explanation on the basis of the four Gricean maxims (those of quantity, quality, relevance, and manner), which we transformed into five explanation aspects (i.e., informativeness, trustworthiness, accuracy, relevance, and readability) for the study participants to more easily understand their evaluation task. The study participants assessed each explanation aspect of the three explanations (one per each explanation method proposed) on the basis of a 7-point Likert scale. In turn, *Survey TS* is a simplified 5-point Likert scale-based variant of *Survey GM* where a single explanation is assessed in terms of trustworthiness and overall satisfaction. Both of the surveys were carried out for a target audience that had sufficient domain knowledge and a high level of expertise. The research and experimental protocols were approved by the ethical committee of the University of Santiago de Compostela.

Whereas a high number of automatically computable metrics have been proposed for estimating quality of automated explanations, those usually serve to assess quality from the algorithmic point of view (e.g., the geometric distance to the test instance). However, automatic metrics that estimate user-oriented aspects of explanation perception remain scarce. To address this issue, we propose the metric of perceived explanation complexity, i.e. an estimate of how complex an explanation appears to be from the user's point of view upon reading it. The newly proposed metric is inspired by the Gunning Fog Index (an indicator of how easy to understand the given piece of text appears for the intended audience). In particular, it relies on two factors that are present in textual rule-based explanations generated using the proposed explanation generation methods: the explanation length and the aggregated ratio of the number of linguistic terms of all the features utilised in the explanation. The findings from the human evaluation studies carried out show that perceived explanation complexity has a moderate positive correlation with user-estimated informativeness and a strong negative correlation with user-estimated relevance and readability whereas the proposed metric scores do not happen to correlate with trustworthiness or accuracy. Hence, it can be concluded that the proposed metric encompasses three of the aforementioned explanation aspects for users that have sufficient domain knowledge and expertise. Further, using the perceived explanation complexity score can be useful to decrease human evaluation costs (for the targeted audience), as it can be calculated to (partly) substitute the corresponding user studies.

It is worth noting that the algorithms proposed in Chapter 7 are limited to the generation of one-shot contrastive selected explanations, i.e. they are void of any direct interaction with the end user and merely represent the most relevant CF data points from the algorithmic point of view. In these settings, the end user has to make a decision on how trustworthy the offered automated explanations are on the basis of a single piece of information. If the explanation is not considered trustworthy or satisfactory enough, the end user may want to discard it even if it is valid. It is therefore indispensable to enable the end user to explore the explanation space if she finds it necessary.

To make the output explanations *social*, we model interaction between the system and the user in the form of argumentative explanatory dialogue where the user is capable to argue over specific pieces of explanation and, in this way, explore the explanation space until she can make an informed decision about the system's prediction. To do so, we extend our explanation generation algorithms with an explanatory dialogue generation module. Designed as a conversational agent, the resulting dialogue model ensures interactive dialogic communication between the system and the user where the latter is enabled to request and process necessary explanations offered in a comprehensible, human-like manner. Further, this extension enhances the proposed explanation generation framework with the option of forming dynamic personalised explanations in contrast to static generic ones.

In Chapter 8, we propose the so-called “explanatory dialogue game”, a formal model of explanatory dialogue based on the corresponding approach to dialogue modelling from argumentation theory. The dialogue model is designed in the top-down manner, i.e. it relies on the predefined dialogue protocol that contains four possible user requests (those for factual or CF explanation, detailisation, clarification, and alternative explanation), with various possible system’s responses mapped to the user’s requests. Further, we generalise the proposed explanatory dialogue model in the form of a context-free dialogue grammar to make it universally applicable to the output of any rule-based classification system enhanced with an explainer that is able to produce textual rule-based explanations.

We validate the resulting dialogue model by carrying out a human evaluation study using three use cases: basketball player position classification, beer style classification, and thyroid disease classification. In the so-called *information-seeking* dialogue settings, one of the dialogue parties (in this case, the user) is initially informed about the data being processed and then she aims to make sense of the information given by the other party (in this case, the system’s prediction). In all three use cases, the dialogues aim to explain a single system’s prediction for a pre-selected correctly classified data instance. As the communicative aspect of explanation generation is addressed in this experiment, crisp decision trees are used as classifiers in this experiment, with the XOR method employed to generate the corresponding explanations, in order to ensure the transparency of the experimental results.

To analyse the collected dialogue transcripts, we apply process mining techniques treating instances of explanatory dialogue as process threads. In particular, we perform the so-called conformance checking to relate the dialogue protocol and the corpus of actually registered explanatory dialogues. It turns out that the study participants make an active use of all the offered types of requests. Hence, the conformance checking procedure confirms the utility of the initially proposed dialogue model in its entirety. Furthermore, we register a high number of requests for alternative (second- and third-best ranked by the system) CF explanations for almost all the CF classes across all the use-cases. This observation further points to the necessity for presenting diverse multiple CFs in information-seeking dialogues.

It is worth noting that all the journal papers and software that form the basis of the present thesis are made publicly available. Finally, we make concluding remarks and outline directions for future work in Chapter 9.

1 Introduction

1.1 Overview

Artificial Intelligence (AI) is living the era of data-driven algorithms whose striking accuracy is often achieved at the expense of explainability of the predictions obtained [10]. As the number of AI applications found in daily life is growing continuously, numerous ongoing discussions raise awareness about the impact of the use of such applications on their users. Thus, the poorly interpretable nature of most of the state-of-the-art data-driven algorithms has led to adoption of novel legal regulations. For example, the European Union's General Data Protection Regulation [28] addresses explainability-related issues of the present-day AI-based models in the context of automated decision-making. Whereas individual AI strategies are developed at the national level worldwide, the European Union is making a consolidated effort on shaping the legal future of AI, which is expected to be discussed further in the so-called AI Act (AIA) [29]. AIA recognises the need to provide end users of AI applications with explanations for their automated predictions or recommendations. Researchers in the field of eXplainable AI (XAI) [1, 11] are actively addressing the aforementioned challenges.

A large body of knowledge testifies that explanations have a diverse nature. In this thesis, we adhere to modelling explanations that satisfy three main properties that automated explanations are claimed to possess to be effective [23]. Namely, effective explanations are expected to be:

- *contrastive*, i.e., the given phenomenon is explained in terms of non-occurring hypothetical alternatives;
- *selected*, i.e. only the most relevant pieces of information are included in the explanation;
- *social*, i.e. the explanation is a product of interaction between the explainer and the explainee.

To explain a fact contrastively means to answer the why-question of the form "Why P rather than Q?" [20] where P is the fact under consideration and Q is a hypothetical, non-occurring alternative (also referred to as a "foil"). In the context of XAI, contrastive explanations are, in general, designed to oppose the predicted outcome to an alternative hypothetical prediction [25].

In this regard, the property of contrastiveness is at the core of the so-called *counterfactual* explanations (or counterfactuals, or CFs, for short), a sub-group of contrastive explanations that suggest minimal changes to the input feature values so that the output changes in the desired way [33]. Contrastive by nature, CFs are found to be inherent to human reasoning [4] and can therefore greatly facilitate explanation processing by end users [5]. For these reasons, explaining predictions counterfactually has become among key explainability issues, especially when explaining "black-box" models [21].

Given an evident lack of transparency in the reasoning of many complex AI models (e.g. neural networks), the use of possibly less accurate or robust but more interpretable models has been actively argued for [32]. In light of this, we explore the potential of rule-based classification systems to provide their end users with automated explanations for their predictions. Indeed, the potential of interpretable models for explanations is left largely underexplored [19]. Enhancing rule-based classification systems with effective methods of CF explanation generation allows them to become self-explanatory while providing their end users with contrastive selected explanations. Further, such self-explanatory rule-based interpretable classifiers can then be used as part of more complex explainers to address the issue of explanability of "black-box" models. In order to preserve the state-of-the-art levels of performance while gaining explainability, "black-box" models can be enhanced with explanation generation modules that make use of (possibly, surrogate) interpretable models [9]. Such interpretable models (e.g., decision trees or fuzzy rule-based classification systems) [2] have shown to effectively explain "black-box" models in a post-hoc manner when, for example, trained on a local neighbourhood around the test instance [41]. In this regard, they can serve as a proxy to approximate given single "black-box"-based predictions.

As the need for explaining decisions made by AI-based systems is recognised legally, various researchers are urging for making a step forward towards responsible, human-centric AI [6]. Whereas several automatic metrics have been designed to estimate the quality of automated CF explanations with respect to their computational aspects [26], human evaluation remains among the key challenges for truly effective CF explanation generation [43]. Indeed, only a limited number of state-of-the-art CF explanation algorithms have undergone assessment by potential beneficiaries of such explanations [16].

Human evaluation of automated explanations is closely connected with the social aspect of explanation. It is often addressed in XAI by means of engaging the end user in explanatory dialogue with the system [42]. Further, insights from humanities and social sciences (e.g., argumentation) allow us to propose explanatory dialogue models that rely on a consolidated body of knowledge about human reasoning and connect it to that of an AI-based agent. In fact, argumentation makes an integrative part of certain explanation theories and therefore appears to be a suitable methodological fit to bridge the gap between the explainer (the explanation generation module) and the explainee (the end user). Despite specific methodological differences, argumentation and explanation are found to greatly completement each other [3]. For example, some theories of explanation conceptualise explanations as arguments [12]. Whereas argumentation theories provide a diverse repertoire of frameworks that is capable of generating explanations for automatic predictions in a wide range of tasks [46], we aim to explore its potential as a communication channel between the end user and the system to enhance the previously designed framework for contrastive-counterfactual selected explanations for interpretable rule-based classification systems with a social dimension.

1.2 Thesis structure

The present thesis contains nine chapters. The remainder of the thesis is structured as follows.

Chapter 2 states the hypothesis tested in this thesis as well as the general and specific objectives. As we list the objectives of the thesis, we refer the reader to the publications where the objectives were reached.

Chapter 3 describes the general methodology applied throughout the thesis and describes specific tools that were used in order to reach the thesis objectives.

Chapter 4 provides the reader with a general discussion on explanation properties in the context of XAI and analyses in detail the strengths and weaknesses of their modelling in this thesis.

Chapter 5 lists the contributions of this thesis, i.e., the software developed to reach the thesis objectives and all the publications that emerged during the doctoral project.

Chapter 6 provides the reader with the background information on contrastive and CF explanations. In addition, it examines theoretical foundations thereof, the related state-of-the-art computational frameworks, and inspects the degree of synergy between the former and the latter.

Chapter 7 introduces three algorithms for CF explanation generation (namely, XOR, EUC, and GEN) used to explain predictions of an FRBCS. In addition to discussing technicalities of the aforementioned algorithms, it evaluates the algorithms via two human evaluation studies. Further, it proposes a novel metric of perceived explanation complexity (PEC) that aims to facilitate evaluation of automatically generated explanations.

Chapter 8 proposes an argumentative framework for communication of automatically generated rule-based explanations. In particular, it formalises explanatory dialogue in form of the so-called “dialogue game” and describes in detail the corresponding dialogue protocol. Further, it additionally represents the protocol in form of context-free dialogue grammar to make the protocol universally applicable to other explainer-classifier pairs that are capable of generating textual rule-based explanations. Last but not least, it reports the results of a human evaluation experiment that serves the purpose of validation of the proposed explanatory dialogue model.

Finally, Chapter 9 presents main conclusions derived from the results of the doctoral project and outlines prospective directions for future work that are relevant to the problems of CF explanation generation, communication, and evaluation.

2 Hypothesis and objectives

In this thesis, we develop an explanation generation framework for interpretable (i.e., "white-box") classifiers (e.g., decision trees) and semi-interpretable (i.e., "grey-box") rule-based classification systems (e.g., fuzzy inference systems). Despite the fact that such models provide predictions that can be easily interpreted factually, their CF potential remains understudied. We formulate the main hypothesis tested in the present thesis as follows: "*By modelling explanations satisfying the properties specifically relevant for XAI and enhancing them with dialogic interactive facilities, we can convey both factual and CF explanations that are appealing for a good number of users in different application domains*".

The general objective of the present doctoral thesis is to advance state-of-the-art XAI technologies for (1) automatic generation of factual and CF explanations for interpretable rule-based classifiers and (2) effective and comprehensive communication of such explanations. The implemented explanation generation framework is expected to output explanations that satisfy the aforementioned requirements to effective explanations (i.e. being contrastive, selected, and social). More precisely, the following specific objectives are considered to achieve the overall goal:

O1. Design, implement, and validate a framework for factual and CF explanation generation applied to given pretrained (semi-)interpretable rule-based classifiers. This objective has been successfully reached in the following publications:

- Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martín Pereira-Fariña. "A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence". *IEEE Access*, vol. 9, pp. 11974–12001, 2021. DOI: 10.1109/ACCESS.2021.3051315;
- Ilia Stepin, Jose M. Alonso-Moral, Alejandro Catala, Martín Pereira-Fariña. "An empirical study on how humans appreciate automated counterfactual explanations which embrace imprecise information". *Information Sciences*, vol. 618, pp. 379–399, 2022. DOI: 10.1016/j.ins.2022.10.098.

O2. Design, implement, and validate a conversational agent endowed with credibility via natural language processing and argumentation technologies in the context of XAI in order to communicate and customise automatically generated explanations. This objective has been successfully reached in the following publication:

- Ilia Stepin, Katarzyna Budzynska, Alejandro Catala, Martín Pereira-Fariña, Jose M. Alonso-Moral. "Information-seeking dialogue for explainable artificial intelligence: Modelling and analytics". *Argument and Computation*, in press, 2023. DOI: 10.3233/AAC-220011.

O3. Develop a human evaluation framework for the purpose of validation of the algorithms designed to achieve O1 and O2. This evaluation framework serves the purpose of estimating various aspects of automatically generated explanations and that of assessing the quality of the process of communication of such explanations, respectively. This objective has been successfully reached in the publications addressing both O1 and O2.

3 Methodology

The methodology employed in this thesis bases on the iterative development approach. In order to achieve the objectives listed in Section 2, the work on each of them presupposes the following consecutive steps:

1. *Requirement specification*: defining the research objective while considering possible limitations of the corresponding state-of-the-art AI techniques and software. Special attention is paid to aspects of automatic explanation generation and communication (i.e., properties of explanation, requirements to the communicative aspects of the language used for modelling argumentative explanatory dialogue, use cases, target audience, etc.);

2. *Literature review*: a bibliographic study that serves to identify the state-of-the-art conceptual, theoretical, and computational frameworks designed to address the goal-specific problems. This study includes an analysis of the advantages and disadvantages of the identified methods;

3. *Implementation:* design and development of conceptual models and algorithms aimed at producing advances with respect to the state-of-the-art. Theoretical contributions are followed by software implementations to be validated empirically at the next step;

4. *Validation*: the process of verification of the software as well as revision of the algorithm or model implemented if the experimental results obtained are not satisfactory;

5. *Integration*: once validated, the new algorithms or prototypes are integrated with those validated previously so that they make part of the unified framework.

In order to achieve objective O1, we first perform a systematic literature review (SLR) of the state-of-the-art methods of contrastive and CF explanation generation in the context of XAI following the guidelines for performing SLRs in software engineering [17, 18]. In particular, we formulate a series of research questions to be addressed, design a search strategy, and extract and synthesise data in accordance with the predefined inclusion and exclusion criteria. In addition, we perform the so-called "snowballing" procedure, i.e. a revision of the bibliography lists of the previously collected studies, following the corresponding guidelines [45]. Subsequently, based on the insights from our SLR, we develop a conceptual framework for factual and CF explanation generation for interpretable rule-based classifiers that outputs contrastive selected explanations. The framework includes one model-agnostic and two model-specific explanation generation algorithms. To generate automatic explanations in natural language, we adapt one of the most commonly used natural language generation (NLG) pipelines [7, 31]. We opted for using a template-based NLG approach instead of an end-to-end neural NLG approach to maximise the fidelity to the data of the generated narratives versus their naturalness. Accordingly, the use of pre-trained large language models for NLG falls outside the scope of this thesis. Consequently, we evaluate the proposed methods in a series of human evaluation experiments contributing to reaching objective O3, as we adhere to the human evaluation guidelines designed specifically for XAI [13].

To accomplish objective O2, we develop an argumentative conversational agent relying on the “dialogue game”-based theoretical approach to argumentative dialogue modelling [30]. More specifically, we design a set of original requests and responses that constitute a newly proposed explanatory dialogue protocol. Similarly to O1, we evaluate the designed argumentative framework by performing a human evaluation study and analyse the collected dialogue transcripts. In addition, we employ concepts from process mining to perform the so-called “conformance checking” treating instances of the collected explanatory dialogues as processes [27].

To reach objective O3, we design a software framework relying on human evaluation guidelines for XAI [13] and implement it as a stand-alone application. Its flexible structure allows us to adapt it to the needs of specific experiments carried out as part of the present thesis. As a result, we use it to evaluate all the CF explanation generation and communication methods proposed in this thesis.

In order to see the connection between the research questions posed and the tools developed to answer them (as well as the publications addressing them), we kindly refer the reader to Fig. 5.1 from Chapter 5 which lists the main contributions of this thesis.

4 General discussion

In this thesis, we explore the potential of interpretable rule-based classification systems to generate effective automated explanations in accordance with recent requirements to the quality of such explanations. Recall that effective explanations are claimed to be contrastive, selected, and social [23]. In this section, we discuss peculiarities of modelling automated explanations ensuring encapsulation of these key properties in the context of the explanation generation framework for interpretable rule-based classifiers developed in this thesis. In particular, we first inspect computational aspects of contrastive and CF explanations for classification problems. Then, we discuss how one-shot selected CF explanations can be generated for predictions made by interpretable rule-based classification systems. Finally, we inspect how a social layer can be added on top of the previously proposed explanation framework. More specifically, we focus on explanatory dialogue modelling as a means of explanation communication. We thus examine (1) the most relevant aspects of formal explanatory dialogue modelling for textual rule-based explanations and (2) a general architecture of the corresponding argumentative conversational agent. As a result, we propose an explanation generation framework whose output embraces all the aforementioned properties of explanations for interpretable rule-based classifiers (see Fig. 4.1).

The general discussion on enhancing automated explanations with the aforementioned properties is structured in the following manner. Section 4.1 discusses computational aspects of CF explanations and their connection with the family of contrastive explanations as defined in the literature. Section 4.2 examines aspects of one-shot, selected CF explanation generation for interpretable rule-based classification systems. Section 4.3 explores how insights from argumentation theory can enhance the social aspect of automated explanations in the context of explanatory dialogue systems.

4.1 Making explanations contrastive(-counterfactual)

Explanations are argued to be contrastive, as the facts that they explain are (sometimes, implicitly) opposed to pieces of information related to a contrast case [23]. They can be defined as answers to the contrastive why-question (i.e., “Why P rather than Q?”) where P is the fact being explained and Q is an alternative non-observed foil. Then, P is said to be explained

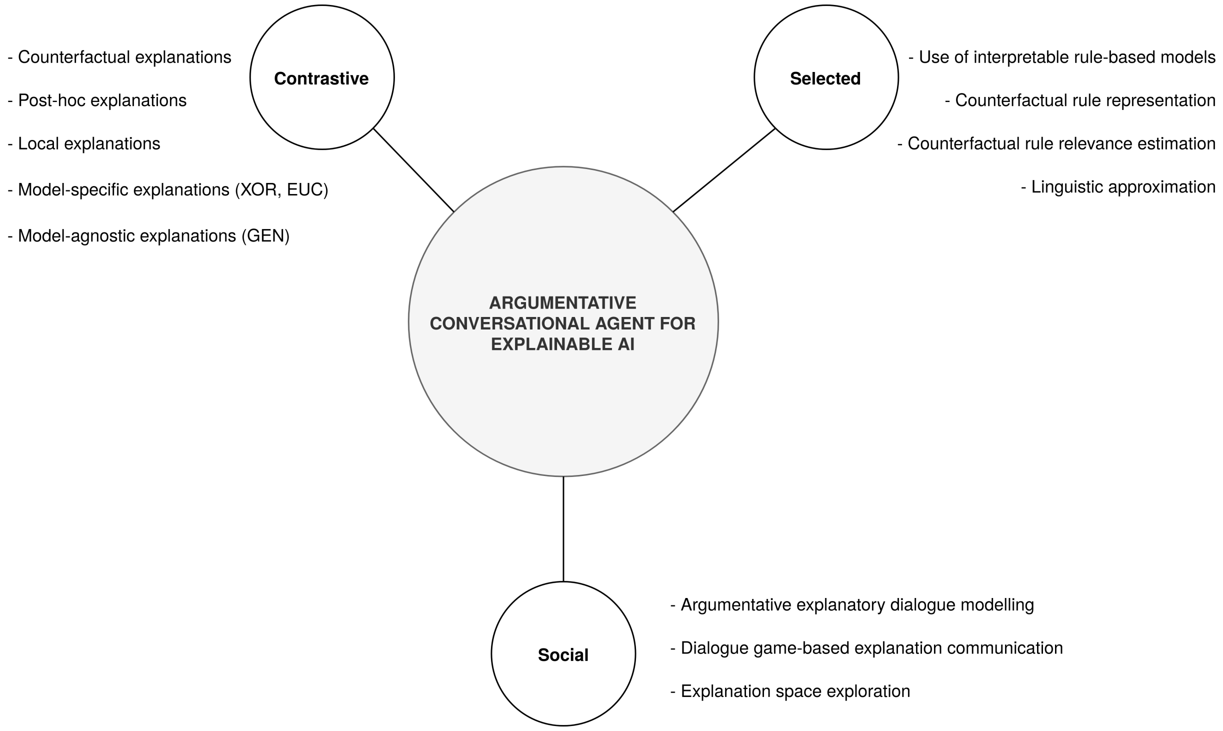

Contrastive
- Counterfactual explanations
- Post-hoc explanations
- Local explanations
- Model-specific explanations (XOR, EUC)
- Model-agnostic explanations (GEN)
Selected
- Use of interpretable rule-based models
- Counterfactual rule representation
- Counterfactual rule relevance estimation
- Linguistic approximation
ARGUMENTATIVE CONVERSATIONAL AGENT FOR EXPLAINABLE AI
Social
- Argumentative explanatory dialogue modelling
- Dialogue game-based explanation communication
- Explanation space exploration

Figure 4.1: Explanation properties that the proposed argumentative conversational agent embraces.

contrastively if a causal difference between *P* and *not-Q* is verified (the so-called "difference condition") [20]. Ensuring that automatically generated explanations are contrastive is claimed to greatly contribute to their readability, as contrastive explanations are said to prune the space of all causal factors aiding finer-grained understanding [15].

Theories of contrastive explanation largely rely on causal accounts of explanation. However, a number of existing explanation generation algorithms offer purely non-causal explanations [25]. Further, the notion of contrastive explanation is found to greatly overlap with that of CF explanation in the XAI community, despite certain methodological differences. Thus, causal accounts of contrastive explanation imply that CF explanations can be deemed contrastive so long as they respect the difference condition [24]. Nevertheless, CF explanations are imposed a number of additional constraints that make them a sub-group of contrastive explanations if causality is not being considered.

In the context of classification problems, contrastive and/or CF explanations differentiate the given piece of factual information (e.g., "Your loan application has been rejected because your monthly income is too low") from some CF information (e.g., "Your loan application would have been approved if you had at least one active loan less"). Altogether, a combination of factual and CF explanations enables the end user to construct a mental representation of the AI-based agent's reasoning for all possible outcomes.

CF explanations are a powerful tool for explaining predictions or decisions made by AI-based agents. In addition to factual explanations, they provide complementary information to explain the AI-based agent's reasoning, so that a combination of factual and CF explanations can explain all possible predictions of the AI-based agent irrespective of whether they took place or not. CF explanations are said to be post-hoc (i.e., they explain predictions of pretrained models) and local (i.e., they are designed to explain individual predictions). In addition, CFs can be model-agnostic (i.e., the corresponding explanation generation method operates only on the input feature values and the output prediction of the classifier) or model-specific (i.e., the explainer also has access to the classifier's internals). Automated CF explanations are claimed to have a number of properties (see [8, 26] for an exhaustive list thereof):

- **Validity.** A CF explanation is said to be valid iff the corresponding CF leads to the desired CF prediction;
- **Proximity.** A CF is said to be proximate iff the distance between the test instance and the CF data point that the given CF explanation is related to is as small as possible;
- **Sparsity.** A CF is said to be sparse iff it contains the minimal number of features in comparison to other valid CFs;
- **Diversity.** CFs are said to be diverse iff they form a set of valid CF data points that are at the same time maximally different from each other so that the explainee is offered a number of legit suggestions on how to change the given prediction to the alternative desired one.

Whereas the quantitative metrics defined to measure the properties listed above are commonly used for evaluating CFs [43], some may be incompatible with others. For example, there exists a trade-off between proximity and diversity: no CF explanation generation method is claimed to maximise both due to their divergent treatment of CFs with respect to the distance from the test instance [26]. In this case, other factors (e.g., the target audience or the application domain) may become crucial to estimate the quality of automated CFs.

In addition, researchers distinguish various other properties that automated CF explanations are desired to have. These include *actionability* (i.e., the mere ability to change the feature values as the given CF explanation suggests), *causality* (i.e., establishing causal relations between the features with respect to a given causal model and ensuring that such relations are maintained in the given CF), and *fairness* (i.e., the given CF explanation is unbiased with respect to specific protected features, e.g., gender or race). Indeed, theoretical accounts of contrastive and CF explanations are found to largely differ from each other in terms of their relation to the causal aspect of explanation. A vast majority of theoretical academic studies appeal to the causal nature of contrastive and/or CF explanation [36]. However, modelling CFs that possess these properties falls outside the scope of this thesis and is left for future work.

CF explanations are inherently contrastive when generated in the classification problem settings, as they oppose the actual output of the AI agent and an alternative hypothetical outcome. Approaches to automated CF explanation generation are shown to form a highly diverse set of algorithms. Our interdisciplinary SLR on state-of-the-art contrastive and CF explanations [36] shows that there exist various dimensions in which CF explanation generation algorithms can be categorised. Thus, we believe that the following four dimensions adequately encompass the state-of-the-art CF explanation generation methods:

- the problem that the corresponding AI agent aims to solve (e.g., classification, regression, knowledge engineering, planning, recommendation ranking, or conflict resolution);
- the generality of the given CF explanation generation method (model-specific or model-agnostic);
- the output representation of the automated CFs (numerical, linguistic, visual, or multi-modal);
- the strategy applied to evaluate the given CF explanation generation algorithm (intrinsic or extrinsic and subjective or objective).

Notably, CF explanation generation algorithms are found helpful at explaining different kinds of AI agents: from those solving such classic AI problems as classification and regression to those addressing recommendation ranking or planning tasks in robotics, with a notable focus on explaining classification algorithms.

Model-specific CF explanation generation methods exploit the internals of the agent that they are trying to explain (provided that they are available). Whereas they can be claimed to specifically target the reasoning of the given model, the generality of this family of methods is restricted to specific AI-based models. On the other hand, model-agnostic CF explanation generation methods oftentimes focus on exploring optimal solutions to the corresponding optimisation problem mainly using the mechanism of feature perturbation.

CF explanations have different modalities. Depending on the application domain and/or the target audience (among other decisive factors), CF explanations can be presented in tabular form, framed in natural language, visualised in form of, for example, heat maps, or be a mixture of the above making multi-modal CFs.

Importantly, existing CF explanation generation methods differ in how their performance is evaluated. Given a diverse nature of CF explanation generation methods, there is no uniform set of metrics consolidated for evaluation of automated CFs. Whereas there exist numerous automatically computable metrics [26], applying properly human evaluation techniques remains among the key challenges for effective CF explanation generation, validation, and communication.

Some of the aforementioned differences (as categorised above) make distinct CF explanation generation methods poorly comparable to each other. It may be impossible to directly compare distinct CF explanation generation algorithms due to, for example, the use of different evaluation techniques or the inability of the given CF explanation generation algorithms to produce explanations in the same settings. As the design of CF explainers is oftentimes domain-dependent, quantitative evaluation metrics for explaining, e.g., classification algorithms may be irrelevant or even non-applicable to explainers for planning tasks.

4.2 Making explanations selected

Whereas CF explanations are designed to describe minimal changes to the test instance's features, this does not necessarily imply that the number of features making part of such an explanation is minimal. Indeed, as shown above, sparsity is known to be one of indicators of quality of automated CF explanations. It is therefore important to find a balance between proximity and sparsity of the output CFs.

In this thesis, we focus on generation and communication of explanations to classification problems. Given a low degree of interpretability of black-box models, a wider use of interpretable models is advocated instead [32]. Indeed, such classifiers as decision trees (DT) or fuzzy rule-based classification systems (FRBCS) are designed to provide interpretable and human-comprehensive output. DTs provide straightforward means for generating factual explanations. Thus, a factual explanation can be generated for the given DT prediction by aggregating the information on the feature-value pairs found along the decision path that is responsible for the given prediction. Since there exists only one decision path to the given prediction in this case, a summary of the feature-value pairs can unambiguously explain the given prediction factually. In the case of FRBCSs, one may inspect all the activated rules leading to the predicted class. A factual explanation can then be generated solely on the basis of the feature-value pairs found in the antecedent of the rule that has the highest activation degree or by aggregating information from all the activated rules [2]. It is also possible considering the use of linguistic quantifiers for signalling how likely each given rule is responsible for the given prediction. For consistency, we choose to rely on single decision paths or rules to generate factual explanations for all the considered classification systems.

However, the problem of CF generation for interpretable rule-based classifiers remains far from trivial, as it poses numerous important algorithmic challenges. First, it is fundamental to determine how the CF space is pruned given a possibly infinite set of CFs taking into consideration that candidate CFs should be formatted to be compatible for comparison with the test instance. Second, distinct node of decision paths or conditions on predicting features of rules may aggregate information on the same feature. Hence, an aggregation of the relevant feature values may form intervals or sets, resulting in the predicting feature being a CF set. Therefore, it becomes essential to quantify the relevance of such CF sets to the given prediction. Third, some data-driven FRBCS learning algorithms (e.g., FURIA [14]) are known to return rules where the predicting values are not grounded semantically (e.g., "*Height is MF0*" where *MF0* stands for membership function 0). In this case, it is essential to transform such feature values into (approximated) linguistic terms. Whereas the previously mentioned research problems barely form a small subset of all CF generation-related problems (even for classification problems alone), we focus on addressing only the aforementioned challenges in this thesis.

We address these challenges by proposing two algorithms for CF explanation generation (i.e., XOR and EUC). Both of them operate on the internals of the given interpretable rule-based classifier and are designed to follow the reasoning of the pretrained rule-based classifiers to output comprehensive explanations in natural language. In this way, resulting explanations present only the most important features (as listed in the decision path or rule) leading to the prediction. Since both algorithms can be universally applied to (crisp or fuzzy) DTs or FRBCSs, we can summarise them in the following pipeline of four steps:

- **CF rule representation**. First of all, the test instance under consideration and the available internals of the classifier are transformed into rules. For example, a DT can be traversed to inspect distinct decision paths from the root to the leaf to aggregate all the information found relevant along the given decision path. In the case of FRBCSs, the rules are straightforwardly retrieved from the rule base. Subsequently, all CF rules and the test instance are vectorised to ensure their compatibility for subsequent calculations. Both CF rules and the test instance can be vectorised in terms of all the flags of presence or absence of all k conditions found in the tree nodes in the case of DTs or membership function values for all m feature-linguistic term pairs in the case of FRBCSs. The same procedure is then applied to the test instance so that a k- or m-dimensional test instance vector is obtained, respectively.

- **Relevance estimation**. Recall that CF explanations describe, by convention, minimal changes to the characteristics of the test instance under consideration to be made for the given prediction to change. Hence, once the test instance and CF rules are vectorised, they can be inspected for closeness with respect to each other. To do so, a measure of vector distance can be calculated for each pair of test instance vector and CF rule. As the test instance vector and all rule vectors for DTs are binary, we suggest that the eXclusive-OR (XOR) function be calculated for all such pairs of vectors. For FRBCSs, measuring the Euclidean distance (EUC) between the test instance vector and the given CF rule vector is suggested instead, as it allows for better capturing (possibly, real-valued) differences in membership function values without information loss. The resulting pairs are sorted by distance. The closest CF rule is deemed to be the most relevant to the test instance under consideration. If two or more CF rules are claimed to be the most relevant, the most sparse of them is selected (i.e., the one whose antecedent has a minimal number of features). If there exists more than one CF rule that are equally distant from the test instance vector being equally sparse, such rules are claimed to have equal explanatory power. In this case, the most relevant of such CF rules can be selected randomly.

- **Linguistic approximation**. To ensure further applicability of the proposed explanation generation method, it may be necessary to introduce an extra-linguistic layer if CF feature values are either desired to be converted from numerical to linguistic format or if they are not semantically meaningful. In the former case, intervals capturing feature values (e.g., "$180cm \leq Height \leq 200cm$") can be mapped to predefined intervals defining the corresponding linguistic terms (e.g., "*Height* is *tall*"). In the latter case, fuzzy sets learnt from the data may not be automatically mapped to the given linguistic terms. Then, the fuzzy set corresponding to the given feature value (e.g., "*Height* is *MF0*") is mapped to the most similar linguistic term (e.g., "*Height* is *short*") given an alpha-cut. In both cases, similarity measures are calculated for the given feature value interval and all intervals corresponding to linguistic terms. Then, the most similar linguistic term is selected to make part of the explanation for the given feature.

- **Surface realisation**. Once the CF rule is selected and all its feature values are defined in a human-comprehensive manner (i.e., they are semantically grounded), the resulting rule is verbalised using NLG techniques. In this thesis, we rely on preselected templates that follow the structure of the rules, as described below.

  Factual explanations that the proposed algorithms produce are designed to follow the template "The test instance is [CLASS] because [FEATURE] is [VALUE]" where [CLASS], [FEATURE], and [VALUE] are variables representing the predicted class, the most relevant features, and the corresponding feature values, respectively. In turn, CF explanations follow the template "The test instance would be [CLASS] if [FEATURE] were [VALUE]." In both examples above, the explanation templates are designed to have only one explanatory feature-value pair. Nevertheless, both automatically generated factual and CF explanations can contain multiple feature-value pairs following the structure of the underlying rule or the decision path. The subject of the main clause of the template (i.e., "The test instance") remains constant but can be preselected arbitrarily depending on the application domain (e.g., "The patient" in health-care settings).

Recall that the proposed explanation generation algorithms (we also refer to them as qualitative) follow the structure of the CF rule that is deemed to be the most relevant to the test instance. On the one hand, this is assumed to lead to the generation of valid CFs, as the underlying CF rule would fire if the feature values of the given test instance were set accordingly.

On the other hand, the resulting CF rules may possibly contain all the features from the dataset preventing them from being truly selected. In this case, such rules can be filtered out from the initial collection of vectorised CF rules in favour of, possibly, less relevant CF explanations but more concise and, possibly, actionable equivalents.

Notably, selected explanations generated on the basis of the internals of interpretable classifiers have several limitations. First, the utility of such explanations can be undermined if the feature space is not interpretable. Second, such explanations are model-specific, i.e. they cannot be directly generated when applied to non-rule-based classifiers using the same explanation generation method. Third, the property of being selected of the given explanation may be questioned if the corresponding explanation contains too many features due to a high complexity of the classifier. There exist various strategies to overcome some of the aforementioned limitations. For instance, model-specific CF explanation generation methods for DTs can be transformed into model-agnostic ones if, for example, a DT is trained on some data around the test instance generated synthetically [9].

The explanation generation methods proposed above rely on the decision paths or rules retrieved from the internals of the classifier. However, a large number of state-of-the-art CF explanation generation algorithms output explanations referring to single CF data points that are minimally different from the test instance under consideration. To perform a comparative human evaluation of the algorithms, we additionally propose a genetic algorithm for CF explanation generation (GEN) that outputs single-point CF explanations (we also refer to it as quantitative) and perform a user study to estimate the utility of both formats of CF explanations. GEN is designed to solve an optimisation problem operating on available numerical feature values and includes the following steps: initial population, fitness function calculation, binary tournament selection, crossover, mutation, and elitist selection. Verbalisation of CF explanations generated using the GEN algorithm follows the same template-based method that was defined above for rule-based CF explanations.

4.3 Making explanations social

The methods discussed in Section 4.2 allow for generating textual one-shot CF selected explanations for interpretable rule-based classification systems. Whereas the resulting individual responses to, e.g., why- or why-not questions can be useful for explaining single predictions of AI-based classifiers, such explanations are void of the social aspect. Indeed, automated (factual or CF) explanations may be questioned by the end user. A lack of the communication channel that the user may want to use to provide his or her feedback concerning the quality of an automatically generated explanation may lead to mistrust in the given prediction, even if the prediction is correct. Conversely, interaction with the explainer in natural language may facilitate the user's willingness to make an informed decision with respect to the system's prediction.

Establishing a communication channel for user-system interaction appears to bring indispensable benefits to the end user. First, it makes it possible to structure and formalise the explanation communication process. Second, the user becomes able to clarify any components of the offered explanation. Third, the user becomes able to examine the entire explanation space. Let us now go into detail with the advantages of the proposed communication channel.

To ensure fair and transparent interaction between the explainer and the user, it is hypothesised that user-explainer communication should rely on a predefined protocol where dialogue states and transitions between them are defined unambiguously. Taking into consideration the known structure of automated explanations, it is essential to establish rules that facilitate understanding of the classifier's reasoning iteratively.

Given explanation templates defined in Section 4.2, every explanation generated by means of the proposed explanation generation methods contains three variable components: [CLASS], [FEATURE], and [VALUE]. The examples above presuppose that a predicting attribute (i.e., an explanation feature) can be defined both numerically and linguistically. Hence, an effective explanation is claimed to include the possibility to question all variable components of the templates. In this manner, the end user becomes able to retrieve not only explanations but also fully understand their nature. Assuming that the feature space is given and interpretable, the variable [FEATURE] can relate to any feature retrieved from the dataset (e.g., *Height*) whereas the variable [VALUE] can be presented in two formats: textual (e.g., *tall*) or numerical (either single, e.g. *185cm*, or interval-based, as in $180cm \leq Height \leq 200cm$).

In the case of factual explanations, the variable [CLASS] refers to the class predicted by the classifier. The user may want to question it by asking a why-not question, e.g. "Why is the test instance not [CLASS$'$]?". Then, the user is assumed to start exploring the set of CF explanations available for the given prediction. In the case of CF explanations, the variable [CLASS] refers to the desired CF class. It can then be questioned by the end user if she submits the same why-not question to the system referring to another CF class, e.g. "Why is the test instance not [CLASS$''$]?" if there exist multiple CF classes. The user may want to question the value of the variable [FEATURE] by asking for a definition of the given feature (e.g., "What is *Height*?"). Questioning the value of the variable [VALUE] can be modelled by providing a shift from the textual modality to the numerical modality of the given feature value ("In what range is *tall* defined?"). Enabling the user to question any of the explanation components can be beneficiary for both expert and lay users. In the former case, the user is able to request only the most necessary details about the explanation under consideration. In the latter case, the user can iteratively request sufficient information to make an informed decision with respect to the system's prediction, especially when she has little domain knowledge (if any).

Importantly, there may exist multiple factual explanations for the given prediction as well as numerous CF explanations for any of CF classes. Further, the explanation that the explainer finds to be the most relevant for the given (factual or CF) class may not coincide with user expectations or preferences, leading to decreased utility of such an explanation. It is therefore of paramount importance to enable the end user to inquire alternative explanations so that the user can prune the explanation space until she is fully satisfied with the information accumulated along the explanatory dialogue.

To ensure that the desired aforementioned advantages of user-system interaction are achieved fully, argumentation theory methods can be used to model a communication channel between the user and the explainer. In this thesis, we design explanatory user-system dialogue applying the so-called "dialogue game" approach to argumentation [22]. This mechanism allows us to (1) formalise and integrate request types described below, (2) enable the end user to iteratively explore the explanation space by arguing over the explanations offered previously, (3) personalise explanations giving to the end user full freedom to request only necessary and sufficient information about the dataset, prediction, or explanation components.

The corresponding dialogue protocol establishes a typology of user's requests, explainer's responses, and transitions between the dialogue states. Thus, the set of the proposed dialogue requests includes the following categories:

- **Factual and CF explanation requests**. These include why and why-not questions to the explainer for factual and CF classes, respectively.

- **Detailisation requests**. These tackle the switch from purely linguistic values of specific features that make part of the given piece of explanation to their numerical counterparts.

- **Clarification requests**. These requests are meant to question definitions of specific features that make part of the given piece of explanation.

- **Alternative explanation requests**. These requests are designed to enable the user to explore the explanation space. Notably, they are made unavailable for factual explanations in the case of DTs, as alternative decision paths could be erroneous with respect to the actual prediction and do not adequately explain the classifier's reasoning for the given test instance.

The proposed formal model of explanatory dialogue has been implemented in form of a task-oriented dialogue system[1]. The dialogue system's main tasks are to (1) communicate to the end user explanations generated automatically by an explainer and (2) handle follow-up user requests concerning explanation-related details. Adapting a classic pipeline for task-oriented

[1] The source code is made publicly available at `https://gitlab.citius.usc.es/ilia.stepin/fcfexpgen`, branches "dialgame" and "dialgame_nlu".

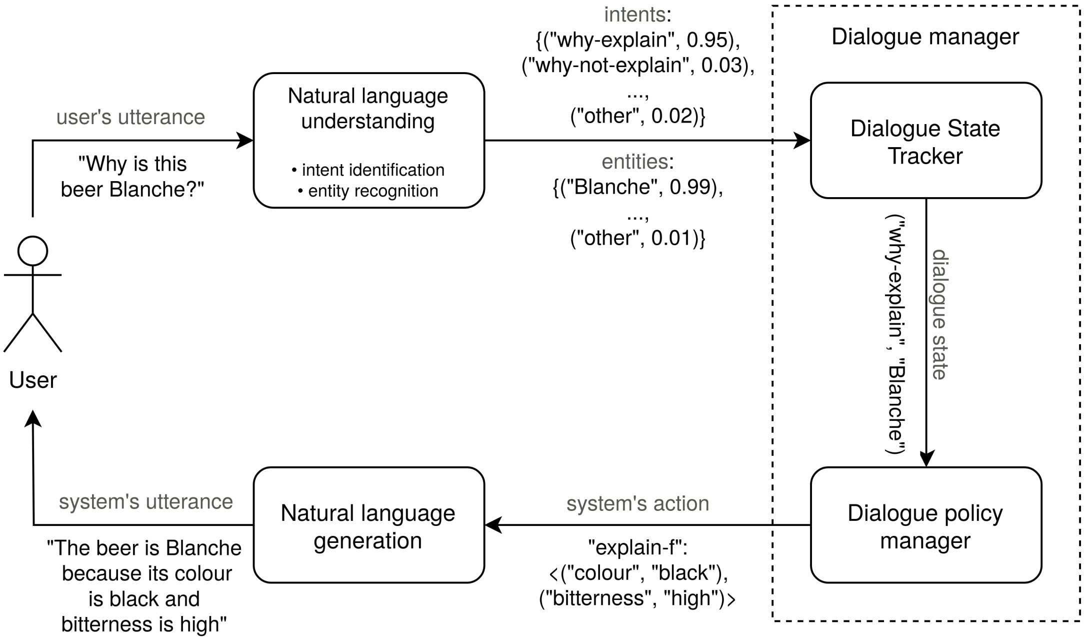


Figure 4.2: The dialogue system pipeline making use of the proposed argumentative explanatory model.

spoken dialogue systems [44] to (textual) explanatory dialogue modelling, we designed an argumentative conversational agent that handles user requests sequentially in accordance with the pipeline of dialogue system components that comprises the following four components:

- **The natural language understanding (NLU) module**. It serves two purposes: (1) to identify user's intent and (2) recognise all entities that the request has (if any).
- **The dialogue state tracker (DST)**. It defines the state that the dialogue is currently in based on the user's intent recognised by the NLU module.
- **The dialogue policy manager (DPM)**. It selects the most appropriate response among those available at the given dialogue state passed by DST.
- **The NLG module**. It generates well-formed, grammatical system's responses based on the information received from DPM.

Altogether, DST and DPM are said to constitute the dialogue manager. The argumentative dialogue protocol serves as the basis of the dialogue manager, as it tracks the state of the dialogue and defines all possible transitions among dialogue states.

Fig. 4.2 illustrates the pipeline of the components of the implemented conversational agent. Let us consider a beer style classification problem to illustrate request processing. Given a pre-trained beer style classifier, the user passes the characteristics of a specific beer (e.g., colour, bitterness, and strength) to the classification system and obtains the classifier's prediction (e.g., "This beer is Blanche"). Then, the user seeks an explanation for the given prediction and submits the corresponding request to the dialogue system (e.g., "Why is this beer Blanche?"). At first, the NLU module analyses the user's request to (1) identify what category it belongs to and (2) recognise all entities that the request contains. In our example, the NLU module first attempts to solve the user intent classification problem where it assigns probabilities to each category of user requests, e.g., ("*why-explain*", 0.95). The same procedure is applied to entity recognition. The intent and entity rankings are passed on to DST. Then, DST selects the most probable intent (in this case, "*why-explain*") and switches the dialogue to the corresponding state (in our example, that of a factual explanation request). Given the dialogue state, DPM seeks the most adequate response among the possible options. In accordance with the dialogue protocol, the system is allowed to respond to a factual explanation request (i.e., "*why-explain*") by either offering a factual explanation to the end user (i.e., "*explain-f*") if it is able to generate it or refusing to offer it, otherwise (i.e., "*no-explain-f*"). In our example, the system finds that the decision path responsible for the given prediction can be summarised in terms of two features (i.e., *colour* and *bitterness*) with the corresponding values (i.e., *black* and *high*, respectively). The explanatory feature-value pairs are then passed on to the NLG module, which generates a well-formed, grammatical utterance. Once the NLG module returns the system's utterance (in this example, "The beer is Blanche because its colour is black and bitterness is high."), it is presented to the end user.

Similarly to the factual explanation request from the example above, the explanatory dialogue system follows the aforementioned pipeline to handle all other types of user requests defined in the argumentative dialogue protocol (i.e., CF explanation, detailisation, clarification, and alternative explanation requests). In each case, DPM makes calls to external modules if necessary (e.g., the explainer when generating explanations or the knowledge base containing domain knowledge when processing clarification requests).

The proposed explanatory dialogue protocol provides a transparent means of explanation communication in information-seeking settings. Being transparent, the proposed model can be aligned with regulatory requirements to automatic explanation generation. Furthermore, it is shown to take into account user preferences, as it favours diversity of the output explanations and allows the user to further question all variable components of such explanations. The proposed argumentative dialogue model can be potentially used as a tool for assessing effectiveness of CF explanations generated by other rule-based CF explanation generation algorithms. To facilitate adaptation to different rule-based CF explainers, the proposed dialogue model is further conceptualised using the formalism of dialogue grammars.

Notably, the proposed dialogue protocol also allows us to quantitatively estimate the necessity in diversity of qualitative CF explanations. Recall that the proposed dialogue protocol represents the explanation space for each (possibly, CF) class as a list of explanations ranked by relevance, as measured by the explainer. Given a corpus of explanatory dialogues collected, we can compare the explainer-measured relevance to the demand in the given explanations based on the data collected from actual users. Statistics of alternative explanation requests for each class can be useful for this purpose. On the one hand, a large number of alternative explanations asked for may be a signal of little utility of the initially offered explanations. On the other hand, the empirical data from end users may not adequately reflect their satisfaction with the explanation space in its entirety, as end users are at all times exposed to only a part of the explanation space unless they sequentially request all the explanations that the system can offer to them. In this regard, a metric of similarity between the set of explanations (at least, potentially) generated and that of actually requested may be useful for assessing automatically the user satisfaction with the explanation space, in general.

Finally, the concept of diversity of CF explanations regarded in terms of a set of single alternative explanations allows us to question the nature of the mere definition of a CF explanation. Indeed, automated CFs generated following the conventional definition search minimally different feature-value pairs that ensure a different classification. However, if there is a strong tendency to disregard minimally different CF data points that form the basis of a CF explanation and it is shown to be consistent for different explainers and audiences, it may be timely to reconsider the definition of a CF explanation or empower it with the property of human-centricity that goes beyond existing automatically computable metrics. Whereas the piece of work presented in this thesis only makes the first step in this direction, it can serve an inspiring source of ideas for elaboration on the user-centric prospects of automated CF explanations.

5 Contributions

The work on the present thesis resulted in (1) several pieces of software developed to reach the thesis objectives and (2) various publications that emerged as a result of the studies carried out during the doctoral project. Section 5.1 details the software produced as part of the thesis. Section 5.2 lists all the publications that this thesis bases upon.

5.1 Software

The following pieces of software have been developed in order to achieve the thesis objectives:

C1: FCFExpGen[1] – a framework for factual and CF explanation generation for interpretable rule-based classification systems. The proposed framework includes the following three algorithms:

- XOR: a model-specific algorithm that selects candidate CF rules and ranks them by relevance to the test instance using the eXclusive-OR (XOR) function. The rule claimed to be the most relevant represents a set of CF data points that are minimally different from the test instance in terms of their features. This CF set forms the basis of the output CF explanation;
- EUC: a model-specific variant of the XOR algorithm that utilises Euclidean distance as a metric of relevance of the candidate CF rule to the test instance (i.e., it measures proximity of the vector representation of the CF rule to the test instance in the selected *n*-dimensional space);
- GEN: a model-agnostic genetic algorithm that solves an optimisation problem looking for the closest single CF data point w.r.t. the test instance under consideration.

C2: DialGame[2] – an argumentative conversational agent for communication of automatically generated textual rule-based factual and CF explanations. Noteworthy, the dialogue man-

[1] `https://gitlab.citius.usc.es/ilia.stepin/fcfexpgen` (branch “xor_euc_gen”)

[2] `https://gitlab.citius.usc.es/ilia.stepin/fcfexpgen` (branches “dialgame” and “dialgame_nlu”)

ager (i.e., an integrative part of the conversational agent) implements a dialogue protocol basing on the argumentation theory-based technique referred to as "dialogue game" [30].

C3: SurveyGenerator[3] – a web-tool for carrying out human evaluation experiments designed for assessing various explanation aspects as well as the quality of explanation communication. Examples of the experiments carried out include:

- *Survey GM*: the explanation evaluation survey that enables the end-user to rate a series of distinct explanations for the same test instance in terms of informativeness, trustworthiness, accuracy, relevance, and readability;
- *Survey TS*: a simplified version of *Survey GM* which welcomes the end-user to rate a single explanation for the given test instance in terms of trustworthiness and satisfaction.

It is worth noting that all the source code, the data used in the human evaluation experiments and the corresponding experimental results are made publicly available and can be reached at a public Gitlab repository.

5.2 Publications

The work on the present doctoral thesis has resulted in three journal papers, three papers presented at international conferences and included in conference proceedings (both main and other tracks), and one book chapter. Namely, the following journal publications cover the algorithms developed and evaluated within the doctoral project [36, 37, 38]:

- Ilia Stepin, Jose M. Alonso, Alejandro Catala, Martín Pereira-Fariña. "A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence". *IEEE Access*, vol. 9, pp. 11974–12001, 2021. DOI: 10.1109/ACCESS.2021.3051315;

- Ilia Stepin, Jose M. Alonso-Moral, Alejandro Catala, Martín Pereira-Fariña. "An empirical study on how humans appreciate automated counterfactual explanations which embrace imprecise information". *Information Sciences*, vol. 618, pp. 379–399, 2022. DOI: 10.1016/j.ins.2022.10.098;

- Ilia Stepin, Katarzyna Budzynska, Alejandro Catala, Martín Pereira-Fariña, Jose M. Alonso-Moral. "Information-seeking dialogue for explainable artificial intelligence: Modelling and analytics". *Argument and Computation*, in press. DOI: 10.3233/AAC-220011.

[3] `https://gitlab.citius.usc.es/jose.alonso/surveygenerator`

Further, the work in progress has been presented at several international conferences (including main tracks, workshops, and doctoral consortia), which resulted in the following publications [34, 35, 39]:

- Ilia Stepin, Alejandro Catala, Jose M. Alonso, Martín Pereira-Fariña. "Paving the way towards counterfactual generation in argumentative conversational agents". In Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI) collated with the Conference on International Natural Language Generation (INLG), pp. 20-25, Tokyo (Japan), 2019. DOI: 10.18653/v1/W19-8405;

- Ilia Stepin, Jose M. Alonso, Alejandro Catala, M. Pereira-Fariña. "Generation and evaluation of factual and counterfactual explanations for decision trees and fuzzy rule-based classifiers". In Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow (UK), 2020. DOI: 10.1109/FUZZ48607.2020.9177629;

- Ilia Stepin. "Argumentation-based interactive factual and counterfactual explanation generation". In Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pp. 61-62, Santiago de Compostela (Spain), 2020.

Notably, further experiments concerning specific technicalities of the XOR algorithm for automated factual and CF explanation generation can be found in the following book chapter [40]:

- Ilia Stepin, Alejandro Catala, Martín Pereira-Fariña, Jose M. Alonso. "Factual and counterfactual explanation of fuzzy information granules". In: Pedrycz, W., Chen, SM. (eds) Interpretable Artificial Intelligence: A Perspective of Granular Computing. Studies in Computational Intelligence, vol 937. Springer, Cham. DOI: 10.1007/978-3-030-64949-4_6

In addition, a proof of concept of the argumentative framework for factual and CF explanation communication was presented orally at the European Conference on Argumentation in Rome (Italy) on September 30, 2022. The full paper is currently under review, pending to be finally published by College Publications in their "Studies in Logic and Argumentation" book series[4] in 2023. Besides, one of the model-specific interpretable fuzzy rule-based explanation generation algorithms (i.e., EUC) is presented in an immersive article entitled "How to build self-explaining fuzzy systems: From interpretability to explainability" and submitted to the special issue "Artificial Intelligence eXplained" (AI-X) of the IEEE Computational Intelligence Magazine. At the moment of writing, the manuscript is undergoing a second review round.

[4] https://www.collegepublications.co.uk/logic/sla/

Fig. 5.1 visualises the main contributions (i.e., software and journal papers) listed in Sections 5.1-5.2, respectively.

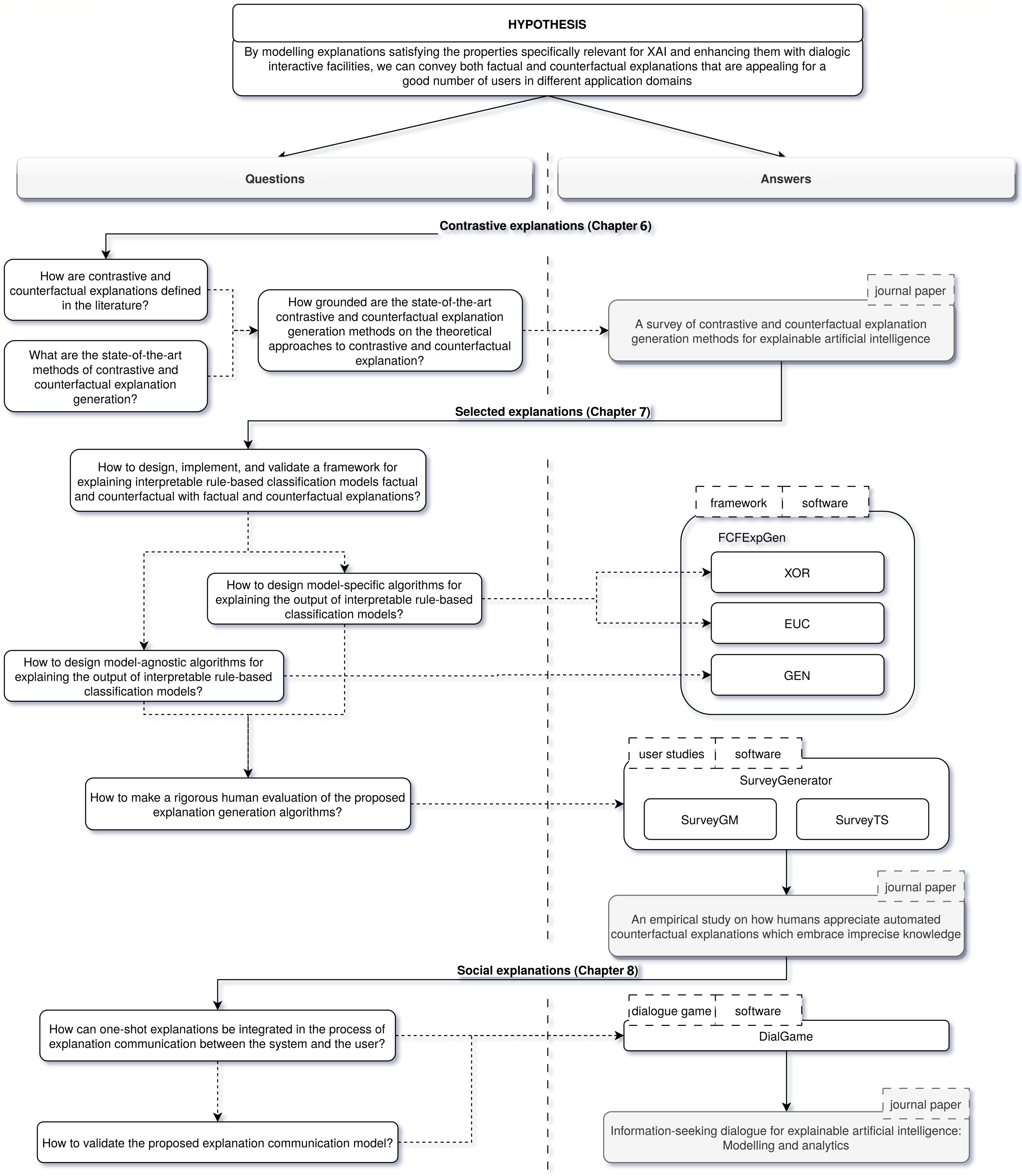

HYPOTHESIS
By modelling explanations satisfying the properties specifically relevant for XAI and enhancing them with dialogic interactive facilities, we can convey both factual and counterfactual explanations that are appealing for a good number of users in different application domains
Questions
Answers
Contrastive explanations (Chapter 6)
How are contrastive and counterfactual explanations defined in the literature?
What are the state-of-the-art methods of contrastive and counterfactual explanation generation?
How grounded are the state-of-the-art contrastive and counterfactual explanation generation methods on the theoretical approaches to contrastive and counterfactual explanation?
journal paper
A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence
Selected explanations (Chapter 7)
How to design, implement, and validate a framework for explaining interpretable rule-based classification models factual and counterfactual with factual and counterfactual explanations?
How to design model-specific algorithms for explaining the output of interpretable rule-based classification models?
How to design model-agnostic algorithms for explaining the output of interpretable rule-based classification models?
How to make a rigorous human evaluation of the proposed explanation generation algorithms?
framework
software
FCFExpGen
XOR
EUC
GEN
user studies
software
SurveyGenerator
SurveyGM
SurveyTS
journal paper
An empirical study on how humans appreciate automated counterfactual explanations which embrace imprecise knowledge
Social explanations (Chapter 8)
How can one-shot explanations be integrated in the process of explanation communication between the system and the user?
How to validate the proposed explanation communication model?
dialogue game
software
DialGame
journal paper
Information-seeking dialogue for explainable artificial intelligence: Modelling and analytics

Figure 5.1: Main research questions and answers (regarding techniques, software and publications).

6 State-of-the-art methods of contrastive and counterfactual explanation generation

As discussed in Chapter 4, the problem of contrastive and CF explanation generation is among the most trending XAI topics [21]. The recent rise of attention to contrastive and CF explanation encourages a discussion about the nature of these concepts and their (mis-)use in XAI. Whereas XAI is a newly emerged research field, the notion of explanation, in general, and its contrastive and CF sub-types, in particular, have long been discussed in social sciences. It is therefore important to know how theoretical concepts related to contrastive and CF explanation can guide XAI researchers in developing effective explanation generation methods.

In this chapter, we (1) inspect theoretical foundations of the concepts of contrastive and CF explanation, (2) explore technical aspects of the state-of-the-art computational frameworks of generation of contrastive and/or CF explanations, and (3) discuss how theoretically grounded the computational frameworks are. We show that a majority of state-of-the-art contrastive and CF explanation generation frameworks only loosely (if at all) rely on theoretical models of explanation from social sciences. Instead, they mainly represent data-driven solutions to optimisation problems oftentimes neglecting a wide body of knowledge about the nature of explanation accumulated over the centuries. As a consequence, it sometimes leads to terminological confusion and makes us strive for standardisation of the explanation-related notions used across sub-fields of XAI.

The results from this chapter are published in the following paper [36]:

Ilia Stepin[a], Jose M. Alonso[a], Alejandro Catala[a], Martín Pereira-Fariña[b]. “A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence”. In *IEEE Access* (Open Access), vol. 9, pp. 11974-12001, ISSN: 2169-3536, **2021**. DOI: 10.1109/ACCESS.2021.3051315

[a] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Rúa de Jenaro de la Fuente Domínguez, s/n, 15782 Santiago de Compostela, Spain

[b] Departamento de Filosofía e Antropoloxía, Universidade de Santiago de Compostela, Plaza de Mazarelos, s/n, 15705 Santiago de Compostela, Spain

Scientific production indicators:

IEEE Access, the journal where Chapter 6 was published, had, in the year of publication, a CiteScore index of 6.7 (calculated by Scopus on 05 May, 2022) and an impact factor of 3.476 (2021 Journal Citation Reports). In addition, it had the following positions in the categories listed below:

- Scopus: Q1 (rank #28/300) in General Engineering (the 90th percentile), Q1 (rank #34/231) in General Computer Science (the 85th percentile), Q1 (rank #124/708) in Electrical and Electronic Engineering (the 82th percentile), Q1 (rank #104/455) (the 77th percentile) in General Materials Science;

- JCR: Q2 (rank #79/164) in Computer Science and Information Systems, Q2 (rank #43/93) in Telecommunications, Q2 (rank #105/276) in Electrical & Electronic Engineering.

As of 29 June 2023, the publication has, according to distinct citation databases, the following number of citations: 199 (GoogleScholar), 94 (Scopus), 65 (WoS).

Personal authorship statement:

In accordance with the Contributor Roles Taxonomy (CRediT), the personal authorship contribution comprises the following roles: investigation, data curation, writing - original draft, visualization.

Publishing rights:

The journal paper where the results of this Chapter are published is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. The individual or entity exercising the licensed rights (hereinafter, “the user”) is free to copy and redistribute the material in any medium or format under the following terms[1]:

- **Attribution:** The user must give appropriate credit, provide a link to the license, and indicate if changes were made. The user may do so in any reasonable manner, but not in any way that suggests the licensor endorses the user or the user’s use.

- **NonCommercial:** The user may not use the material for commercial purposes.

- **NoDerivatives:** If the user remixes, transforms, or builds upon the material, the user may not distribute the modified material.

[1] For more information, see `https://creativecommons.org/licenses/by-nc-nd/4.0/`

7 Counterfactual explanation generation for rule-based classification systems

As follows from Chapter 6, a variety of CF explanation generation algorithms explain the output of ML classifiers. However, they mainly focus on explaining “black-box” classification models leaving the explanatory potential of interpretable models largely unexplored [19]. Whereas some interpretable rule-based models can be trivially explained factually, CF explanation generation methods for such classifiers remain scarce due to a number of open challenges. In this chapter, we propose two algorithms of a model-specific CF explanation generation method for explaining the output of interpretable rule-based classification systems. We show how both algorithms can be applied to interpretable rule-based classifiers. Operating on the internals of the classifier, they generate rule-based CFs in natural language using imprecise knowledge that is codified in the corresponding rules. In addition, we propose a genetic model-agnostic CF explanation algorithm that outputs textual explanations that refer to a single CF data point that is minimally different from the test instance under consideration.

Human evaluation remains one of the greatest challenges for CF explanation generation [43]. Despite being costly and difficult to design, human evaluation experiments are indispensable to ensure the utility of automated CFs for their end users. In this chapter, we evaluate the proposed CF explanation generation methods, as we perform two human evaluation studies. Further, we propose a metric of perceived explanation complexity for textual rule-based explanations. We show that it correlates with several explanation aspects (such as informativeness, relevance, and readability) for the selected target audience (in this case, expert users that have a high degree of expertise in XAI or related fields). Consequently, the proposed metric can be used in future studies where it can help to reduce (some of) human evaluation costs.

The results from this chapter are published in the following paper [37]:

Ilia Stepin[a,b], Jose M. Alonso-Moral[a,b], Alejandro Catala[a,b], Martín Pereira-Fariña[c]. “An empirical study on how humans appreciate automated counterfactual explanations which embrace imprecise information”. In *Information Sciences* (Open Access), vol. 618, pp. 379-399. ISSN: 0020-0255. Elsevier, **2022**. DOI: 10.1016/j.ins.2022.10.098

[a] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Rúa de Jenaro de la Fuente Domínguez, s/n, 15782 Santiago de Compostela, Spain.
[b] Departamento de Electrónica e Computación, Universidade de Santiago de Compostela, Rúa Lope Gómez de Marzoa, s/n, 15782 Santiago de Compostela, Spain.
[c] Departamento de Filosofía e Antropoloxía, Universidade de Santiago de Compostela, Plaza de Mazarelos, s/n, 15705 Santiago de Compostela, Spain.

Scientific production indicators:

Information Sciences, the journal where Chapter 7 was published, had, in the year of publication (i.e., 2022), a CiteScore index of 13.4 (calculated by Scopus on 05 May, 2023) and an impact factor of 8.1 (2022 Journal Citation Reports). In addition, it had the following positions in the categories listed below:

- Scopus: Q1 (rank #35/792) in Computer Science Applications (the 95th percentile), Q1 (rank #8/127) in Theoretical Computer Science (the 94th percentile), Q1 (rank #17/286) in Control and Systems Engineering (the 94th percentile), Q1 (rank #9/140) in Information Systems and Management (the 93rd percentile), Q1 (rank #32/404) in Software (the 92nd percentile), and Q1 (rank #26/301) in Artificial Intelligence (the 91st percentile);

- JCR: Q1 (rank #13/158) in Computer Science and Information Systems.

As of 29 June 2023, the publication has, according to distinct citation databases, the following number of citations: 4 (GoogleScholar), 1 (Scopus), 0 (WoS).

Personal authorship statement:

In accordance with the Contributor Roles Taxonomy (CRediT), the personal authorship contribution comprises the following roles: methodology, software, validation, investigation, data curation, writing - original draft, visualization.

Publishing rights:

The journal paper where the results of this Chapter are published is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. The individual or entity exercising the licensed rights (hereinafter, "the user") is free to copy and redistribute the material in any medium or format under the following terms[1]:

[1] For more information, see `https://creativecommons.org/licenses/by-nc-nd/4.0/`

- **Attribution:** The user must give appropriate credit, provide a link to the license, and indicate if changes were made. The user may do so in any reasonable manner, but not in any way that suggests the licensor endorses the user or the user's use.
- **NonCommercial:** The user may not use the material for commercial purposes.
- **NoDerivatives:** If the user remixes, transforms, or builds upon the material, the user may not distribute the modified material.

8 Argumentative explanation communication for rule-based classification systems

The CF explanation generation algorithms proposed in Chapter 7 offer one-shot textual CFs. Whereas such explanations can be generated for a variety of CF classes, they are limited to providing only static information about the underlying reasoning of the classifier. Being void of any interaction with the end user, such explanations run the risk of losing their utility (possibly, partly), as one-shot explanations may ignore the user's needs and/or preferences.

In this chapter, we propose to address this issue by establishing an argumentative communication channel between the explainer (or, more generally, the system) and the user. To do so, we rely on the formalism of so-called "dialogue games" [22]. Based on this well-known argumentation mechanism, our explanatory dialogue model allows for generating interactive explanations. The end user is allowed to request necessary additional information while ensuring the balance between the system's capacity to offer specific details about the prediction and user's needs to make an informed decision with respect to that prediction.

The proposed argumentative dialogue model has been implemented in form of an argumentative conversational agent. This allows for transparent human evaluation of the formal explanatory dialogue model. Hence, we carried out an experiment to estimate how useful the proposed model is in the information-seeking explanatory dialogue settings. The results show that users actively use all types of the modelled requests to make a decision (i.e., acceptance or rejection) with respect to the classifier's prediction. Further, a high number of requests for alternative CFs demonstrates the need to enable the user to explore the explanation space by arguing over the offered pieces of information.

The results from this chapter are published in the following paper [38]:

Ilia Stepin[a,c], Katarzyna Budzynska[b], Alejandro Catala[a,c], Martín Pereira-Fariña[d], Jose M. Alonso-Moral[a,c]. "Information-seeking dialogue for explainable artificial intelligence: Modelling and analytics". In *Argument and Computation* (Open Access), ISSN print: 1946-2166; ISSN online: 1946-2174. IOS Press, in press. DOI: 10.3233/AAC-220011

[a] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Rúa de Jenaro de la Fuente Domínguez, s/n, 15782 Santiago de Compostela, A Coruña, Spain
[b] Laboratory of The New Ethos, Warsaw University of Technology, plac Politechniki 1, 00-661, Warsaw, Poland
[c] Departamento de Electrónica e Computación, Universidade de Santiago de Compostela, Rúa Lope Gómez de Marzoa, s/n, 15782 Santiago de Compostela, A Coruña, Spain
[d] Departamento de Filosofía e Antropoloxía, Universidade de Santiago de Compostela, Plaza de Mazarelos s/n, 15705 Santiago de Compostela, A Coruña, Spain

Scientific production indicators:

At the moment of writing, the scientific production indicators for the year of publication (i.e., 2023) are unavailable for *Argument and Computation*, the journal where Chapter 8 was published. In the year immediately preceding that of publication (i.e., 2022), the journal had a CiteScore index of 3.6 (calculated by Scopus on 05 May, 2023) and an impact factor of 1.4 (2022 Journal Citation Reports). In addition, it had the following positions in the categories listed below:

- Scopus: Q1 (rank #95/1078) in Linguistics and Language (the 91st percentile), Q2 (rank #57/172) in Computational Mathematics (the 67th percentile), Q2 (rank #366/792) in Computer Science Applications (the 53rd percentile), Q3 (rank #159/301) in Artificial Intelligence (the 47th percentile);

- JCR: Q4 (rank #155/192) in Computer Science and Artificial Intelligence.

As of 29 June 2023, the publication has not been cited yet.

Personal authorship statement:

In accordance with the Contributor Roles Taxonomy (CRediT), the personal authorship contribution comprises the following roles: methodology, software, validation, investigation, data curation, writing - original draft, visualization.

Publishing rights:

The journal paper where the results of this Chapter are published is licensed under a Creative Commons Attribution-NonCommercial 4.0 License. The individual or entity exercising the licensed rights (hereinafter, "the user") is free to copy and redistribute the material in any medium or format under the following terms[1]:

[1] For more information, see `https://creativecommons.org/licenses/by-nc/4.0/`

- **Attribution:** The user must give appropriate credit, provide a link to the license, and indicate if changes were made. The user may do so in any reasonable manner, but not in any way that suggests the licensor endorses the user or the user's use.
- **NonCommercial:** The user may not use the material for commercial purposes.

9 Conclusion

In this thesis, we addressed the problem of explanation generation and communication for interpretable rule-based systems. More specifically, we focused on the task of generation of interactive CF explanations, which meet theoretically grounded requirements to quality explanations for XAI. To address this challenge, we first performed a literature review of theoretical foundations of the family of contrastive and CF explanations and the state-of-the-art methods of their automatic generation, which resulted in a two-level taxonomy of contrastive and CF explanations. Taking into consideration the insights from the review, we then designed, implemented, and validated a computational framework for generating factual and CF explanations associated to interpretable rule-based classification systems. It includes one model-agnostic and two model-specific algorithms, all of which offer human-comprehensive explanations in natural language. Further, we enhanced the framework with an argumentative dialogue generation module, which allows for interactive explanations in agreement with end user's needs. All in all, the generated explanations have been shown to be contrastive, selected, and social.

In what follows, we summarise main lessons learned from the work carried out during this doctoral project. Section 9.1 encapsulates our concluding remarks for each piece of research reported in Chapters 6-8. Section 9.2 outlines prospective directions for future work.

9.1 Concluding remarks

The literature review on contrastive and CF explanations revealed several gaps in the inspected sub-field of XAI. First, the state-of-the-art computational frameworks of contrastive and CF explanation generation are scarcely grounded on explanation theories from social sciences. This is, in part, due to the fact that the existing theories and computational frameworks mainly address distinct aspects of explanation generation. Thus, theories of contrastive explanation often discuss products of the explanatory process in terms of cause-and-effect relationships whereas a large number of computational contrastive and/or CF explanation generation methods focus on non-causal (e.g., spacial) relations between the test point whose prediction is to be explained and potential CF data points. Second, it turns out that the terms "contrastive" and "counterfactual" are often used interchangeably in the XAI community despite certain methodological differences. This observation calls for standartisation of the terminology used in the field. In order to unify the terminology (where applicable), we suggest that the term “contfactual” be promoted for contrastive-CF explanations. We believe that this term adequately encompasses the aspect of contrastiveness in CF explanations and vice versa. Third, the state-of-the-art computational frameworks have been observed to greatly lack human evaluation support. In fact, a vast majority of contrastive and CF explanation generation methods have only been evaluated using data-driven automatically computed metrics. Nevertheless, human evaluation studies are indispensable for shifting towards human-centric AI despite being expensive and difficult to design.

In order to reduce the gap between the automatic data-driven and human evaluation-based metrics, we proposed the metric of perceived explanation complexity (PEC), i.e. a measure of how complex the given piece of textual explanation seems to be for the end user to process it. For the target audience (in this case, users who have a high degree of expertise in XAI or related fields), human evaluation experiments have shown that the computed PEC scores correlate with informativeness, relevance, and readability of automated generations. Thus, the proposed metric can effectively replace human evaluation experiments measuring the aforementioned explanation aspects for domain experts or highly qualified specialists. Overall, the quality of the explanations generated by all the proposed algorithms was positively evaluated in the human evaluation studies that we carried out in this thesis. In terms of all the assessed explanation aspects (i.e. informativeness, trustworthiness, accuracy, relevance, readability, and satisfaction), the resulting scores were above average for all the proposed CF explanation generation methods. However, none of these methods has been found to consistently outperform the others in all the explanation aspects. In this regard, we conclude that the methods modelling imprecise knowledge (e.g., XOR and EUC) and those making use of precise feature values (e.g., GEN) are best used complementarily to each other to satisfy the needs of a wider audience of users.

Whereas such complementary information can be aggregated in a single piece of text, the social aspect of explanation remains unaddressed in case of one-shot explanations. To overcome this issue, we proposed an argumentative dialogue protocol to model information-seeking explanatory dialogues and developed a corresponding conversational agent. The human evaluation results of the dialogue protocol validation prove the necessity for all the proposed types of requests and responses for effective explanation communication. Further, a large number of requests for alternative CF explanations testify that the most relevant CF explanations from the algorithmic point of view may oftentimes not seem optimal from the user’s point of view. Whereas further comparative studies are necessary to analyse explanatory power of distinct explanation generation algorithms from the cognitive point of view, it can be concluded at this stage that the best-ranked CFs (i.e., most relevant or minimally different CFs from the test instance) may have to be combined with one or more alternatives given a set of multiple candidate CFs.

Striving for enabling the end user to play a decisive role in the process of explanation communication, we believe that it is indispensable to further emphasise the social aspect of automated explanations, e.g., by designing additional metrics of the quality of explanation communication similarly to the PEC score proposed in this thesis.

In light of the statements made above, we conclude that the explanation generation framework proposed in this thesis offers factual and CF explanations that turn out to be appealing to end users. Thus, textual explanations generated in both modalities (those modelling imprecise knowledge and those outputting specific numerical values or intervals thereof) received higher-than-average estimates from the target audiences in all the experiments carried out. In addition, end users appreciated their interaction with the argumentative conversational agent which was carefully designed to communicate automated factual and CF explanations.

9.2 Future work

The research results presented in this thesis indicate several directions for future work. From the theoretical point of view, the proposed framework can be extended to introduce causal relations between the predicted data and related features. In fact, this could further bridge the gap between theoretical and computational paradigms of contrastive and CF explanation generation and therefore appears highly desirable in light of the results obtained from the performed literature review. From the algorithmic point of view, the proposed explanation generation framework should be further extended with a surrogation approach to handle other types of classifiers (including those non-interpretable and non-rule-based). Nevertheless, in the absence of access to the internals of the classifier and/or the feature space, it may be essential to transform the proposed model-specific explanation generation algorithms into their model-agnostic equivalents. Further, different settings may require changes in the dialogue protocol that models the communication process between the explainer and the user for the sake of deeper customisation. In addition, further extension is required to guarantee properties of CF explanations not addressed in this thesis. For example, the presented algorithms do not allow to assess straightforwardly how actionable the generated CFs are. Hence, enhancing output CF explanations with other desired CF properties is believed to further increase effectiveness of such explanations.

Importantly, it seems impossible to achieve the state of human-centric AI without formalising and modelling ethical relations on the basis of the data being processed. Thus, bias mitigation, yet another highly relevant line of research in the XAI community, is another algorithmic challenge to address. Finally, it is of our particular interest to further adapt the designed human evaluation framework for future experiments on explanation, trustworthiness, and satisfaction. Altogether, the prospective extensions of the work presented in this thesis are believed to have great potential for moving forward from XAI to Trustworthy AI.

Ethical considerations

The experiments whose results are reported in the present thesis were designed to involve human evaluation. Therefore, a prior permission to carry them out had been obtained from the Ethics Committee of the University of Santiago de Compostela (a copy of the corresponding certificate is attached below).

All the information collected from the human evaluation study participants was in agreement with the European Union's General Data Protection Regulation (GDPR). Further, human evaluation was based solely on non-personal or anonymous data. In addition, all the participants gave informed consent confirming the following:

- the participant reached the age of majority;
- participation in the study was completely voluntary;
- participation in the study could be terminated at any time;
- participant's anonymous responses would be used for research purposes in accordance with the GDPR.

None of the experimental results could anyhow be used against the study participants. Risks of misuse of the collected results were minimal. In light of an increasing use of AI-based applications for automated text and/or image generation, it is important to mention that no piece of text or pictures from the present thesis or any of the published or accepted articles was generated using any generative AI-based applications (e.g., ChatGPT).

VICERREITORÍA DE INVESTIGACIÓN
E INNOVACIÓN
Oficina de Investigación e Tecnoloxía
Servizo de Convocatorias e Recursos Humanos de I+S
Edificio CACTUS – Campus Vida
15782 Santiago de Compostela
Tel. 981 547 040 - Fax 981 547 077
Correo electrónico: cittinfo@usc.es
http://imaisd.usc.es

JOSÉ MANUEL CIFUENTES MARTÍNEZ, PRESIDENTE DO COMITÉ DE BIOÉTICA DA UNIVERSIDADE DE SANTIAGO DE COMPOSTELA,

INFORMA:

Que o proxecto de investigación con rexistro **USC-23/2021** titulado **"Agentes conversacionales argumentativos para la inteligencia artificial explicable"**, do que é investigador responsable D. **José María Alonso Moral,** ten sido examinado por o Comité de Bioética desta Universidade, cumprindo o seu protocolo experimental os requisitos éticos esixidos.

Este documento non exime da obtención de permisos ou autorizacións e do cumprimento de outras normativas de aplicación.

Lugo, 17 de maio de 2021.

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
VICERREITORÍA DE INVESTIGACIÓN
OFICINA DE INVESTIGACIÓN E TECNOLOXÍA

Bibliography

[1] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, Jose M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera. Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, in press, 2023. DOI: 10.1016/j.inffus.2023.101805.

[2] J. M. Alonso, C. Castiello, L. Magdalena, and C. Mencar. *Explainable Fuzzy Systems - Paving the Way from Interpretable Fuzzy Systems to Explainable AI Systems*, volume 970. Springer International Publishing, 2021. DOI: 10.1007/978-3-030-71098-9.

[3] F. Bex and D. Walton. Combining explanation and argumentation in dialogue. *Argument & Computation*, 7(1):55–68, 2016. DOI: 10.3233/AAC-160001.

[4] R. M. J. Byrne. Counterfactual thought. *Annual Review of Psychology*, 67:135–157, 2016. DOI: 10.1146/annurev-psych-122414-033249.

[5] R. M. J. Byrne. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 6276–6282, 2019. DOI: 10.24963/ijcai.2019/876.

[6] V. Dignum. *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer, Cham, 2019. DOI: 10.1007/978-3-030-30371-6.

[7] A. Gatt and E. Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018. DOI: 10.1613/jair.5477.

[8] R. Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55, 2022. DOI: 10.1007/s10618-022-00831-6.

[9] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini. Factual and Counterfactual Explanations for Black Box Decision Making. *IEEE Intelligent Systems*, 34(6):14–23, 2019. DOI: 10.1109/MIS.2019.2957223.

[10] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, 51(5), 2018. DOI: 10.1145/3236009.

[11] D. Gunning, E. Vorm, J. Y. Wang, and M. Turek. DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters*, 2(4):e61, 2021. DOI: 10.1002/ail2.61.

[12] C. G. Hempel. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: The Free Press, 1965.

[13] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Metrics for explainable AI: Challenges and prospects. Technical report, DARPA Explainable AI Program, 2018. DOI: 10.48550/arXiv.1812.04608.

[14] J. Hühn and E. Hüllermeier. FURIA: an algorithm for unordered fuzzy rule induction. *Data Mining and Knowledge Discovery*, 19(3):293–319, 2009. DOI: 10.1007/s10618-009-0131-8.

[15] A. Jacovi, S. Swayamdipta, S. Ravfogel, Y. Elazar, Y. Choi, and Y. Goldberg. Contrastive explanations for model interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1597–1611. Association for Computational Linguistics, 2021. DOI: 10.18653/v1/2021.emnlp-main.120.

[16] M. T. Keane, E. M. Kenny, E. Delaney, and B. Smyth. If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, page 4466–4474, 2021. DOI: 10.24963/ijcai.2021/609.

[17] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman. Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology*, 51(1):7–15, 2009. DOI: 10.1016/j.infsof.2008.09.009.

[18] B. Kitchenham and S. Charters. Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report, 2007.

[19] P. Lertvittayakumjorn and F. Toni. Argumentative explanations for pattern-based text classifiers. *Argument & Computation*, 14(2):163–234, 2023. DOI: 10.3233/AAC-220004.

[20] P. Lipton. Contrastive explanation. *Royal Institute of Philosophy Supplement*, 27:247–266, 1990. DOI: 10.1017/S1358246100005130.

[21] N. Maslej, L. Fattorini, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, H. Ngo, J. C. Niebles, V. Parli, Y. Shoham, R. Wald, J. Clark, and R. Perrault. The AI Index 2023 Annual Report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2023. `https://aiindex.stanford.edu/report/`.

[22] P. McBurney and S. Parsons. Dialogue games for agent argumentation. *Argumentation in artificial intelligence*, pages 261–280, 2009.

[23] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. DOI: 10.1016/j.artint.2018.07.007.

[24] T. Miller. Contrastive explanation: a structural-model approach. *The Knowledge Engineering Review*, 36:e14, 2021. DOI: 10.1017/S0269888921000102.

[25] C. Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. Leanpub, 2 edition, 2022. `https://christophmolnar.com/books/interpretable-machine-learning/`.

[26] R. K. Mothilal, A. Sharma, and C. Tan. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT)*, page 607–617. Association for Computing Machinery, 2020. DOI: 10.1145/3351095.3372850.

[27] M. Ocaña, D. Chapela-Campa, P. Alvarez, N. Hernández, M. Mucientes, J. Fabra, Á. Llamazares, M. Lama, P. A. Revenga, A. Bugarín, M. García-Garrido, and Jose M. Alonso. Automatic linguistic reporting of customer activity patterns in open malls. *Multimedia Tools and Applications*, 81:3369–3395, 2021. DOI: 10.1007/s11042-021-11186-3.

[28] Official Journal of the European Union L119. Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA, 2016. `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3AOJ.L_.2016.119.01.0089.01.ENG&toc=OJ%3AL%3A201%3A119%3ATOC`.

[29] Parliament and Council of the European Union. Proposal for laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2021. `https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN`.

[30] H. Prakken. Coherence and flexibility in dialogue games for argumentation. *Journal of logic and computation*, 15(6):1009–1040, 2005. DOI: 10.1093/logcom/exi046.

[31] E. Reiter and R. Dale. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press, 2000. DOI: 10.1017/CBO9780511519857.

[32] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. DOI: 10.1038/s42256-019-0048-x.

[33] K. Sokol and P. Flach. One explanation does not fit all: The promise of interactive explanations for machine learning transparency. *KI - Künstliche Intelligenz*, 34:235–250, 2020. DOI: 10.1007/s13218-020-00637-y.

[34] I. Stepin. Argumentation-based interactive factual and counterfactual explanation generation. In *1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), Santiago de Compostela (Spain)*, pages 61–62, 2020.

[35] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Fariña. Generation and evaluation of factual and counterfactual explanations for decision trees and fuzzy rule-based classifiers. In *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2020. DOI: 10.1109/FUZZ48607.2020.9177629.

[36] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Fariña. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021. DOI: 10.1109/ACCESS.2021.3051315.

[37] I. Stepin, J. M. Alonso-Moral, A. Catala, and M. Pereira-Fariña. An empirical study on how humans appreciate automated counterfactual explanations which embrace imprecise information. *Information Sciences*, 618:379–399, 2022. DOI: 10.1016/j.ins.2022.10.098.

[38] I. Stepin, K. Budzynska, A. Catala, M. Pereira-Fariña, and J. M. Alonso-Moral. Information-seeking dialogue for explainable artificial intelligence: Modelling and analytics. *Argument and Computation*, in press. DOI: 10.3233/AAC-220011.

[39] I. Stepin, A. Catala, M. Pereira-Fariña, and J. M. Alonso. Paving the way towards counterfactual generation in argumentative conversational agents. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019) collated with the International Conference on Natural Language Generation (INLG)*, pages 20–25. Association for Computational Linguistics, 2019. DOI: 10.18653/v1/W19-8405.

[40] I. Stepin, A. Catala, M. Pereira-Fariña, and J. M. Alonso. Factual and counterfactual explanation of fuzzy information granules. In *Interpretable Artificial Intelligence: A Perspective of Granular Computing*, pages 153–185. Springer International Publishing, 2021. DOI: 10.1007/978-3-030-64949-4_6.

[41] J. van der Waa, E. Nieuwburg, A. Cremers, and M. Neerincx. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291:103404:1–19, 2021. DOI: 10.1016/j.artint.2020.103404.

[42] A. Vassiliades, N. Bassiliades, and T. Patkos. Argumentation and explainable artificial intelligence: A survey. *The Knowledge Engineering Review*, 36(e5), 2021. DOI: 10.1017/S0269888921000011.

[43] S. Verma, J. Dickerson, and K. Hines. Counterfactual explanations for machine learning: A review. In *Proceedings of the Machine Learning Retrospectives, Surveys & Meta-Analyses (ML-RSA) Workshop at the Conference on Neural Information Processing Systems (NeurIPS)*, 2020. DOI: 10.48550/arXiv.2010.10596.

[44] J. D. Williams, A. Raux, and M. Henderson. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33, 2016. DOI: 10.5087/dad.2016.301.

[45] C. Wohlin. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE)*, 2014. DOI: 10.1145/2601248.2601268.

[46] K. Čyras, A. Rago, E. Albini, P. Baroni, and F. Toni. Argumentative XAI: A Survey. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4392–4399, 2021. Survey Track. DOI: 10.24963/ijcai.2021/600.

List of Figures

List of Acronyms

AI	artificial intelligence
AIA	artificial intelligence act
CF	counterfactual
DPM	dialogue policy manager
DST	dialogue state tracker
DT	decision tree
FRBCS	fuzzy rule-based classification system
FURIA	fuzzy unordered rule induction algorithm
GDPR	general data protection regulation
ML	machine learning
NLU	natural language understanding
NLG	natural language generation
XAI	explainable artificial intelligence

APPENDIX. PUBLISHED OR ACCEPTED ARTICLES

This appendix contains three journal papers that form the basis of the present thesis. All of them are Open Access publications, making them publicly available to the interested reader. In this appendix, they appear in the following order:

- Ilia Stepin, Jose M. Alonso, Alejandro Catala, Martín Pereira-Fariña. “A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence”. In *IEEE Access*, vol. 9, pp. 11974-12001, ISSN: 2169-3536. IEEE Inc. (Open Access), **2021**. DOI: 10.1109/ACCESS.2021.3051315

- Ilia Stepin, Jose M. Alonso-Moral, Alejandro Catala, Martín Pereira-Fariña. “An empirical study on how humans appreciate automated counterfactual explanations which embrace imprecise information”. In *Information Sciences*, vol. 618, pp. 379-399. ISSN: 0020-0255. Elsevier (Open Access), **2022**. DOI: 10.1016/j.ins.2022.10.098

- Ilia Stepin, Katarzyna Budzynska, Alejandro Catala, Martín Pereira-Fariña, Jose M. Alonso-Moral. “Information-seeking dialogue for explainable artificial intelligence: Modelling and analytics”. In *Argument and Computation*, ISSN print: 1946-2166; ISSN online: 1946-2174, in press. IOS Press (Open Access). DOI: 10.3233/AAC-220011

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 16, 2020, accepted January 3, 2021, date of publication January 13, 2021, date of current version January 22, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3051315

A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence

ILIA STEPIN[1], JOSE M. ALONSO[1], (Member, IEEE), ALEJANDRO CATALA[1], AND MARTÍN PEREIRA-FARIÑA[2]

[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain
[2]Departamento de Filosofía e Antropoloxía, Universidade de Santiago de Compostela, 15705 Santiago de Compostela, Spain

Corresponding author: Ilia Stepin (ilia.stepin@usc.es)

This work was supported in part by the Spanish Ministry of Science, Innovation and Universities under Grant RTI2018-099646-B-I00 and Grant RED2018-102641-T, in part by the Galician Ministry of Education, University and Professional Training under Grant ED431F 2018/02, Grant ED431C 2018/29, Grant ED431G/08, and Grant ED431G2019/04; and in part by the European Regional Development Fund (ERDF/FEDER Program).

ABSTRACT A number of algorithms in the field of artificial intelligence offer poorly interpretable decisions. To disclose the reasoning behind such algorithms, their output can be explained by means of so-called evidence-based (or factual) explanations. Alternatively, contrastive and counterfactual explanations justify why the output of the algorithms is not any different and how it could be changed, respectively. It is of crucial importance to bridge the gap between theoretical approaches to contrastive and counterfactual explanation and the corresponding computational frameworks. In this work we conduct a systematic literature review which provides readers with a thorough and reproducible analysis of the interdisciplinary research field under study. We first examine theoretical foundations of contrastive and counterfactual accounts of explanation. Then, we report the state-of-the-art computational frameworks for contrastive and counterfactual explanation generation. In addition, we analyze how grounded such frameworks are on the insights from the inspected theoretical approaches. As a result, we highlight a variety of properties of the approaches under study and reveal a number of shortcomings thereof. Moreover, we define a taxonomy regarding both theoretical and practical approaches to contrastive and counterfactual explanation.

INDEX TERMS Computational intelligence, contrastive explanations, counterfactuals, explainable artificial intelligence, systematic literature review.

I. INTRODUCTION

In the last few decades, the field of Artificial Intelligence (AI) has witnessed major changes. As available computational resources have grown significantly, AI algorithms are attracting a significant amount of attention in industry and research [1]. While a great number of such algorithms present strikingly accurate decisions, their decision-making apparatus is frequently left unclear to users of such applications. In particular, a number of Machine Learning (ML)-based algorithms are often perceived as "black-box" algorithms because they are overloaded with millions of hardly interpretable parameters to be optimized at the training stage. This fact makes the algorithm's output hard to explain. A lack of the ability to explain such automatic decisions undermines users' trust and hence decreases usability of such systems [2]. Furthermore, it prevents users from a responsible exploitation of their decisions [3]. In addition, many of the existing eXplainable AI (XAI[1]) methods provide summaries of automatically made predictions rather than true explanations [4]. As a result, the need to motivate automatic decisions with a clear explanation of why the algorithm outputs a particular decision has made the XAI research field grow quickly [5].

Since the number of high-stakes AI applications found in daily life increases, the requirements to their explana-

The associate editor coordinating the review of this manuscript and approving it for publication was Francesco Piccialli.

[1] XAI stands for eXplainable Artificial Intelligence. This acronym was made popular by the USA Defense Advanced Research Projects Agency when launching to the research community the challenge of designing self-explanatory AI systems (`https://www.darpa.mil/program/explainable-artificial-intelligence`).

tory capacity increase accordingly. This also provokes the introduction of regulations and laws concerned with explanation requirements for AI-based applications. For instance, the need for explaining reasoning mechanisms behind such applications is now legally regulated in the European Union by means of the General Data Protection Regulation.[2] According to these legal provisions, the data subject must be provided with ''meaningful information about the logic involved'' in the automatic decision making process, which is commonly referred to as the ''right to explanation'' [6]. Thus, an AI application is expected not only to provide accurate decisions but also to justify them in a comprehensive manner to end-users.

The goal of approaching human-centric AI has led towards a deeper research on the nature of explanation. However, no agreement about a definition of explanation has been reached despite the fact that explanation has called a significant amount of attention in, e.g., philosophy of science [7], [8]. In its most general form, explanation is normally treated as ''an answer to the question of why something is the case'' [9]. In the context of AI, it often bases on judgments about why a certain outcome is predicted by an AI algorithm and hypotheses about causes with respect to given effects [10].

The need of generating more human-like explanations has attracted AI researchers' attention to particular properties of explanation as well as its sub-types [11]. Thus, it appears particularly challenging to explain a given algorithm's output in terms of reasonable yet non-occurring alternatives given a possibly infinite set of such options. Furthermore, this can be enhanced with the ability of suggesting relevant changes in the input so that the algorithm outputs a different decision.

Given a rising interest towards these types of explanation (referred to as contrastive and counterfactual, respectively) within the XAI community, it is of crucial importance to review the existing theoretical accounts of contrastive and counterfactual explanation as well as state-of-the-art computational frameworks for automatic generation thereof. Thus, the aim of this study is to fulfill the next three objectives: (1) to scrutinize theoretical works on the contrastive and counterfactual accounts of explanation; (2) to summarize state-of-the-art methods in the field of automatic explanation generation thereof; and (3) to discuss a degree of synergy between the revised theories and their related up-to-date implementations.

The rest of the manuscript is organized as follows. Section II introduces the notions of contrastive and counterfactual explanation as well as their main application areas. Section III presents the terminology used throughout the review, poses the research questions, and describes the methodology employed to address the given questions. Section IV presents the main findings collected within the present survey and the emerging taxonomy thereof. Section V discusses peculiarities of the existing theoretical and computational frameworks of contrastive and counterfactual explanation. Finally, we conclude in Section VI.

II. BACKGROUND

A. CONTRASTIVE EXPLANATION

Findings on explanation accumulated in humanities and social sciences show that it is intrinsically contrastive [11]. The property of contrastiveness presupposes that an explanation answers the given why-question regarding the cause of the event in question (''Why did P happen?'') in terms of hypothesized non-occurring alternatives (''Why did P happen rather than Q?'') [12]. Thus, supporters of the pragmatic approach to explanation argue that it is exactly the ability to distinguish the answer to an explanatory question from a set of contrastive hypothesized alternatives that provides the explainee with sufficiently comprehensive information on the reasoning behind the question [13]. This approach is also claimed to set a minimum criterion that an explanation must fulfill: it must favor the probability of the observed event P to all the hypothetical alternatives ($Q_1, Q_2, \ldots, Q_n$) [14].

Contrastive explanation is among influential topics in cognitive science [15]–[17]. Thus, contrastive explanations are claimed to be inherent to human cognition [16]. Indeed, we are used to question those decisions that we once made, especially if such decisions or coinciding circumstances resulted in tragic events [18].

In addition, contrastive reasoning forms the basis of abductive inference [19], i.e., the process of inferring certain facts that render some observation plausible [20]. In other words, a given observation can be explained on the basis of the most likely among a pool of competing hypotheses [21].

B. COUNTERFACTUAL EXPLANATION

Given the property of contrastiveness, it is possible to imagine explanatory alternatives to how things would stand if a different decision had been made at some point. They can serve to explain potential consequences of such contrastive non-taken alternative decisions. In this case, the mind is assumed to construct and compare mental representations of an actually happened event and that of some event alternative to it [22]. Cognitive scientists refer to such mental representations of alternatives to past events as counterfactuals (''contrary-to-fact'') [15]. The process of ''thinking about past possibilities and past or present impossibilities'' is therefore called counterfactual thinking [23]. Alternatively, the combination of imagining an alternative scenario in relation to the one that actually happened and the exploration of its consequences is referred to as counterfactual reasoning [24]. In addition, counterfactual reasoning is claimed to be a key mechanism for explaining adaptive behavior in a changing environment [25], [26].

Counterfactuals describe events or states of the world that did not occur and implicitly or explicitly contradict factual world knowledge [27]. Formulated in natural language, counterfactuals are usually presented in the form of conditional

[2] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02016R0679-20160504

statements. Broadly speaking, they contain: (1) an antecedent describing an outcome alternative to an actual event; (2) a consequent describing (a set of) consequences, had the antecedent been the case; and (3) a binary counterfactual dependency relation between them. Thus, Grahne defines a counterfactual to be a conditional statement where the antecedent "can contradict the current state of affairs, or our current knowledge thereof" [28]. However, despite a general agreement on structural properties of counterfactuals, existing interpretations of counterfactual conditionals still compete. As such, further constraints imposed on their structure differ depending on the approach adopted. According to Ginsberg [29], a counterfactual is a conditional statement of the form "If P, then Q" where P is "expected to be false". Aumann limits a counterfactual to be a conditional with a false antecedent only [30]. In contrast, Spohn argues that both the antecedent and the consequent of a counterfactual must be false [31]. All in all, counterfactual conditional statements are claimed to enable people to produce utterances that are factually false yet truthful irrespective of the interpretation adopted [32].

A line of research devoted to modeling human counterfactual reasoning has been thoroughly investigated in computer science. Thus, counterfactual reasoning in computer science is defined as the process of evaluating conditional claims about alternative possibilities and their consequences [33]. It is argued to be valid arising from antecedents that are true in a hypothetical model but false in reality [34]. In this setting, the truth of a counterfactually inferred statement is resolved by: (1) modeling a situation where the smallest possible change in features of the actual world (as set in the antecedent) leads to a different (possibly, desired) state of things (the so-called "closest" or "nearest" possible world); and (2) estimating what is true in that setting [35].

Moreover, counterfactuality is among the most fundamental concepts in theories of causation [36], [37]. Indeed, counterfactuals are argued to represent a causal relation between the event happened in reality and its imaginary counterpart. A counterfactual definition of a cause of an arbitrary event traces back to Hume [38]. According to him, a cause is an object (antecedent) that justifies the existence of another object (consequent) which it is followed by: "If the first object had not been, then the second never had existed". Therefore, once a causal connection between the antecedent and the consequent is established, a counterfactual conditional can be generalized to be a conditional claim about an alternate possibility and its consequences of the form "If X were to occur, then Y would (or might) occur" [33]. Similarly, Kment applies a similarity-based approach between possible worlds to formulate a general account of counterfactuals [39] driven by a non-epistemic interpretation of explanation (i.e., factors that serve as reasons for some fact to obtain are responsible for that fact).

The conditional structure of counterfactual statements gave rise to a probabilistic account of such statements. Thus, Pearl extended the definition of the causal counterfactual to estimate the probability of the truth of the consequent caused by the antecedent ("a probability statement about the truth of y, had x been true, when it is known that y had been false when x was false") [37]. This approach to counterfactuals motivated a number of experiments on the existence of the relation between counterfactuals and conditional probability. In support of this assumption, Over *et al.* [40] showed the existence of connection between counterfactuals and conditional probability, as they experimented with probability judgments about counterfactuals. Thus, they proposed that the subjective probability of the counterfactual at the present time is the same as the conditional probability $P(y|x)$ at some earlier time. Twenty-six subjects were asked to estimate the probability of truth of thirty-two counterfactual conditionals with both affirmative and negative antecedents and consequents. Their findings point to a strong correlation between the probability of the counterfactual conditional and causal strength judgments. On a similar note, Edgington regarded counterfactual judgments as uncertain conditional statements and therefore evaluated them by estimating their conditional probability given some endorsing event [41].

C. DISTINCTION BETWEEN CONTRASTIVE AND COUNTERFACTUAL EXPLANATION

It is important to note that some researchers tend to either collapse or intentionally distinguish contrastive reasoning from counterfactual reasoning despite their conceptual similarity. For instance, Lombrozo treated counterfactual and contrastive explanations as equivalent assuming hypothesized events non-occurred in reality to be "counterfactual cases" where a subset of these cases forms a contrastive explanation [10]. In contrast, McGill and Klein distinguished contrastive reasoning from its counterfactual counterpart [42]. According to them, contrastive reasoning is concerned with situations where different target situations are analyzed ("What made the difference between the employee who failed and the employees who did not fail?"). On the other hand, counterfactual reasoning is claimed to deal with cases where the antecedent is altered to account for changes in the outcome ("Would the employee have failed had she not been a woman?"). Alternatively, Fang *et al.* [43] referred to contrastive reasoning as a procedure operating on "but-statements", as in "all cars are polluting, but hybrid cars are not polluting", which serves a principally different explanation generation task in comparison with the other aforementioned approaches.

D. CONTRASTIVE AND COUNTERFACTUAL EXPLANATION IN THE CONTEXT OF XAI

The stochastic nature of predictions made by various AI algorithms is claimed to be among the main obstacles in reaching a true explanation [44]. Research on automatic contrastive and counterfactual explanation generation shows a number of considerable observations that help overcome this issue. Thus, empirical studies prove that incorporating contrastiveness improves the quality of explanations offered to the end-user [45]. Furthermore, contrastive explanations can be used to personalize human-machine interaction when a user is engaged in an explanatory dialogue with an AI application. Thus, they can be employed with the aim of adjusting the contents of the explanation for the algorithm's output in accordance with the user's preferences [46]. Finally, the ability to explain a decision contrastively is claimed to lead to responsible decision-making [47].

It is important to note that contrastive explanations point to the difference between the actual and a hypothetical decision. On the other hand, counterfactual explanations specify necessary minimal changes in the input so that a contrastive output is obtained. However, these terms are sometimes used interchangeably in the context of XAI [48], [49].

Various families of techniques have been proposed to generate contrastive and counterfactual explanations of AI algorithm output. In the context of XAI, an explanation for an automatic decision or prediction, treated as an observation, can be obtained abductively by attempting the search problem over the set of the known information concerning that observation [50]. Alternatively, counterfactual explanation is widely addressed in the paradigm of case-based reasoning, i.e., a family of problem solving methods based on appeals to precedent solutions. In this setting, generating the most suitable counterfactual may be viewed as a search problem where the most similar precedent is looked for among those making part of the case database [14]. Furthermore, Keane *et al.* argue that applying case-based reasoning techniques for generating counterfactuals increases their explanatory competence [51].

Counterfactual explanations are normally considered contrastive by nature and therefore present a source of valuable complementary information to a given automatic prediction [52]. For instance, a counterfactual explanation of an ML-based algorithm prediction may describe "the smallest change to the feature values that changes the prediction to a predefined output" [53]. An important advantage of counterfactual explanations over their non-counterfactual analogs is that they are devoid of any prerequisites to the data or model. Indeed, counterfactual explanations are data-agnostic as they can be based on the features of the neighbouring data examples extracted from the same training set and/or on the data generated synthetically around the data instance in question. In addition, counterfactual explanations are, in principle, model-agnostic, as they are suitable to explain the output of any black-box algorithm in a post-hoc manner.

Whereas counterfactual explanation generation is concerned with a number of technical challenges, it also requires to take into account several ethical aspects. For instance, their use is expected to be safe (revealing model's internals through counterfactuals may lead to model stealing) [54], fair (discriminatory explanations should be avoided) [55], actionable (suggested changes in the input should be feasible) [56], and accountable (ensuring responsibility for the explanations provided) [57].

III. METHODOLOGY

The present survey has been undertaken as a systematic literature review following the guidelines by Kitchenham and Charters [58], Kitchenham *et al.* [59], and Wohlin [60]. The background notation necessary to follow the findings of the review is specified in Section III-A.

In short, the study comprises three phases as established in the research method by Kitchenham and Charters [58]: (1) planning the review procedure; (2) conducting the review; and (3) reporting the results. During the first phase, three research questions (RQ_1, RQ_2, and RQ_3) were specified (see Section III-B). Subsequently, we determined a search strategy to retrieve primary studies, i.e., we collected all the relevant publications investigating the research questions (see Section III-C). Then, we developed inclusion and exclusion criteria (see Section III-D) in order to select the studies relevant for this article. When the same publication was retrieved from multiple sources, all-but-one instances of the publication (duplicates) were discarded. In addition, we identified and added manually other relevant publications extracted from the bibliography lists of the previously selected manuscripts to ensure a maximum coverage of the related subject areas. It is worth noting that this additional procedure is informally known as snowballing [60]. Finally, we extracted and synthesized the data necessary to address the research questions (see Section III-E).

A. PRELIMINARY TERMINOLOGY

As has been shown in Section II, contrastive and counterfactual explanations presuppose a diverse nature across various application domains. Hence, let us now define the general terms used henceforth in this manuscript. As we are primarily concerned with explainability of AI algorithms, we define explanation in terms of the observed output of such an algorithm. Thus, we regard an *explanation* as a non-empty set of pieces of information justifying the given algorithm's output for an input data instance. The explanation for the given output on the basis of the features of the input data instance is deemed as *factual*. An explanation opposing the actual outcome to one of possible other outcomes is considered to be *contrastive* (e.g., "The data instance is of class A and not B because . . . "). An explanation containing instructions on how the output could have been changed constitutes a *counterfactual* explanation (e.g., "The data instance would be of class B if . . . "). Explanations exhibiting patterns of both contrastive and counterfactual explanation are deemed to be *contrastive-counterfactual* explanations (e.g., "The data instance is of class A and not B because . . . . However, it would be of class B if . . . ").

We distinguish between contrastive and counterfactual explanation throughout the rest of the manuscript if and only if only one of these two terms is used in the given primary study. In contrast, we unify the notions of counterfactual and contrastive explanation introducing the term *"contfactual explanation"* or *"contfactual"* to identify potential similarities and differences of both types of explanation within a broader scope of literature. This term is used hereafter wherever both terms for contrastive and counterfactual explanation can be used interchangeably. The terms "contrastive explanation" and "counterfactual explanation" are only used when they are found in the corresponding study and cannot be used interchangeably in the given context. Notice that the term "contfactual explanation" is not equivalent to "contrastive-counterfactual explanation" but covers both independently used types of explanation as well as their fusion.

A theoretical framework providing justification and a reasoning mechanism for obtaining a contfactual explanation is regarded as a *theory of contfactual explanation*. Altogether, we use the term *contfactual explanation generation* to refer to the process of automatic composition of contfactual explanations for a given output of an AI algorithm in the form of a complementary piece of information associated to a factual explanation.

B. RESEARCH QUESTIONS

In order to reach the three objectives of the study as formulated in Section I, the following three research questions were specified:

- RQ_1: How are contfactual explanations defined in the literature?
- RQ_2: What are the state-of-the-art methods of contfactual explanation generation?
- RQ_3: How grounded are the state-of-the-art contfactual explanation generation methods on the theoretical approaches to contfactual explanation?

C. SEARCH STRATEGY

We selected the digital libraries Scopus and Web of Science (WoS) to retrieve relevant publications from. These libraries do not only include research publications in computing but also index studies across all scientific fields, which allows for an objective analysis of the interdisciplinary literature relevant to the research questions posed.

Subsequently, we performed six queries over the title, abstract, and author keywords in the aforementioned libraries (see the overall structure of the query pipeline in Fig. 1). It is worth noting that the proximity operator *NEAR* is used following the WoS notation whereas the equivalent proximity operator *W* is used for the same queries in Scopus. The following search strings were used for querying the digital libraries:

q_1 = counterfactual* W/3 expla*

q_2 = contrastive* W/3 expla*

$q_3 = q_1$ OR q_2

$q_4 = q_3$ AND (defin* OR theor* OR infer* OR implic*)

$q_5 = q_3$ AND (generat* OR implement* OR framework* OR develop* OR software* OR model* OR artificial intelligence OR AI) AND SUBJAREA(Computer Science OR Mathematics OR Engineering)

$q_6 = q_4$ AND q_5

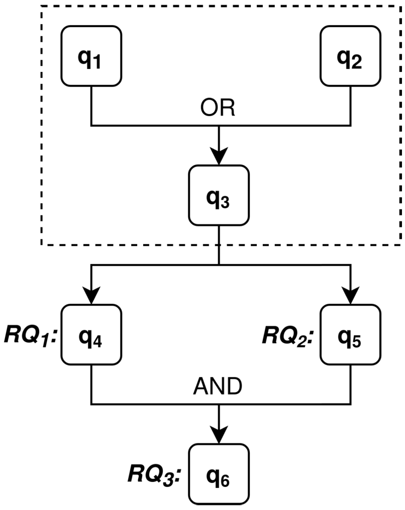


FIGURE 1. A pipeline of the queries executed. The queries found in the dashed area are considered preparatory to those directly addressing the research questions.

The search was performed on October 2nd, 2020. The search web tools of the selected digital libraries allow researchers to reproduce the original study. Furthermore, their use guarantees performing equivalent queries across both libraries. In order to capture all relevant publications, we only used the corresponding word-stems to allow for maximal diversity of the retrieved papers. For instance, the search item "expla*" was used to cover all publications containing such word-forms as "explanation", "explaining", "explanatory", and so on and so forth.

Queries q_1 and q_2 embrace all the up-to-date publications containing mentions of counterfactual and contrastive explanation, respectively, found across all subject areas. In addition, we used a window span of three words (i.e., "NEAR/3") to ensure that the attributes "counterfactual" and "contrastive" relate to explanation. The resulting sets of publications were then unified (q_3).

Subsequently, the preprocessed collection of publications was split into two overlapping subsets aiming to distinguish the publications covering theoretical accounts of contfactual explanation with the aim of extracting the related definitions, theories (or their inferences or implications) (q_4) and existing computational frameworks for contfactual explanation generation (q_5). The terms "definition", "theory", "inference", and "implication" as well as their corresponding word-forms (q_4) were expected to appropriately limit the pool of the unified set of publications with the aim of retrieving definitions as required for addressing RQ_1. Similarly, we used the terms "generation", "implementation", "framework", "development", "software", "model", and their corresponding word-forms (q_5) to retrieve publications concerning contfactual explanation generation frameworks. In addition, the terms "artificial intelligence" and "AI" were used to ensure retrieving relevant AI-related publications. Since RQ_2 addresses purely technical issues of state-of-the-art implementations of such tools, we further imposed an additional restriction on q_5 so that it would return only publications from such subject areas as computer science, mathematics, and engineering. Last but not least, the findings from q_4 and q_5 were merged to examine the connection between the existing theories of contfactual explanation and frameworks for automatic contfactual explanation generation in the context of XAI (q_6).

It is important to note that publications retrieved as a result of q_4 form an exhaustive set of papers addressing RQ_1. Similarly, publications obtained as a result of q_5 address RQ_2. Finally, the papers that q_6 returned address RQ_3.

D. INCLUSION AND EXCLUSION CRITERIA

The publications retrieved during the initial search were subsequently inspected on the basis of the following inclusion and exclusion criteria. To address the epistemology of contfactual explanation, we filtered the retrieved publications to include in the collection of primary studies only those satisfying the following criteria: (1) a publication proposes a contrastive or counterfactual or contrastive-counterfactual approach to explanation or (2) it contains a clearly formulated definition of counterfactual or contrastive explanation referring to other publications in the corresponding field. In order to capture existing computational frameworks for contfactual explanation generation, we included publications that: (1) present a novel approach, method, or framework for contfactual explanation generation whose output can serve to explain the reasoning of an AI algorithm and (2) are found in such subject areas as computer science, mathematics, engineering as well as in their sub-fields.

In contrast, we excluded duplicate reports of the same studies appeared in both Scopus and WoS. As for the publications related to RQ_1, we also removed: (1) the studies whose contents did not introduce any contfactual theory of explanation or (2) those containing no formal or informal definition of contrastive or counterfactual or contrastive-counterfactual explanation. As for the publications related to RQ_2, we discarded: (1) the publications which were not related to AI algorithms or applications as well as (2) those where the proposed framework did not provide any human-comprehensible contfactual explanations as output.

E. DATA EXTRACTION AND SYNTHESIS

Table 1 shows the number of publications retrieved after each independent query, duplicates found among them in Scopus and WoS, as well as Candidate Primary Studies (CPS). Note that the numbers of duplicates indicated in Table 1 refer only to within-query duplicates, i.e., the same publications retrieved from Scopus and WoS for the given single query. Recall that q_4 and q_5 exhaustively cover all the three research questions. Hence, the numbers of CPS are calculated as a sum of the publications retrieved after q_4 and q_5. Furthermore, CPS are reduced by the number of publications addressing RQ_3 because they are found in both sets of publications collected for RQ_1 and RQ_2 and are therefore duplicates.

TABLE 1. Numbers of publications retrieved after each single query as well as those forming the pool of candidate primary studies. The numbers of publications making part of the primary studies are highlighted in bold.

	q_1	q_2	q_3	q_4	q_5	q_6	CPS
Scopus	169	125	288	**140**	**67**	22	185
WoS	122	94	212	**109**	**22**	7	124
In total (including duplicates)	291	219	500	**249**	**89**	29	309
Duplicates	108	80	184	**92**	**21**	6	107
In total (excluding duplicates)	183	139	316	**157**	**68**	23	202

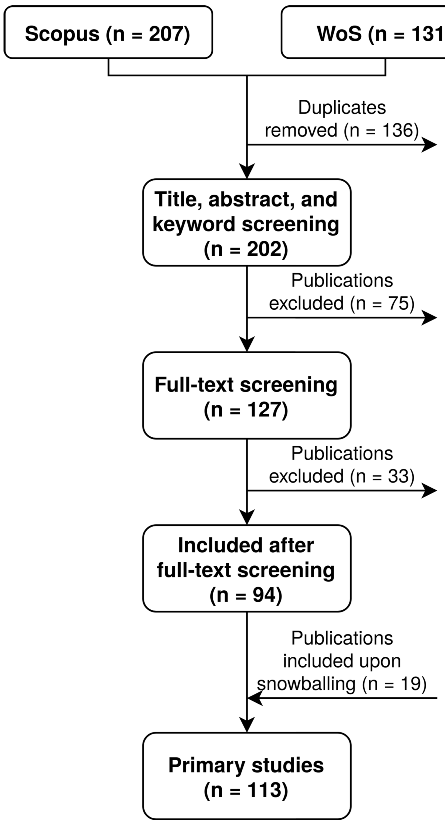

Scopus (n = 207)
WoS (n = 131)
Duplicates removed (n = 136)
Title, abstract, and keyword screening (n = 202)
Publications excluded (n = 75)
Full-text screening (n = 127)
Publications excluded (n = 33)
Included after full-text screening (n = 94)
Publications included upon snowballing (n = 19)
Primary studies (n = 113)

FIGURE 2. A flow diagram of the primary study selection on the basis of queries q_4 and q_5 (*n* is the number of publications at each stage).

Fig. 2 displays the flow diagram of the primary study selection. A sum of 338 publications (207 from Scopus and 131 from WoS) made up the collection of CPS addressing the research questions. 107 within-query duplicates were identified and removed from further analysis. In addition, 29 more duplicates were excluded when merging the sets of publications retrieved after q_4 and q_5. All in all, 136 duplicates were removed.

The title, abstract, and author keywords of each candidate primary study were screened to discard the studies irrelevant to the research questions posed. As shown in Fig. 2, 75 publications were deemed irrelevant and filtered out at this stage. A deeper analysis of the remaining 127 publications

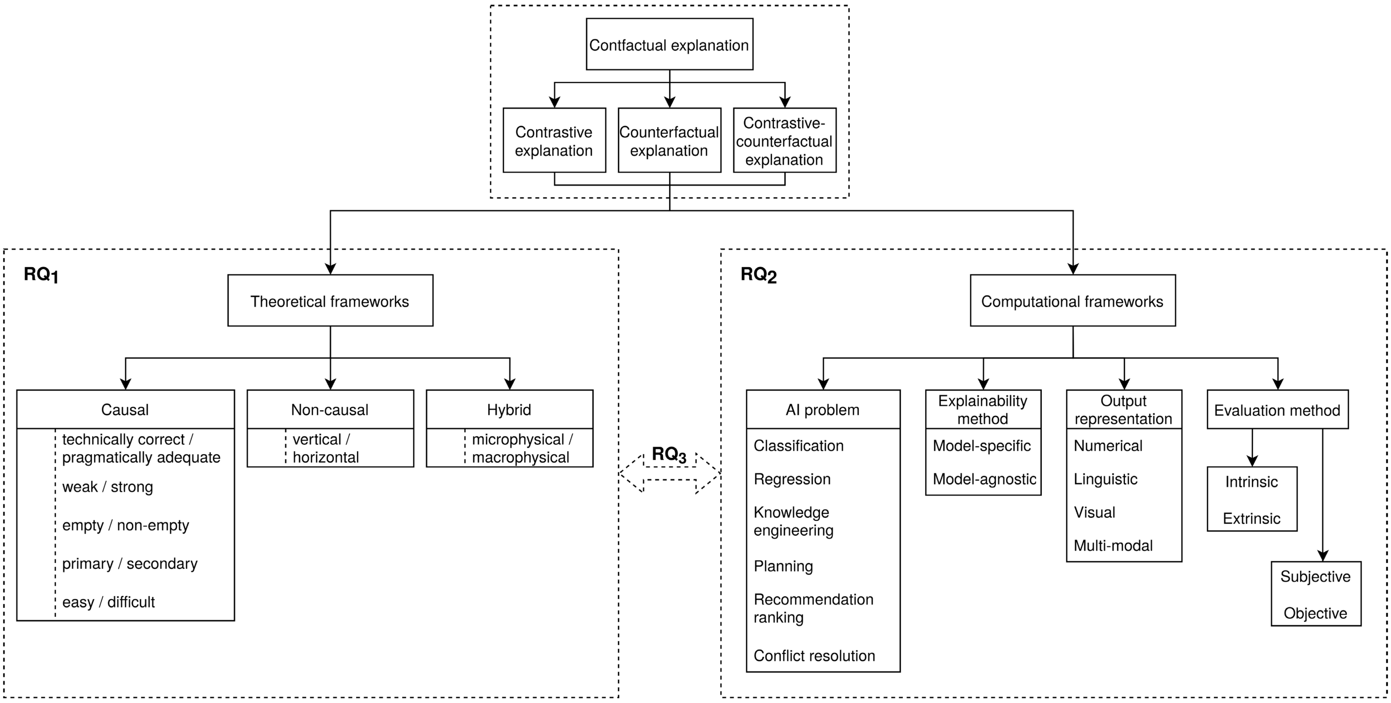


FIGURE 3. A taxonomy of contfactual explanation emerging from our systematic literature review.

TABLE 2. The exhaustive list of all the primary studies in relation to each research question.

Research question	Primary studies
RQ_1	[13], [17], [21], [22], [36], [37], [49], [61]–[127]
RQ_2	[6], [46], [48], [49], [55], [56], [61], [67], [80], [81], [84], [86], [91], [93], [94], [100], [105], [122], [128]–[161]
RQ_3	[49], [61], [67], [80], [81], [84], [86], [91], [93], [94], [100], [105], [122]

enforced us to discard 33 studies which did not satisfy the inclusion criteria. Finally, 19 papers were added to the review upon inspecting the bibliography of the primary studies. As a result, 113 unique publications formed the exhaustive pool of primary studies.

Table 2 presents the list of primary studies selected for the review. Thus, a collection of 74 out of 113 (65.49%) original primary studies were found to formulate definitions for contfactual explanation and/or address theoretical accounts thereof (RQ_1). In addition, 52 out of 113 (46.02%) publications describe frameworks (or extensions of other frameworks) for contfactual explanation generation (RQ_2). Note that 13 out of 113 (11.50%) primary studies were found to address both RQ_1 and RQ_2 and therefore answer RQ_3.

The following data were extracted from each primary study: title, authors, year of publication, author keywords. In addition, all publications related to RQ_1 were read to analyze contfactual theories of explanation and, subsequently, extract the sought-for definitions of contfactual explanation. As RQ_2 concerned a broader number of technical characteristics of contfactual explanation frameworks, we additionally extracted the following information: (1) the problem that the retrieved framework aims to solve; (2) the method proposed for contfactual explanation generation; (3) the form of output explanation (for instance, textual or visual); and (4) the corresponding evaluation methods. Based on the data extracted from the primary studies, the publications were grouped and classified in accordance with the aforementioned criteria.

IV. RESULTS

Prior to answering the research questions, we carried out a bibliometric analysis over the results of the general independent queries on counterfactual and contrastive explanation (q_1 and q_2, respectively) as well as their union (q_3). We report the results of the bibliometric analysis in Section IV-A. The findings related to the theoretical accounts of contfactual explanation (RQ_1) are presented in Section IV-B. The analysis of the computational frameworks for contfactual explanation generation (RQ_2) can be found in Section IV-C. Finally, the publications describing theoretically grounded computational frameworks (RQ_3) are reported in Section IV-D.

An emerging taxonomy of contfactual explanation frameworks is depicted in Fig. 3 and forms the core of the results discussed in the rest of the manuscript.

A. BIBLIOMETRIC ANALYSIS

The bibliometric analysis over the queries q_1, q_2, and q_3 allows us to obtain a big picture of the research area of contfactual explanation generation and spot its key characteristics. To illustrate the state of affairs within the field, we report annual scientific production and maps of author keywords revealing the main problem-specific notions. The reference

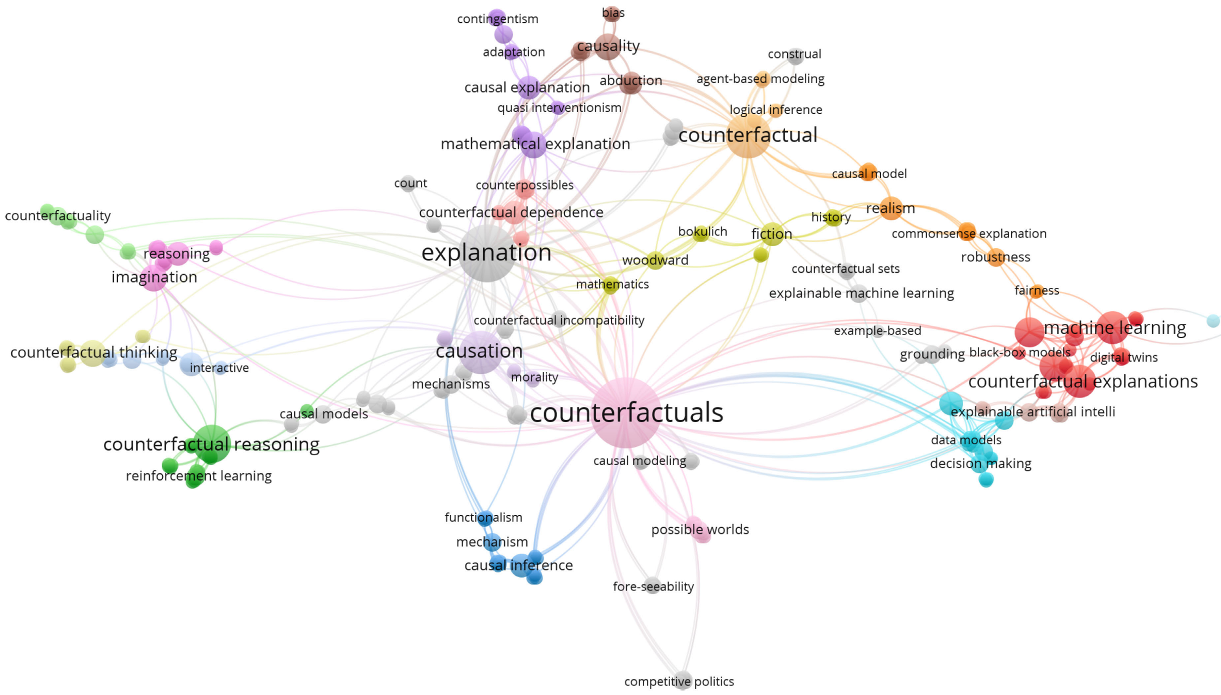


FIGURE 4. The map of author keywords for the q_1 publications.

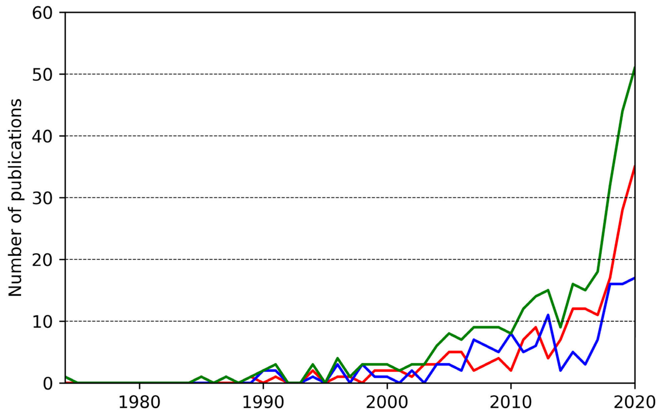


FIGURE 5. Annual scientific production for the publications retrieved after q_1 (the red line), q_2 (the blue line), and q_3 (the green line).

manager *Mendeley* was used to filter out duplicate publications. In addition, we utilized the tool *VOSViewer* [162] to generate the author keyword maps.

It can be seen that contfactual explanation appears to attract an increasing attention across all subject areas in the past two decades. Furthermore, Fig. 5 shows a rapid rise in the number of publications in the past three years. It is worth noting that the number of publications in 2020 is limited to the search date.

Author keyword maps allow us to present an overview of the terms most relevant to those specified in the preparatory queries (q_1, q_2, and q_3). For illustrative purposes, non-linked keywords were deemed to be outliers and filtered out from the analysis. Table 3 shows the overall number of keywords as well as that of linked keywords for each preparatory query.

TABLE 3. Numbers of linked and non-linked keywords in the preparatory query results (q_1, q_2 and q_3).

Query	All keywords	Linked keywords
q_1	422	323
q_2	341	228
q_3	702	530

Fig. 4 shows a graph containing the most popular author keywords for counterfactual explanation. It can be concluded that counterfactual explanation is often investigated in the context of causation (pay attention to such keywords as "causation", "causal inference" or "causal models") as well as cognitive science (as reflected by the keywords "imagination", "reasoning", etc.) and AI ("machine learning", "data models", "black-box models"). Similar notions are observed to be essential for contrastive explanation (see Fig. 6). However, a distinction between different clusters in the latter case is visible more clearly. This is hypothesized to be due to a more diverse usage of the term "contrastive explanation" across various scientific areas.

A stronger impact of counterfactual explanation in the results of the joint query q_3 appears to affect significantly the overall allocation of the related keywords in the corresponding keyword map (see Fig. 7). The keywords identified in the studies related to q_3 testify that the issue of contfactual explanation is highly interdisciplinary and finds application in both humanities and natural sciences.

B. CONTFACTUAL EXPLANATION AS DEFINED IN RELATED THEORIES (ANSWER TO RQ$_1$)

As presented in Section II, the surface form of contfactual explanation is found to preserve the same syntactic structure

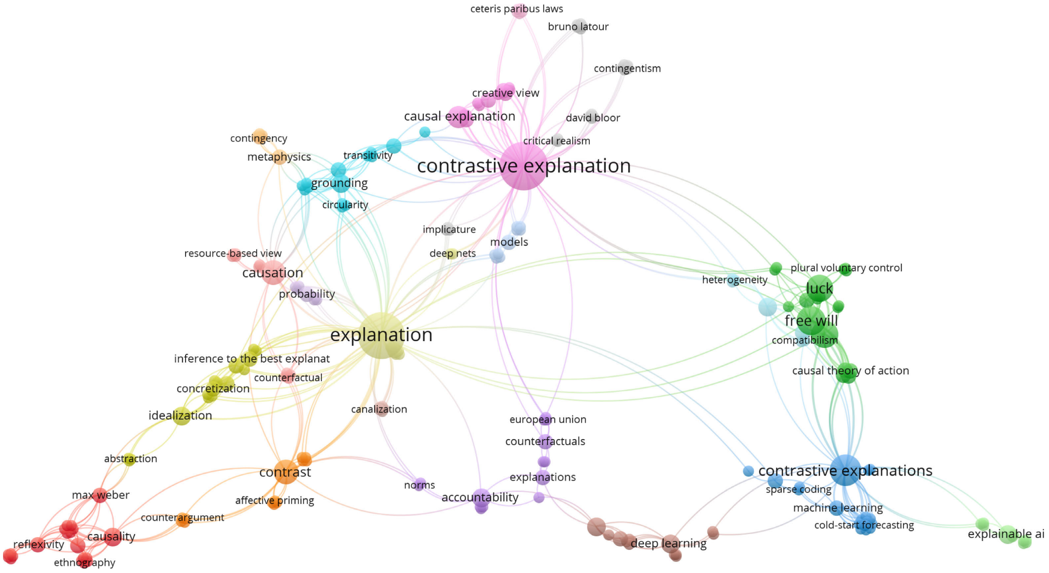


FIGURE 6. The map of author keywords for the q_2 publications.

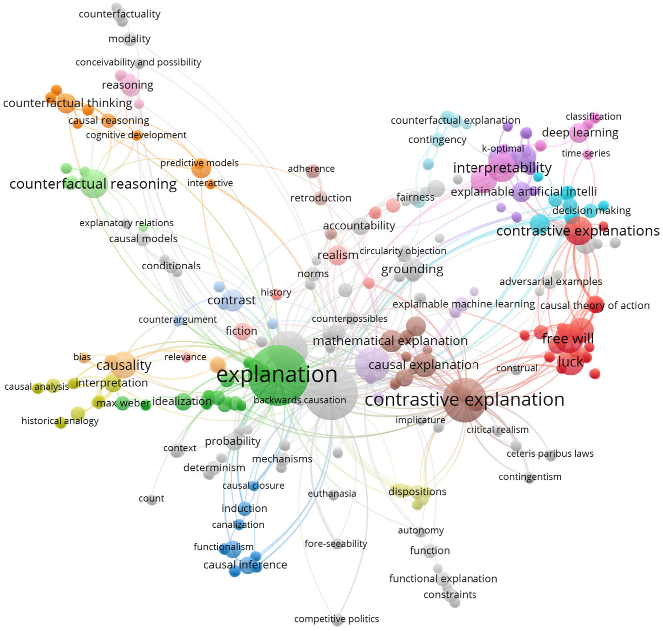


FIGURE 7. The map of author keywords for the q_3 publications.

TABLE 4. A classification of approaches defining contfactual explanation.

Approach	Publications
Causal	[13], [21], [36], [37], [61]–[63], [66]–[68], [70]–[73], [75], [77]–[79], [82]–[84], [88]–[90], [92]–[94], [97], [99], [101]–[103], [106], [107], [109], [113]–[121], [123]–[127]
Non-causal	[49], [64], [65], [69], [74], [80], [81], [86], [87], [91], [96], [100], [110], [111], [122]
Hybrid	[17], [22], [76], [85], [95], [98], [104], [105], [108], [112]

in general. However, a major factor discerning theoretical approaches to contfactual explanation is found to consist in their relation to causation. As will be shown below, several accounts of contfactual explanation presuppose a causal nature and establish a causal contfactual dependency between the phenomenon to be explained and the explanation itself. In contrast, other theoretical frameworks seek purely non-causal dependencies in explanation. In addition, several researchers attempt to unify causal and non-causal contfactual explanation under the same paradigm. Hence, we distinguish three main groups of contfactual explanation that encompass all the retrieved primary studies: causal, non-causal, and hybrid. The exhaustive list of the retrieved publications in accordance with the suggested taxonomy is presented in Table 4 and discussed further in the following sections.

Remind that contfactual explanation embraces contrastive, counterfactual, and contrastive-counterfactual explanation. Each type of contfactual explanation is present in the findings, causal counterfactual making up a majority of the considered theoretical frameworks (see Fig. 8). Hence, we analyze each contfactual explanation type independently in terms of causality in this section to draw a comparison between different approaches. In addition, we consider (1) the issue of quantitative evaluation of causality for causal contfactuals as reflected in specific primary studies and (2) different subcategorizations of causal, non-causal, and hybrid contfactual explanation.

1) CAUSAL CONTFACTUAL EXPLANATION

- **Causal contrastive explanation** is frequently found to be designed as an answer to a why-question of the following canonical form: "Why P rather than Q?" where P is an explanandum (i.e., the fact to be explained), Q being a foil (i.e., one of alternative non-occurring options) [21]. Lipton introduces the notion of

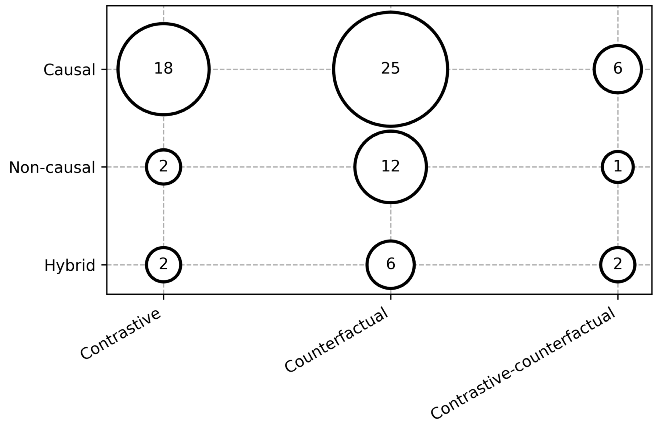


FIGURE 8. Numbers of identified theoretical contfactual explanation frameworks with respect to causality.

TABLE 5. Counterfactual explanation theories reflected in the primary studies.

Theory	Author(s)	Primary studies concerned
“Closest possible worlds”	Lewis [36], Stalnaker [120]	[66], [72], [99], [109] [114], [117], [123]
SCM	Pearl [37]	[71], [113]
CTE	Woodward [125]	[62], [67], [68], [79], [83], [116]
NRCM	Neyman [107], Rubin [115]	[90]

“difference condition”: the contrast between the fact and the foil is explained by identifying the cause of the fact and proving the absence of the corresponding cause of foil [103]. Following Lipton [103], Kean redefines a contrastive explanation to be the difference between the causal explanations for the question and the contrast [93]. Barnes further requires that “P and Q be culminating events of a single type of natural causal process” [63]. Similarly, Day and Botterill introduce the concept of differential inference, i.e., a form of inference based on contrastive explanation that “can be used in order to generate causal hypotheses” [77].

Van Fraassen formalizes a contrastive question to be the triple $\langle P, X, R\rangle$ where P is a topic (explanandum), X is a contrast space or contrast-class (i.e., a set of alternative answers to the given question), and R being the corresponding relevance criteria [13]. Then, the answer to a why-question must differentiate the topic from the contrast space. Contrarily, Chien excludes the contrast class when taking contrastive explanation as a model for scalar implicature [75].

Hitchcock generalizes the notion of contrastive explanation over all explanations bearing contrastive stress irrespective of their syntactic structure [88]. Conversely, Aguilar-Palacios *et al.* oppose the alternative explanation-seeking question “Why P rather than Q?” to the congruent question “Why P but Q?” [61]. This formulation of the question generalizes the explanation to justify why some fact P occurs in the current situation whereas some foil Q occurred in different circumstances with the aim of establishing a cause and effect relation between the fact and the foil. Similarly, Tsang and Ellsaesser claim that a contrastive explanation should point to the importance of identifying the most relevant factors differing the causal histories of the fact and the foil where both fact and foil must be true [124].

Contrastive explanations may not only concern the explanandum but also the answers to contrastive questions (often referred to as explanans). Thus, Sober stresses that the question of whether some hypothesis H explains why a non-contrastive proposition E is true is “incomplete until H is contrasted with an alternative hypothesis” [119]. In addition, the canonical forms of the contrastive question and the corresponding explanatory answer (i.e., the core of contrastive explanation) have raised a number of epistemological concerns in philosophy of science. For instance, Dickenson reformulates the contrastive explanation-seeking question to be: “What explains how it is possible that an agent can act on R_1 other than R_2, given that R_2 is present?” [78] (where reason R_1 is the cause of some action and R_2 is not). The notion of contrastive explanation is further developed in the agent-causal theory of free will. Thus, contrastive explanation is applied to agent's decision-making (i.e., why the agent makes a choice refraining from an alternative choice) [82], [101].

Campbell redefines a contrastive explanation in terms of the so-called “structuring causes”, i.e., the traits of the structure of the causal system that trigger actual causes of some event to happen [73]. A cause of this kind is responsible for the connection between the types C and M in a system S. A contrastive explanation thus explains why a system S is claimed to be “wired” in such a way that an internal state of type C regularly causes a movement of type M. Similarly, Kim *et al.* regard contrastive explanation as a constraint for a system to be satisfied by a specific set of plan traces [94].

Finally, Boulter illustrates the use of contrastive explanations to distinguish between actual and non-actual biological forms [70]. Claiming all explanations in biology to be causal, the researcher introduces the following template for a causal relation in contrastive explanation: “c_1 rather than c_2 or c_3 or c_n causes e_1 rather than e_2, or e_3 or e_n” leaving contrasting causes implicit.

- **Causal counterfactual explanation.** Most of the considered studies on causal counterfactual explanation relate to either of the four theoretical milestones: Lewis-Stalnaker's theory of closest possible worlds [36], [120], Pearl's Structural Causal Models (SCM) [37], Woodward's Counterfactual Theory of Explanation (CTE) [125], or the Neyman-Rubin Causal Model (NRCM) [107], [115] (see Table 5).

  The Lewis-Stalnaker approach codifies a counterfactual conditional as a logical proposition where the antecedent and the consequent are connected by means of the "might"- or "would"-conditional operator. Exploiting the mechanism of possible world semantics, the truthfulness of a counterfactual is assessed by assigning to it a binary truth-value in accordance with its proximity to the world in question.

Following this approach, Kutach defines counterfactuals in natural language to be "propositions obeying a logic whose semantics is given in terms of a comparative similarity relation among possible worlds" [99]. Similarly, Strohminger and Yli-Vakkuri assume that the counterfactual modus ponens preserves truth-functional possibility ("If it is possible that p and p counterfactually implies q, then it is possible that q") [123] where p is a logical proposition and q is a conjunction of such propositions.

Briggs extends the Lewis-Stalnaker model by applying causal modeling language to comprise logically complex antecedents [72]. Schweder considers a counterfactual to be an implicit claim within the explanatory answer to an explanation-seeking question [117]. In addition, Pruss and Rasmussen take into account antecedents that are not necessarily "contrary-to-fact" and define a counterfactual to be a contingent proposition establishing a causal connection between a specific description of the circumstances of a choice and a report of an action in such circumstances [109].

Pearl's SCM operates on a predefined causal model $M = \langle U, V, F\rangle$ consisting of sets of background variables determined by factors outside the model (U) and within it (V) and a set of functions $F = \{f_i \mid 1 \leq i \leq n\}$ mapping from $U \cup (V \setminus V_i)$ to V_i, that associates each variable V_i with all the variables from U and V. Given a set of variables $X \in V$ and a causal submodel $M_x = \langle U, V, F_x\rangle$ so that $F_x = \{f_i : V_i \notin X\} \cup \{X = x\}$ and by defining a minimal change in M required to make a selected variable $X = x$ ($X \in V$) hold true under any $u \in U$, a causal counterfactual is formally defined as the solution for some subset $Y \in V$ on the set of equations F_x [37]. Counterfactuals are thus pruned by interventions on the antecedent component [113], which leads to interpreting counterfactuals as non-observable hypothetical contrasts [71].

Similarly to Pearl's SCM, Woodward's CTE establishes the counterfactual dependence between the two variables by means of the intervention mechanism. Thus, for two variables X and Y taking on some values x and y, respectively, to explain the value of y counterfactually is to show that Y would have taken on some value y' if X had taken some counterfactual value x' [125]. In other words, some small enough change in the value of X from x to x' would cause a change in Y from y to y' in the absence of changes in values of other variables.

Following Woodward's theory, Schneider and Rohlfing define a counterfactual as "a theoretically relevant manipulation of the observed case in order to ascertain whether this manipulation would make a difference to the outcome" [116]. Further, Bertossi defines a causal counterfactual explanation to be a set of the original feature values in the given data instance that are affected by a minimal counterfactual intervention [67] (where minimality is assumed to be based on a partial order relation on counterfactual interventions).

Conversely, Andreas and Casini reconsider explanatory counterfactuals to be "hypothetical assumptions about the values of quantities or the values of propositions". They argue that Woodward's interventionist account of explanation cannot handle the cases where interventions are physically impossible (e.g., due to violations of laws of nature) [62]. Applied to theorem proving, Gijsbers leaves out the mechanism of intervention from Woodward's CTE. He states that a mathematical proof has explanatory power only when the explanandum is complemented with a contrasting claim that shows how the mathematical object in question varies in the process of theorem proving. Also, Fang infers counterfactual dependencies in the form of counterfactual claims: "in the model M, had the variable X taken such-and-such a value x_i, then the variable Y would have taken such-and-such a value y_j" [79].

Last but not least, Holland argues that causal counterfactuals are highly relevant to research in social sciences. Thus, he follows NRCM interpreting counterfactuals in terms of potential outcomes of a dependent causal variable given some intervention with respect to that variable [90].

- **Causal contrastive-counterfactual explanation** is sometimes considered to include "all kinds of subjunctive conditionals, regardless of whether the antecedent is true in the actual world or not" [126]. Thus, Kuorikoski and Ylikoski elaborate on a contrastive counterfactual theory of explanation claiming that the property of contrastiveness helps to resolve linguistic ambiguity inherent in explanation [97]. In this setting, interventions (or manipulations) specify the truth conditions of such explanations: "c [$c*$] causes e [$e*$] if we can bring about $e*$ [e] by bringing about $c*$ [c]" [127] (where c and $c*$ are causes, e and $e*$ being the corresponding effects).

Following Kuorikoski and Ylikoski [97], Northcott examines explanatory relevance of counterfactuals placed in a contrastive framework [106]. Similarly, Hohwy regards causal counterfactuals as an integrative part of causal contrastive explanations. Thus, he claims counterfactuals supported by laws are able to "go into contrastive explanations even though unfavourable conditions ensure that the forces they describe are not actually occurring in the way described by any law taken alone" [89]. On a similar note, Steglich-Petersen [121] proposes two-level semantics of contrastive causal statements requiring specific semantically complete counterfactual justifications.

- **Degree of causality in causal contfactuals.** Causal relations between the variables in explanation are not always considered binary (i.e., in the presence/absence of a cause). There have been several attempts to measure a degree of causality in contfactual explanation. Since causal contrastive explanations describe a certain aspect of the explanandum, Rips and Edwards claim them to be partial by nature [113]. In other words, the explanatory power of such explanations can be quantified and compared with that of others. In light of this assumption, Northcott defines the degree of causation (i.e., causal strength of a cause variable) to be the difference between the values that the effect variables take on in the actual and counterfactual cases [106]. In this regard, he determines a counterfactual to be the value of the effect variable. A counterfactual can thus be measured quantitatively as the distance between the target levels of the causal effect variables. Similarly, Ylikoski and Kuorikoski distinguish five dimensions of explanatory power of contrastive explanations: (1) non-sensitivity (i.e., how sensitive the explanation is to background conditions); (2) precision (i.e., how precisely the explanation characterizes the explanandum); (3) factual accuracy (i.e., a proportion of true facts captured by the given explanation in comparison with another); (4) degree of integration (i.e., unification to a larger theoretical framework); and (5) cognitive salience (i.e., "the ease with which the reasoning behind the explanation can be followed") [126].
- **Subtypes of causal contfactuals.** It is worth noting that several subcategorizations of causal contfactuals have been suggested within some of the aforementioned theoretical frameworks.

  As for contrastive explanation, Franklin follows Hitchcock [88] differentiating "technically correct contrastive explanation" (the explanation citing explanatory relevant information) and "pragmatically adequate/defective contrastive explanations" (the explanation providing more information than explanatory relevant) [82]. Levy distinguishes between weak contrastive explanation (if the agent is not able to explain how the agent-causal power was exercised for reasons) and strong contrastive explanations (otherwise) [101].

  As for counterfactual explanation, Holland points to a deceptive use of "empty" counterfactuals, i.e., counterfactuals whose antecedent "could never occur in any real sense" [90]. Steglich-Petersen distinguishes between primary counterfactuals (i.e., those that relate two events A and B as the cause and the effect) and secondary ones (i.e., those that establish the fact that it is event A that causes B to happen) [121]. Finally, Schneider and Rohlfing claim counterfactuals to be either easy or difficult [116]. From this perspective, easy counterfactuals are "the assumptions about the outcome of logical remainders" that simplify theoretical expectations. In contrast, the assumptions that simplify the solution "but run counter to our theoretical expectations about whether single conditions involved in a remainder should or should not contribute to the outcome" are assumed to be difficult.

2) NON-CAUSAL CONTFACTUAL EXPLANATION

- **Non-causal contrastive explanation.** Notably, non-causal contrastive explanations can address the physical nature of a modeled system. Hence, they can be used to explain the properties and relations inherent to such systems. Thus, Chakravartty extends the concept of contrastive explanation to answering non-causal what-questions, e.g.: "What dispositions of p are relevant to circumstances x as opposed to y?", where p is the object whose traits require an explanation and x and y are the circumstances determined by the question-dependent context [74].

  In contrast to Dickenson [78] (see Sect. IV-B1), Botterill appeals to a non-causal nature of contrastive explanations [69]. Thus, the researcher argues that "the fact that, in the absence of R_2 but with R_1 still present the agent would perform an action of some kind does not show that when both R_1 and R_2 are present an agent does not act in that way because of both those reasons" (where reason R_1 is the cause of some action and R_2 is not).
- **Non-causal counterfactual explanation.** Reutlinger develops a non-causal counterfactual theory of explanation to apply it to Euler's explanation[3] and the renormalization group theory[4] [110]. This counterfactual theoretical framework is subsequently extended to capture non-causal explanations in metaphysics [111].

  Driven by the assumption that physical facts and mathematical models share certain features, Baron *et al.* apply a structural equation modelling framework to model counterfactuals that could explain physical facts in terms of non-causal mathematical explanations [64]. Further, Baron introduces the concept of the so-called "counterfactual scheme" applied to mathematical explanation [65]. A counterfactual scheme is thus defined as a triple containing (1) a counterfactual statement with non-logical expressions substituted with variables, (2) instructions stating which parts of the statement can be substituted to produce a counterfactual, and (3) a classification for evaluating the given counterfactual. A counterfactual is then claimed explanatory if all the instances of a counterfactual scheme are true and at least two counterfactual schemes are distinct so that the corresponding physical laws relevant for evaluation of the given counterfactuals are different. Also, Hird uses

[3] Reutlinger refers to the phenomenon found in the city of Königsberg where no-one succeeded to cross the seven bridges located in four different parts of the city exactly once. Euler provided a non-causal explanation for this phenomenon in terms of graph theory.

[4] According to Reutlinger, renormalization group explanations are intended "to provide understanding of why microscopically different physical systems display the same macrobehavior when undergoing phase transitions" [110].

the term ‘‘counterfactual’’ to define projects that have been funded in the absence of congressional committee influence [87].

In addition, a number of definitions for non-causal counterfactual explanation come from AI. In the context of automatic decision-making, counterfactuals are found to be most generally defined as counterarguments for an alternative prediction [86]. Fernández *et al.* refer to a counterfactual as an effective type of explainable ML technique that explains predictions by describing the changes needed in a sample to flip the outcome of the prediction [81]. More precisely, Fernández *et al.* define a counterfactual for classification tasks as a ‘‘hypothetical instance similar to an example whose explanation is of interest but with different predicted class’’ [80]. Kanehira *et al.* attempt to explain counterfactually video classification output framing a (visual-linguistic) counterfactual explanation in the form of the conditional statement ‘‘X would be classified as B and not A if C and D are not in X’’ [91] (where X is the data example requiring an explanation, A is the class predicted for X, B is the contrast-class in question, C and D are specific visual patterns present or absent in the given video frame X). On a similar note, Laugel *et al.* treat a counterfactual explanation as a specific data instance, close to the observation whose prediction is explained, but predicted to belong to a different class [100]. Kostic defines a counterfactual to be a statement describing a hypothetically different situation to the actual state of affairs [96]. He distinguishes between vertical and horizontal counterfactuals. Thus, a counterfactual is considered vertical if ‘‘a global topological property determines certain general properties of the real-world system’’. In contrast, a counterfactual is deemed horizontal if ‘‘a local topological property determines certain local dynamical properties of the real-world system’’. Finally, Stepin *et al.* point that a counterfactual explanation should refer to a set of features ‘‘minimally different from those inherent to the original data point’’ [122].

- **Non-causal contrastive-counterfactual explanation.** Poyiadzi *et al.* do not distinguish between counterfactual and contrastive explanations assuming counterfactuals to be the new state of the considered object [49].

3) HYBRID CONTFACTUAL EXPLANATION

- **Hybrid contrastive explanation.** Chin-Parker and Bradner [17] as well as Chin-Parker and Cantelon [76] provide a unified theoretical framework for causal and non-causal contrastive explanation for category learning. Emphasizing the crucial importance of context for an explanation, they consider a contrast class to be a set of non-occurring alternates that delimits the set of potentially relevant information irrespective of the inherent causal relations.
- **Hybrid counterfactual explanation.** Explanatory pluralism is as well recognized in the research on counterfactual explanation. Thus, Byrne states that ‘‘not all counterfactuals are about causes, and counterfactuals that imply a causal relation differ in systematic ways from counterfactuals that identify other sorts of relations, such as intentions’’ [22]. Indeed, a large body of research on both causal and non-causal counterfactual explanation testifies that counterfactuals have a diverse nature with respect to causality [104]. Thus, Lowe claims counterfactuals to be causal ‘‘when the modality involved is evidently natural or causal necessity’’. Contrarily, other explanation cases such as those arising in mathematics ‘‘clearly do not involve this sort of necessity, but instead something like logical necessity’’ [104]. Further, Knowles and Saatsi discuss the notion of explanatory generality presuming both causal and non-causal nature of counterfactuals arguing that ‘‘explanatory counterfactuals are appropriately directed and change-relating, capturing objective, mind-independent modal connections that show how the value of the explanandum variable depends on the value of the relevant explanans variables’’ [95].

  In light of this, there have been several attempts to unify causal and non-causal counterfactuals within one framework. Hence, a hybrid approach, originating from monism,[5] has been adopted to unify causal and non-causal counterfactual explanation. Following this approach, Reutlinger introduces a unified explanation framework consisting of the following elements: a statement about the explanandum E, a set of generalizations (or explanans) $G_1,\ldots,G_m$, and a set of auxiliary statements setting initial conditions for the explanatory system [112]. A relation between an explanandum and a set of explanans is claimed to be explanatory if and only if at least one of the explanans supports the counterfactual statement ‘‘had $S_1,\ldots,S_n$ been different than they actually are (in at least one way deemed possible in the light of the generalizations), then E or the conditional probability of E would have been different as well’’. At the same time, the generalizations and auxiliary statements must logically entail the explanatory statement in question. As such, both causal and non-causal explanations are argued to be captured because they ‘‘reveal counterfactual dependencies between the explanandum and the explanans’’. Following Reutlinger’s account of explanation, Held argues that the notion of counterfactuals can hardly be supported only by generalizations [85]. Furthermore, true generalizations (e.g., ‘‘all ravens are black’’) might not allow for counterfactual situations at all. Instead, he weakens the counterfactual dependency to shift from generalizations to plain counterfactuals.

  Mothilal *et al.* suggest a feature-based counterfactual explanation generation framework where importance of independent features is evaluated [105]. Nevertheless,

[5] Monism is a philosophical account of explanation that captures both causal and non-causal explanations reducing them to a single entity [112].

TABLE 6. A classification of the contfactual explanation generators by AI problem.

AI problem	Publications
Classification	[6], [46], [48], [49], [55], [67], [80], [81], [86], [91], [100], [105], [122], [128]–[131], [133]–[148], [150]–[155], [158]–[160]
Regression	[56], [61], [129]
Knowledge engineering	[93]
Planning	[94], [132], [156], [157], [161]
Recommendation	[84]
Conflict resolution	[149]

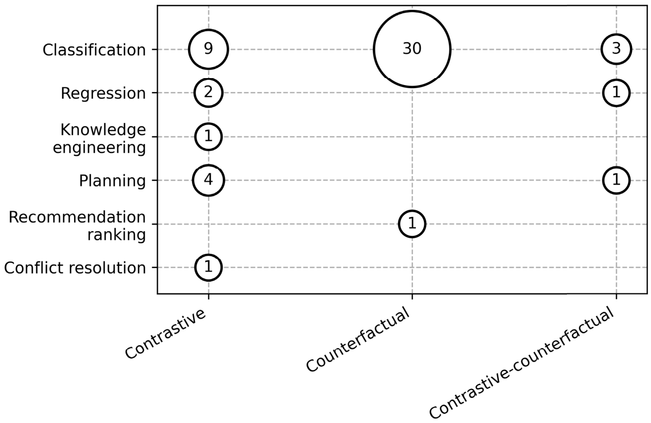


FIGURE 9. Numbers of frameworks grouped by AI problem with respect to the type of contfactual explanation generated.

they emphasize the need for causal attribution, as ignoring causal relations may lead to generating unfeasible counterfactuals. Therefore, they suggest a hybrid framework for counterfactual explanation generation.

- **Hybrid contrastive-counterfactual explanation.** Kuorikoski and Ylikoski point to the multifaceted nature of contrastive-counterfactual explanation. They argue ''there exist constitutive and possibly formal counterfactual dependencies as well as combinations of these'' [98]. Similarly, Pexton suggests a two-level hierarchy of explanation [108]: microphysical explanations are non-causal and form the lower-level of the hierarchy whereas manipulable causal explanations are placed at the higher-level.

C. CONTFACTUAL EXPLANATIONS AS DEFINED IN AUTOMATIC GENERATION FRAMEWORKS (ANSWER TO RQ_2)

The analysis of the primary studies related to RQ_2 allows us to categorize the state-of-the-art contfactual explanation generation frameworks in accordance with the following criteria: (1) the problem the solution for which is to be explained (i.e., the AI problem); (2) the method employed to generate such an explanation (i.e., the explainability method); (3) the output representation of the explanation; and (4) the evaluation method thereof.

1) AI PROBLEM

Contfactual explanations are used to justify automatic decisions obtained for a variety of AI-related problems. Table 6 provides the reader with a taxonomy of the state-of-the-art frameworks from the primary studies. It is derived from the considered publications in terms of the domain tasks that these frameworks are used for. As depicted in Fig. 9, most contfactual explanation generation frameworks deal with counterfactual explanation (31 out of 52 frameworks; 59.62%). In contrast, 17 out of 52 (32.69%) generate contrastive explanations. Only four studies (7.69%) fuse contrastive and counterfactual explanations. One of these studies [129] deals with both classification and regression.

- **Contfactuals for classification.** A vast majority of state-of-the-art AI applications that generate contfactuals (42 out of 52; 80.77%) are used to explain the outcome of ML-based classifiers, i.e., algorithms that learn a mapping function $f : X \longrightarrow Y$ from a training dataset of n labeled examples $X = \{x_i \mid 1 \leq i \leq n\}$ to a discrete output variable (class) $Y = \{y_j \mid 1 \leq j \leq m\}$ where m is the number of classes. Indeed, contfactuals are particularly suitable for informing the end-user why a given data example is assigned a particular class label. Thus, the outlined classification-oriented frameworks are evaluated on classifiers based on logistic regression [55], [136], [153], [158], decision trees [46], [80], [122], [140], [150], [155], [159], gradient boosted decision trees [147], support vector machines [131], [138], [146], random forests [81], [86], [142]–[144], neural networks [6], [48], [49], [91], [129], [130], [133], [135], [139], [141], [145], [148], [151], or combinations of these [100], [105], [134], [152], [154], [160]. In three studies [67], [128], [137], the classifiers used in the experiments are not specified.
- **Contfactuals for regression.** One of the classification-oriented frameworks [129] is extended to also handle the regression problem, i.e., learning a mapping function f from a training dataset X to a continuous output variable Y. However, the continuous output is, in this case, subsequently converted to a lower-scale discrete value mapped to a textual description similar to that typical of a classification problem. The other frameworks addressing the regression problem aim to leverage gradient-boosted decision trees [61] and indicate how large errors in regression tasks could be overcome [56].
- **Contfactuals for knowledge engineering.** The first of the considered frameworks (in chronological order) [93] offers explanations by reasoning abductively over the information extracted from a given knowledge base to answer a specific contrastive question. In this setting, an explanation is considered to be a consistent set of disjunctive literals for the explanation-seeking question. It is worth noting that the framework is not designed to provide explanations for ML algorithms.

TABLE 7. A classification of the contfactual explanation generators by explainability method.

Explainability method	Publications
Model-specific	[46], [48], [56], [61], [80], [84], [86], [91], [93], [94], [122], [128], [132], [138], [139], [145], [147], [149], [153], [155]–[158], [161]
Model-agnostic	[6], [49], [55], [67], [81], [100], [105], [129]–[131], [133]–[137], [140]–[144], [146], [148], [150]–[152], [154], [159], [160]

- **Contfactuals for planning.** Contfactual explanation generation appears highly relevant to sequential tasks in robotics such as automatic planning [94], [157], [161]. Moreover, some of the robotics-related frameworks found in reinforcement learning settings provide explanations for policies that a robot selects at a given time step [132], [156].
- **Contfactuals for recommendation.** Ghazimatin *et al.* propose a graph-based recommendation system in a counterfactual setup [84]. They obtain a counterfactual explanation by removing a minimal set of user actions so that the output recommendation changes.
- **Contfactuals for conflict resolution.** Mosca *et al.* introduce an argumentation-based framework for social network management [149]. They use contrastive explanations to answer critical questions about agent actions in the context of multi-user privacy conflict.

2) EXPLAINABILITY METHOD

All the frameworks generating contfactual explanations can be classified by their explainability method as either model-specific or model-agnostic. The former type of implementations is meant for explaining decisions of particular AI algorithms. The frameworks of the latter type generate explanations irrespective of the nature of the underlying algorithm. Table 7 presents the publications under study grouped in terms of the explainability method that they apply. The distribution of model-specific and model-agnostic explainability methods for generation of different types of contfactual explanation is shown in Fig. 10. Most frameworks deal with counterfactual model-agnostic methods.

- **Model-specific contfactual explanation generators.** Several model-specific frameworks generate counterfactuals to explain the output of decision trees [46], [80], [122], [155]. For instance, Fernández *et al.* [80] present a recursive algorithm which extracts counterfactuals in the form of contrast-class decision tree nodes. The relevance of the generated counterfactuals is then measured by calculating a variant of the Gower distance. The proposed metric penalizes the number of feature changes when traversing the tree so that sparsity is promoted. Alternatively, Sokol and Flash rely on the Manhattan distance measuring leaf-to-leaf distance in the tree to retrieve the most relevant counterfactuals [46], [155]. Designed specifically for decision trees, their "Glass-box" framework is argued to be easily extendable to capture the output of other logical (rule-based) models. Aguilar-Palacios *et al.* generate contrastive explanations using gradient boosted decision trees to forecast promotional sales [61]. The researchers make use of the weighted Euclidean distance to present the forecast as a contrast to the neighbouring vectorized promotions. Stepin *et al.* retrieve counterfactuals from a rule matrix where each rule is encoded in terms of all possible feature values [122]. Subsequently, the generated counterfactuals are ranked using a XOR-based distance to find the most relevant counterfactual pertinent to the given contrast class. This method is further extended to generating counterfactuals for fuzzy decision trees.

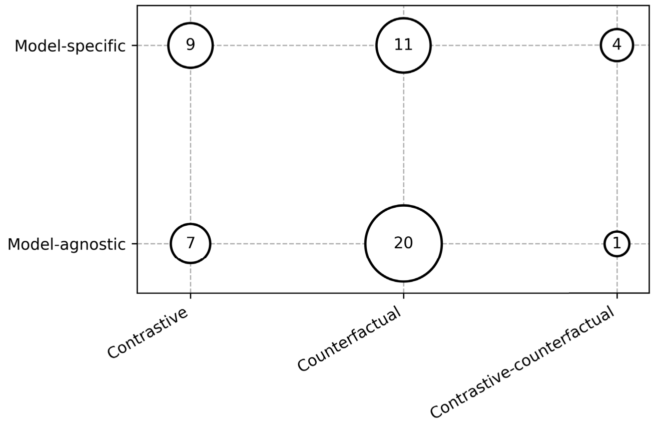


FIGURE 10. Numbers of frameworks grouped by explainability method with respect to the type of contfactual explanation generated.

A number of frameworks address specific properties of counterfactuals. Thus, Ustun *et al.* tackle the problem of actionability, i.e., constraining the generated counterfactuals in such a manner that the imposed changes "do not alter immutable features" and that they "do not alter mutable features in an infeasible way" [158]. To approach this problem, a mixed integer programming method is employed. Russell *et al.* adopt a similar approach to encompass continuous and discrete variables as well as the combination of the two [153]. The main focus of the work is however placed on assessing coherence and diversity of generated counterfactuals. In order to guarantee the coherence of the counterfactual data example used for explanation, an integer programming-based method is proposed. In addition, the generated counterfactual explanations are claimed to be diverse, as diversity constraints are applied iteratively to a set of candidate counterfactuals. However, this framework is limited to: (1) explaining predictions of only linear classifiers and (2) a simple structure of the textual explanation template.

A large number of frameworks are limited to explaining the output of particular models due to task-specific constraints. For instance, several explanation generators address computer vision tasks. Hendricks *et al.* bind an input visual image with a paired textual counterfactual explanation generated by a recurrent neural network [86]. In their framework, a number of candidate explanations (both image-relevant and non-relevant) are generated, paired, and ranked. The best counterfactual explanation is then selected to be the most class-specific to the counterfactual image while being the most relevant to the input image. Goyal *et al.* argue that their model is more faithful by design, as it generates visual explanations directly from "the target model based on the receptive field of the model's neurons" [139].

Two model-specific frameworks are found in the context of video processing. For instance, Akula *et al.* [128] present an empirical study where input video frames are paired with the corresponding AND-OR graphs, i.e., compositional recursively defined graph-based knowledge representations capturing contextual information. The explanations based on such graphs are passed on to human subjects to evaluate the contrastive answers to the predefined questions. Alternatively, Kanehira *et al.* train a post-hoc explanatory model to justify a video classifier's output [91]. A counterfactual explanation is, in this case, dependent on how likely a selected region in the given frame is classified positive and not negative, hence all such regions are scored and normalized.

In accordance with the findings in the previous Section IV-C1, contfactual explanations have a great potential for automatic planning-related tasks. Most explanation generators meant for planning-based tasks are model-specific due to the problem- and approach-specific restrictions preventing them from being used for other AI challenges. For instance, Kim *et al.* employ a Bayesian probabilistic model for generating contrastive explanations [94]. Thus, the framework operates on a pair of plan traces defined in terms of linear temporal logic templates. The problem of obtaining contrastive explanations is designed as a Bayesian inference problem, with the posterior distribution to be maximized defined as the probability of a contrastive explanation given a set of positive and negative plan traces. Conversely, Sreedharan *et al.* consider the task of automatic analysis of counterfactual explanations in their "Hierarchical Expertise-Level Modeling" framework [156]. A robot provides a user with a plan for the next action to take. Then, the robot expects the user to respond with a set of foils. The robot's task is then to convincingly refute the foils by offering a minimal explanation for why the foils are not acceptable under the given circumstances. In addition, Chakraborti *et al.* formulate the multi-model planning problem as a tuple consisting of the planner's model of the problem and the corresponding human approximation thereof [132]. As plan explicability is reformulated in terms of its comprehensibility by an end-user. The robot's model is adapted to the updates of human's model of the problem.

The problem of contrastive explanation generation for planning is also found to be framed in the reinforcement learning setting. For instance, Sukkerd *et al.* formulate the planning problem as the shortest stochastic path problem and develop the corresponding problem solver to obtain a contrastive explanation [157]. Hence, their objective is to find an optimal policy "that minimizes the expected cumulative cost of reaching a goal state over all closed policies". The explanation is believed to justify the rejection of the policies alternative to the optimal one. In addition, Zhao and Sukkerd explain an autonomous system's behaviour modeling it as a Markov decision process [161]. Thus, a contrastive explanation is presented as a product of the analysis of the optimal policy at the next time step and an opposing policy on the basis of the objective values.

- **Model-agnostic contfactual explanation generators.** A large number of model-agnostic frameworks treat contfactual explanation generation as an optimization problem in a post-hoc manner. Wachter *et al.* design a generic counterfactual explanation framework to find the closest point to the test data example [6]. Fixing the optimal set of weights of a trained classifier, the objective function minimizes the distance between the nearest data points of opposing classes. Note that counterfactual data points can be synthesized artificially. The researchers suggest the use of the Manhattan distance weighted by the inverse median absolute deviation to calculate the proximity of a counterfactual to the input data example.

  Another case of counterfactual explanation generation regarded as an optimization problem is the "Constrained Adversarial Examples" framework [148]. Adversarial examples that could serve as the basis for the counterfactual explanation of the output of deep learning models are searched for with the aim of minimizing the loss with respect to the attributes (features) between the original and counterfactual data examples. The researchers attempt to find the best counterfactual explanation by minimizing the number of attributes changed. Furthermore, the gradient direction is constrained to ensure the ethical adequacy of the explanation generated. Dandl *et al.* [134] formulate counterfactual search as a multi-objective optimization problem using a distance metric for mixed feature spaces aiming to obtain sparse and most plausible counterfactuals. Labaien *et al.* generate contrastive explanations for time-series data [141]. The explanation generation is considered a two-fold optimization problem of finding pertinent positives and negatives. Pawelczyk *et al.* make use of an autoencoder architecture for a pretrained classifier performing counterfactual search in the nearest neighbor style [151].

  Model-agnostic frameworks are largely found to use decision trees as part of the reasoning mechanism instead of explaining their output. In contrast to the model-specific frameworks operating on decision tree output, Guidotti *et al.* employ decision trees as part of reconstructing the reasoning behind any arbitrary classifier in a post-hoc fashion [140]. In their ''Local Rule-based Explanation'' framework, they generate a local neighbourhood for the given pre-classified data example using a genetic algorithm and subsequently train a decision tree on that newly obtained dataset to select a minimally distant foil within that local neighbourhood. Similarly, van der Waa *et al.* randomly sample or generate a data set in the neighbourhood local to the data point in question [159]. A decision tree is then trained to select the foil based on the minimum number of nodes between the original data point and the candidate foils. Furthermore, their ''Foil-Trees'' framework provides the methodological basis for perceptual-level contrastive explanation generation within the ''Perceptual-Cognitive Explanation'' framework [150]. Subsequently, the generated contrastive explanations are attributed to a specific group of users by means of ontology engineering at the cognitive level of the framework to make the explanations adaptive. In contrast, Martens and Provost argue that decision trees are an inadequate tool for representing, e.g., large documents [146]. Hence, they suggest the model-agnostic ''Search for Explanations for Document Classification'' algorithm for retrieving counterfactual explanations. However, it is only directly applicable to binary linear classifiers, whereas heuristics are proposed for non-linear models.

Several model-agnostic frameworks aim at measuring specific properties of contfactuals. Anjomshoae *et al.* [129] focus on contrastive explanations that maximize contextual importance and contextual utility. On the one hand, contextual importance measures the extent to which the input feature values affect the black-box algorithm's output. On the other hand, contextual utility testifies how favorable the values of the selected features are for a given decision. Thus, the context-based values are calculated for each feature used by a black-box model observing the changes in the output as the input varies across the range of all possible input values. Being based on model-agnostic and problem-independent concepts, this framework is shown to be universally applicable to various classification and regression algorithms. However, the scalability of such an algorithm is limited to the use-cases operating on a small number of features. A similar limitation is observed due to possibly high variability of the input. Laugel *et al.* raise the issue of justification for counterfactual explanation [144]. They argue that a synthesized counterfactual data point must be connected to the training data. Counterfactuals are selected from a local neighbourhood circling around the test example with the radius of the distance to the closest correctly predicted data point of a contrast-class. The candidate counterfactuals are then clustered, as the initial local neighbourhood is updated to become a more extensive hyperspherical layer, until it can no longer be extended. Laugel *et al.* [100] enhance the work on justified counterfactual explanations. They argue that the distance from the test instance to a counterfactual does not sufficiently measure counterfactual's relevance, as the counterfactual in question may appear disconnected from the ground-truth data. Thus, a counterfactual is deemed justified if it can be connected to an associated ground-truth data instance without crossing the decision boundary. Fernández *et al.* introduce the notion of counterfactual sets to enhance counterfactual diversity [81]. They explain random forest predictions by fusing different tree predictors so that the resulting counterfactual set contains the most relevant counterfactual. The other neighboring counterfactuals serve to diversify the output explanation. Mothilal *et al.* are also concerned with counterfactual diversity [105]. They design a loss function with a diversity metric over the generated counterfactuals to provide end-users with multiple relevant counterfactual explanations. Kusner *et al.* propose a causal model to assess the so-called ''counterfactual fairness'' [55]. It is worth noting that counterfactuals are presented in the form of conditional distributions and not structural equations despite the fact that the causal model employed follows Pearl's formalism [37].

Similarly to the model-specific frameworks, numerous model-agnostic explanation generators are found to be task-specific. In computer vision-related classification tasks, Chang *et al.* find the smallest region in the image whose substitution would change the classifier's prediction [133]. They employ a generative model to construct a saliency map while masking the other regions of the input image. Similarly, Dhurandhar *et al.* address an optimization problem over a perturbation variable to produce a contrastive explanation for the image classification task [135]. However, the proximity of the selected counterexample to the test point is, in this case, guaranteed by using an autoencoder.

3) OUTPUT REPRESENTATION

The considered frameworks output contfactual explanations in several ways. Depending on the problem considered, contfactuals are presented in the form of: (1) intervals or specific values of the appropriate feature values whose alteration would have changed the output (i.e., numerical or feature-based output); (2) single- or multiple-sentence coherent text (i.e., linguistic output); (3) specific regions in the input image (i.e., visual output); or (4) a multi-modal combination of (some of) the above (see Table 8). As depicted in Fig. 11, most frameworks focus on numerical counterfactual output.

- **Numerical (feature-based) contfactual explanation.** Numerical values (or intervals of values) associated to the most relevant features usually explain the behavior of AI algorithms. They can be represented as logical formulas [67], [93], [94] or in tabular form reflecting necessary changes to affect the decision [55], [56], [61],

TABLE 8. A classification of the contfactual explanation frameworks by output representation.

Output representation	Publications
Numerical (feature-based)	[55], [56], [61], [67], [80], [81], [93], [94], [100], [105], [132], [136], [140], [147], [148], [151], [152], [154], [158], [160]
Linguistic	[6], [48], [84], [122], [128], [131], [146], [149], [153], [156], [157], [159], [161]
Visual	[49], [130], [133]–[135], [139], [141], [142], [144]
Multi-modal	[46], [86], [91], [129], [137], [138], [143], [145], [150], [155]

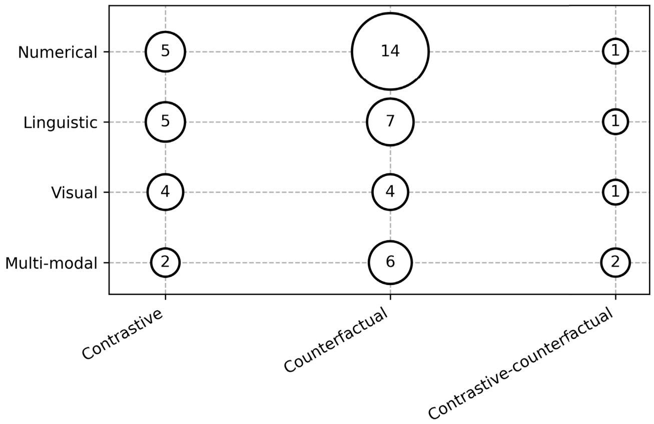


FIGURE 11. Numbers of frameworks grouped by output representation with respect to the type of contfactual explanation generated.

[81], [105], [136], [147], [148], [151], [154], [158], [160]. They can be extracted from interpretable feature-value pairs as a result of pruning in the search space [132]. In addition, they can replicate the internals of the classifier's structure, e.g., in the form of decision tree nodes or rules [80], [140], [152].

- **Linguistic contfactual explanation** is a piece of grammatical single- or multiple-sentence text in natural language. Single-sentence textual explanations combine a textual description with explicitly stated numerical feature values [6]. Such explanations suggest feature-value based instructions [48], [122], [131] or alternative actions for a possible output change [84], [149]. They also answer end-user's inquiries with respect to the automatic decision in question [128], [159]. In contrast, multiple-sentence explanations provide end-users with specific details for the given decision [153], [156], [157], [161] or explain multiple decisions at once [146].
- **Visual contfactual explanation.** On the one hand, visual explanations for non-visual input data (i.e., datasets containing continuous or categorical feature-value pairs) plot feature-value pair dependencies [134], [141], [142]. On the other hand, visual input data (i.e., images) are associated with saliency maps [133] or explained by contrastive patterns between the given data example and that of an opposing class in one iteration [130], [144] or a series thereof [49]; by depicting critical regions absent in the input data example that determine what lacks in the image to be classified differently [135]; or by visualizing spatial regions associated to data examples of opposed classes [139].
- **Multi-modal contfactual explanation** is a combination of numerical and/or linguistic and/or visual explanations. Multi-modal explanations are claimed to enhance human-robot interaction [46]. They are often selected to be the most appropriate where the problem addressed is concerned with pairing a computer vision problem with a natural language processing task such as object detection and language grounding [86]. In addition, visual-linguistic explanations identify counterfactuality in videos [91] and allow for dialogic interaction [150]. Such hybrid counterfactuals (in terms of their output representation) may as well complement each other while addressing the same task. For instance, Gomez *et al.* visualize the generated explanations in the form of bar plots combining them with explicitly stated numerical values [138]. Alternatively, Liu *et al.* combine feature importance bar plots with visual input and output [145]. While a textual explanation summarizes the degree of importance of the selected features, a visual explanation may present contextual in-method metrics that justify the classifier's reasoning [129]. Contfactual explanations, as a mixture of tabular and visual output representations, appear also in an augmented reality framework [137]. Nevertheless, explanations of different modalities are not necessarily merged. To ensure the universality of the proposed approaches, specific feature values are presented for tasks with datasets containing only continuous features, i.e., where the same method is used to output images for a handwritten digit classification problem [143]. Finally, interaction with users can be enhanced by means of voice-based explanations combined with textual explanations [46], [155].

4) EVALUATION METHOD

Evaluation of generated contfactual explanations is an issue of main concern. Unfortunately, despite an increasingly expanding use of contfactual explanations, no uniform set of evaluation methods has been adopted so far. Hence, it is worth taking a look at evaluation methods from other generation-oriented sub-areas of AI. For instance, it is common to distinguish between intrinsic and extrinsic evaluation methods in natural language generation [163]. Intrinsic evaluation implies assessing the performance of a natural language generation system (or its modules) as an isolated unit. In contrast, extrinsic (task-based) methods are designed to estimate how successfully the system performs with respect to an external task. In addition, Gatt and Krahmer make a distinction between "objective" (automatic, corpus-based) and "subjective" (human judgements) metrics [164]. Objective metrics include (but are not limited to) precision- and/or recall-oriented scores, number of insertions/deletions/substitutions,

TABLE 9. A classification of the contfactual explanation generators by evaluation method.

Evaluation method	Publications
None	[6], [46], [49], [67], [93], [129]–[131], [136], [145], [149], [150], [153], [155], [157], [158], [161]
Subjective	[56], [128], [137]
Objective	[55], [61], [80], [81], [86], [91], [94], [100], [105], [122], [132]–[135], [138]–[144], [146]–[148], [151], [152], [154], [156], [159], [160]
Hybrid	[48], [84]

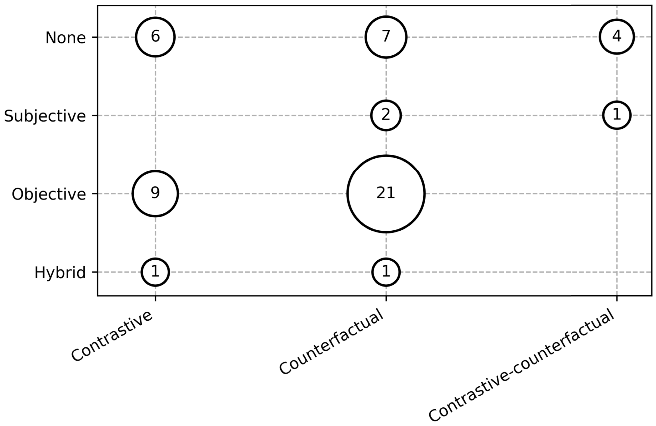


FIGURE 12. Numbers of frameworks grouped by evaluation method with respect to the type of contfactual explanation generated.

etc. In turn, subjective metrics measure readability, accuracy, relevance of the generated text, as perceived by humans. Thanks to their methodological universality, they can be extrapolated to other (non-linguistic) modalities of generated explanations (e.g., numerical, visual, or multi-modal) and are therefore used to form the basis of the evaluation method classification in this review. It is worth noting that all the considered frameworks are evaluated by means of intrinsic (either subjective or objective) metrics. Hence, we only make a clear distinction between subjective and objective evaluation methods in this study (see Table 9 for details). A distinction between the use of different types of evaluation metrics can be seen in Fig. 12. It is easy to appreciate how most frameworks deal with objective evaluation of counterfactuals. Let us give further details below, regarding the four groups of publications in Table 9.

- **No evaluation details provided.** 17 out of 52 (32.69%) of the considered publications do not evaluate their frameworks, i.e., neither automatic metrics for contfactual explanation generation are suggested nor a human evaluation survey is presented in such publications. However, whereas certain publications do not provide any specific evaluation method, some do stress that human evaluation should be encouraged to estimate the quality and effectiveness of the generated counterfactuals [46], [150], [153].
- **Subjective evaluation.** The subjective methods include human preferences for certain types of contfactual explanation over others. For instance, Akula *et al.* show that that contrastive explanation-seeking questions are in general better answered by means of contfactual explanations [128]. They classify contrastive questions in the following 10 categories suggesting the template questions for arbitrary objects x, x_1, and x_2 (all being of some class X) and y, y_1, and y_2 (all being of some other class Y):

  - WH-X: "Why x rather than not x?";
  - WH-X-NOT-Y: "Why x rather than y?";
  - WH-X1-NOT-X2: "Why x_1 rather than x_2?";
  - WH-NOT-Y: "Why not y?";
  - NOT-X: "Is it x rather than not x?";
  - NOT-X1-BUT-X2: "Is it x_1 rather than x_2?";
  - NOT-X-BUT-Y: "Is it x rather than y?";
  - DO-X-NOT-Y: "What if it is x rather than y?";
  - DO-NOT-X: "What if it is not x?".
  - DO-X1-NOT-X2: "What if it is x_1 and not x_2?"

  It is worth noting that 6 out of 10 question types (WH-NOT-Y, NOT-X, NOT-X1-BUT-X2, NOT-X-BUT-Y, DO-NOT-X, and DO-X1-NOT-X2) matched with automatically generated contfactuals are shown to be highly preferred to factual explanations.

  In addition, Ferrario *et al.* propose an augmented reality-based setting to favor interactivity and facilitate explaining ML algorithm output to non-experts [137].

  However, the two aforementioned studies [128], [137] lack an evaluation of the quality of the generated contfactual explanations themselves.

  Lucic *et al.* asked 75 subjects to judge interpretability, actionability, and trustworthiness of the generated contfactual explanations [56]. They concluded contfactual explanations are highly interpretable and actionable. In addition, they help users understand why the model makes large errors while solving a regression problem but do not support users' trust in the model's output.

  In addition, Hendricks *et al.* provide results of human evaluation for the generated explanations [86]. However, these only include evaluations for factual explanations and are therefore excluded from the taxonomy group being discussed.
- **Objective evaluation.** A majority of the researchers propose objective (automatic) methods for evaluating automatically generated contfactuals. A number of the frameworks are evaluated by means of accuracy-based metrics [61], [91], [94], [159]. Kanehira *et al.* propose one accuracy-based evaluation metric for visual and linguistic explanations each: negative class accuracy and concept accuracy, respectively [91]. Negative class accuracy estimates the quality of the visual explanation as the ratio of the probability of the contrast class after the image region in question is masked out. In turn, concept accuracy estimates how compatible the output linguistic explanation is to its visual counterpart. It is calculated as the intersection over union between a given region

and all bounding boxes in the image. Kim *et al.* define the domain-specific accuracy for the automatic planning problem of unique contrastive explanations as a sum of the number of traces in a positive set of traces satisfying a contraint for a contrastive explanation and those of the negative set where the constraint is unsatisfied over all plan traces [94]. The consistency of output explanations is otherwise shown by measuring their accuracy on the basis of mean error, mean absolute error, and mean absolute percentage error [61].

An extensive number of evaluation methods are found to be strictly task- or approach-specific. Hendricks *et al.* measure word detection (i.e., which words are not image relevant by holding out one word at a time from the sentence to determine the least relevant word in the explanation) and word correction (i.e., a number of replacements of the foiled word with words from a set of target words) [86]. Similarly, Martens and Provost estimate explanation complexity by calculating the average number of words in the shortest explanation and problem complexity according to the overall number of generated explanations [146]. Fernández *et al.* evaluate the relevance of the generated counterfactuals measuring a Gower distance-based metric in comparison with the number of feature changes and minimum distance (in terms of decision tree nodes) between the leaves in the given decision tree classifier [80]. Van der Waa *et al.* evaluate the generated explanations by means of such model-specific metrics as the average length of the explanation in terms of decision tree nodes and the F_1-score of the foil-tree on the test set compared to the model's output [159]. Kusner *et al.* estimate counterfactual fairness on the basis of the density of the predicted data for their causal models [55]. Laugel *et al.* claim that understandability of the generated explanations can be estimated by means of their sparsity defined as the number of non-zero coordinates of the explanation vector [143]. Moore *et al.* measure the number of solutions, the distances to the nearest training set data points, and the transferability of the generated counterfactuals to other datasets and classifiers [148]. Sreedharan *et al.* calculate the number of predicates that are used to generate the model lattice [156]. Similarly, Chakraborti *et al.* calculate the number of nodes in the search space remaining after pruning [132]. Goyal *et al.* report how often the discriminative regions lie inside the test data example segmentations as well as relevant specific key regions [139]. Labaien *et al.* calculate the number of changes to switch from the original to the selected contrastive sample following the dataset constraints [141]. To estimate faithfulness of the generated counterfactuals, Pawelczyk *et al.* suggest calculating the so-called degree of difficulty of a counterfactual suggestion to measure how costly it is to achieve the state of the given suggestion [151]. Aiming to provide realistic counterfactuals, Sharma *et al.* introduce the counterfactual explanation robustness-based score defined as the expected distance between the input instances and their corresponding counterfactuals [154]. In addition, the generated counterfactuals are inspected in terms of fairness which is calculated as the expected distance between the input and a counterfactual over distinct values for a specified feature set. Merrick and Taly evaluate output explanations in terms of mean feature attributions to show the importance of relevant references [147]. Gomez *et al.* evaluate counterfactuals in terms of data distribution, feature importance, as well as possible and actionable changes to the input [138]. Dandl *et al.* use the hypervolume indicator metric to estimate the quality of the estimated Pareto front during counterfactual search [134]. In addition, Chang *et al.* measure the weakly supervised localization error for an image detection task – the intersection-over-union ratio over 0.5 with any of the ground truth bounding boxes and the saliency metric, i.e., "the log ratio between the bounding box area and the in-class classifier probability after upscaling" [133].

Several metrics can be extended to be applied to other approaches. Lash *et al.* estimate how much the probability of a given prediction reduces given a feature perturbation as determined by a contrastive explanation [142]. Dhurandhar *et al.* employ the concept of pertinent positives (i.e., "factors whose presence is minimally sufficient in justifying the final classification" [135]) and pertinent negatives (i.e., "factors whose absence is necessary in asserting the final classification") to evaluate factuals and counterfactuals, respectively, for a given classification task. Both types of evaluation methods highlight the features supporting evidence as formulated in the contrastive explanation on the basis of the values that a perturbation variable takes on. Fernández *et al.* evaluate counterfactuals in terms of the average of the pairwise distances based on the feature type and the percentage of valid counterfactuals [81].

Mothilal *et al.* stress that counterfactuals should be evaluated in terms of validity (i.e., whether a generated counterfactual really leads to a different outcome), proximity (i.e., feature-wise distance between the original and counterfactual samples), sparsity (i.e., number of features differing in the original and counterfactual samples), and diversity (i.e., feature-wise distance between each pair of counterfactuals) [105]. Similarly, Stepin *et al.* calculate factual and counterfactual explanation length to estimate conciseness of the generated explanations [122]. They also compute the number of counterfactuals and their best minimal distance to the factual explanation to assess the relevance of counterfactuals. Rajapaksha *et al.* consider coverage (as an indicator of representativeness of a rule for a given dataset), confidence (i.e., the percentage of instances in the dataset which contain the consequent and antecedent together over the number of instances which only contain the antecedent), lift (i.e., an association between antecedent and consequent), leverage (i.e., the observed frequency between the antecedent and consequent), and the number of features in explanation for evaluating their framework against other rule-based methods [152]. Also, White and Garcez reintroduce fidelity to the underlying classifier on the basis of distance to the decision boundary [160].

In addition, some of the model-agnostic frameworks [140], [159] allow for measuring how well the output of black boxes (i.e., actual output to be explained) and grey boxes (i.e. interpretable intermediate predictors) mimic the local neighbourhood (i.e., fidelity) and the data example to be explained (i.e., hit). Laugel *et al.* measure how justified counterfactuals are by averaging a binary score (one if the explanation is justified following the proposed definition, zero otherwise) over all the generated explanations [100], [144].

It is worth noting that the run-time of explanation generation algorithms is reported in addition to the evaluation metrics for several frameworks [132], [139], [146], [152], [156], [159].

- **Hybrid evaluation.** Two frameworks are evaluated in terms of both automatic metrics and human judgments. Ghazimatin *et al.* calculate explanation length to discuss comprehensiveness of explanations as well as estimate their usefulness and credibility by surveying 500 subjects [84]. In addition, Le *et al.* compute fidelity, conciseness, information gain, and influence [48]. Automatic metrics are complemented with a user study on intuitiveness, friendliness, comprehensibility, and understandability of generated explanations.

D. LINKS BETWEEN THEORETICAL AND PRACTICAL CONTRIBUTIONS TO CONTFACTUAL EXPLANATION GENERATION (ANSWER TO RQ_3)

We find that only few of the existing computational frameworks are grounded on theories of contfactual explanation. Indeed, only 13 out of 113 studies (11.50%) were present in both of the pools of primary studies related to RQ_1 and RQ_2. Table 10 summarizes the characteristics of such theoretically grounded contfactual explanation generation frameworks.

Moreover, only 3 out of the 13 (23.08%) studies interpret the insights from the theoretical foundations to propose their own contfactual explanation definition for problem-oriented purposes. Kean states that "explanation in artificial intelligence is based on the inference of deduction" [93]. He complements a deductive evidence-based explanation with a redefined abductive contrastive explanation drawing parallels to the "inference to the best explanation" [103]. He models Lipton's theoretical framework distinguishing two types of contrastive explanation: non-preclusive (i.e., non-restrictive) and preclusive. The key aspect distinguishing the two types of contrastive explanation is in regard to how a model explains the contrast given an explanation-seeking question. Thus, a non-preclusive contrastive explanation is "irrelevant to the model of explaining the contrast" being "necessary in the model of explaining the question". On the contrary, a preclusive contrastive explanation is assumed to be restricted by a negated model of the contrast.

Aguilar-Palacios *et al.* [61] refer to Lipton's definition of contrastive explanation [12]. Referring to Pearl [37], Bertossi redefines causal explanation in the context of XAI to be "a set of feature values for the entity under classification that is most responsible for the outcome" [67].

The rest of works redefine contfactual explanation on the basis of the problem-specific constraints without explicitly referring to the theoretical foundations described in Section IV-B. Driven by the task of automatic planning, Kim *et al.* define a contrastive explanation to be a constraint satisfied by a specific set plan traces [94]. Fernández *et al.* define a counterfactual to be a set of feature changes that turn the given data example to be classified differently [80]. Whereas this definition is applicable to the classification problem in general, the applicability of the framework is restricted to decision trees only. Similarly, Hendricks *et al.* explain visual concepts for the image classification task on the basis of the so-called counterfactual evidence (i.e., an attribute discriminative enough for another class of objects in the image absent in the given image) [86]. Ghazimatin *et al.* [84] define a counterfactual on the basis of their model's internal structure: an explanation is deemed counterfactual if after removing the edges from the recommendation graph, the user receives a different top-ranked recommendation. In addition, Kanehira *et al.* only specify the linguistic form of a counterfactual explanation without defining it explicitly [91].

Finally, there is a number of marginal interpretations of contfactuals among the RQ_3-related studies. Laugel *et al.* denote a counterfactual as a specific data instance that changes the algorithm's prediction [100]. Poyiadzi *et al.* denote a counterfactual to be "the new state of the object" [49]. Nevertheless, the most commonly acceptable definition of a contfactual in the observed RQ_3-related studies states that a contfactual explanation is a set of minimal feature modifications that makes the model change the prediction [81], [105], [122].

V. DISCUSSION

The findings show that a large body of research has been elaborated on theoretical accounts of contrastive, counterfactual, and contrastive-counterfactual explanation. In addition, the topic has recently attracted attention from researchers in XAI (see Fig. 13). Thus, 50 out of the 52 considered state-of-the-art contfactual explanation generation frameworks (96.15%; RQ2) have been developed from 2017 to 2020.

The results of the study, in relation to RQ_1, show that a majority of the considered theoretical accounts of contfactual explanation (49 out of 74; 66.22%) speculate on the causal nature of explanation. However, whereas most researchers in philosophy of science have mainly used the concept of counterfactuality to explain causal relations between entities in question, causal inference is poorly addressed in the

TABLE 10. A summary of characteristics of theoretically grounded computational frameworks for contfactual explanation generation. CT stands for contrastive explanation, CF means counterfactual explanation, and CT-CF is contrastive-counterfactual explanation.

Reference	Type of contfactual explanation			Relatedness to causality			Computational framework properties			
	CT	CF	CT-CF	Causal	Non-causal	Hybrid	AI problem	Explainability method	Output representation	Evaluation method
Aguilar-Palacios et al. (2020) [61]	✓			✓			regression	model-specific	numerical	objective
Bertossi (2020) [67]		✓		✓			classification	model-agnostic	numerical	none
Fernández et al. (2019) [80]		✓			✓		classification	model-specific	numerical	objective
Fernández et al. (2020) [81]		✓			✓		classification	model-agnostic	numerical	objective
Ghazimatin et al. (2020) [84]		✓		✓			recommendation	model-specific	linguistic	hybrid
Hendricks et al. (2018) [86]		✓			✓		classification	model-specific	multi-modal	objective
Kanehira et al. (2019) [91]		✓			✓		classification	model-specific	multi-modal	objective
Kean (1998) [93]	✓			✓			knowledge engineering	model-specific	numerical	none
Kim et al. (2019) [94]	✓			✓			planning	model-specific	numerical	objective
Laugel et al. (2020) [100]		✓			✓		classification	model-agnostic	numerical	objective
Mothilal et al. (2020) [105]		✓				✓	classification	model-agnostic	numerical	objective
Poyiadzi et al. (2020) [49]			✓		✓		classification	model-agnostic	visual	none
Stepin et al. (2020) [122]		✓			✓		classification	model-specific	linguistic	objective

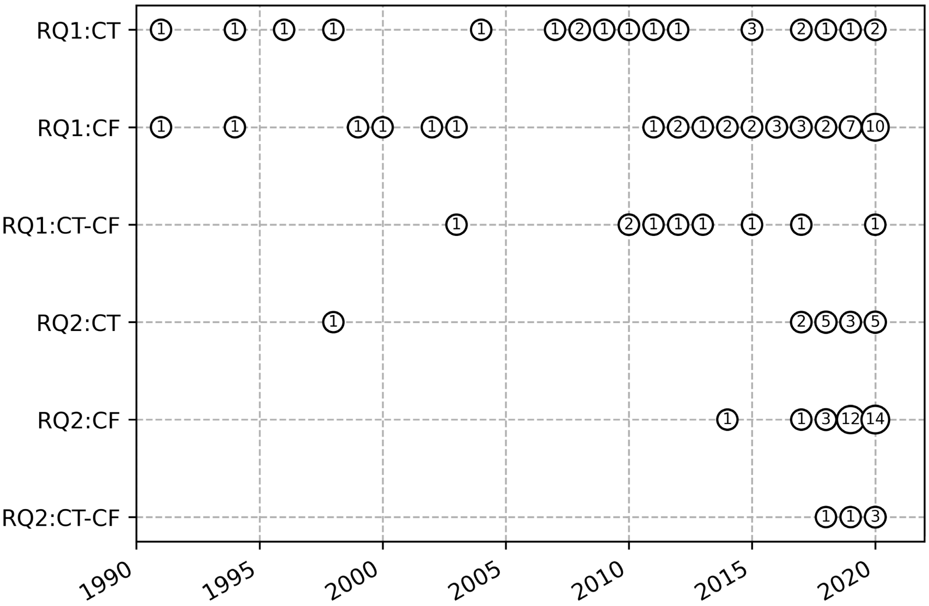


FIGURE 13. Numbers of theoretical and computational contfactual explanation generation frameworks grouped by year of publication. For illustrative purposes, only the studies published from January 1990 to September 2020 are displayed.

pool of publications concerning computational frameworks of contfactual explanation generation. Kean directly refers to a causal account of contrastive explanation to address the problem of abductive reasoning [93]. In addition, Lucic *et al.* [56] explicitly specify that their method is based on previous work on philosophical accounts of contrastive explanation [12] as well as on causal attribution [165], [166]. Kusner *et al.* [55] make use of causal inference models and the corresponding tools provided by Pearl [37]. They assess how discriminatory the generated counterfactual explanations are for the given classification task output. On the other hand, Bertossi redefines the concept of causal explanation [67]. Following a causal account of contfactual explanation, Fernández *et al.* introduce weakly causal irreducible counterfactual explanation [136]. As most of the current ML-tasks are centered around singling meaningful patterns out from unstructured data, establishing causal relations appears to be among the AI problems that are yet to attract global attention. This partly explains why most of the modern contfactual explanation generators focus on feature perturbation when searching for the most relevant contfactuals and not establishing causal relations between them.

At the same time, the other computational frameworks are primarily non-causal. Furthermore, a strikingly low number of such frameworks appear to be rooted in theoretical accounts of explanation due to an imbalance in favor of causality-oriented theoretical accounts. However, the amount of publications for RQ_3 may be somewhat misleading, as contfactual explanations are often redefined without specifically referring to theoretical contributors in explanation. This is hypothesized to be due to the problem-specific necessities ignored in previous theoretical works from different branches of science. For instance, Laugel *et al.* emphasize that the minimal perturbations required to change the predicted class of a given observation enable a user to understand which features locally impact the prediction and therefore how it can be changed [144]. In this interpretation, counterfactual explanations are conceptually most similar to counterfactuals as defined by Lewis [36]. Indeed, while the concept of ''the closest possible world'' does not always appear in the related publications, it turns out to be implicitly wired in almost all works. Instead, other considered frameworks do not appeal directly to any of the theoretical accounts of explanation addressed in RQ_1. Hence, it is worthwhile taking a look at how contrastive and counterfactual explanations are redefined in the frameworks not appearing in Section IV-D.

In general, a consensus among researchers has been observed on how contfactual explanations are defined irrespective of the theoretical framework proposed by individual authors. In humanities and social sciences, a major difference in various contfactual theories of explanation is observed to concern the causal nature of explanation and its extrapolation to non-causal cases. In computer science and AI, notions of counterfactual and contrastive explanation are found to be the most dissimilar when applied to non-overlapping problems. Nevertheless, the corresponding line of research in AI makes little use of the rich theoretical background accumulated by now. While some rule-based approaches used in expert systems are justified theoretically (e.g., see [93]), newly emerging tasks present novel challenges for theorists and call for updating the theories developed so far.

More precisely, ML-specific contfactual explanations are designed to answer the question: ''Why was the outcome Y observed instead of Y'?'' [148]. Anjomshoae *et al.* define finding a contrastive explanation as ''contrasting instance against the instance of interest'' [129]. Fernández *et al.* specify that a counterfactual is generally regarded as a hypothetical instance similar to an example whose explanation is of interest but with a different predicted class [80]. Also, counterfactual explanations ''show a difference in a particular scenario that causes an algorithm to change its mind'' [155].

As a majority of the considered frameworks are designed for tackling classification problems, contfactual explanations operate on the notion of a contrast-class (e.g., see [155]) answering the question: ''How is the prediction altered when the observation changes, given a classifier and an observation?'' Furthermore, these changes are normally expected to be minimal [143].

However, certain application domains as well as the selection of a classifier require researchers to redefine contfactuals imposing task-dependent constraints, which makes it nearly impossible to connect them to any of the existing theories of contfactual explanation. For instance, Martens and Provost define a contrastive explanation for a document classification task to be a minimal set of words such that removing all words within this set from the document changes the predicted class from the class of interest [146]. In addition, Guidotti *et al.* reformulate a counterfactual to be a set of split conditions of a decision tree describing the minimal number of changes in the feature values of a test example [140]. In image classification, it is found necessary to detect specific regions in the given test image. For this type of tasks, the contrastive explanation-seeking question is formulated as follows: ''Which parts of the image, if they were not seen by the classifier, would most change its decision? or which inputs, when replaced by an uninformative reference value, maximally change the classifier output?'' [133], [139]. Similarly, Dhurandhar *et al.* ask what should be minimally and necessarily present and absent in the given image to justify its classification [135]. Alternatively, counterfactuals are viewed as ''solutions that are guaranteed to map back onto the underlying data structure'' [153]. Redefined contrastive explanations are also found in the domain of robotics and automatic planning. According to Sukkerd *et al.*, a contrastive explanation answers the question why a generated behavior is optimal with respect to the planning objectives of an autonomous system [157]. Alternatively, contrastive explanations are used to answer why-not questions about the system's behavior in which the consequences of the counterfactuals in question are pointed out [161].

In addition, the nature of the explanation-seeking questions for computational frameworks deserves further discussion. Sokol and Flash distinguish three types of counterfactual explanations: (1) a plain counterfactual (''Why?'') generated as the shortest possible class-contrastive counterfactual; (2) a counterfactual explanation not conditioned on the indicated feature(s) (''Why despite?''); and (3) a (partially) fixed counterfactual explanation (''Why given?'') which is conditioned on a predetermined set of features [46]. Hilton proposes different types of contrastive questions such as: (1) ''Why X rather than not X?''; (2) ''Why X rather than the default value for X?'' and (3) ''Why X rather than Y?'' [166]. Following this distinction, Akula *et al.* extend this set of contrastive questions to formulate ten contrastive question types for counterfactual explanation generation [128] (see Section IV-C4). Alternatively, only linguistic templates for such explanations are defined without any theoretical grounding in accordance with any accounts described in Section IV-B. For instance, Sokol and Flash define a counterfactual explanation to be a piece of text following the template: ''The prediction is ⟨prediction⟩. Had a small subset of features been different ⟨foil⟩, the prediction would have been ⟨counterfactual prediction⟩ instead'' [155].

Remarkably, a wide range of frameworks favor automatic evaluation methods. Thus, they rarely place the end-user in the center of the explanation evaluation process. However, we find an increasing number of interactive frameworks that attempt not only to present the automatically generated explanations to the end-user but also interact with him or her [46], [150], [155]. Promoting interactivity (e.g., by engaging the end-user to participate in an explanatory dialogue with the system) is expected to make explanation social and further increase user's trust in the system's output.

VI. CONCLUDING REMARKS

In this work, we made two main contributions. First, we provided readers with an overview on theoretical accounts of contrastive, counterfactual, and contrastive-counterfactual explanation as well as frameworks of automatic generation thereof. This overview was based on a systematic literature review. This research methodology allowed us to carry out an unbiased reproducible study from an interdisciplinary topic-specific search in reputable and trustworthy sources. Second, we proposed a two-level taxonomy of the aforementioned types of explanation with the aim of providing a well-established tool that allows us to jointly analyze different proposals in this research field. As a result, this taxonomy facilitates the comparison of approaches and publications. We expect that it raises awareness in researchers in the community about main categories (definitions, practical frameworks, etc.) and subcategories (causal, non-causal, etc.) in the taxonomy. Moreover, we hope that it helps properly characterize the body of work and leverages a deeper collaboration and citation among similar related work.

The findings allow us to draw the following remarks. Contrastive and counterfactual explanations are found to be in great demand across various sub-fields of AI. Mainly applied to a wide range of tasks in computer vision and natural language processing, they present a powerful tool that enhances human-machine interaction and allows for further personalization of the output generated by various AI algorithms, including ML-based black-box algorithms.

In our systematic review, we introduced the term "contfactual explanation" to unify the aforementioned families of explanation to subsequently analyse the existing approaches to them from three points of view.

First, we investigated theoretical accounts of contfactual explanation to infer the similarities and differences among the existing theoretical approaches. Contfactuals are found to address both causal and non-causal dependencies. Hence, being a significant challenge, unification of causal and non-causal explanatory engines within a contfactually-driven framework opens new perspectives for the XAI community.

Second, existing computational frameworks for contfactual explanation generation have been inspected. Despite the fact that the notion of contfactual explanation is found to be highly task- and domain-specific, the most commonly accepted definition of a contfactual explanation in the context of XAI refers to a minimal set of feature modifications that makes the model change the prediction. A crucial shortcoming relevant to the inspected frameworks is a lack of standardization with respect to the evaluation methods. While designing a set of standard evaluation metrics is particularly complicated due to a different nature of the tasks that these explanations serve, this is hypothesized to be among major factors preventing researchers from making faster progress in solving the problem of contfactual explanation generation, as it complicates a fair evaluation of newly developed frameworks against the state-of-the-art equivalents. Furthermore, as automatically generated explanations are meant to be user-oriented, more effort is needed to include end-users in the process of assessing the generated explanations.

Third, a synergy between the related theories and computational frameworks has been investigated. We find that a gap between philosophical accounts of contfactual explanation to scientific modeling and ML-related concepts makes the theoretical frameworks poorly applicable to XAI. In addition, the existing methodological differences affect greatly the definition of contfactual explanation found across various approaches. In fact, definitions vary depending on domains of science and even approaches used for solving specific tasks.

Finally, we believe a joint interdisciplinary effort of researchers from both humanities and computational sciences can be particularly fruitful for further progress in contfactual explanation generation.

ACKNOWLEDGMENT

Ilia Stepin is an *FPI* Researcher (PRE2019-090153). Jose M. Alonso is a *Ramon y Cajal* Researcher (RYC-2016-19802). Alejandro Catala is a *Juan de la Cierva* Researcher (IJC2018-037522-I).

REFERENCES

[1] M. Attaran and P. Deb, "Machine learning: The new 'big thing' for competitive advantage," *Int. J. Knowl. Eng. Data Mining*, vol. 5, no. 4, pp. 277–305, 2018.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': Explaining the predictions of any classifier," in *Proc. 22nd Int. Conf. Knowl. Discovery Data Mining (ACM SIGKDD)*. New York, NY, USA: Association Computing Machinery, 2016, pp. 1135–1144.

[3] N. Wang, D. V. Pynadath, and S. G. Hill, "Trust calibration within a human-robot team: Comparing automatically generated explanations," in *Proc. 11th ACM/IEEE Int. Conf. Hum.-Robot Interact. (HRI)*, Piscataway, NJ, USA: IEEE Press, Mar. 2016, pp. 109–116.

[4] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019.

[5] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable agents and robots: Results from a systematic literature review," in *Proc. 18th Int. Conf. Auto. Agents MultiAgent Syst. (AAMAS)*. Richland, SC, USA: International Foundation Autonomous Agents Multiagent Systems, 2019, pp. 1078–1088.

[6] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard J. Law Technol.*, vol. 31, no. 2, pp. 841–887, 2018.

[7] G. Schurz, "Scientific explanation: A critical survey," *Found. Sci.*, vol. 1, no. 3, pp. 429–465, Sep. 1995.

[8] J. C. Pitt, *Theories of Explanation*. Oxford, U.K.: Oxford Univ. Press, 1988.

[9] C. B. Cross, "Explanation and the theory of questions," *Erkenntnis*, vol. 34, no. 2, pp. 237–260, Mar. 1991.

[10] T. Lombrozo, "Explanation and abductive inference," in *Oxford Handbook of Thinking and Reasoning*. Oxford, U.K.: Oxford Univ. Press, 2012, pp. 260–276.

[11] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2019.

[12] P. Lipton, "Contrastive explanation," *Roy. Inst. Philosophy Suppl.*, vol. 27, pp. 247–266, Mar. 1990.

[13] C. V. F. Bas, *The Scientific Image*. Oxford, U.K.: Oxford Univ. Press, 1980.

[14] F. Sørmo, J. Cassens, and A. Aamodt, "Explanation in case-based reasoning—Perspectives and goals," *Artif. Intell. Rev.*, vol. 24, no. 2, pp. 109–143, 2005.

[15] N. J. Roese, "Counterfactual thinking," *Psychol. Bull.*, vol. 121, no. 1, pp. 133–148, 1997.

[16] R. M. J. Byrne, "Mental models and counterfactual thoughts about what might have been," *Trends Cognit. Sci.*, vol. 6, no. 10, pp. 426–431, Oct. 2002.

[17] S. Chin-Parker and A. Bradner, "A contrastive account of explanation generation," *Psychonomic Bull. Rev.*, vol. 24, no. 5, pp. 1387–1397, Oct. 2017.

[18] R. Wenzlhuemer, "Counterfactual thinking as a scientific method," *Historical Social Res.*, vol. 34, no. 2, pp. 27–56, 2009.

[19] R. Folger and C. Stein, "Abduction 101: Reasoning processes to aid discovery," *Hum. Resour. Manage. Rev.*, vol. 27, no. 2, pp. 306–315, Jun. 2017.

[20] C. Boutilier and V. Beche, "Abduction as belief revision," *Artif. Intell.*, vol. 77, no. 1, pp. 43–94, Aug. 1995.

[21] P. Lipton, *Inference to the Best Explanation*, 2nd ed. Evanston, IL, USA: Routledge, 2004.

[22] R. M. J. Byrne, "Counterfactual thought," *Annu. Rev. Psychol.*, vol. 67, pp. 135–157, Jan. 2016.

[23] R. M. J. Byrne, "Cognitive processes in counterfactual thinking about what might have been," *Psychol. Learn. Motivat.-Adv. Res. Theory*, vol. 37, pp. 105–154, Oct. 1997.

[24] L. M. Pereira and A. B. Lopes, "Cognitive prerequisites: The special case of counterfactual reasoning," in *Machine Ethics* (Studies in Applied Philosophy, Epistemology and Rational Ethics), vol. 53. Cham, Switzerland: Springer, 2020, pp. 97–102.

[25] J. Paik, Y. Zhang, and P. Pirolli, "Counterfactual reasoning as a key for explaining adaptive behavior in a changing environment," *Biologically Inspired Cognit. Archit.*, vol. 10, pp. 24–29, Oct. 2014.

[26] Y. Zhang, J. Paik, and P. Pirolli, "Reinforcement learning and counterfactual reasoning explain adaptive behavior in a changing environment," *Topics Cognit. Sci.*, vol. 7, no. 2, pp. 368–381, Apr. 2015.

[27] E. Kulakova, M. Aichhorn, M. Schurz, M. Kronbichler, and J. Perner, "Processing counterfactual and hypothetical conditionals: An fMRI investigation," *NeuroImage*, vol. 72, pp. 265–271, May 2013.

[28] G. Grahne, "Update and counterfactuals," *J. Log. Comput.*, vol. 8, no. 1, pp. 87–117, 1998.

[29] M. Ginsberg, "Counterfactuals," *Artif. Intell.*, vol. 30, no. 1, pp. 35–79, 1986.

[30] R. J. Aumann, "Backward induction and common knowledge of rationality," *Games Econ. Behav.*, vol. 8, no. 1, pp. 6–19, Jan. 1995.

[31] W. Spohn, "A ranking-theoretic approach to conditionals," *Cognit. Sci.*, vol. 37, no. 6, pp. 1074–1106, Aug. 2013.

[32] E. Kulakova and M. S. Nieuwland, "Understanding counterfactuality: A review of experimental evidence for the dual meaning of counterfactuals," *Lang. Linguistics Compass*, vol. 10, no. 2, pp. 49–65, Feb. 2016.

[33] N. Hendrickson, *Counterfactual Reasoning: A Basic Guide for Analysts, Strategists, and Decision Makers*. Morrisville, NC, USA: LULU Press, 2011.

[34] H. J. Ferguson and A. J. Sanford, "Anomalies in real and counterfactual worlds: An eye-movement investigation," *J. Memory Lang.*, vol. 58, no. 3, pp. 609–626, Apr. 2008.

[35] D. K. Lewis, *On the Plurality of Worlds*. Oxford, U.K.: Blackwell, 1986.

[36] D. K. Lewis, *Counterfactuals*. Oxford, U.K.: Blackwell, 1973.

[37] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[38] D. Hume, *Enquiry Concerning Human Understanding*. Oxford, U.K.: Clarendon, 1904.

[39] B. Kment, "Counterfactuals and explanation," *Mind*, vol. 115, no. 458, pp. 261–310, Apr. 2006.

[40] D. E. Over, C. Hadjichristidis, J. S. B. T. Evans, S. J. Handley, and S. A. Sloman, "The probability of causal conditionals," *Cognit. Psychol.*, vol. 54, no. 1, pp. 62–97, Feb. 2007.

[41] D. Edgington, "Estimating conditional chances and evaluating counterfactuals," *Studia Logica*, vol. 102, no. 4, pp. 691–707, Aug. 2014.

[42] A. L. McGill and J. G. Klein, "Contrastive and counterfactual reasoning in causal judgment," *J. Personality Social Psychol.*, vol. 64, no. 6, pp. 897–905, 1993.

[43] J. Fang, Z. Huang, and F. van Harmelen, "A method of contrastive reasoning with inconsistent ontologies," in *The Semantic Web*. Berlin, Germany: Springer, 2012, pp. 1–16.

[44] A. Páez, "The pragmatic turn in explainable artificial intelligence (XAI)," *Minds Mach.*, vol. 29, no. 3, pp. 441–459, Sep. 2019.

[45] J. Kunkel, T. Donkers, L. Michael, C.-M. Barbu, and J. Ziegler, "Let me explain: Impact of personal and impersonal explanations on trust in recommender systems," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2019, pp. 1–487.

[46] K. Sokol and P. Flach, "One explanation does not fit all: The promise of interactive explanations for machine learning transparency," in *KI-Künstliche Intelligenz*, no. 2. Berlin, Germany: Springer, 2020, pp. 235–250.

[47] N. Elzein, "The demand for contrastive explanations," *Philos. Stud.*, vol. 176, no. 5, pp. 1325–1339, 2019.

[48] T. Le, S. Wang, and D. Lee, "GRACE: Generating concise and informative contrastive sample to explain neural network model's prediction," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 238–248.

[49] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. D. Bie, and P. Flach, "FACE: Feasible and actionable counterfactual explanations," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Feb. 2020, pp. 344–350.

[50] F. Bergadano, V. Cutello, and D. Gunetti, "Abduction in machine learning," in *Handbook of Defeasible Reasoning and Uncertainty Management Systems: Volume 4 Abductive Reasoning and Learning*. Norwell, MA, USA: Kluwer Academic, 2000, pp. 197–229.

[51] M. T. Keane and B. Smyth, "Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI)," in *Case-Based Reasoning Research and Development*. Cham, Switzerland: Springer, 2020, pp. 163–178.

[52] R. M. J. Byrne, "Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 6276–6282.

[53] C. Molnar, *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*, 1st ed. Fletcher, NC, USA: LULU, Feb. 2019.

[54] K. Sokol and P. Flach, "Counterfactual explanations of machine learning predictions: Opportunities and challenges for AI safety," in *Proc. AAAI Workshop Artif. Intell. Saf.*, 2019, pp. 1–4.

[55] M. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, 2017, pp. 4069–4079.

[56] A. Lucic, H. Haned, and M. de Rijke, "Why does my model fail?: Contrastive local explanations for retail forecasting," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2020, pp. 90–98.

[57] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[58] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Softw. Eng. Group, School Comput. Sci. Math., Keele Univ., Keele, U.K., Dept. Comput. Sci., Durham Univ., Durham, U.K., Tech. Rep. EBSE 2007-001, 2007.

[59] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering—A systematic literature review," *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, Jan. 2009.

[60] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proc. 18th Int. Conf. Eval. Assessment Softw. Eng. (EASE)*, 2014, pp. 1–10.

[61] C. Aguilar-Palacios, S. Munoz-Romero, and J. L. Rojo-Alvarez, "Cold-start promotional sales forecasting through gradient boosted-based contrastive explanations," *IEEE Access*, vol. 8, pp. 137574–137586, 2020.

[62] H. Andreas and L. Casini, "Hypothetical interventions and belief changes," *Found. Sci.*, vol. 24, no. 4, pp. 681–704, Dec. 2019.

[63] E. Barnes, "Why P rather than Q? The curiosities of fact and foil," *Phil. Stud.*, vol. 73, no. 1, pp. 35–53, Jan. 1994.

[64] S. Baron, M. Colyvan, and D. Ripley, "How mathematics can make a difference," *Philosophers Imprint*, vol. 17, no. 3, pp. 1–29, 2017.

[65] S. Baron, "Counterfactual scheming," *Mind*, vol. 129, no. 514, pp. 535–562, Apr. 2020.

[66] S. Baron, M. Colyvan, and D. Ripley, "A counterfactual approach to explanation in mathematics," *Philosophia Mathematica*, vol. 28, no. 1, pp. 1–34, Feb. 2020.

[67] L. Bertossi, "An ASP-based approach to counterfactual explanations for classification," *Rules and Reasoning* (Lecture Notes in Computer Science), vol. 12173. Cham, Switzerland: Springer, 2020, 70–81.

[68] A. Bokulich, "How scientific models can explain," *Synthese*, vol. 180, no. 1, pp. 33–45, May 2011.

[69] G. Botterill, "Right and wrong reasons in folk-psychological explanation," *Int. J. Phil. Stud.*, vol. 17, no. 4, pp. 463–488, Oct. 2009.

[70] S. Boulter, ''Contrastive explanations in evolutionary biology,'' *Ratio*, vol. 25, no. 4, pp. 425–441, 2012.

[71] M. J. L. Bours, ''A nontechnical explanation of the counterfactual definition of confounding,'' *J. Clin. Epidemiol.*, vol. 121, pp. 91–100, May 2020.

[72] R. Briggs, ''Interventionist counterfactuals,'' *Phil. Stud.*, vol. 160, no. 1, pp. 139–166, Aug. 2012.

[73] N. Campbell, ''Self-forming actions, contrastive explanations, and the structure of the will,'' *Synthese*, vol. 197, pp. 1225–1240, Mar. 2018.

[74] A. Chakravartty, ''Perspectivism, inconsistent models, and contrastive explanation,'' *Stud. Hist. Philosophy Sci. A*, vol. 41, no. 4, pp. 405–412, Dec. 2010.

[75] A. Chien, ''Scalar implicature and contrastive explanation,'' *Synthese*, vol. 161, no. 1, pp. 47–66, Mar. 2008.

[76] S. Chin-Parker and J. Cantelon, ''Contrastive constraints guide explanation-based category learning,'' *Cognit. Sci.*, vol. 41, no. 6, pp. 1645–1655, Aug. 2017.

[77] M. Day and G. S. Botterill, ''Contrast, inference and scientific realism,'' *Synthese*, vol. 160, no. 2, pp. 249–267, Jan. 2008.

[78] J. Dickenson, ''Reasons, causes, and contrasts,'' *Pacific Phil. Quart.*, vol. 88, no. 1, pp. 1–23, Mar. 2007.

[79] W. Fang, ''An inferential account of model explanation,'' *Philosophia*, vol. 47, no. 1, pp. 99–116, Mar. 2019.

[80] R. R. Fernández, I. M. de Diego, V. Ace na, J. M. Moguerza, and A. Fernández-Isabel, ''Relevance metric for counterfactuals selection in decision trees,'' in *Intelligent Data Engineering and Automated Learning—IDEAL 2019* (Lecture Notes in Computer Science), vol. 11871. Cham, Switzerland: Springer, 2019, pp. 85–93.

[81] R. R. Fernández, I. M. D. Diego, V. Aceña, A. Fernández-Isabel, and J. M. Moguerza, ''Random forest explainability using counterfactual sets,'' *Inf. Fusion*, vol. 63, pp. 196–207, Nov. 2020.

[82] C. E. Franklin, ''Agent-causation, explanation, and Akrasia: A reply to Levy's hard luck,'' *Criminal Law Philosophy*, vol. 9, no. 4, pp. 753–770, Dec. 2015.

[83] V. Gijsbers, ''A quasi-interventionist theory of mathematical explanation,'' *Logique et Analyse*, vol. 60, no. 237, pp. 47–66, 2017.

[84] A. Ghazimatin, O. Balalau, R. S. Roy, and G. Weikum, ''PRINCE: Provider-side interpretability with counterfactual explanations in recommender systems,'' in *Proc. 13th Int. Conf. Web Search Data Mining*, Jan. 2020, pp. 196–204.

[85] C. Held, ''Towards a monist theory of explanation,'' *J. Gen. Philosophy Sci.*, vol. 50, no. 4, pp. 447–475, Dec. 2019.

[86] L. A. Hendricks, R. Hu, T. Darrell, and Z. Akata, ''Grounding visual explanations,'' in *Computer Vision—ECCV 2018* (Lecture Notes in Computer Science), vol. 11206. Cham, Switzerland: Springer, 2018, pp. 269–286.

[87] J. A. Hird, ''The political economy of pork: Project selection at the U.S. Army corps of engineers,'' *Amer. Political Sci. Rev.*, vol. 85, no. 2, pp. 429–456, Jun. 1991.

[88] C. R. Hitchcock, ''The role of contrast in causal and explanatory claims,'' *Synthese*, vol. 107, no. 3, pp. 395–419, Jun. 1996.

[89] J. Hohwy, ''Capacities, explanation and the possibility of disunity,'' *Int. Stud. Philosophy Sci.*, vol. 17, no. 2, pp. 179–190, Jul. 2003.

[90] P. W. Holland, ''Causal counterfactuals in social science research,'' in *International Encyclopedia of the Social Behavioral Sciences*, 2nd ed. Oxford, U.K.: Elsevier, 2015, pp. 251–254.

[91] A. Kanehira, K. Takemoto, S. Inayoshi, and T. Harada, ''Multimodal explanations by predicting counterfactuality in videos,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8586–8594.

[92] J. Katz, ''Situational evidence: Strategies for causal reasoning from observational field notes,'' *Sociol. Methods Res.*, vol. 44, no. 1, pp. 108–144, Feb. 2015.

[93] A. Kean, ''A characterization of contrastive explanations computation,'' in *PRICAI'98: Topics in Artificial Intelligence* (Lecture Notes in Computer Science), vol. 1531. Berlin, Germany: Springer, 1998, pp. 559–610.

[94] J. Kim, C. Muise, A. Shah, S. Agarwal, and J. Shah, ''Bayesian inference of linear temporal logic specifications for contrastive explanations,'' in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 5591–5598.

[95] R. Knowles and J. Saatsi, ''Mathematics and explanatory generality: Nothing but cognitive salience,'' in *Erkenntnis*. Dordrecht, The Netherlands: Springer, Aug. 2019, pp. 1–19, doi: 10.1007/s10670-019-00146-x.

[96] D. Kostic, ''General theory of topological explanations and explanatory asymmetry,'' *Philos. Trans. Roy. Soc. B, Biol. Sci.*, vol. 375, no. 1796, 2020, Art. no. 20190321.

[97] J. Kuorikoski and P. Ylikoski, ''Explanatory relevance across disciplinary boundaries: The case of neuroeconomics,'' *J. Econ. Methodol.*, vol. 17, no. 2, pp. 219–228, Jun. 2010.

[98] J. Kuorikoski and P. Ylikoski, ''External representations and scientific understanding,'' *Synthese*, vol. 192, no. 12, pp. 3817–3837, Dec. 2015.

[99] D. N. Kutach, ''The entropy theory of counterfactuals,'' *Philosophy Sci.*, vol. 69, no. 1, pp. 82–104, Mar. 2002.

[100] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, ''Unjustified classification regions and counterfactual explanations in machine learning,'' in *Machine Learning and Knowledge Discovery in Databases* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence), vol. 11907. Cham, Switzerland: Springer, 2020, pp. 37–54.

[101] N. Levy, ''Luck and agent-causation: A response to Franklin,'' *Criminal Law Philosophy*, vol. 9, no. 4, pp. 779–784, Dec. 2015.

[102] P. Lichterman and I. A. Reed, ''Theory and contrastive explanation in ethnography,'' *Sociol. Methods Res.*, vol. 44, no. 4, pp. 585–635, Nov. 2015.

[103] P. Lipton, *Inference to the Best Explanation* (Philosophical Issues in Science). Evanston, IL, USA: Routledge, 1991.

[104] E. J. Lowe, ''What is the source of our knowledge of modal truths?'' *Mind*, vol. 121, no. 484, pp. 919–950, 2012.

[105] R. K. Mothilal, A. Sharma, and C. Tan, ''Explaining machine learning classifiers through diverse counterfactual explanations,'' in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2020, pp. 607–617.

[106] R. Northcott, ''Degree of explanation,'' *Synthese*, vol. 190, no. 15, pp. 3087–3105, Oct. 2013.

[107] J. Neyman and K. Iwaszkiewicz, ''Statistical problems in agricultural experimentation,'' *Suppl. J. Roy. Stat. Soc.*, vol. 2, no. 2, pp. 107–180, 1935.

[108] M. Pexton, ''Manipulationism and causal exclusion,'' *Philosophica*, vol. 92, no. 2, pp. 13–51, 2017.

[109] A. R. Pruss and J. L. Rasmussen, ''Explaining counterfactuals of freedom,'' *Religious Stud.*, vol. 50, no. 2, pp. 193–198, Jun. 2014.

[110] A. Reutlinger, ''Is there a monist theory of causal and noncausal explanations? The counterfactual theory of scientific explanation,'' *Philosophy Sci.*, vol. 83, no. 5, pp. 733–745, Dec. 2016.

[111] A. Reutlinger, ''Does the counterfactual theory of explanation apply to non-causal explanations in metaphysics?'' *Eur. J. Philosophy Sci.*, vol. 7, no. 2, pp. 239–256, 2017.

[112] A. Reutlinger, ''Extending the counterfactual theory of explanation,'' in *Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanations*. Oxford, U.K.: Oxford Univ. Press, 2018, pp. 74–95.

[113] L. J. Rips and B. J. Edwards, ''Inference and explanation in counterfactual reasoning,'' *Cognit. Sci.*, vol. 37, no. 6, pp. 1107–1135, Aug. 2013.

[114] D.-H. Ruben, ''A counterfactual theory of causal explanation,'' *Nous*, vol. 28, no. 4, pp. 465–481, 1994.

[115] D. B. Rubin, ''Bayesian inference for causal effects: The role of randomization,'' *Ann. Statist.*, vol. 6, no. 1, pp. 34–58, Jan. 1978.

[116] C. Q. Schneider and I. Rohlfing, ''Case studies nested in fuzzy-set QCA on sufficiency: Formalizing case selection and causal inference,'' *Sociol. Methods Res.*, vol. 45, no. 3, pp. 526–568, 2016.

[117] R. Schweder, ''Causal explanation and explanatory selection,'' *Synthese*, vol. 120, no. 1, pp. 115–124, 1999.

[118] C. Seelos and J. Mair, ''Organizational closure competencies and scaling: A realist approach to theorizing social enteprise,'' in *Social Entrepreneurship and Research Methods*, vol. 9. Bingley, U.K.: Emerald Group Publishing Limited, 2014, pp. 147–187.

[119] E. Sober, ''A theory of contrastive causal explanation and its implications concerning the explanatoriness of deterministic and probabilistic hypotheses,'' *Eur. J. Philosophy Sci.*, vol. 10, no. 3, pp. 1–15, Oct. 2020.

[120] R. Stalnaker, ''A theory of conditionals,'' in *Studies in Logical Theory (American Philosophical Quarterly Monographs 2)*. Oxford, U.K.: Blackwell, 1968, pp. 98–112.

[121] A. Steglich-Petersen, ''Against the contrastive account of singular causation,'' *Brit. J. Philosophy Sci.*, vol. 63, no. 1, pp. 115–143, Mar. 2012.

[122] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Farina, ''Generation and evaluation of factual and counterfactual explanations for decision trees and fuzzy rule-based classifiers,'' in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jul. 2020, pp. 1–8.

[123] M. Strohminger and J. Yli-Vakkuri, ''Knowledge of objective modality,'' *Phil. Stud.*, vol. 176, no. 5, pp. 1155–1175, May 2019.

[124] E. W. K. Tsang and F. Ellsaesser, "How contrastive explanation facilitates theory building," *Acad. Manage. Rev.*, vol. 36, no. 2, pp. 404–419, Apr. 2011.

[125] J. Woodward, *Making Things Happen: A Theory of Causal Explanation*. Oxford, U.K.: Oxford Univ. Press, 2003.

[126] P. Ylikoski and J. Kuorikoski, "Dissecting explanatory power," *Phil. Stud.*, vol. 148, no. 2, pp. 201–219, Mar. 2010.

[127] P. Ylikoski, *Social Mechanisms and Explanatory Relevance*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[128] A. R. Akula, S. Todorovic, J. Y. Chai, and S.-C. Zhu, "Natural language interaction with explainable AI models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jul. 2019, pp. 87–90.

[129] S. Anjomshoae, K. Främling, and A. Najjar, "Explanations of black-box model predictions by contextual importance and utility," in *Explainable, Transparent Autonomous Agents and Multi-Agent Systems* (Lecture Notes in Artificial Intelligence), vol. 11763. Cham, Switzerland: Springer, 2019, pp. 95–109.

[130] A. Apicella, F. Isgrò, R. Prevete, and G. Tamburrini, "Contrastive explanations to classification systems using sparse dictionaries," *Image Analysis and Processing—ICIAP 2019* (Lecture Notes in Artificial Intelligence), vol. 11751. Cham, Switzerland: Springer, 2019, pp. 207–218.

[131] B. R. Barricelli, E. Casiraghi, J. Gliozzo, A. Petrini, and S. Valtolina, "Human digital twin for fitness management," *IEEE Access*, vol. 8, pp. 26637–26664, 2020.

[132] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati, "Plan explanations as model reconciliation: Moving beyond explanation as soliloquy," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 156–163.

[133] C.-H. Chang, E. Creager, A. Goldenberg, and D. Duvenaud, "Explaining image classifiers by counterfactual generation," in *Proc. 7th Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 1–13.

[134] S. Dandl, C. Molnar, M. Binder, and B. Bischl, "Multi-objective counterfactual explanations," *Parallel Problem Solving from Nature—PPSN XVI* (Lecture Notes in Computer Science), vol. 12269. Berlin, Germany: Springer, 2020, pp. 448–469.

[135] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives,"in *Proc. Neural Inf. Process. Syst. Found.*, 2018, pp. 592–603.

[136] C. Fernandez, F. Provost, and X. Han, "Counterfactual explanations for data-driven decisions," in *Proc. 40th Int. Conf. Inf. Syst. (ICIS)*, 2019, pp. 1–10.

[137] A. Ferrario, R. Weibel, and S. Feuerriegel, "ALEEDSA: Augmented reality for interactive machine learning," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2020, pp. 1–8.

[138] O. Gomez, S. Holter, J. Yuan, and E. Bertini, "ViCE," in *Proc. Int. Conf. Intell. User Interfaces (IUI)*, 2020, pp. 531–535.

[139] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 4254–4262.

[140] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini, "Factual and counterfactual explanations for black box decision making," *IEEE Intell. Syst.*, vol. 34, no. 6, pp. 14–23, Nov. 2019.

[141] J. Labaien, E. Zugasti, and X. D. Carlos, "Contrastive explanations for a deep learning model on time-series data," in *Big Data Analytics and Knowledge Discovery* (Lecture Notes in Computer Science), vol. 12393. Berlin, Germany: Springer, 2020, pp. 235–244.

[142] M. Lash, Q. Lin, N. Street, J. Robinson, and J. Ohlmann, "Generalized inverse classification," in *Proc. Int. Conf. Data Mining (SDM)*, 2017, pp. 162–170.

[143] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "Comparison-based inverse classification for interpretability in machine learning," in *Proc. 17th Int. Conf. Inf. Process. Manage. Uncertainty Knowl.-Based Syst. (IPMU)*. New York, NY, USA: Springer-Verlag, 2018, pp. 100–111.

[144] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "The dangers of post-hoc interpretability: Unjustified counterfactual explanations," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 2801–2807.

[145] S. Liu, B. Kailkhura, D. Loveland, and Y. Han, "Generative counterfactual introspection for explainable deep learning," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2019, pp. 1–5.

[146] D. Martens and F. Provost, "Explaining data-driven document classifications," *MIS Quart.*, vol. 38, no. 1, pp. 73–100, 2014.

[147] L. Merrick and A. Taly, "The explanation game: Explaining machine learning models using shapley values," in *Machine Learning and Knowledge Extraction* (Lecture Notes in Computer Science), vol. 12279. Cham, Switzerland: Springer, 2020, pp. 17–38.

[148] J. Moore, N. Hammerla, and C. Watkins, "Explaining deep learning models with constrained adversarial examples," in *PRICAI 2019: Trends in Artificial Intelligence* (Lecture Notes in Computer Science), vol. 11670. Cham, Switzerland: Springer, 2019, pp. 43–56.

[149] F. Mosca, S. Sarkadi, J. M. Such, and P. McBurney, "Agent EXPRI: Licence to explain," *Explainable, Transparent Autonomous Agents and Multi-Agent Systems* (Lecture Notes in Computer Science), vol. 12175. Cham, Switzerland: Springer, 2020, pp. 21–38.

[150] M. A. Neerincx, J. van der Waa, F. Kaptein, and J. van Diggelen, "Using perceptual and cognitive explanations for enhanced human-agent team performance," in *Engineering Psychology and Cognitive Ergonomics* (Lecture Notes in Computer Science), vol. 10906. Cham, Switzerland: Springer, 2018, pp. 204–214.

[151] M. Pawelczyk, K. Broelemann, and G. Kasneci, "Learning model-agnostic counterfactual explanations for tabular data," in *Proc. Web Conf.*, Apr. 2020, pp. 3126–3132.

[152] D. Rajapaksha, C. Bergmeir, and W. Buntine, "LoRMIkA: Local rule-based model interpretability with K-optimal associations," *Inf. Sci.*, vol. 540, pp. 221–241, Nov. 2020.

[153] C. Russell, "Efficient search for diverse coherent explanations," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2019, pp. 20–28.

[154] S. Sharma, J. Henderson, and J. Ghosh, "CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Feb. 2020, pp. 166–172.

[155] K. Sokol and P. Flach, "Glass-box: Explaining AI decisions with counterfactual statements through conversation with a voice-enabled virtual assistant," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 5868–5870.

[156] S. Sreedharan, S. Srivastava, and S. Kambhampati, "Hierarchical expertise level modeling for user specific contrastive explanations," in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2018, pp. 4829–4836.

[157] R. Sukkerd, R. Simmons, and D. Garlan, "Towards explainable multi-objective probabilistic planning," in *Proc. 4th Int. Workshop Softw. Eng. Smart Cyber-Phys. Syst.* Washington, DC, USA: IEEE Computer Society, May 2018, pp. 19–25.

[158] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2019, pp. 10–19.

[159] J. van der Waa, M. Robeer, J. van Diggelen, M. Brinkhuis, and M. Neerincx, "Contrastive explanations with local foil trees," in *Proc. Workshop Hum. Interpretability Mach. Learn. (WHI)*, 2018, pp. 1–7.

[160] A. White and A. S. D. A. Garcez, "Measurable counterfactual local explanations for any classifier," in *Proc. 24th Eur. Conf. Artif. Intell. (ECAI)*, 2020, pp. 2529–2535.

[161] E. Zhao and R. Sukkerd, "Interactive explanation for planning-based systems," in *Proc. 10th ACM/IEEE Int. Conf. Cyber-Phys. Syst.*, Apr. 2019, pp. 322–323.

[162] N. J. van Eck and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping," *Scientometrics*, vol. 84, no. 2, pp. 523–538, Aug. 2010.

[163] K. S. Jones and J. R. Galliers, *Evaluating Natural Language Processing Systems: An Analysis and Review*. New York, NY, USA: Springer-Verlag, 1996.

[164] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," *J. Artif. Intell. Res.*, vol. 61, pp. 65–170, Jan. 2018.

[165] D. J. Hilton and B. R. Slugoski, "Knowledge-based causal attribution: The abnormal conditions focus model," *Psychol. Rev.*, vol. 93, no. 1, p. 75, 1986.

[166] D. J. Hilton, "Conversational processes and causal explanation," *Psychol. Bull.*, vol. 107, no. 1, pp. 65–81, 1990.

ILIA STEPIN received the Engineer degree (Hons.) in software engineering from the Moscow State University of Instrument Engineering and Computer Science, Russia, in 2012, the bachelor's degree in linguistics from Moscow State Linguistic University, Russia, in 2016, and the M.Sc. degree in computational linguistics from the University of Stuttgart, Germany, in 2018. He is currently pursuing the Ph.D. degree with the University of Santiago de Compostela, Spain. He is currently a Research Assistant with the Research Center in Intelligent Technologies (CiTIUS), Santiago de Compostela, Spain. He is also working on a doctoral project entitled ''Argumentative Conversational Agents for Explainable Artificial Intelligence'' with an emphasis on counterfactual explanation generation. His research interests include (but are not limited to) natural language processing, argumentation theory, and human–machine interaction.

JOSE M. ALONSO (Member, IEEE) received the M.Sc. and Ph.D. degrees in telecommunication engineering from the Technical University of Madrid (UPM), Spain, in 2003 and 2007, respectively. Since June 2016, he has been a Postdoctoral Researcher with the Research Centre in Intelligent Technologies (CiTIUS), University of Santiago de Compostela (USC). He is member of the CiTIUS-USC Research Group on Intelligent Systems. He has published more than 140 papers in international journals, book chapters, and in peer-review conferences. His research interests include explainable artificial intelligence, computational intelligence, interpretable fuzzy systems, natural language generation, and development of software tools. He is also the Deputy Coordinator of the H2020-MSCA-ITN-2019 Project titled ''Interactive Natural Language Technology for Explainable Artificial Intelligence'' (NL4XAI). He is recognized as the Ramon y Cajal Researcher (RYC-2016-19802), the Chair of the Task Force on Explainable Fuzzy Systems in the Fuzzy Systems Technical Committee of the IEEE Computational Intelligence Society (IEEE-CIS), a member of the IEEE-CIS Task Force on Explainable Machine Learning, the IEEE-CIS Task Force on Fuzzy Systems Software, the IEEE-CIS Content Curation Subcommittee, and the IEEE 1855 Working Group for the maintenance and update of the IEEE Standard for Fuzzy Markup Language IEEE Std 1855TM-2016, an Associate Editor of *IEEE Computational Intelligence Magazine*, and a Secretary of the ACL Special Interest Group on Natural Language Generation.

ALEJANDRO CATALA received the Ing., M.Sc., and Ph.D. degrees from the Universitat Politécnica de València (UPV), in 2006, 2008, and 2012, respectively. He is currently a Postdoctoral Researcher awarded with the Juan de la Cierva Fellowship at CiTIUS. He carried out his coBOTnity Project at the Human Media Interaction Laboratory, University of Twente, The Netherlands. His research interests include human–computer interaction, artificial intelligence, and intelligent user interfaces. Along his research career, he has been awarded with several personal grants by the Spanish and Valencian governments as well as the European Union's Horizon 2020 Research and Innovation Program under the Marie Sklodowska-Curie Individual Fellowship Grant (MSCA-IF). He has contributed with more than 70 peer-review publications, and regularly serves as a Reviewer and a Program Committee Member in conferences related to software engineering, artificial intelligence, and human–computer interaction.

MARTÍN PEREIRA-FARIÑA received the B.A. and Ph.D. degrees in philosophy and computational logic from the University of Santiago de Compostela, Spain, in 2007 and 2014, respectively. He is currently a Lecturer with the Department of Philosophy and Anthropology, University of Santiago de Compostela. His research interests include digital humanities, philosophy of language, argumentation theory, and computational models for it.

Information Sciences 618 (2022) 379–399

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

An empirical study on how humans appreciate automated counterfactual explanations which embrace imprecise information

Ilia Stepin [a,b,*], Jose M. Alonso-Moral [a,b], Alejandro Catala [a,b], Martín Pereira-Fariña [c]

[a] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Rúa de Jenaro de la Fuente Domínguez, s/n, 15782 Santiago de Compostela, A Coruña, Spain
[b] Departamento de Electrónica e Computación, Universidade de Santiago de Compostela, Rúa Lope Gómez de Marzoa, s/n, 15782 Santiago de Compostela, A Coruña, Spain
[c] Departamento de Filosofía e Antropoloxía, Universidade de Santiago de Compostela, Plaza de Mazarelos, s/n, 15705 Santiago de Compostela, A Coruña, Spain

ARTICLE INFO

Article history:
Received 15 February 2022
Received in revised form 19 October 2022
Accepted 21 October 2022
Available online 4 November 2022

2020 MSC:
94D05
03B52
03E72
68T30
68T37

Keywords:
Explainable artificial intelligence
Interpretable fuzzy modeling
Fuzzy rule-based classification
Counterfactual explanation
Human evaluation

ABSTRACT

The explanatory capacity of interpretable fuzzy rule-based classifiers is usually limited to offering explanations for the predicted class only. A lack of potentially useful explanations for non-predicted alternatives can be overcome by designing methods for the so-called counterfactual reasoning. Nevertheless, state-of-the-art methods for counterfactual explanation generation require special attention to human evaluation aspects, as the final decision upon the classification under consideration is left for the end user. In this paper, we first introduce novel methods for qualitative and quantitative counterfactual explanation generation. Then, we carry out a comparative analysis of qualitative explanation generation methods operating on (combinations of) linguistic terms as well as a quantitative method suggesting precise changes in feature values. Then, we propose a new metric for assessing the perceived complexity of the generated explanations. Further, we design and carry out two human evaluation experiments to assess the explanatory power of the aforementioned methods. As a major result, we show that the estimated explanation complexity correlates well with the informativeness, relevance, and readability of explanations perceived by the targeted study participants. This fact opens the door to using the new automatic complexity metric for guiding multi-objective evolutionary explainable fuzzy modeling in the near future.

1. Introduction

Artificial intelligence (AI)-based algorithms show striking accuracy in a wide range of domains and applications [1]. However, the most accurate models are known to produce scarcely explainable decisions [2]. This lack of explainability may damage the overall trust in AI [36]. In the light of possible negative consequences of following such automatic decisions without

[*] Corresponding author at: Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Rúa de Jenaro de la Fuente Domínguez, s/n, 15782 Santiago de Compostela, A Coruña, Spain.

E-mail addresses: ilia.stepin@usc.es (I. Stepin), josemaria.alonso.moral@usc.es (J.M. Alonso-Moral), alejandro.catala@usc.es (A. Catala), martin.pereira@usc.es (M. Pereira-Fariña).

https://doi.org/10.1016/j.ins.2022.10.098
0020-0255/© 2022 The Author(s). Published by Elsevier Inc.

having them explained, legal regulations concerning data processing are becoming widely adopted, e.g. the General Data Protection Regulation (GDPR) in the European Union [33]. Moreover, a new European regulation on AI is in progress and highlights the importance of preserving the European values by promoting trustworthy and responsible human-centric AI [9,34].

The gap between obscurity of automatic decisions and their explainability can be overcome by using interpretable models [37]. Among all AI tools, such soft computing techniques as fuzzy sets and systems have been shown to be not only interpretable but also explainable [3]. Thus, two key advantages are distinguished when relating the properties of interpretability and explainability of fuzzy systems. First, their transparent (i.e., interpretable) structure allows for making unambiguous inferences of why the given output was produced. Second, the use of linguistic variables and rules enables such systems to be explainable, i.e., to produce comprehensible explanations in natural language.

Nevertheless, the ability to demonstrate evidence on why specific output is produced (i.e., explain the factual output) may not be sufficient to display the underlying reasoning to the end user. Therefore, a factual explanation may need to be complemented with an explanation of why some other output was not produced. Opposed to factual explanations justifying the given prediction, counterfactual (CF) explanations (or counterfactuals) inform the end user about minimally different alterations to the input features for the outcome to change [41]. In the context of classification problems, CF explanations are typically designed as answers to the template question "Why was P predicted rather than Q?" where P is the output (factual) class and Q is a non-predicted hypothesized alternative CF class [29].

CF explanation generation is often regarded as an optimization problem in search of the data point of another class which represents the closest data point alternative to the test instance in an n-dimensional Euclidean space [46]. In the context of fuzzy sets and systems, however, such minimal changes may be described not only by means of a continuous variable representing numerical feature values (which we call "quantitative CFs" in this paper) but also by a discrete linguistic variable whose values are linguistic terms (which we refer to as "qualitative CFs" in this paper). In the former case, distinctive (numerical) features point to specific values, which are minimally different from those the test instance has, that should be set for the outcome to change. In the latter case, linguistic terms represent sets of suitable CF feature values in form of text and conceal the underlying numerical intervals.

The difference in end user's perception of these types of CF explanations remains unclear [45]. On the one hand, it may be affected by peculiarities of the structure of explanation, such as the number of explanatory features or explanation length. On the other hand, user's perception may be influenced by a degree of precision of the explanation content. Thus, qualitative CFs may be regarded as pieces of imprecise information which can facilitate understanding of the communicated explanation but may, however, be underinformative or even misleading to the end user. Conversely, quantitative CFs specify fine-grained changes to values of features. Last but not least, existing metrics for measuring quality of CF explanations (e.g., validity, proximity, diversity, among others) are strongly related to the data used for explanation generation [31]. However, those metrics ignore perceptual skills of the explanation's recipient and may not be sufficient for assessing the overall explanation effectiveness. In order to make another step towards human-centric AI, it therefore appears necessary to propose novel means of capturing and assessing human perception of explanations.

As part of previous work [41], we introduced a method for generating qualitative CF explanations applied to decision trees (DT). Then, we generalized this method to fuzzy information granules [43]. In this paper, our contribution is fourfold. First, we extend our previous work with a generalized Euclidean distance-based metric for CF explanation generation which better grasps membership function values. Second, we propose a novel genetic-based quantitative CF explanation generation method. Third, we define a new metric for assessing the complexity of automated explanations. Fourth, we carefully validate both qualitative and quantitative CF explanations via human evaluation in agreement with the best known practices for fair and sound evaluation of Natural Language Generation (NLG) and analyze the findings in terms of explanation complexity as expected to be perceived by the end user.

The rest of the manuscript is structured as follows. Section 2 presents a brief overview of existing methods for quantitative and qualitative CF explanation generation. Section 3 introduces our methods for generating CF explanations associated to fuzzy rule-based classification systems (FRBCS). Section 4 describes the key characteristics of the experimental design for subsequent human evaluation studies. Section 5 goes in detail with the analysis of the data collected in two evaluation surveys. Section 6 discusses the findings and offers suggestions on how they can be exploited. Finally, we outline directions for future work and conclude in Section 7.

2. Related work

CF explanation generation has in recent years attracted increasing attention from researchers in the AI field. As CFs oppose actual and potential outcomes, they are most widely used to explain the output of various classifiers, from linear machine learning models to deep neural networks [42]. Further, they are extensively found across different application domains. For example, CFs are found applicable in healthcare where they, e.g., serve to provide a patient with a bigger picture of the risk of developing diabetic retinopathy [26] or in banking where CFs suggest recommendations on necessary changes to have a loan application approved if previously rejected [16]. In addition, CF explanations are as well extensively used in robotics (e.g., in planning – to justify the choice of a robot over other feasible but unfavored possible solutions [44]). Despite numerous potential application domains, the use of CFs is advised to be controled due to possible malicious implications. As such, they have been misused or misinterpreted (what may lead to data breaches) in cases of, e.g. password masking or e-voting [20]. Other privacy concerns include inferring sensitive patterns of the training data or manipulations with the revealed internals of the model [40].

In the context of qualitative CFs, a number of generation methods output CF sets to support diversity. For example, Sokol and Flash inspect the internal structure of DTs in their "Glass-Box" framework for generating CF sets [40]. Thus, the authors retrieve CF sets from the decision paths ranking them by their leaf-to-leaf distance to the actual prediction. On a similar note, Stepin et al. generate set-based (i.e., qualitative) CFs from either crisp or fuzzy DTs [41] but also regarding fuzzy information granules [43] while introducing an extra-linguistic layer to approximate numerical intervals or membership function values, respectively, using predefined linguistic terms.

Whereas the aforementioned methods are model-specific, i.e., they only allow for explaining counterfactually the given output of the DT itself, DT-based approaches are also used for model-agnostic methods. In their LOcal Rule-based Explanation (LORE) method, Guidotti et al. employ a genetic algorithm to first synthesize a local neighborhood around the test instance which is subsequently used to train a DT and generate CF sets [17]. The collection of CF sets is then reconstructed from the decision paths. Then, the minimally different CF set is selected on the basis of the (minimal) number of Boolean split conditions of the DT that the given CF path does not satisfy. Maaroof et al. extend LORE to fuzzy logic-based applications by proposing Contextualised LORE for Fuzzy attributes (C-LORE-F) [26]. Alternatively to LORE, the researchers formulate a local neighborhood generation approach for solving the uniform cost search problem. Potential neighbors are generated by applying iterative changes over a single feature taking into account intersections between two corresponding fuzzy sets. Further, the authors propose to induce the rules instead of building up a DT using the Dominance-based Rough Set Approach (DRSA) where the decision rules take into consideration the preference directions of the input variables. In addition, Fernández et al. extract CF sets from a random forest classifier by partly fusing individual tree predictors [12]. Further, their Random Forest Optimal Counterfactual Set Extractor (RF-OCSE) prunes the search space of candidate CFs using the minimum observable approach to filter out CFs whose distance to the test instance exceeds the best up-to-now distance.

On the other hand, quantitative (i.e., single-point-output) CF explanation generation methods address the optimization problem searching for an individual data point found to be minimally different from the test point under consideration in accordance with the selected distance function, e.g., Manhattan distance weighted by the inverse median absolute deviation [46]. Similarly, Moore et al. use a differentiable model on the basis of a gradient-based method over the cross entropy loss function to identify a single minimally distant CF data point [30].

Alternatively, genetic algorithms are also frequently used to generate CFs [39]. Model-agnostic genetic algorithms are used not only to generate a local neighborhood but also to identify a specific optimal CF data point. In addition to the standard genetic algorithm, Lash et al. apply local search to non-mutated children so that the best solution is preserved for the next generation [24]. Sharma et al. propose another approach called Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence (CERTIFAI) where a genetic algorithm based on natural selection, mutation, and crossover appeals to user feedback (regarding feature mutation, feature range specification, and enquiries for a specific number of explanations) [39]. Whereas these user constraints allow for generating actionable human-centric explanations, imposing too severe restrictions may overreduce the search space resulting in generating null explanations. In addition, Schleich et al. make use of a complete search space in their GeCo framework [38]. Thus, the authors present a customizable genetic algorithm enhanced with two optimization techniques to reduce memory costs and running time. The compressed δ-representation of the input features reduces the memory storage required for mutation-related calculations whereas the so-called partial evaluation optimizes the evaluation of the classifier, as static components of the classifier can be pre-evaluated using an equivalent sub-model of the same classifier [38].

Finally, both qualitative and quantitative generation methods are primarily evaluated with automatically computable metrics (e.g., fidelity, validity, proximity, or diversity) [12,17,31]. Unfortunately, empirical studies involving human evaluation for assessing the goodness of automated CFs are scarcely found in the literature. Baaj and Poli show that explanations based on the use of linguistic terms appear rather satisfactory and convincing despite being overly repetitive for a general audience [5]. Wang and Yin state that CFs increase understanding for users who have sufficient domain knowledge but fail to calibrate trust in the model [47]. Further, Lucic et al. demonstrate that CFs help users understand why a model makes large errors [25]. Olson et al. show that CFs can be also effective for non-expert users in the identification of flawed agents [32]. In addition, Woodcock et al. stress that lay users trust CFs only if the information gap in the existing domain knowledge between them and expert users is not significant, specifically in the healthcare domain [48]. Nevertheless, unlike our work, none of the aforementioned studies contrasts the output of single-point-output quantitative generation methods and set-based qualitative ones.

3. Explanation generation methods

3.1. Notation

The methods proposed in this study address a multi-class classification problem, i.e., learning a mapping function $h : X \longrightarrow Y$ from a dataset $X = \{x_i\}|_{i=1}^{n}$ containing n labeled instances to a discrete output variable (class) $Y = \{y_j\}|_{j=1}^{m}$ where m is the number of classes. The dataset is characterized by the set of p numerical[1] features $F = \{f_k\}|_{k=1}^{p}$, which are mapped to the corresponding linguistic variables. By definition [49], each feature is a tuple $f_k = \left\langle L^{f_k}, T_L^{f_k}, U^{f_k}, G^{f_k}, M^{f_k} \right\rangle, \forall f_k \in F$ where L^{f_k} is the name of the feature $f_k, T_L^{f_k} = \left\{ t_l^{f_k} \right\}|_{l=1}^{s}$ is the set of linguistic terms defined in the universe of discourse U^{f_k}, G^{f_k} and M^{f_k} being syntactic and semantic rules, respectively. Let $V_T = \bigcup T_L^{f_k}, \forall f_k \in F$ denote the set of all linguistic terms.

In our experiments (see Sections 4 and 5), we aim to explain (both factually and counterfactually) the output of an FRBCS [23] which is defined by the following components:

- a knowledge base containing a set of input and output variables and a rule base which represents a set $R = \{r_i(w_i)\}|_{i=1}^{|R|}$ of weighted fuzzy rules of the form $r_i(w_i) : \text{IF}\, L^{f_1}\, \text{is}\, t_1^{f_1} \left[\text{AND} \ldots L^{f_k}\, \text{is}\, t_k^{f_k} \ldots \text{AND} \ldots \right] \text{THEN}\, y\, \text{IS}\, y_i$, where $r_i \in R, w_i \in [0, 1]$ is the rule weight (i.e., the higher w_i the more relevant r_i), $t_k^{f_k} \in T_L^{f_k}, f_k \in F, y_i \in Y$;
- a fuzzy processing structure containing fuzzification and defuzzification interfaces as well as a fuzzy reasoning mechanism. Given an input vector $\mathbf{x} = [x_1, \ldots, x_p]$ and a rule $r_i \in R$, its activation degree a_i is computed as $a_i(\mathbf{x}) = \mu_{t_1^{f_1}}(x_1) \otimes \ldots \otimes \mu_{t_k^{f_k}}(x_k) \otimes \ldots \otimes \mu_{t_p^{f_p}}(x_p)$, being $\mu_{t_k^{f_k}}(x_k)$ the membership degree of the value x_k for the linguistic term t_k associated to feature f_k, and $\otimes$ is a t-norm such as minimum or product.

Any rule r_i can be denoted as a tuple $r_i(w_i) = \langle AC_i, cq_i \rangle$ where AC_i is an antecedent (i.e., a non-empty set of feature-value pairs) and cq_i is a consequent (i.e., a class label).

The output class $y_{FAC} \in Y$ predicted by an FRBCS is said to be the factual explanation class. All the rules from the rule base that lead to the predicted outcome form a set of factual explanation rules $R_{FAC} = \bigcup_{r_j \in R} \{r_j | cq_j = y_{FAC}\}$, being $R_{FAC} \subseteq R$. Similarly, all the non-predicted classes form a set of CF classes, with a collection of the corresponding rules mapped to each of them: $R_{CF} = \bigcup_{r_j \in R} \left\{ r_j | cq_{r_j} = y_{CF} \right\}, Y_{CF} = \{y_{CF} | y_{CF} \in Y \setminus y_{FAC}\}$.

Given an FRBCS s, a data instance $\mathbf{x} \in X$, and the classification output y_{FAC} predicted by s, each class $y_j \in Y$ is associated with a single explanation of why $\mathbf{x}$ is classified in the given way. Hence, there exists only one factual explanation $E_{FAC}(s, \mathbf{x}, y_{FAC})$. In addition, there is a non-empty set of CF explanations $E_{CF}(s, \mathbf{x}, Y_{CF}) = \bigcup_{y_{CF} \in Y_{CF}} E_{CF}(s, \mathbf{x}, y_{CF})$ for each non-predicted class $y_{CF} \in Y_{CF}$.

Throughout the manuscript, we assume that the output is explained in its entirety if the corresponding explanation contains a factual explanation specifying why the given decision is made as well as $|Y| - 1$ CF explanations indicating why all the alternative classification options are discarded. Therefore, a (full) explanation for a data instance $\mathbf{x} \in X$ is assumed to contain one factual explanation and a non-empty set of CF explanations: $E(s, \mathbf{x}, Y) = E_{FAC}(s, \mathbf{x}, y_{FAC}) \cup E_{CF}(s, \mathbf{x}, Y_{CF})$. Accordingly, explanation generation methods aim to produce (1) a factual explanation for the test instance and (2) the most relevant CF explanations for all the CF classes.

3.2. Factual explanation generation

We design the process of explanation generation to include three main stages (text planning, sentence planning, and surface text realization) as in the NLG pipeline proposed by Reiter and Dale [35]. We selected this NLG pipeline because it is by far the most commonly used in the scientific community [14]. It is worth noting that we apply the same NLG pipeline no matter if we consider either factual or CF explanations:

- **Text planning**, where the information to be conveyed in the text is identified (content determination), as well as some order and general structure of the text is planned. In the case of CF explanations, content determination relies on relevance estimation (as described in the next section).
- **Sentence planning**, which includes grouping of messages when needed (sentence aggregation) and decisions about the words/expressions to be used (referring expression generation and/or lexicalization). This stage is crucial to avoid repetitions and make the output text more natural.

[1] The use of categorical features is out of the scope of this work.

- **Surface text realization**, which consists of generating a syntactically, morphologically, and orthographically correct text. This last stage is implemented using a pool of templates dynamically instantiated, populated and mixed with a Python wrapper of the SimpleNLG library [6].

Specifically, the factual explanation generation process presupposes the following steps: factual explanation rule selection, linguistic approximation of the feature values used in the antecedent (optionally), and linguistic realization. First, the factual explanation rule is selected from all the rules whose consequent is the predicted class. To do so, we calculate the product of the activation degree a_j of each rule $r_j \in R$ and its associated rule weight w_j, s.t. $argmax\ w_j \cdot a_j$, i.e., the factual explanation rule has the maximum product of the activation degree a_j and rule weight w_j. Second, if the rules are semantically grounded, i.e., if they use meaningful strong fuzzy partitions (SFP), the feature values in the factual explanation are readily available and mapped to the corresponding linguistic terms (e.g., “IF Color IS *Pale* AND Strength IS *Standard* THEN Beer style IS Blanche” where *Pale* and *Standard* are expert-defined linguistic terms). Otherwise, i.e., if only local semantics are available (e.g., “IF Color IS *MF0* AND Strength IS *MF1* THEN Beer-style IS Blanche” where *MF0* and *MF1* are two membership functions with local semantics), linguistic approximation is necessary to generate a meaningful explanation. Notice that the mechanism of linguistic approximation is also used for qualitative CF explanation generation and will be described in detail in the next section. Finally, once the relevant pieces of information are identified, linguistic realization is performed.

3.3. Qualitative counterfactual explanation generation

In this section, we introduce a new method for generating qualitative CF explanations (hereinafter denoted as *EUC*). This method can be regarded as an extension of our previously proposed method (hereinafter denoted as *XOR*) [43]. The *EUC* method aims to be more sensitive than *XOR* to variations in membership functions. Despite certain methodological differences, both methods form a pipeline containing the following steps to be described in detail below (see Fig. 1): CF rule representation, relevance estimation, linguistic approximation (optional in terms of the local/global semantics attached to the FRBCS), and textual explanation generation.

CF rule representation. First of all, the test instance (as well as all the CF candidates) must be represented in a compatible form. Both *EUC* and *XOR* methods reason over the information retrieved from the rule base. Multiple candidates form CF sets which are labeled in accordance with the selected linguistic terms for the given features. Thus, we regard CF sets as collections of data instances covered by the rules leading to the desired CF class. In this sense, there exist as many potential CFs as there are rules that lead to the desired CF class.

For a given FRBCS, a test instance $\mathbf{x} \in X$ can be represented as a vector $\mathbf{x} = \bar{x}_{1\times|V_T|} = \left[\mu_x(t_i)|_{i=1}^{|V_T|}\right]$ of membership function values of each linguistic variable. Similarly, each CF rule can be regarded in terms of the membership function values that the linguistic variables take on. Therefore, each CF rule $r_{CF} \in R_{CF}$ is vectorized over V_T for compatibility purposes so that the collection of such vectorized rules makes up a rule-term matrix $M_{|R_{CF}|\times|V_T|}$ where the i-th row corresponds to a CF rule and the j-th column corresponds to the given linguistic term $t_j \in V_T$. Hence, the rule-term matrix is populated with such membership values as functions of a given linguistic term $M_{ij} = \mu_x(t_{ij})$.

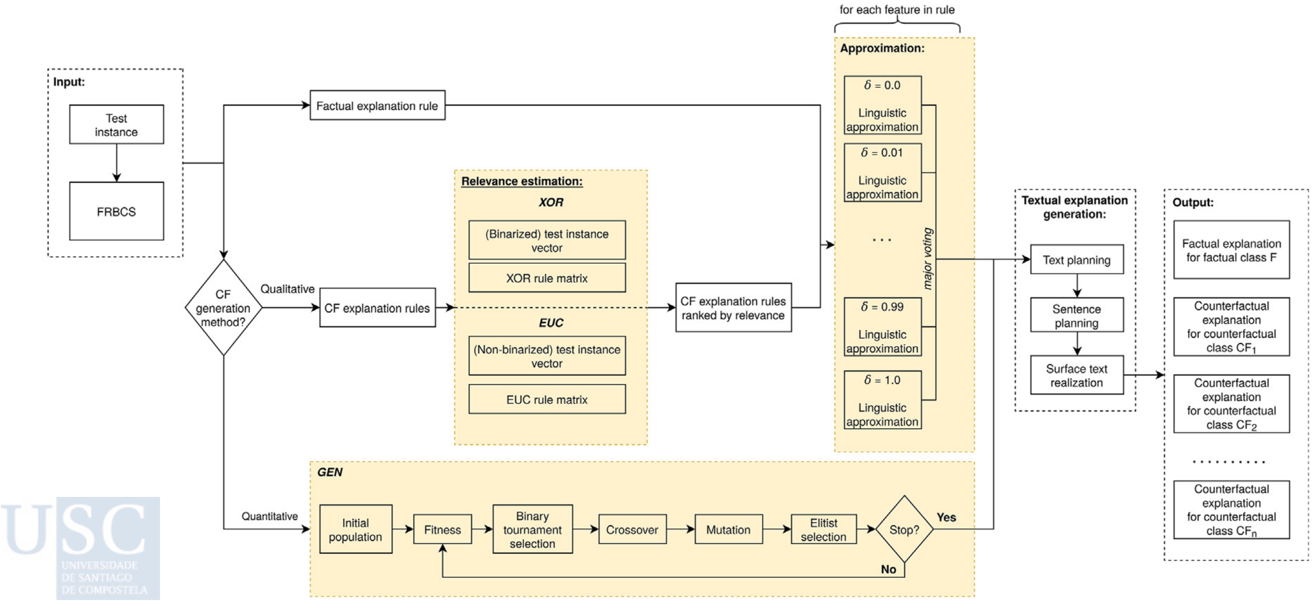


Fig. 1. CF explanation generation pipeline. The shadowed building blocks influence the surface realization of the output explanation.

It is worth noting that the *XOR* method additionally binarizes both the test instance vector and the rule-term matrix, at the cost of information loss because of the test instance and rule vectors being approximated. Instead, the *EUC* method represents the original information without further approximation. This is claimed to better capture fuzzy variable ambiguity and avoid potential information loss.

Relevance estimation. Given vector representations of the candidate CF rules, it becomes essential to identify the CF set that is minimally different from (and therefore most relevant to) the test instance. Whereas *XOR* calculates relevance by minimizing the number of different bits, *EUC* relates each vectorized CF rule to the test instance vector in a $|V_T|$-dimensional space and measures CF relevance as the Euclidean distance d between pairs of vectors $\langle \overline{x}, \overline{r_{CF_i}} \rangle, 1 \leqslant i \leqslant |R_{CF}|$; being $\overline{r_{CF_i}} = \overline{M_{i,*}}$ the vector associated to row i in matrix M, i.e., the vector which corresponds to CF rule i.

- $d_{XOR}(\overline{x}, \overline{r_{CF_i}}) = \frac{\sum_j |x^j - r^j_{CF_i}|}{|V_T|} \in [0, 1]$;
- $d_{EUC}(\overline{x}, \overline{r_{CF_i}}) = \sqrt{\sum_j \left(x^j - r^j_{CF_i}\right)^2} \in [0, \infty)$.

where x^j and $r^j_{CF_i}$ are the j-th elements in vectors $\overline{x}$ and $\overline{r_{CF_i}}$, respectively.

The candidate CF rules are then ranked in accordance with the given distance metric. Subsequently, we include the minimally distant (or most relevant) CF rules for each CF class in the pool E_{CF} of the resulting CF explanations for the given test instance $\mathbf{x}$. If multiple CF rules are equally minimally distant from $\mathbf{x}$, such rules are deemed equally explanatory. In this case, the most relevant CF is selected randomly. Representing the test instance and CF rules in a Euclidean $|V_T|$-dimensional space is hypothesized to better capture fuzzy-specific properties of an FRBCS. For example, the Euclidean distance appears more sensitive to changes in membership function values. The number of unique values that the *XOR*-based distance can take on is limited by $|V_T|$. In consequence, several CF rules may result in having the same relevance score while being distinct in the number of features or their labeling. On the contrary, *EUC* provides a more flexible and diverse measure of relevance of different CF rules and therefore gives a better insight into the fuzzy system's behavior.

Linguistic approximation. If the linguistic terms are not based on a SFP and therefore not semantically grounded, the selected CF rule must be enhanced with an additional linguistic layer so that the output explanation is meaningful to the end user. Once the CF rules are ranked by relevance and the most relevant CF is identified, it must therefore be linguistically approximated. To do so, each fuzzy set corresponding to the linguistic term of the selected CF rule is mapped to the gold standard annotations. Note that this mapping is actionable if the α-cut is applied to such a fuzzy set given some threshold value δ. To illustrate the process of linguistic approximation, consider a fuzzy set *FS* characterized by a trapezoidal membership function and three linguistic terms ($T = \{t_1, t_2, t_3\}$) which are candidates to be associated with *FS* (see Fig. 2 for details). Given some cut-off threshold value δ_1, the fuzzy set *FS* can be projected to an interval of numerical values $L = \left[v_{\delta_1}, v_{\delta_2}\right]$. In addition, each linguistic term $t_i \in T$ can be projected to an interval $t_{i\delta 1} (1 \leqslant i \leqslant |T|)$. Then, the interval L can be compared with the intervals $t_{i\delta 1}$ using the Jaccard Similarity Index [13]:

$$\forall L \approx t^f_a \in V_T : S(t_{i\delta 1}, L) = \frac{t_{i\delta 1} \cap L}{t_{i\delta 1} \cup L} \in [0, 1], \tag{1}$$

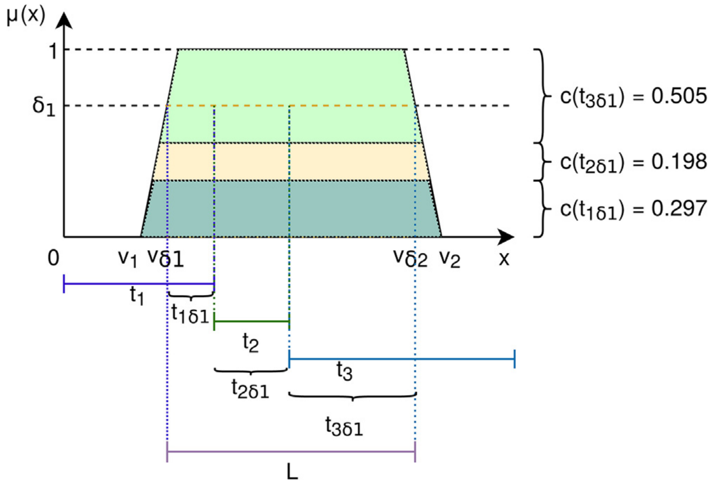


Fig. 2. Illustrative example of the linguistic approximation mechanism.

Table 1
Approximation confidence score calculation.

Term	δ	Approximation confidence
t_1	[0.0, 0.3)	30/101 = 0.297
t_2	[0.3, 0.5)	20/101 = 0.198
t_3	[0.5, 1.0]	51/101 = 0.505

where $t_{i\delta 1}$ is the numerical interval closer to the linguistic term t_i, and L is the numerical interval associated to the selected α-cut. As follows from Fig. 2, $S(t_{3\delta 1}, L) > S(t_{2\delta 1}, L) > S(t_{1\delta 1}, L)$. Hence, the feature f_j characterized by fuzzy set *FS* is verbalized as "f_j is t_3" in this case.

Note that the threshold value δ for the α-cut serves as a hyperparameter. The previously proposed *XOR* method uses heuristics to specify δ manually. Instead, both qualitative CF generation methods now use major voting in order to reduce possible approximation error. Thus, given some small enough *step*, we inspect all the approximated linguistic terms over the cut-off interval $[0, 1]$ for each term in the given CF rule and assign a confidence score to each term t_i as follows: $c(t_i) = \frac{\#t_i}{1+\frac{1}{step}}$, being $\#t_i$ the number of times t_i is the winner.

For each feature f_j involved in the classification and considered in the output explanation, we apply major voting to identify which linguistic term is covered by the widest range of the inspected approximations using the approximation confidence score $c(t_i)$ as a reference, so that the selected linguistic term is $t_j \in V_T | argmax\ c(t_j)$. Considering the example in Fig. 2, let *step* be 0.01. We therefore perform $n = 1 + 1/0.01 = 101$ linguistic approximations. Suppose that the term under consideration is mapped to the set of linguistic terms as indicated in Table 1.

Approximation confidence scores are calculated for all the competing linguistic terms. Since we aim to use the most frequently found term among all the considered threshold values, the linguistic term that has the highest score (in this case, t_3) is selected for the output explanation. It is worth noting that in this illustrative example, the selected linguistic term is the same as the one selected when considering only δ_1. However, in the general case they may be different. Therefore, it is recommended to follow the major voting approach instead of relying only on a single δ value selected heuristically.

As only two building blocks (relevance estimation and linguistic approximation) influence the output explanation (see the shadowed blocks in Fig. 1), *XOR* and *EUC* generate CFs following one of the three scenarios below:

- the two methods select the same rule to be the most relevant, the approximation algorithm gets the same semantically grounded linguistic terms;
- the two methods select two different CF rules (e.g., "IF f_1 IS MF_0 and f_2 IS MF_0 THEN y_{cf}" and "IF f_1 IS MF_1 and f_2 IS MF_1 THEN y_{cf}") which nevertheless generate identical CF explanations due to a large enough overlap between the corresponding fuzzy sets. This scenario is possible when all the features used in both rules are identical and their non-semantically grounded values overlap to a large enough extent;
- the two methods select two different CF rules (e.g., "IF f_1 IS MF_0 and f_2 IS MF_0 THEN y_{cf}" and "IF f_1 IS MF_2 and f_3 IS MF_4 THEN y_{cf}") where feature values are approximated to different linguistic terms.

Textual explanation realization. At the last stage, the selected factual and CF pieces of information are converted to explanations in natural language while applying the NLG pipeline introduced in the previous section. It is worth noting that the text and sentence planning along with text realization for a factual explanation follow the structure of the corresponding winner rule from the rule base. Thus, a factual explanation is assumed to include a subordinate clause of cause (e.g., "The data instance x is of class y_f because f_1 is v_1 and f_2 is v_2"), which lists the features and the corresponding values or linguistic terms that influenced the actual decision. On the other hand, a CF explanation is verbalized in natural language as a complex conditional sentence that adopts the structure of the rule, e.g., "x would be of class y_{cf} if f_1 were v_2 and f_3 were v_4" for the given CF class y_{cf}.

Implementation details. The *XOR* and *EUC* methods are implemented as open source software in Python and are made publicly available at a Gitlab repository[2].

3.4. Quantitative counterfactual explanation generation

In this section, we present a new method for CF explanation generation which is grounded in evolutionary and bio-inspired computation algorithms for explainable AI [11]. More precisely, we have implemented a Genetic Algorithm (hereafter denoted as *GEN*) which takes as the starting point the genetic fuzzy tuning approach previously proposed by Alonso et al. [4]. Indeed, the original algorithm was first introduced by Cordon and Herrera [7] and later adapted to explainable SFP tuning in [4].

GEN manages a population P with N individuals which evolve in g generations. The given test instance $\mathbf{x}$ is used for building the first individual of the population. Each individual is associated to a real-coded chromosome which is made up of p

[2] https://gitlab.citius.usc.es/ilia.stepin/fcfexpgen (branch "xor_euc_gen")

genes, with each gene representing one of the features in F. Since all the features are numerical, gene $i \in [1, p]$ encodes the double value associated to feature i. The rest of the population is generated randomly. Thus, a random value is assigned to each gene i within its variation interval which is determined by the numerical range associated to feature i. The pseudocode of the developed algorithm is as follows (see the *GEN* shadowed block in Fig. 1):

```
1. Initialize the generation counter, $g = 0$, and evaluate the initial population, $P^{(0)}$. Evaluating a population means computing Fitness for each individual in the population. Here, Fitness is computed as the Euclidean distance between the data instance $\hat{x}$ associated to the current chromosome and the original test instance $\mathbf{x}$, if the inferred output is in agreement with the target CF class. Otherwise, Fitness equals the maximum distance which comes out from the Euclidean distance between the two vectors representing the extreme values (min/max) for the variation intervals associated to each feature. Hence, the smaller Fitness, the better.
2. while $g < MaxGener$ and $Fitness \geqslant StopThres$ and $Nbest \leqslant NrepThres$
       $g := g + 1$
       $Select\ P^{(g)}\ from\ P^{(g-1)}$
       $Crossover\ P^{(g)}$
       $Mutate\ P^{(g)}$
       $Elitist\ selection\ P^{(g-1)}$
       $Evaluate\ P^{(g)}$
   end while
```

The procedure ends either when the maximum number of generations (*MaxGener*) is reached, or *Fitness* is under the predefined threshold (*StopThres*), or the number of consecutive generations for which the best fitness value remains the same (*Nbest*) is greater than the predefined threshold (*NrepThres*). On the one hand, *MaxGener* should be defined empirically in terms of the complexity of the dataset under consideration. It must be large enough to guarantee that *GEN* converges to a good enough solution. On the other hand, *StopThres* and *NrepThres* are threshold values to speed up the procedure, so that the algorithm stops before *MaxGener* is reached in case *Fitness* is small enough or becomes constant for a large enough number of generations. For each generation, the following steps are repeated:

- The selection of $P^{(g)}$ from $P^{(g-1)}$ is made as a deterministic tournament selection procedure. Each individual in the new population, $P^{(g)}$, is chosen from the previous one, $P^{(g-1)}$, after making a tournament that involves TS individuals randomly selected from $P^{(g-1)}$. The best individual is selected in any tournament. The selection pressure can be adjusted by changing $TS \leqslant N$. The larger TS, the smaller the chance of weak individuals to be selected. For example, if $TS = N$, then all the individuals in $P^{(g)}$ are equal to the best one in $P^{(g-1)}$, what is unsatisfactory from the point of view of diversity in the population.
- The $BLX - \alpha$ crossover operator [10] is applied to $P^{(g)}$. The parents, i.e., the selected chromosomes in the current population, are crossed over in pairs. Each pair of parents, $dad = (d_1, \cdots, d_p)$ and $mom = (m_1, \cdots, m_p)$, is replaced in the new population by two offsprings, $O_d = (o_{d1}, \cdots, o_{dp})$ and $O_m = (o_{m1}, \cdots, o_{mp})$, where o_{dj} and o_{mj} are random values from the intervals $[min_{dj}, max_{dj}]$ and $[min_{mj}, max_{mj}]$, respectively. I_j=$[I_j^l, I_j^u]$ is the variation interval of gene j. According to the taxonomy for the crossover operator presented by [21], $\alpha = 0.3$ is a suitable value for letting $BLX - \alpha$ exploit the nature of real coding as follows:

$$min_{dj} = maximum\ \left(I_j^l, d_j - \alpha \cdot |d_j - m_j|\right)$$
$$max_{dj} = minimum\ \left(d_j + \alpha \cdot |d_j - m_j|, I_j^u\right)$$
$$min_{mj} = maximum\ \left(I_j^l, m_j - \alpha \cdot |m_j - d_j|\right)$$
$$max_{mj} = minimum\ \left(m_j + \alpha \cdot |m_j - d_j|, I_j^u\right)$$

- A uniform mutation operator is considered. The value of the selected gene is changed by another one generated randomly within its variation interval.
- The elitist selection ensures perpetuating the best individual from the given generation to the next one. If the best individual, B_i in $P^{(g-1)}$, is not included in $P^{(g)}$, then the worst individual in $P^{(g)}$ is replaced by B_i.

Once *GEN* ends, we have identified a new data instance $\hat{x}$ that is assumed to minimally change the original test instance $\mathbf{x}$ while making the FRBCS infer the desired CF output[3]. Then, it is time for generating the related CF explanation in natural language. To do so, we once again apply the NLG pipeline described previously. First of all, we compute the percentage of modification $D_j = 100 \ast \frac{\hat{x}_j - x_j}{I_j}$ associated to each feature j to go from $\mathbf{x}$ to $\hat{x}$. The text which describes D_j is as follows: x_j is [*slightly*] *increased* | *decreased*; where *increased* appears if $D_j > 0$. On the contrary, *decreased* is used if $D_j < 0$. In addition, the linguistic modifier *slightly* appears only in case of small modifications, i.e., only if $0.9 \leqslant D_j \leqslant 5$, which means the percentage of modification is smaller or equal than 5%. Notice that nothing is said about feature j if $D_j < 0.9$. In this case, we consider the feature j to remain the same assuming that such a small change (less than 0.9%) does not have sufficient explanatory power for the recipient of the explanation. This assumption is made heuristically in accordance with our previous experience with designing NLG systems while keeping in mind the limited processing capability of human beings [28]. As a result, the generated textual explanations are shorter and easier to process while referring only to relevant changes.

Afterwards, at the sentence planning stage, for the sake of simplicity and naturalness, we aggregate those pieces of information associated to different features which are affected by the same type of modification (e.g., "f_1 and f_2 are *slightly increased*" replaces to "f_1 is *slightly increased* and f_2 is *slightly increased*"). We also apply lexicalization for each feature to be described in a fully meaningful way. Therefore, *increased* and *decreased* are replaced by more meaningful terms (e.g., *strength is bigger* or *color is darker*).

Finally, text realization is done again using the following template and the SimpleNLG library with the aim of ensuring syntactically, morphologically and orthographically correct final text: "[Output Class Name] *would be* [CF Class Name] *if* [Name of the most Relevant Feature$_j$] *were* [linguistic description of D_j] (new data value) [AND...]". Notice that the new values for the features associated with the most relevant changes are given in brackets.

Implementation details. The *GEN* method is implemented as a piece of open source software in Python and is made publicly available at a Gitlab repository[4]. It is also integrated with the open source software GUAJE[5] which is devoted to facilitating the design of explainable fuzzy systems [3]. The following *GEN* parameters are considered when generating the quantitative CF explanations under evaluation in the rest of the paper: population length ($N = 30$), tournament size ($TS = 2$), mutation probability ($mprob = 0.1$), crossover probability ($cprob = 0.8$), α-crossover ($\alpha = 0.3$), *MaxGener* = 1000, *StopThres* = 0, *NrepThres* = 30. The interested reader is kindly referred to Appendix A for further details about how such parameters were selected.

4. Evaluation design

In this section, we specify some of the key features that subsequent human evaluation studies rely upon. Section 4.1 introduces the dataset and FRBCS whose classifications are explained. Then, Section 4.2 presents a novel metric for measuring the complexity of automated explanations.

4.1. Dataset and fuzzy inference system

The experiments have been carried out using the BEER dataset[6]. It contains characteristics of 400 instances of beer each of which belongs to one of 8 classes (Blanche, Lager, Pilsner, IPA, Stout, Barleywine, Porter, or Belgian Strong Ale). All data instances are described in terms of three features: color, strength, and bitterness. The corresponding linguistic terms and their ranges of values are displayed in Table 2. It is worth noting that all linguistic terms are commonsense and fully meaningful because they were provided by expert brewers.

In our experiments, we generate explanations for an FRBCS associated with the Fuzzy Unordered Rule Induction Algorithm (FURIA) [22]. The min–max inference mechanism [27] is applied so that both conjunction (AND) and implication (THEN) are implemented by the t-norm minimum, and the output accumulation is done by the t-conorm maximum. All membership functions are trapezoidal. All rule weights are set to the default value of 1. In addition, it is necessary to apply linguistic approximation as part of the explanation generation pipeline because FURIA rules are endowed only with local semantics. It is worth noting that such a linguistic approximation makes use of meaningful SFP-based linguistic terms as well as their combinations. Thus, explanations may contain combinations of adjacent terms (e.g., "*Feature*$_1$ is *Term*$_1$ or *Term*$_2$") with the aim of enhancing further their explanatory capacity. Fig. 3 illustrates the SFP associated to color.

In this work, we use the same FRBCS that was previously designed and evaluated in [43] with 10-fold cross-validation, achieving 95.5% of correctly classified instances and F1-score equals 0.954 (see the confusion matrix in Table 3 for further details). Notice that, with the aim of avoiding generation of misleading explanations and mainly because the present work focuses on the intended human evaluation, the misclassified test instances are excluded from further analysis in the rest of this manuscript. Whereas explaining misclassification is a challenging problem, it falls outside the scope of this work.

[3] Due to the well-known random heuristic nature of genetic algorithms, they avoid stacking in a local minimum but they can not always guarantee the convergence to the global minimum. Anyway, as shown in Appendix A, GEN succeeds to be effective in the search of "sub-optimal" solutions which are expected to be close enough to the optimal one.

[4] https://gitlab.citius.usc.es/ilia.stepin/fcfexpgen (branch "xor_euc_gen")

[5] https://gitlab.citius.usc.es/jose.alonso/guaje/

[6] The BEER dataset is publicly available athttps://dx.doi.org/10.13140/RG.2.2.20313.67680

Table 2
Numerical intervals associated to each SFP-based linguistic term.

Feature	Linguistic term	Range of values
Color	Pale	[0.0, 3.0]
	Straw	[3.0, 7.5]
	Amber	[7.5, 19.0]
	Brown	[19.0, 29.0]
	Black	[29.0, 45.0]
Bitterness	Low	[7.0, 21.0]
	Low-medium	[21.0, 32.5]
	Medium–high	[32.5, 47.5]
	High	[47.5, 250.0]
Strength	Session	[0.035, 0.052]
	Standard	[0.052, 0.067]
	High	[0.067, 0.090]
	Very high	[0.090, 0.136]

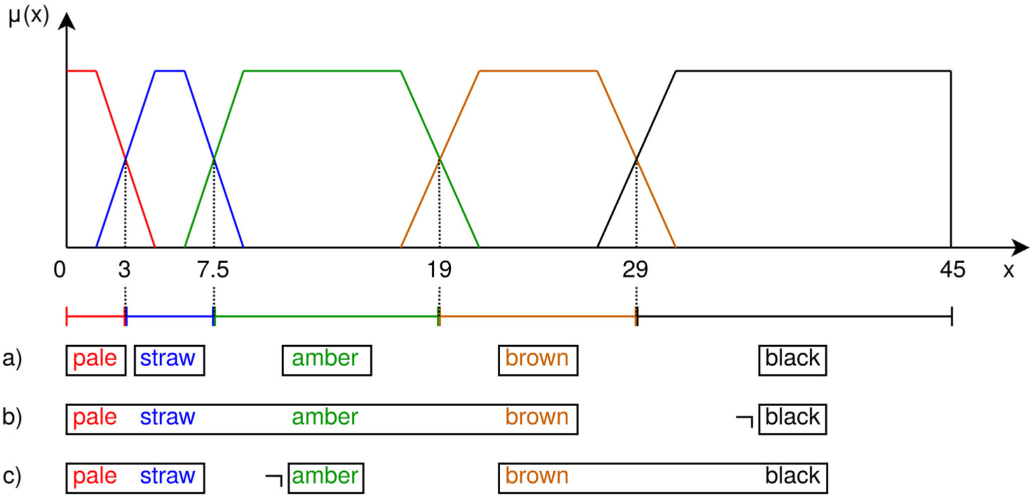


Fig. 3. Interpretation of SFP-based linguistic terms associated to Color.

Table 3
FURIA confusion matrix. UC stands for Unclassified instances.

	Predicted class								
Observed class	**BLA**	**LAG**	**PIL**	**IPA**	**STO**	**BAR**	**POR**	**BSA**	**UC**
Blanche (BLA)	50								
Lager (LAG)		48	1				1		
Pilsner (PIL)		1	49						
IPA			1	43		5			1
Stout (STO)					50				
Barleywine (BAR)				5		43	1		1
Porter (POR)		1			1		47		1
Belgian Strong Ale (BSA)					1	1	1	47	

4.2. Perceived explanation complexity

The use of explanations in natural language poses the problem of adequate estimation of explanation complexity. For example, it remains unclear whether the use of adjacent linguistic terms in an explanation (e.g., "…*if color were pale or straw*") increases or decreases understandability (and therefore effectiveness and usability) of such an explanation.

As the starting inspiring point for our proposal of automatic calculation of explanation complexity, we refer to existing readability tests in linguistics, which estimate how easily a text can be read by the intended audience. More precisely, the well-known *Gunning Fog Index* [19] is the weighted average of the normalized sentence length and the percentage of complex words in the text. Similarly, an estimate of complexity of a feature-based linguistic explanation (as perceived by the end user) may rely on the explanation length as well as on the number of features and linguistic terms used in the explanation.

In light of the above, we formally define the perceived explanation complexity (*PEC*) of an automated explanation e as follows:

$$PEC(e) = \lambda * \frac{min(l(e), \sigma)}{\sigma} + (1 - \lambda) * \frac{1}{|F|} \sum_{i=1}^{F_e} \frac{t^{f_i}}{|T_L^{f_i}|} \tag{2}$$

where $\lambda \in [0, 1]$ is the weight regularizing the impact of the explanation length and number of features and terms used in the explanation, $l(e)$ is the explanation length in characters, σ is a normalization hyperparameter over the explanation length, $|F|$ is the total number of features in the dataset, F_e is the number of unique features used in the given explanation, t^{f_i} is the number of terms associated with the i-th feature used in the explanation, $|T_L^{f_i}|$ is the power of the set of linguistic terms of the i-th feature.

In the case of the qualitative methods *XOR* and *EUC*, the basic linguistic terms to take into account are those already described in Table 2. However, in order to guarantee a fair comparison between quantitative and qualitative CF explanations, it is necessary to linguistically represent numerical feature value changes suggested by the quantitative method *GEN*. The sets of linguistic terms associated to each feature by the *GEN* method are the following:

$T_L(Color) = \{darker,\ slightly\ darker,\ lighter,\ slightly\ lighter\}$.

$T_L(Bitterness) = \{smaller,\ slightly\ smaller,\ bigger,\ slightly\ bigger\}$.

$T_L(Strength) = \{smaller,\ slightly\ smaller,\ bigger,\ slightly\ bigger\}$.

To illustrate computation of *PEC(e)*, let us consider the following example: given a data instance, $\lambda = 0.5$ and $\sigma = 150$, we have three alternative CF explanations with their corresponding complexity scores.

- *XOR*: "Beer style would be Stout if color were black."
  $PEC\ (e) = 0.5\ *\ \frac{46}{150} + 0.5\ *\ \frac{1}{3}\ *\ \frac{1}{5} = 0.153 + 0.033 = 0.186$
- *EUC*: "Beer style would be Stout if bitterness were low or low-medium, color were black, and strength were standard or high or very high."

  $PEC(e) = 0.5\ *\ \frac{130}{150} + 0.5\ *\ \frac{1}{3}\ *\ \left(\frac{2}{4} + \frac{1}{5} + \frac{3}{4}\right) = 0.433 + 0.242 = 0.675$

- *GEN*: "Beer style would be Stout if color were bigger (30.501) and strength were smaller (0.078)."

  $PEC(e) = 0.5\ *\ \frac{90}{150} + 0.5\ *\ \frac{1}{3}\ *\ \left(\frac{1}{4} + \frac{1}{4}\right) = 0.300 + 0.083 = 0.383$

Noteworthy, it always holds that $PEC(e) \in [0, 1]$. *PEC(e)* is null only if the explanation is empty and the associated weight $\lambda = 1$. On the contrary, the highest value of *PEC(e)* is obtained when the explanation length is equal to the normalization hyperparameter σ or all the dataset features and all the linguistic terms are included in the explanation. However, both of these special cases are of no interest, as the empty explanation has got null explanatory power whereas explanation including all the possible categories of features is clearly misleading.

5. Human evaluation

The human evaluation study consisted of two online questionnaires that allowed us to assess how the metric *PEC* is related to different explanation aspects. Section 5.1 presents the instruments and design of the first questionnaire (hereinafter referred to as *Survey GM* because the items to rate are associated to the so-called *Gricean Maxims* [15] as we will show below) as well as the analysis of collected data and the discussion of main results. In the light of lessons learned from this survey, we developed a subsequent one (hereinafter referred to as *Survey TS* because the focus is on assessing *Trustworthiness* and *Satisfaction* of the given explanations) whose experimental design and main discoveries are described in Section 5.2. In both surveys, all the subjects participated voluntarily and anonymously. This research obtained ethics approval from the University Ethics committee.

5.1. Survey GM: Evaluating CF explanations in terms of Gricean Maxims

5.1.1. Experimental settings

The first experiment was designed as a within-subject study. In order to perform a comparative analysis of qualitative and quantitative CF explanations, we considered only those test instances for which the qualitative methods (*XOR* and *EUC*) generated distinct explanations (thus avoiding misleading repetitions).

Since the BEER dataset has 8 classes, given a test instance we have 1 factual class and 7 alternative CF classes. Because the FURIA rules were trained and evaluated with 10-fold cross-validation, the 400 data instances in the BEER dataset were split 10 times into training set (90%) and test set (10%). As a result, we built 10 sets of FURIA rules. They were used to make predictions for all test instances in each fold (see details in Table 4). Then, we filtered out unclassified and misclassified test instances with the aim of avoiding the inclusion of void or misleading explanations to be evaluated in the survey. Notewor-

Table 4
Screening of test instances for defining the survey stimuli in terms of *XOR* and *EUC* CF explanations generated fold by fold. Unclassified instances are those for which no rule was activated. Misclassified instances are those where the FRBCS prediction does not match the ground-truth class label. Wrong factuals correspond to test instances for which wrong factual explanations were generated.

Fold	CV0	CV1	CV2	CV3	CV4	CV5	CV6	CV7	CV8	CV9
Test instances	40	40	40	40	40	40	40	40	40	40
Unclassified instances	–	–	2	–	–	1	–	–	–	–
Misclassified instances	2	2	3	1	3	1	2	–	3	3
Wrong factuals	1	–	1	–	1	2	–	1	–	2
Screened instances	37	38	34	39	36	36	38	39	37	35
CF explanations	259	266	238	273	252	252	266	273	259	245
Distinct CF pairs	25	28	29	24	20	36	32	74	28	38
Unique CF pairs	6	5	6	4	2	7	8	10	3	5

Table 5
Test instances and the corresponding CF explanations under study.

Task	Feature			Factual class	CF class	CF explanations		
	Color	Bitterness	Strength			XOR	EUC	*GEN*
1	17	87	0.096	Barleywine	IPA	if strength were high (*PEC=0.195*)	Beer style would be IPA if color were pale or straw or amber and strength were session or standard or high (*PEC=0.582*)	if strength were smaller (0.085) (*PEC=0.232*)
2	8	69	0.083	IPA	Barleywine	if strength were very high (*PEC=0.235*)	Beer style would be Barleywine if color were amber or brown or black and strength were very high (*PEC=0.465*)	if strength were bigger (0.096) (*PEC=0.252*)
3	5	34	0.068	Pilsner	Lager	if bitterness were low or low-medium or medium-high and color were amber (*PEC=0.488*)	Beer style would be Lager if bitterness were low or low-medium or medium-high and color were straw or amber (*PEC=0.552*)	if color were slightly darker (6.500) (*PEC=0.255*)
4	28	38	0.091	Belgian Strong Ale	Stout	if color were black (*PEC=0.186*)	Beer style would be Stout if bitterness were low or low-medium, color were black, and strength were standard or high or very high (*PEC=0.675*)	if color were darker (30.501) and strength were smaller (0.078) (*PEC=0.383*)
5	3	16	0.054	Blanche	Porter	if color were brown and strength were session or standard (*PEC=0.400*)	Beer style would be Porter if bitterness were low-medium or medium-high or high, color were brown, and strength were session or standard (*PEC=0.699*)	if color were darker (16.001) and strength were slightly smaller (0.052) (*PEC=0.416*)

thy, only 3 out of the 400 (0.75%) test instances (across all the folds) were unclassified, whereas 20 out of the 400 (5%) test instances were misclassified. Then, we generated CF explanations for each given prediction using the qualitative methods *XOR* and *EUC*. All in all, after careful screening, we identified all unique pairs of distinct CFs to exclude pieces of repeated explanations from the survey. Then, we picked 5 test instances representing illustrative cases (among the instances associated to all the previously identified unique CF pairs) that would be used as stimuli in the human evaluation study. Afterwards, we generated quantitative CF explanations for the selected stimuli using the *GEN* method.

Hence, *Survey GM* includes the following 5 tasks which were presented in a randomized order to each subject (see Table 5 for details). *Task 1* (predicted class: Barleywine, CF class: IPA) and *Task 2* (predicted class: IPA, CF class: Barleywine) represent pairs of classes where the classifier predicted the greatest number of incorrect results (see the confusion matrix from Table 3 in the previous section for details). Hereinafter we therefore refer to the first two tasks as "confusing" (CONF) while the rest of the tasks are deemed "non-confusing" (NON-CONF). In addition, the last three stimuli were selected by their relation to color. In *Task 3* (predicted class: Pilsner, CF class: Lager) both the predicted and CF classes are characterized by low values of color (i.e., from Pale to Amber in Table 2). In contrast, *Task 4* (predicted class: Belgian Strong Ale, CF class: Stout) the corresponding classes presume high values of color (i.e., Brown or Black in Table 2) for both factual and CF classes. Finally, the stimulus for *Task 5* (predicted class: Blanche, CF class: Porter) was selected to have contrastive values of color for the predicted and CF classes (i.e., Pale for Blanche versus Black for Porter).

Table 6
Explanation aspects under evaluation in *Survey GM*.

Related maxim of	Evaluation aspect	Description
Quantity	Informativeness	An estimate of how complete a CF explanation is perceived to be
Quality	Trustworthiness	An estimate of how credible a CF explanation is perceived to be
	Accuracy	An estimate of how precisely a CF explanation describes the CF class instances
Relevance	Relevance	An estimate of how pertinent the CF explanation details are in order to make a minimal change in feature values
Manner	Readability	An estimate of how grammatical a CF explanation is perceived to be

Table 7
Self-reported demographic data (*Survey GM*). The number of subjects comes along with the percentage in brackets for each category.

Demographic parameter	Number of participants
(a) Age	
18–25	3 (16.67%)
26–35	7 (38.89%)
36–45	5 (27.78%)
46–55	3 (16.67%)
(b) Gender	
Male	15 (83.33%)
Female	2 (11.12%)
Preferred not to say	1 (5.55%)
(c) Education	
Doctorate (Ph.D)	10 (55.56%)
Master's (M.A./M.Sc.)	7 (38.89%)
Bachelor's (B.A./B.Sc.)	1 (5.55%)
(d) English proficiency	
Native speaker	3 (16.67%)
Proficient (C2)	7 (38.89%)
Advanced (C1)	4 (22.22%)
Upper intermediate (B2)	4 (22.22%)
(e) Areas of expertise	
Explainable AI	12 (66.67%)
Fuzzy logic	9 (50.00%)
Mathematics	6 (33.33%)
Engineering	8 (44.44%)
Computer science	14 (77.78%)
Computational linguistics	4 (22.22%)
Social sciences	1 (5.56%)

The survey was implemented as an online questionnaire[7] which was developed in Python[8]. Each task screen included two panels. On the left panel, the factual explanation was given in the upper-left corner (for reference only) followed by three different CF explanations, each corresponding to one of the methods under study. The given test instance was depicted as a parallel coordinates plot below the explanations. On the right panel, the subjects were asked to rate each CF explanation on a 7-point Likert scale regarding several explanation aspects which are linked to the following Gricean Maxims [8,15]: *Maxim of quantity* (make your contribution as informative as is required without making it more informative than required); *Maxim of quality* (do not give information that is untruthful or lacks evidence); *Maxim of relevance* (present information pertinent to the discussion); and *Maxim of manner* (be clear and orderly, avoid ambiguity and obscurity).

It is worth noting that these four maxims were transformed into five explanation aspects (see Table 6). First, *informativeness* is related to the maxim of quantity and estimates whether the information present in the explanation sufficiently describes a necessary feature perturbation and whether it contains any unnecessary information. Then, the maxim of quality is represented by two explanation aspects. On the one hand, *trustworthiness* measures how credible the suggested changes are perceived (without them necessarily being accurate). On the other hand, *accuracy* indicates whether the suggestions found in the explanation are perceived to be correct and truly leading to the desired different outcome. In addition, the aspect of *relevance* aims to estimate how adequate the suggested changes are with respect to the test instance characteristics. Further, the aspect of *readability* estimates how grammatical and easy to read the given explanation is. Finally, in order

[7] https://tec.citius.usc.es/qxaisurvey1/

[8] https://gitlab.citius.usc.es/jose.alonso/surveygenerator

to estimate a degree of association between *PEC* and the estimated explanation aspects, Spearman's rank correlation coefficients (ρ) were calculated pairwise for the *PEC* scores and mean human evaluation scores of each explanation aspect. The threshold of $p = 0.05$ was used to confirm whether the correlation between *PEC* and the given explanation aspect exists.

5.1.2. Results

A total of 18 subjects participated in the *Survey GM*, each evaluating all the three explanation generation methods. All the demographic data collected from participants in *Survey GM* as well as their self-reported areas of expertise can be found in Table 7. To sum it up, 15 participants were males (83.33%), two were females (11.12%), and one person (5.55%) did not disclose its gender. In addition, all the participants held at least a Bachelor degree and had expertise in a wide range of sciences. Further, all the participants had at least the B2 level of English proficiency and represented various areas of expertise. Note that participants were allowed to select multiple areas.

Table 8 shows the mean and median human evaluation scores in *Survey GM* as well as the corresponding standard deviation (St.dev.). On average, the *EUC* explanations are perceived more informative than *GEN* or *XOR* explanations. However, the quantitative *GEN* method is perceived to generate more trustworthy explanations, *XOR* explanations being the second most credible, and the *EUC* method offering the least trustworthy explanations among the three methods. The *GEN* explanations are found more accurate than those generated by *XOR* and *EUC* methods. The *GEN* method also appears to generate more relevant explanations than *XOR* and *EUC*. Nevertheless, *XOR* explanations are perceived as grammatical as those offered by *GEN*, with *EUC* offering the least readable explanations, possibly due to their increased length.

As we consider all the sample explanations collectively, we observe important correlations between *PEC* and averaged scores for several explanation aspects. Thus, explanation complexity is found to moderately correlate with informativeness ($\rho = 0.545, p = 0.036$). In addition, strong negative correlations are observed between *PEC* and relevance ($\rho = -0.688, p = 0.005$) but also between *PEC* and readability ($\rho = -0.871, p < 0.001$). On the other hand, no conclusion can be made regarding the correlation either between *PEC* and trustworthiness ($\rho = -0.3, p = 0.278$) or between *PEC* and accuracy ($\rho = 0.07, p = 0.804$).

As for the "confusing" tasks alone, a strong negative correlation is found only between *PEC* and trustworthiness ($\rho = -0.87, p = 0.024$). The human evaluation scores for the other explanation aspects do not allow us to draw any other significant conclusions on their association with *PEC*. As for the "non-confusing" tasks alone, the findings testify that more complex explanations are perceived less readable ($\rho = -0.983, p < 0.001$).

The main lessons learned from this survey are as follows: (1) most participants agreed that the online questionnaire was long because it involved many different evaluation aspects for the three different methods; and (2) *PEC* turned out to be a good estimate for some of the explanation aspects under study. Then, we may take profit from these facts when designing future surveys: provide subjects with short questionnaires that regard only those specific aspects which cannot be inferred from *PEC*.

5.2. Survey TS: Evaluating Trustworthiness and Satisfaction of explanations

5.2.1. Experimental settings

In the light of lessons learned from previous survey, we defined a subsequent one. *Survey TS* was designed to have a simplified structure and follow a between-subject design where each subject would assess only one given explanation generation method. We considered the same stimuli as in the previous survey but focused only on trustworthiness and satisfaction of explanations instead. In the new questionnaire[9], the subjects were asked to evaluate the given CF explanation only in terms of trustworthiness and satisfaction. In addition, we adhered to the DARPA[10] [18] guidelines for assessing these explanation aspects on a 5-point Likert scale.

As designed previously, the task screens were presented in a randomized order to each subject. Similarly to *Survey GM*, Spearman's rank correlation coefficients (ρ) were calculated to estimate the association between *PEC* scores and human evaluation scores for trustworthiness and satisfaction. The same threshold value of $p = 0.05$ was used to verify whether such correlations existed.

5.2.2. Results

Sixty subjects participated in *Survey TS*. Each method was assessed by 20 participants independently. All the demographic data collected from participants are detailed in Table 9. Out of all the participants, a total of 57 (95 %) disclosed their demographic data. Thus, 32 of all the participants reported to be males (56.14%), 21 participants were females (36.84%) whereas 4 people (7.02%) preferred not to indicate their gender. Similarly to *Survey GM*, all the participants self-assessed their English language proficiency to be of at least the B2 level, and 53 out of 57 subjects disclosed their area of expertise.

Table 10 summarizes the human evaluation scores in *Survey TS*. Regarding trustworthy, *XOR* and *GEN* explanations are on average perceived nearly the same, the *EUC* explanations slightly falling behind. A similar pattern is observed for satisfaction. The quantitative *GEN* explanations are, in general, found to be the most satisfying. Nevertheless, the qualitative *XOR* expla-

[9] https://tec.citius.usc.es/cfsurvey/

[10] The acronym DARPA stands for Defense Advanced Research Projects Agency, which is the research and development agency of the USA.

Table 8
Survey GM results. ALL corresponds to the average for the five tasks. CONF averages only confusing tasks (1 and 2). NON-CONF averages only non-confusing tasks (3, 4 and 5). The highest average values for each (group of) task(s) and explanation aspect are highlighted in bold. Notice that, *PEC* values for ALL, CONF, and NON-CONF are averaged scores for the corresponding groups of tasks.

Task	Method	*PEC*	Informativeness			Trustworthiness			Accuracy			Relevance			Readability		
			Mean	**Median**	**St.dev.**	**Mean**	**Median**	**St.dev.**	**Mean**	**Median**	**St.dev.**	**Mean**	**Median**	**St.dev.**	**Mean**	**Median**	**St.dev.**
1	*XOR*	0.195	4.667	4.500	1.152	4.889	5.000	1.451	4.667	5.000	1.609	4.722	5.000	1.447	**5.778**	7.000	1.592
	EUC	0.582	**5.333**	5.000	1.188	4.333	5.000	1.749	4.667	5.000	1.814	3.611	3.500	1.754	4.500	4.500	1.917
	GEN	0.232	5.222	6.000	1.517	**5.222**	5.000	1.166	**5.500**	6.000	1.581	**5.056**	6.000	1.697	5.556	6.000	1.580
2	*XOR*	0.235	4.611	4.500	1.614	**4.778**	5.000	1.396	**4.778**	5.000	1.592	5.111	5.000	1.323	**6.222**	7.000	1.003
	EUC	0.465	**5.000**	5.000	1.414	4.222	5.000	1.865	3.889	4.500	2.083	4.611	5.000	1.685	4.333	4.500	1.815
	GEN	0.252	4.722	4.500	1.742	**4.778**	5.000	1.353	4.667	5.000	1.715	**5.278**	5.500	1.487	5.611	6.000	1.501
3	*XOR*	0.488	4.722	5.000	1.526	4.111	4.000	1.676	**4.444**	4.000	1.688	4.389	5.000	1.335	4.556	4.500	1.617
	EUC	0.552	**4.833**	5.000	1.339	**4.333**	5.000	1.414	4.167	4.000	1.618	4.167	4.000	1.383	4.333	4.500	2.058
	GEN	0.255	4.333	4.500	1.572	4.167	4.500	1.791	4.278	4.000	1.447	**4.667**	5.000	1.572	**6.000**	6.000	1.237
4	*XOR*	0.186	4.500	4.000	1.581	4.389	4.500	1.819	3.889	3.500	1.676	4.667	4.500	1.879	**6.278**	6.000	0.826
	EUC	0.675	5.389	6.000	1.501	4.944	5.000	1.211	4.778	5.000	1.734	4.278	5.000	1.487	3.833	3.000	2.121
	GEN	0.383	**5.556**	6.000	1.338	**5.833**	6.000	1.200	**5.833**	6.000	1.339	**5.611**	5.500	1.290	5.778	6.000	1.166
5	*XOR*	0.400	4.722	5.000	1.638	**5.000**	5.000	1.495	5.000	5.000	1.283	**4.889**	5.000	1.451	**5.667**	6.000	1.609
	EUC	0.699	**4.889**	5.000	1.231	4.667	5.000	1.534	**5.056**	5.000	1.474	4.389	4.000	1.787	4.056	4.000	1.893
	GEN	0.416	**4.889**	5.000	1.323	4.333	5.000	1.749	4.833	5.000	1.791	**4.889**	5.500	1.676	5.389	6.000	1.539
ALL	*XOR*	0.301	4.644	5.000	1.553	4.633	5.000	1.575	4.556	5.000	1.187	4.756	5.000	1.486	**5.700**	6.000	1.480
	EUC	0.595	**5.089**	5.000	4.944	4.500	5.000	1.560	4.511	5.000	1.769	4.211	4.000	1.625	4.211	4.000	1.934
	GEN	0.308	4.944	5.000	1.531	**4.867**	5.000	1.567	**5.022**	5.000	1.649	**5.100**	5.000	1.551	5.667	6.000	1.398
CONF	*XOR*	0.215	4.639	4.500	1.570	4.833	5.000	1.404	4.722	5.000	1.579	4.917	5.000	1.381	**6.000**	7.000	1.331
	EUC	0.524	**5.167**	5.000	1.298	4.278	5.000	1.783	4.278	5.000	1.966	4.111	4.500	1.769	4.417	4.500	1.842
	GEN	0.242	4.972	5.000	1.630	**5.000**	5.000	1.265	**5.083**	5.500	1.680	**5.167**	6.000	1.577	5.583	6.000	1.519
NON-CONF	*XOR*	0.358	4.648	5.000	1.556	4.500	5.000	1.680	4.444	5.000	1.598	4.648	5.000	1.556	5.500	6.000	1.551
	EUC	0.642	**5.037**	5.000	1.359	4.648	5.000	1.389	4.667	5.000	1.625	4.278	4.000	1.535	4.167	4.000	1.979
	GEN	0.351	4.926	5.000	1.478	**4.778**	5.000	1.745	**4.981**	5.000	1.642	**5.056**	5.000	1.547	**5.722**	6.000	1.433

Table 9
Self-reported demographic data (*Survey TS*). The number of subjects comes along with the percentage in brackets for each category.

Demographic parameter	Number of participants
(a) Age	
18–25	9 (15.79%)
26–35	19 (33.33%)
36–45	10 (17.54%)
46–55	10 (17.54%)
56–65	7 (12.28%)
66+	2 (3.52%)
(b) Gender	
Male	32 (56.14%)
Female	21 (36.84%)
Preferred not to say	4 (7.02%)
(c) Education	
Doctorate (Ph.D)	33 (57.89%)
Master's (M.A./M.Sc.)	17 (29.82%)
Bachelor's (B.A./B.Sc.)	5 (8.77%)
Short-cycle terciary	1 (1.76%)
Post-secondary non-terciary	1 (1.76%)
(d) English proficiency	
Native speaker	9 (15.79%)
Proficient (C2)	20 (35.09%)
Advanced (C1)	21 (36.84%)
Upper intermediate (B2)	7 (12.28%)
(e) Areas of expertise	
Explainable AI	29 (54.72%)
Fuzzy logic	14 (26.42%)
Mathematics	6 (11.32%)
Engineering	11 (20.75%)
Computer science	35 (66.04%)
Computational linguistics	22 (41.51%)
Social sciences	5 (9.43%)

Table 10
Survey TS results. ALL corresponds to the average for the five tasks. CONF averages only confusing tasks (1 and 2). NON-CONF averages only non-confusing tasks (3, 4 and 5). The highest average values for each (group of) task(s) and explanation aspect are highlighted in bold. Notice that, *PEC* values for ALL, CONF, and NON-CONF are averaged for the corresponding groups of tasks.

Task	Method	*PEC*	Trustworthiness			Satisfaction		
			Mean	Median	St.dev.	Mean	Median	St.dev.
1	*XOR*	0.195	**3.100**	3.000	1.221	2.750	2.000	1.545
	EUC	0.582	3.000	3.000	1.183	2.550	2.000	1.161
	GEN	0.232	**3.100**	3.500	1.338	**3.100**	3.000	1.375
2	*XOR*	0.235	3.100	3.000	1.179	2.900	3.000	1.261
	EUC	0.465	2.850	3.000	1.108	2.800	2.000	1.288
	GEN	0.252	**3.300**	3.500	1.308	**2.950**	3.000	1.117
3	*XOR*	0.488	**3.700**	4.000	1.345	**3.150**	3.500	1.352
	EUC	0.552	2.850	3.000	1.108	2.650	2.000	1.108
	GEN	0.255	3.150	3.000	1.152	2.950	3.000	1.203
4	*XOR*	0.186	3.300	3.000	1.382	3.150	3.000	1.424
	EUC	0.675	3.300	3.500	1.145	2.900	3.000	1.044
	GEN	0.383	**3.800**	4.000	1.030	**3.900**	4.000	1.179
5	*XOR*	0.400	**3.700**	4.000	1.100	**3.550**	4.000	1.203
	EUC	0.699	3.600	4.000	1.020	3.400	4.000	1.158
	GEN	0.416	3.400	3.000	1.020	3.200	3.000	1.077
ALL	*XOR*	0.301	**3.380**	4.000	1.279	3.100	3.000	1.389
	EUC	0.595	3.120	3.000	1.151	2.860	2.500	1.192
	GEN	0.308	3.350	4.000	1.203	**3.220**	3.000	1.246
CONF	*XOR*	0.215	3.100	3.000	1.200	2.825	3.000	1.412
	EUC	0.524	2.925	3.000	1.149	2.675	2.000	1.233
	GEN	0.242	**3.200**	3.500	1.327	**3.025**	3.000	1.255
NON-CONF	*XOR*	0.358	**3.567**	4.000	1.296	3.283	3.500	1.343
	EUC	0.642	3.250	3.000	1.135	2.983	3.000	1.147
	GEN	0.351	3.450	4.000	1.102	**3.350**	3.000	1.222

nations turn out more satisfying for certain tasks (3 and 5) whereas the *EUC* explanations appear less favorable in 4 out of the 5 tasks as well as on average.

Considering all the methods and tasks together, the findings from *Survey TS* do not allow us to make any conclusion regarding the correlation either between *PEC* and trustworthiness ($\rho = 0.07, p = 0.803$) or between *PEC* and satisfaction ($\rho = -0.081, p = 0.775$). The same situation is observed irrespective of the "confusing" nature of the tasks. On the one hand, there is a negative correlation but not enough statistical evidence for making distinctive conclusions in the case of "confusing" tasks: *PEC* versus trustworthiness ($\rho = -0.516, p = 0.295$); and *PEC* versus satisfaction ($\rho = -0.371, p = 0.468$). On the other hand, correlation coefficients are smaller but, once again, lack evidence in case of "non-confusing" tasks: *PEC* versus trustworthiness ($\rho = -0.025, p = 0.949$); *PEC* versus satisfaction ($\rho = -0.209, p = 0.589$).

6. Discussion

The explanation generation methods under study have a number of strengths and weaknesses. The qualitative methods favor two essential properties of CFs. First, the output CFs turn out to be diverse, as they can be mapped to a set of individual data points that are all equally minimally different on a categorical scale. Second, these methods are expected to maximize the validity of the generated CFs, as the corresponding explanations mimic the rules from the rule base. Therefore, following such explanations maximizes the probability of the corresponding CF rule to fire. On the other hand, the proposed qualitative methods may generate explanations that include a high number of features, some of them possibly being irrelevant or poorly explanatory.

The human evaluation study testifies that more complex explanations are perceived to be more informative, whereas increasing complexity jeopardizes readability and relevance. These findings specify the necessity of a careful design of automated explanations for specific tasks and/or application domains and/or intended audience. Thus, high-stakes decisions may require the corresponding explanations to be more informative and therefore encourage the use of methods that guarantee higher *PEC* scores of their output explanations (*EUC*). On the other hand, if the intended audience involved only lay users, more readable and therefore less complex explanations (*XOR* or *GEN*) may be preferred.

PEC scores allow us not only to quantify the perceived complexity of automatically computed CFs but also discern the most favorable of them. Lower *PEC* values appear to represent lower explanation complexity from user's point of view and therefore be more comprehensive. It can be seen that explanation length has a major impact on explanation complexity if the number of explanation features is low or if the linguistic terms used for such features are selected from a wider range of terms. Indeed, the terms covering narrower intervals appear more characteristic for the corresponding features and therefore more comprehensive. Further, the use of the proposed metric favors shorter but more informative (in terms of the number of features and/or linguistic terms used) explanations. Hence, driven by a complexity-oriented approach to evaluating CFs, a better understanding of a feature-based explanation can be reached by finding a balance between short enough explanation length and the number of unique features and/or linguistic terms used in the explanation.

Importantly, *PEC* can help to choose among alternative but semantically equivalent explanations. For example, the piece of explanation "if color were pale or straw or amber or brown" can be replaced by the shorter "if color were not black" (see Fig. 3). Then, it becomes essential to define how many linguistic terms are necessary to be properly understood to guarantee a consistent use of the metric. We thus suggest two strategies to calculate the number of terms associated to a feature if the term under consideration is negated. On the one hand, it may be sufficient to calculate the sum of the non-negated terms. In this case, the number of linguistic terms in the aforementioned explanation $t^{Color} = |\{pale, straw, amber, brown\}| = 4$ (see Fig. 3b). On the other hand, it may be argued that, to fully understand the meaning of the negated term, it is only necessary to understand the meaning of the negated term itself (*black*, in this case) as well as that of the collective linguistic terms covering all the contrasting linguistic terms (i.e., lighter/darker than black). Thus, if the negated linguistic variable takes on either of the extreme values (e.g., *pale* or *black*), the number of terms associated with the given explanation for feature t^{f_i} always equals 2. Moreover, if the fuzzy partition presumes that both lower and higher values can be captured by other linguistic terms with respect to the negated term (e.g., "*...if color were not amber*"), the number of the associated terms always includes the negated term as well as the values from the extended set of terms covering both smaller and higher corresponding intervals (see Fig. 3c).

7. Concluding remarks and future work

In this paper, we presented one quantitative method (*GEN*) of CF explanation generation and two methods (*XOR* and *EUC*) of qualitative CF explanation generation for FRBCSs. As all of them provide the end user with output of different kinds, they can be used solely or complementarily to offer explanations on demand and customized for different user profiles. In addition, we proposed the new metric *PEC* for estimating the complexity of a given explanation (as expected to be perceived by an end user).

To evaluate the proposed methods, we collected human evaluation scores in an empirical study which comprised two online questionnaires. In addition, we computed *PEC* scores for each of the explanations under consideration in the study. We observed that a more complexly structured within-subject questionnaire (*Survey GM*) appears to provide a better insight into the goodness of automated explanations given an equivalent number of participants. However, collecting data in such a survey is costly as it requires higher cognitive load and more time from the participants. Therefore, calculating *PEC* automatically allows the survey designer to set up and deploy a shorter questionnaire and thus easier to fill (*Survey TS*). It is worth noting that *PEC* strongly correlates with several explanation aspects but does so in different directions, so an FRBCS designer is advised to carefully select the method of explanation generation based on the peculiarities of the application domain and/or intended audience.

All in all, the insights from this work are expected to advance methods of generation and evaluation for various explanation approaches. As such, they are expected to be helpful for designing future human evaluation surveys in the area of explainable AI. Moreover, as part of future work, we will go deeper with selecting and fusing CF explanations with the aim of customizing them for users having different profiles in different application scenarios. Further research is therefore necessary: (1) to extend the proposed CF explanation generation methods beyond numerical features; (2) to better assess the impact of the *PEC* hyperparameters (σ and λ); and (3) to better understand the connection between complexity and trustworthiness of automated explanations. Notice that, the conclusions derived from the current study are only applicable to the target population under consideration. As part of future work, for the sake of generalization, we intend to design and carry out other similar experiments with a larger and wider panel of respondents, including non-expert lay users. Finally, we plan to use *PEC* as one of the criteria to optimize when designing explainable multi-objective evolutionary fuzzy systems.

Funding

Ilia Stepin is an *FPI* researcher (grant PRE2019-090153). Jose M. Alonso-Moral is a *Ramon y Cajal* researcher (grant RYC-2016–19802). This work was supported by the Spanish Ministry of Science and Innovation (grants RTI2018-099646-B-I00, PID2021-123152OB-C21, and TED2021-130295B-C33) and the Galician Ministry of Culture, Education, Professional Training and University (grants ED431F2018/02, ED431G2019/04, and ED431C2022/19). All the grants were co-funded by the European Regional Development Fund (ERDF/FEDER program).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

In addition to the human evaluation study on the automatically generated CFs, we performed three independent experiments on the genetic algorithm hyperparameter fine-tuning. In particular, we estimated the impact of the following hyperparameters associated to the *GEN* method: (i) the size of the population, (ii) the crossover probability and the corresponding alpha value, and (iii) the mutation probability. All the experiments were run for the five survey stimuli where both the predicted classes and the CF classes were known. The experimental results were assessed in terms of the best achieved fitness scores.

Fig. 4 summarizes the impact of the population size (10, 20, 30, 40, 50). It can be observed that the default population size (30) provides good results, on average, for all the test instances under consideration.

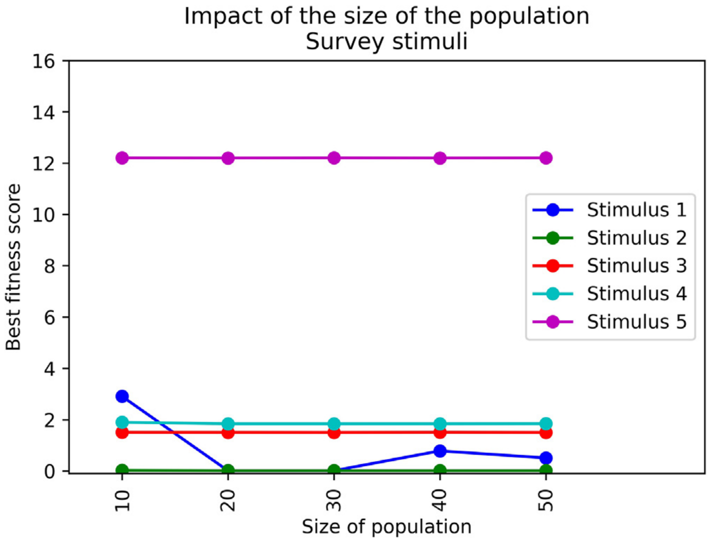


Fig. 4. An empirical assessment of the impact of the population size in the GEN method.

Fig. 5 shows the results of the experiment on the crossover probability values (0.7, 0.8, 0.9), considering different α values (0.2, 0.3, 0.4). In short, the combination of the crossover probability (0.8) and $\alpha = 0.3$ yields the best results for the considered CF data points.

Fig. 6 illustrates the impact of the selected mutation probability values (0.05, 0.1, 0.15, 0.2). It can be seen that doubling the default mutation probability value may result in worsened performance of the algorithm.

To sum it up, the analysis carried out allows us to conclude that the selected hyperparameter values do not only agree with the guidelines found in the literature (e.g., [21]) but also prove to be effective in the given experiments and can indeed be recommended for future use. All the detailed calculations as well as additional plots and the source code for replicating this experimental analysis can be found in our Gitlab repository:https://gitlab.citius.usc.es/ilia.stepin/fcfexpgen (branch “xor_euc_gen”).

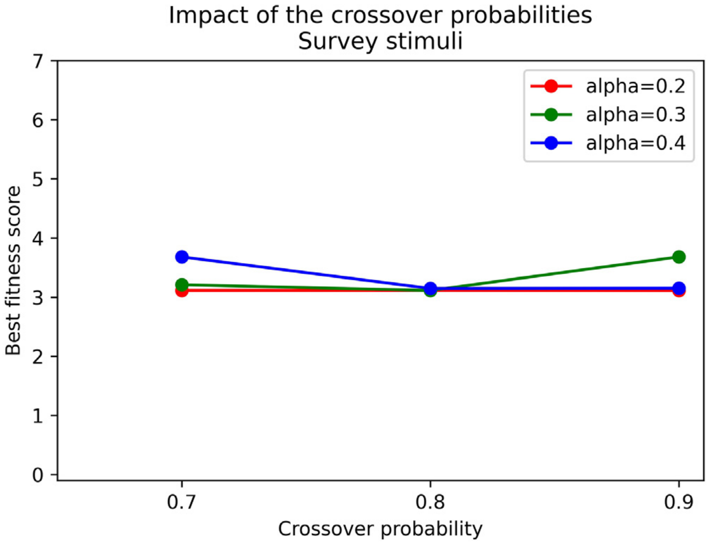


Fig. 5. An empirical assessment of the impact of the crossover hyperparameters (the crossover probability and the α crossover operator) in the GEN method.

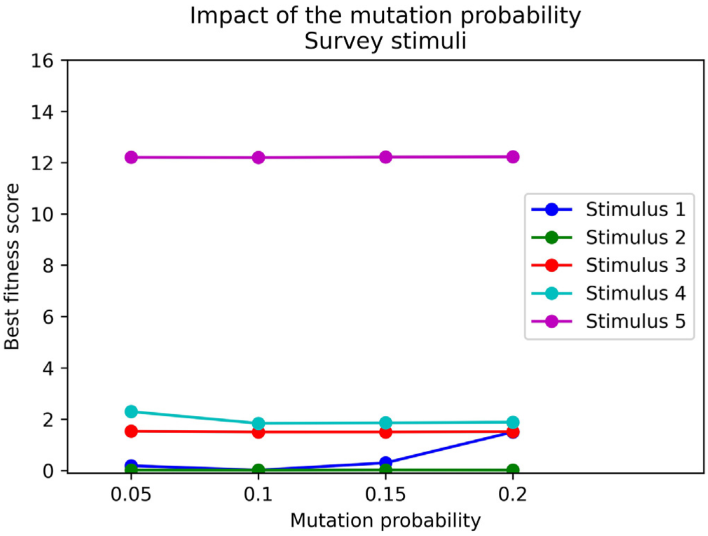


Fig. 6. An empirical assessment of the impact of the mutation probability in the GEN method.

References

[1] A. Abdul, J. Vermeulen, D. Wang, B.Y. Lim, and M. Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In Proceedings of the Conference on Human Factors in Computing Systems (CHI), pages 1–18, Montreal QC, Canada, 2018. Association for Computing Machinery. https://doi.org/10.1145/3173574.3174156.

[2] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160, https://doi.org/10.1109/ACCESS.2018.2870052.

[3] J.M. Alonso, C. Castiello, L. Magdalena, C. Mencar, Explainable Fuzzy Systems - Paving the Way from Interpretable Fuzzy Systems to Explainable AI Systems, volume 970, Springer International Publishing (2021), https://doi.org/10.1007/978-3-030-71098-9.

[4] J.M. Alonso, O. Cordón, S. Guillaume, and L. Magdalena. Highly interpretable linguistic knowledge bases optimization: Genetic tuning versus solis-wetts. Looking for a good interpretability-accuracy trade-off. In Proceedings of the IEEE International Conference on Fuzzy Systems, pages 901–906, London, UK, 2007. https://doi.org/10.1109/FUZZY.2007.4295485.

[5] I. Baaj and J.-P. Poli. Natural language generation of explanations of fuzzy inference decisions. In Proceedings of the IEEE International Conference on Fuzzy Systems, pages 1–6, New Orleans, LA, USA, 2019. https://doi.org/10.1109/FUZZ-IEEE.2019.8858994.

[6] A. Cascallar-Fuentes, A. Ramos-Soto, A. Bugarín, Adapting SimpleNLG to Galician Language, in: In Proceedings of the International Conference on Natural Language Generation, Association for Computational Linguistics (ACL), 2018, https://doi.org/10.18653/v1/W18-6507.

[7] O. Cordón, F. Herrera, A Three-Stage Evolutionary Process for Learning Descriptive and Approximate Fuzzy Logic Controller Knowledge Bases from Examples, International Journal of Approximate Reasoning 17 (4) (1997) 369–407, https://doi.org/10.1016/S0888-613X(96)00133-8.

[8] R. Dale, E. Reiter, Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions, Cognitive science 19 (2) (1995) 233–263, https://doi.org/10.1016/0364-0213(95)90018-7.

[9] V. Dignum, Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Artificial Intelligence: Foundations, Theory, and Algorithms, Springer, Cham, 2019, https://doi.org/10.1007/978-3-030-30371-6.

[10] L.J. Eshelman and J.D. Schaffer. Real-Coded Genetic Algorithms and Interval-Schemata. In L. Darrell Whitley, editor, Foundations of Genetic Algorithms, volume 2 of Foundations of Genetic Algorithms, pages 187–202. Elsevier, 1993. https://doi.org/10.1016/B978-0-08-094832-4.50018-0.

[11] A. Fernandez, F. Herrera, O. Cordon, M. Jose del Jesus, F. Marcelloni, Evolutionary Fuzzy Systems for Explainable Artificial Intelligence: Why, When, What for, and Where to?, IEEE Computational Intelligence Magazine 14 (1) (2019) 69–81, https://doi.org/10.1109/MCI.2018.2881645.

[12] R.R. Fernández, I.M. de Diego, V. Aceña, A. Fernández-Isabel, J.M. Moguerza, Random forest explainability using counterfactual sets, Information Fusion 63 (2020) 196–207, https://doi.org/10.1016/j.inffus.2020.07.001.

[13] S. Fletcher, M.Z. Islam, Comparing sets of patterns with the Jaccard index, Australasian Journal of Information Systems 22 (2018), https://doi.org/10.3127/ajis.v22i0.1538.

[14] A. Gatt, E. Krahmer, Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation, Journal of Artificial Intelligence Research 61 (2018) 65–170, https://doi.org/10.1613/jair.5477.

[15] H.P. Grice, Logic and Conversation, in: P. Cole, J.L. Morgan (Eds.), Syntax and Semantics: Speech Acts, Academic Press, 1975, pp. 41–58, https://doi.org/10.1163/9789004368811_003.

[16] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, Data Mining and Knowledge Discovery (2022) 1–55, https://doi.org/10.1007/s10618-022-00831-6.

[17] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, F. Turini, Factual and Counterfactual Explanations for Black Box Decision Making, IEEE Intelligent Systems 34 (6) (2019) 14–23, https://doi.org/10.1109/MIS.2019.2957223.

[18] D. Gunning, E. Vorm, J.Y. Wang, M. Turek, DARPA's explainable AI (XAI) program: A retrospective, Applied AI Letters 2 (4) (2021), https://doi.org/10.1002/ail2.61 e61.

[19] R. Gunning, Technique of clear writing, McGraw-Hill, 1968.

[20] C. Herley and W. Pieters. If You Were Attacked, You'd Be Sorry: Counterfactuals as Security Arguments. In Proceedings of the 2015 New Security Paradigms Workshop, NSPW '15, pages 112–123, New York, NY, USA, 2015. Association for Computing Machinery. https://doi.org/10.1145/2841113.2841122.

[21] F. Herrera, M. Lozano, A.M. Sánchez, A Taxonomy for the Crossover Operator for Real-Coded Genetic algorithms: An Experimental Study, International Journal of Intelligent Systems 18 (3) (2003) 309–338, https://doi.org/10.1002/int.10091.

[22] J. Hühn, E. Hüllermeier, FURIA: an algorithm for unordered fuzzy rule induction, Data Mining and Knowledge Discovery 19 (3) (2009) 293–319, https://doi.org/10.1007/s10618-009-0131-8.

[23] H. Ishibuchi, T. Nakashima, M. Nii, Classification and modeling with linguistic information granules: Advanced approaches to linguistic Data Mining, Springer Science & Business Media (2004), https://doi.org/10.1007/b138232.

[24] M. Lash, Q. Lin, N. Street, J. Robinson, and J. Ohlmann. Generalized Inverse Classification. In Proceedings of the International Conference on Data Mining (SDM), pages 162–170. Society for Industrial and Applied Mathematics, 2017. https://doi.org/10.1137/1.9781611974973.19.

[25] A. Lucic, H. Haned, and M. de Rijke. Why Does My Model Fail? Contrastive Local Explanations for Retail Forecasting. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, pages 90–98, Barcelona, Spain, 2020. Association for Computing Machinery. https://doi.org/10.1145/3351095.3372824.

[26] N. Maaroof, A. Moreno, A. Valls, M. Jabreel, M. Szelag, A Comparative Study of Two Rule-Based Explanation Methods for Diabetic Retinopathy Risk Assessment, Applied Sciences 12 (7) (2022) 1–18, https://doi.org/10.3390/app12073358.

[27] E.H. Mamdani, Application of Fuzzy Logic to Approximate Reasoning Using Linguistic Systems, IEEE Transactions on Computers 26 (12) (1977) 1182–1191, https://doi.org/10.1109/TC.1977.1674779.

[28] G.A. Miller, The magical number seven, plus or minus two: Some limits on our capacity for processing information, Psychological Review 63 (2) (1956) 81–97, https://doi.org/10.1037/0033-295x.101.2.343.

[29] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38, https://doi.org/10.1016/j.artint.2018.07.007.

[30] J. Moore, N. Hammerla, and C. Watkins. Explaining deep learning models with constrained adversarial examples. In Proceedings of the Pacific Rim International Conference on Artificial Intelligence (PRICAI), pages 43–56. Springer, 2019. https://doi.org/10.1007/978-3-030-29908-8_4.

[31] R.K. Mothilal, A. Sharma, and C. Tan. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, pages 607–617, Barcelona, Spain, 2020. Association for Computing Machinery. https://doi.org/10.1145/3351095.3372850.

[32] M.L. Olson, R. Khanna, L. Neal, F. Li, W.-K. Wong, Counterfactual state explanations for reinforcement learning agents via generative deep learning, Artificial Intelligence 295 (2021) 1–29, https://doi.org/10.1016/j.artint.2021.103455.

[33] Parliament and Council of the European Union. General Data Protection Regulation (GDPR), 2016. URL:http://data.europa.eu/eli/reg/2016/679/oj.

[34] Parliament and Council of the European Union. A European Approach to Artificial Intelligence, 2022. URL:https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence.

[35] E. Reiter, R. Dale, Building Natural Language Generation Systems, in: Studies in Natural Language Processing, Cambridge University Press, 2000, https://doi.org/10.1017/CBO9780511519857.

[36] M.T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 1135–1144, San Francisco, California, USA, 2016. Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778.

[37] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (5) (2019) 206–215, https://doi.org/10.1038/s42256-019-0048-x.

[38] M. Schleich, Z. Geng, Y. Zhang, and D. Suciu. GeCo: Quality Counterfactual Explanations in Real Time. In Proceedings of the Very Large Data Bases (VLDB) Endowment, volume 14(9), pages 1681–1693, 2021. https://doi.org/10.14778/3461535.3461555.

[39] S. Sharma, J. Henderson, J. Ghosh. CERTIFAI, A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-Box Models, Association for Computing Machinery, 2020, pp. 166–172, https://doi.org/10.1145/3375627.3375812.

[40] K. Sokol and P. Flach. One Explanation Does Not Fit All: The Promise of Interactive Explanations for Machine Learning Transparency. KI – Künstliche Intelligenz, 2020. https://doi.org/10.1007/s13218-020-00637-y.

[41] I. Stepin, J.M. Alonso, A. Catala, and M. Pereira-Fariña. Generation and evaluation of factual and counterfactual explanations for decision trees and fuzzy rule-based classifiers. In Proceedings of the IEEE World Congress on Computational Intelligence (WCCI), Glasgow, UK, 2020. https://doi.org/10.1109/FUZZ48607.2020.9177629.

[42] I. Stepin, J.M. Alonso, A. Catala, M. Pereira-Fariña, A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence, IEEE Access 9 (2021) 11974–12001, https://doi.org/10.1109/ACCESS.2021.3051315.

[43] I. Stepin, A. Catala, M. Pereira-Fariña, J.M. Alonso, Factual and Counterfactual Explanation of Fuzzy Information Granules, in: Interpretable Artificial Intelligence: A Perspective of Granular Computing, Springer International Publishing, 2021, pp. 153–185, https://doi.org/10.1007/978-3-030-64949-4_6.

[44] R. Sukkerd, R. Simmons, D. Garlan, Toward Explainable Multi-Objective Probabilistic Planning, in: IEEE/ACM 4th International Workshop on Software Engineering for Smart Cyber-Physical Systems (SEsCPS), Gothenburg, Sweden, 2018, pp. 19–25.

[45] S. Verma, J. Dickerson, and K. Hines. Counterfactual explanations for machine learning: A review. In Proceedings of the Machine Learning: Retrospectives, Surveys and meta-Analyses (ML-RSA) Workshop at the Conference on Neural Information Processing Systems (NeurIPS), 2020.

[46] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR, Harvard Journal of Law & Technology 31 (2) (2018) 841–887, https://doi.org/10.2139/ssrn.3063289.

[47] X. Wang, M. Yin, Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making, in: 26th International Conference on Intelligent User Interfaces, IUI '21, Association for Computing Machinery, 2021, pp. 318–328, https://doi.org/10.1145/3397481.3450650.

[48] C. Woodcock, B. Mittelstadt, D. Busbridge, G. Blank, et al, The Impact of Explanations on Layperson Trust in Artificial Intelligence-Driven Symptom Checker Apps: Experimental Study, Journal of Medical Internet Research 23 (11) (2021), https://doi.org/10.2196/29386 e29386.

[49] L.A. Zadeh, The concept of a linguistic variable and its application to approximate reasoning, Information Sciences 8 (3) (1975) 199–249, https://doi.org/10.1016/0020-0255(75)90036-5.

Argument & Computation -1 (2023) 1–59
DOI 10.3233/AAC-220011
IOS Press

Information-seeking dialogue for explainable artificial intelligence: Modelling and analytics

Ilia Stepin [a,c,*], Katarzyna Budzynska [b], Alejandro Catala [a,c], Martín Pereira-Fariña [d] and Jose M. Alonso-Moral [a,c]

[a] *Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Rúa de Jenaro de la Fuente Domínguez s/n, 15782 Santiago de Compostela, A Coruña, Spain*
E-mails: ilia.stepin@usc.es, alejandro.catala@usc.es, josemaria.alonso.moral@usc.es

[b] *Laboratory of The New Ethos, Warsaw University of Technology, plac Politechniki 1, 00-661, Warsaw, Poland*
E-mail: katarzyna.budzynska@gmail.com

[c] *Departamento de Electrónica e Computación, Universidade de Santiago de Compostela, Rúa Lope Gómez de Marzoa, s/n, 15782 Santiago de Compostela, A Coruña, Spain*

[d] *Departamento de Filosofía e Antropoloxía, Universidade de Santiago de Compostela, Plaza de Mazarelos s/n, 15705 Santiago de Compostela, A Coruña, Spain*
E-mail: martin.pereira@usc.es

Abstract. Explainable artificial intelligence has become a vitally important research field aiming, among other tasks, to justify predictions made by intelligent classifiers automatically learned from data. Importantly, efficiency of automated explanations may be undermined if the end user does not have sufficient domain knowledge or lacks information about the data used for training. To address the issue of effective explanation communication, we propose a novel information-seeking explanatory dialogue game following the most recent requirements to automatically generated explanations. Further, we generalise our dialogue model in form of an explanatory dialogue grammar which makes it applicable to interpretable rule-based classifiers that are enhanced with the capability to provide textual explanations. Finally, we carry out an exploratory user study to validate the corresponding dialogue protocol and analyse the experimental results using insights from process mining and argument analytics. A high number of requests for alternative explanations testifies the need for ensuring diversity in the context of automated explanations.

Keywords: Explainable Artificial Intelligence, information-seeking dialogue game, explanation locutions, counterfactual explanation, process mining analytics, argument analytics

1. Introduction

Explainability in the context of Artificial Intelligence (AI) has long attracted attention of researchers from computer science [57] and argumentation [21]. The first explanation generation methods turned up

[*] Corresponding author. Tel.: +34 8818 16394; E-mail: ilia.stepin@usc.es.

in the 1980s along with the so-called Expert Systems [74]. More precisely, the first explainers addressed the challenge of explaining the output of expert systems and logic programs [7], which eventually led to the emergence of the research field that we now call Computational Argumentation. Recent years have witnessed a new boost of interest in developing eXplainable AI (XAI), as novel machine learning (ML) algorithms produce highly accurate yet oftentimes poorly explainable predictions [1]. As defined at present, XAI aims to (1) generate explainable models preserving a high level of accuracy and (2) enable the end user, e.g., a client of a bank or a patient of a hospital, with the opportunity to understand, trust, and manage the given AI-based systems [2,29] (e.g., querying a bank loan management system to identify reasons for the loan application being rejected or a hospital information system to receive treatment-related recommendations).

The obscure nature of the underlying reasoning of the state-of-the-art predictive algorithms has given way to the so-called "right to explanation" [80]. The corresponding legal regulations are being increasingly adopted worldwide [87]. For example, the European Union (EU)'s General Data Protection Regulation (GDPR) acknowledges the right of the user "not to be subject to a decision evaluating personal aspects relating to him or her which is based solely on automated processing and which produces adverse legal effects concerning, or significantly affects, him or her" [51]. In addition, current EU's legal regulations in, for example, the financial domain require that algorithmic transparency be provided for automatic trading techniques (see the Directive 2014/65/EU on Markets in Financial Instruments, commonly known as MiFID II [52] for details). Being a controversial topic of primary importance for numerous stakeholders, its juridical basis is constantly updated. Thus, the newly proposed EU's AI Act (AIA) [53] establishes a taxonomy of AI-based systems and requires that high-risk AI applications offer explanations for their decisions or recommendations to their end users.

In order to mitigate algorithmic transparency issues of the state-of-the-art AI algorithms, a use of interpretable models is advised [59]. Interpretable rule-based models (such as, e.g., decision trees (DT) or decision rules) are known to provide user-friendly explanations [47]. Remarkably, DTs can be used as part of more complex model-agnostic explainers that are able to justify predictions of other arbitrary classifiers if they are, for example, trained on a local synthetically generated neighbourhood around the test instance [28]. Despite the fact that only few XAI frameworks offer explanations in natural language [12], DTs have also been shown to be a powerful tool for communicating textual explanations to end users, e.g., by engaging the user in an explanatory dialogue [70,79].

Explanations are claimed to have to necessarily be embedded in a dialogical interaction so that the end user is able to challenge the aspects of an explanation that have not been understood [63]. For illustrative purposes, let us consider a beer style classification problem (see Table 1 for details). Given a number of predefined classes (i.e., beer styles) and an instance of beer, the classifier (System) makes a prediction on what beer style the test instance is (move m_1). The end user (User) engages in a dialogue with the System to inquire an explanation for the given prediction (m_2). Then, the System provides the User with a factual explanation based on the most distinctive features, pertinent to the test instance, that led to the given prediction (m_3). Then, given the factual explanation, the User may want to clarify specific features (e.g., bitterness) that the offered explanation contains (m_4). In response, the System offers a definition for the requested feature (m_5). Alternatively, the User may inquire the System to explain why the given test instance is not predicted to be of another class (m_6, m_{10}). The pieces of such counterfactual explanations offered in response suggest minimal changes to the input so that its classification changes in the desired manner (m_7, m_{11}). Given the initial explanation, the User may want to specify a more precise description of the given feature (m_8). Then, the System may present detailed information over the corresponding numerical interval (subject to availability) over which the feature is defined (m_9). Further,

Table 1

An illustrative explanatory dialogue

Move	*Agent*	*Utterance*
m_1	System	The present beer is Blanche.
m_2	User	Could you explain why you think so?
m_3	System	It is Blanche because colour is pale and bitterness is low.
m_4	User	What do you mean by bitterness?
m_5	System	Bitterness is a harsh, acrid taste that is one of the four basic taste sensations.
m_6	User	But is the beer not stout?
m_7	System	It would be Stout if colour were black and bitterness were high.
m_8	User	Could you specify how high bitterness is defined?
m_9	System	Sure! I define bitterness as high because it lies in the range from 47 to 250.
m_{10}	User	But why is it not Porter?
m_{11}	System	It would be Porter if colour were brown.
m_{12}	User	I am not quite satisfied with your explanation. Could you offer me another one?
m_{13}	System	Sure! It would be of class Porter if colour were brown and strength were session.
m_{14}	User	Ok, now I trust your prediction.
m_{15}	System	Thank you for your trust in me. Bye!

the User may disagree with the explanation offered and argue over it (m_{12}). The System should then offer an alternative explanation that would satisfy the User's needs (m_{13}). When the User is sufficiently informed about the reasons that led to the given prediction, he or she makes an informed decision on whether the System's prediction should be trusted or not (m_{14}). The explanatory dialogue ends with the System's farewell locution (m_{15}).

As follows from Table 1, we consider two types of explanations: factual and counterfactual. Assuming knowledge of the feature space, factual explanations (illustrated with move m_3 in Table 1) aim to explain the given classifier's prediction in terms of the most relevant feature values that led to that prediction. On the contrary, counterfactuals (illustrated with moves m_7, m_{11}, and m_{13} in Table 1) are post-hoc example-based explanations that suggest a minimal change in feature values to those of the given data instance so that the system's prediction changes as desired [71].

This paper introduces an explanatory dialogue game for communicating factual and counterfactual explanations for interpretable rule-based classifiers. We assume that the classifier is associated with an explainer that is capable of providing textual (rule-based) explanations. Based on the dialogue typology proposed by Walton and Krabbe [82], we model the information-seeking type of explanatory dialogue equipping it with a specific collection of locutions tailored for the aforementioned types of explanation that the user may ask the system. As a starting point, we consider the typology of dialogue moves proposed by Budzynska et al. [9]. In our work, we extend this typology of dialogue moves with a repertoire of locutions allowing for communication of factual and counterfactual explanations to enable the end user to interactively explore the explanation space. Then, we propose a context-free dialogue grammar to generalise the formal structure of the resulting dialogue model. Despite an empirically shown strong need in both factual and counterfactual explanations [41] and at least a hundred of counterfactual explanation generation methods proposed by now in the context of XAI, less than a third of these methods are evaluated in user studies [37]. To address this issue, we subsequently perform a pilot user study to evaluate the proposed dialogue model. Moreover, we analyse the collected dialogue transcripts treating instances of explanatory dialogue as processes using the state-of-the-art techniques from process mining and argument analytics [43].

As a result, we bridge the gap between ML practitioners and the argumentation community by making the following contributions:

- we model information-seeking explanatory dialogue based on the fundamental notions from the argumentation theory and apply the dialogue model in the context of XAI;
- we propose a set of original dialogue locution types that are found specifically suitable for effective communication of factual and counterfactual explanations;
- we demonstrate the explanatory utility of the proposed dialogue protocol via a human evaluation study based on three use cases for an interpretable rule-based classifier leaving open-source implementations of the dialogue game and the human evaluation toolkit available for public use;
- we suggest formal means for extending the proposed protocol to make it applicable to modelling dialogic human-machine interaction for classification tasks in other applications.

The rest of the manuscript is structured as follows. Section 2 introduces the classification problem formally and outlines the common properties of explanations claimed to be essential for explaining solutions to such a problem. In addition, we subsequently discuss possible discrepancy between automatically generated explanations and user-preferred explanations. Section 3 defines an explanatory dialogue game as an interface between an explanation generation module and the end user. Section 4 introduces essential process mining concepts and shows how we apply them to explanatory dialogue analysis. Section 5 presents the experimental settings of the human evaluation study carried out to assess the utility of the proposed dialogue protocol. Section 6 reports the experimental results obtained from the human evaluation study. Section 7 discusses the dialogue model validation results. Section 8 presents an overview of related work regarding formal explanatory dialogue models as well as recent argumentation-based techniques for explanatory dialogue modelling. Finally, we outline prospective directions for future work and conclude in Section 9.

2. Preliminaries

In this section, we first outline a definition of the classification problem and assumptions about the nature of classifiers and explainers that we are driven by (see Section 2.1 for details). Then, we formally define essential explanation-related concepts that we utilise throughout the manuscript in Section 2.2. Finally, we draw reader's attention to possible discrepancies between the user-preferred explanations and those offered to him or her by the explainer in Section 2.3.

2.1. The classification problem

As outlined in Section 1, we focus on communicating to the end user automated explanations for the output of an interpretable rule-based ML classifier. Figure 1 depicts a general architecture of the modelled explanation communication process. The System is assumed to include, at least, the following core components: an interpretable rule-based classifier, an explainer, a knowledge base, and a dataset that the classifier is trained on. The User starts the communication process by sending a classification request for a specific test instance to the System in form of the test instance's characteristics (i.e., features). The classifier is pretrained on a given dataset $X = \{x_i\}|_{i=1}^{n}$ containing n labelled instances to learn a mapping function $c : X \longrightarrow Y$ where $Y = \{y_j\}|_{j=1}^{m}$ is a discrete output variable (class), m being the number of classes.

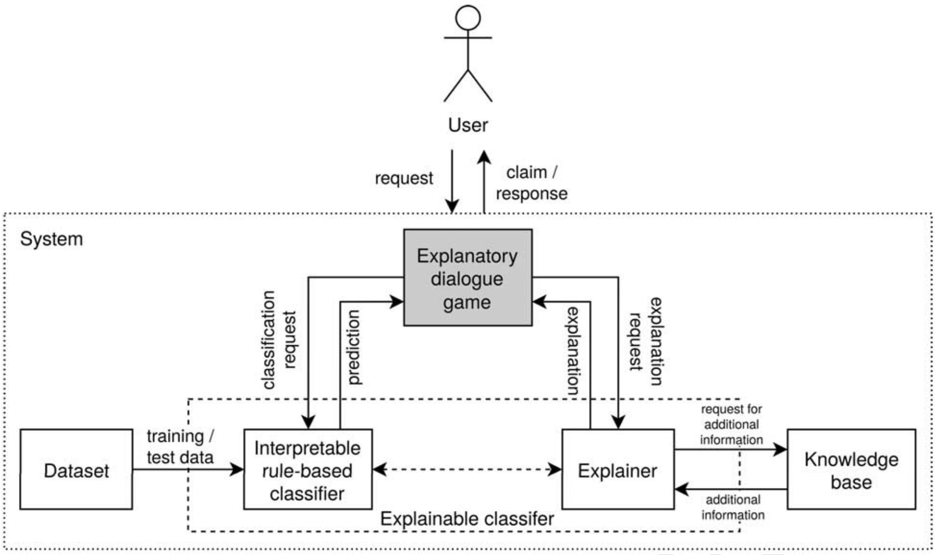


Fig. 1. A schema of the modelled system-user explanation communication process. This paper focuses on designing an explanatory dialogue game for communication of factual and counterfactual explanations for interpretable rule-based classifiers (the shaded block).

In this work, we assume knowledge of the feature space: the dataset is said to contain linearly scaled numerical features. In addition, all the numerical feature values are said to be mapped to the corresponding feature-dependent linguistic variable [86]. Therefore, each data instance $x_i \in X = \langle F_i, y_i \rangle$ is associated to class $y_i \in Y$ and defined over the set of p 3-tuple features $F_i = \{\langle f^k, v^k, t^k \rangle\}|_{k=1}^{p}$ where each feature f^k is assigned to the corresponding numerical value v^k and linguistic term t^k (e.g., ⟨age, 20, young⟩). The values of the linguistic variables (i.e., the so-called linguistic terms) may be defined by an expert. In this case, they are mapped to expert knowledge-based numerical intervals covering all the values of the corresponding feature. Otherwise, the linguistic variable is assigned to a set of textual values and mapped to equal-size numerical intervals. In this respect, the set of textual values that the linguistic variable can take on is of arbitrary cardinality.

The classifier predicts the class label $\hat{y}$ for the given test instance $x_{\text{test}} = \langle F_{\text{test}}, y_{\text{test}} \rangle$ on the basis of the learned mapping function c. The test instance classification is predicted correctly if the predicted class label and the actual test instance class label are the same (i.e., $\hat{y} = y_{\text{test}}$). Otherwise, the predicted class is deemed wrong (i.e., $\hat{y} \neq y_{\text{test}}$). Altogether, the interpretable rule-based classifier and the explainer are said to form an explainable classifier. Once the classifier outputs a prediction, the associated explainer attempts to generate an explanation in natural language for that prediction. Upon request, the explanation is passed to the User via the explanatory dialogue game, which serves as a communication channel between the explainable classifier and the User. During their intercourse, the User is assumed to be able to submit further explanation-related requests and receive responses processed by the dialogue game module whereas the dialogue game module can query the explainer for further explanation-related information.

2.2. *Explanation to the classification*

The upsurging need for explaining a classifier's output is raising interest in the mere nature of the explanation. For instance, social sciences testify that explanations are expected to be contrastive, selected, and social [45]. First, the property of *contrastiveness* implies establishing a relation not only between the cause and effect of the phenomenon under consideration but also another relation between the cause and a given non-observed effect (i.e., another alternative effect). Second, explanations are as well argued to be *selected*, i.e., only the most relevant causes should make part of a specific explanation. Third, explanations are claimed to be *social*, i.e., they are a product of interaction between the explainer and the explainee.

Contrastiveness plays an important role when explaining a solution to the classification problem, as different classes are opposed to the others on the basis of distinctive feature values. Further, contrastiveness is inherent to counterfactual (CF) explanations (or counterfactuals, for short). In the context of XAI, counterfactuals suggest minimal changes in feature-value pairs for a different outcome to be obtained [71]. CFs are said to be post-hoc (i.e., they are generated for pretrained classifiers) and local (i.e., they explain the classifier's output w.r.t. a specific test instance) [27]. CFs may be (1) model-agnostic if they operate only on the given input (i.e., a test instance) and output (i.e., a prediction) of the classifier or (2) model-specific if they utilise the internals of the classifier to explain the given output [47,71].

CF explanations are claimed to have a number of desired properties against which CF explanation methods can be evaluated [27]. For example, CFs should be *valid* (i.e., CFs should truly lead to the desired hypothetical outcome), *proximate* (i.e., CFs should suggest only minimal changes to the test instance w.r.t. the selected distance metric), *sparse* (i.e., CFs should minimise the number of features whose values are to be changed), *actionable* (i.e., CFs should suggest feasible changes), and *diverse* (i.e., CFs should offer multiple alternatives). An exhaustive list of such properties can be found in recent surveys on CF explanation generation and evaluation [27,49,78]).

A large number of explanation generation methods are evaluated using automatically computable metrics that assess the aforementioned properties of CF explanations [49]. However, such metrics oftentimes do not take into consideration user feedback at all. Whereas considering the social factor may not be necessary when, e.g., measuring validity, estimating CF diversity may have to directly involve capturing effects of the interaction between the system and the user. Indeed, CF explanations suggesting minimal changes in feature values may not always be equally appreciated by end users. Given a variety of potential CFs, different users may prefer distinct CFs for the same hypothetical output. Further, the social aspect of explanation becomes crucially important when two alternative automatically generated pieces of explanation are deemed equally explanatory (e.g., when the distances from the test instance to two or more closest CF data points are the same or when two CF sets have the same coverage). As the state-of-the-art AI technologies are shifting towards being user-centric [83], it appears indispensable to enhance existing explanation generation modules with a system-user communication interface that would allow end users to produce such inquiries for alternative CFs in the course of an explanatory dialogue, even if the user is not aware of the dataset-related peculiarities.

Various state-of-the-art CF explanation generation frameworks are known to offer diverse CFs ([15,17,35,49,60,62,75,85], among others). However, the format of such CFs raises several important concerns. First, most of such frameworks lack any interaction with end users leaving the users without further guidance when interpreting the generated explanations. Second, some explainers output a set of distinct CFs altogether [49,60]. In these settings, the Grice's maxim of quantity [25] may be violated, as only a subset of the offered explanations can be sufficient for the end user. Third, a large number of diverse CF explanation generation frameworks provide their output in tabular form [15,17,35,49,62,75]. Whereas natural language generation tools can be used to transform tabular data into text, a taxonomy of necessary explanation-related requests and responses remains missing. To address these issues, we propose a transparent explanatory dialogue model for diverse factual and counterfactual explanation communication that allows the end user to explore the explanation space iteratively until he or she can make an informed decision on whether the system's prediction can be trusted.

In light of the aforementioned considerations, a classifier's prediction can be explained factually and/or counterfactually. As we focus on the social factor of explanation generation in this paper, we assume that an explainer provides us with automatically generated textual factual and CF explanations operating in the settings described in Section 2.1. Below, we define both aforementioned types of explanation in terms of their linguistic realisation.

Driven by the assumptions above, both factual and CF explanations can be represented in two forms: using linguistic terms or numerical values (intervals). On the one hand, a purely textual explanation may be more intuitive and comprehensive to the explainee (e.g., "The test instance is of class Blanche because colour is pale and bitterness is low" or "The test instance would be of class Porter if colour were brown and strength were high"). On the other hand, explanations that incorporate numerical information may offer more detailed (and, perhaps, more precise) information while possibly requiring additional domain knowledge (e.g., "The test instance is of class Blanche because $0 \leqslant$ colour $\leqslant 3$ and $2 \leqslant$ bitterness $\leqslant 5$" or "The test instance would be of class Porter if colour ranged between 20 and 30 and strength ranged between 100 and 200"). In this work, we refer to explanations of both modalities as "high-level" and "low-level" explanations, respectively.

Definition 1. A *high-level explanation* $e^h(\hat{y}, [y'])$[1] is a set of feature-value pairs that explains the classifier's prediction $\hat{y}$ for the given test instance either factually or counterfactually in terms of the linguistic terms associated to the corresponding linguistic variable.

Definition 2. A *low-level explanation* $e^l(\hat{y}, [y'])$ is a set of feature-value pairs that explains the classifier's prediction $\hat{y}$ for the given test instance either factually or counterfactually in terms of the corresponding numerical values (intervals).

Paired explanations of both modalities may be found complementary to each other, as they may target different groups of end users. High-level explanations may facilitate understanding thereof by lay users. In turn, low-level explanations may be necessary for expert users to be able to further verify the validity of the offered explanation without linguistic ambiguity. Hereinafter, we assume that both factual and CF explanations to be paired two-level structures. To meet the requirement of being selective [45], all such explanations should be designed to reflect only the most characteristic features of the test instance that influence the classifier's prediction or its hypothetical counterpart. Let us now define factual and CF explanations in terms of their high- and low-level components.

Definition 3. A *factual explanation* $e_f(\hat{y}) = \langle e^h_f(\hat{y}), e^l_f(\hat{y})\rangle$ is a 2-tuple of affirmative sentences answering the question "Why is the test instance predicted to be of class $\hat{y}$?" where $e^h_f(\hat{y})$ and $e^l_f(\hat{y})$ are the corresponding high- and low-level explanations, respectively.

The given test instance's prediction can be explained in a (possibly, infinite) number of ways. At the same time, different explanations for the same phenomenon may have distinct degrees of explanatory power. Hence, all possible factual explanations are assumed to be ranked by an explainer in terms of their relevance to the test instance. Importantly, the notion of relevance in Definition 3 is determined by

[1] Hereinafter, $[y']$ is used as an optional parameter to refer to the requested CF class whenever a CF explanation is being processed. This parameter is omitted for the same request when a piece of factual explanation is being considered.

peculiarities of the explanation generation method, which falls outside the scope of this paper. The set of all factual explanations E_f for the predicted class $\hat{y}$ is defined as follows:

$$E_f(\hat{y}) = \bigcup_{i=1}^{\infty} e_{fi}(\hat{y}) \tag{1}$$

where e_{f1} is the most relevant factual explanation for the test instance's prediction, $e_{f|E_f|}$ is the least relevant one, i being the rank of the given piece of explanation. On the other hand, a CF explanation is assumed to suggest a minimal set of feature value changes that lead to a different desired classification. Then, a CF explanation is defined as follows.

Definition 4. A *counterfactual (or, shortly, CF) explanation* $e_{cf}(\hat{y}, y') = \langle e^h_{cf}(\hat{y}, y'), e^l_{cf}(\hat{y}, y')\rangle$ for the given CF class $y' \in Y \setminus \{\hat{y}\}$ is a 2-tuple of conditional sentences answering the question "Why is the test instance not predicted to be of class y' instead of $\hat{y}$?" where $e^h_{cf}(\hat{y}, y')$ and $e^l_{cf}(\hat{y}, y')$ are the corresponding high- and low-level explanations, respectively.

Similarly to factual explanations, all possible CFs are assumed to be ranked by their relevance to the test instance in accordance with a preselected criterion (for example, the distance metric from the test instance to the closest data point that the explanation includes). Then, the set of all the CF explanations for the given CF class is defined as follows:

$$E_{cf}\big(\hat{y}, y'\big) = \bigcup_{i=1}^{\infty} e_{cfi}\big(\hat{y}, y'\big) \tag{2}$$

where $e_{cf1}(\hat{y}, y')$ is the most relevant counterfactual explanation to the test instance's prediction $\hat{y}$ for the given CF class y', $e_{cf|E_{cf}|}$ is the least relevant one, i being the rank of an explanation.

Altogether, all ranked candidate factual and CF explanations for the given prediction are assumed to be unique and said to constitute an explanation space for the given prediction. The explanation space therefore contains all the pieces of factual and CF explanations that the system can offer to the end user w.r.t. the given test instance. Consequently, a given classifier's prediction cannot be explained by any piece of explanation that the explanation space does not contain.

Definition 5. An *explanation space* $E_{\text{space}}(\hat{y})$ is the union of all possible factual and CF explanations that an explainer can generate for the given prediction $\hat{y}$, s.t. $E_{\text{space}}(\hat{y}) = E_f(\hat{y}) \cup E_{cf}(\hat{y}, y'), \forall y' \in Y \setminus \{\hat{y}\}$.

2.3. *Explainer-preferred vs. explainee-preferred explanations*

Whereas any single piece of explanation may be satisfactory for the given user, it may have to be combined with other explanation instances for other users. For example, the end user may (1) request and be satisfied with the offered (factual and/or counterfactual) piece of explanation, (2) request and not be satisfied with the offered explanation, or (3) not request any explanation for, e.g. an alternative CF class, at all. In addition, not all the most relevant pieces of explanation from the system's point of view may seem as relevant to the user. To inspect the differences between such combinations of explanations, we therefore introduce the notions of explainer-preferred and explainee-preferred explanation. Explanation rankings provided by the explainer allow us to single out the most relevant pieces of CFs for each CF class from the system's point of view:

$$E_{cf1}(\hat{y}) = \cup e_{cf1}\big(\hat{y}, y'\big), \quad \forall y' \in Y \setminus \{\hat{y}\} \tag{3}$$

Then, an explainer-preferred explanation is said to comprise all the most relevant (both factual and counterfactual) pieces of explanation from the explainer's point of view.

Definition 6. An *explainer-preferred explanation* is the union of the most relevant automatically generated factual explanation for the predicted class and the most relevant explanations for each of the CF classes:

$$E_{\text{explainer}}(x_{\text{test}}, \hat{y}) = e_{f1}(\hat{y}) \cup E_{cf1}(\hat{y}) \tag{4}$$

An explainer-preferred explanation may be claimed to comprehensively explain the output of the given classifier to any end user. Given a set of multiple candidate factual and/or counterfactual explanations from the explanation space, the explanation generation module ranks them by relevance to the test instance (e.g., a distance metric) and subsequently presents the most relevant pieces of explanation to the end user. However, the explanation generation module output ignores end user preferences in these settings. Therefore, we define an explainee-preferred explanation as follows.

Definition 7. An *explainee-preferred explanation* is the union of all the pieces of explanation that the explainee finds the most satisfactory, as he or she explores the explanation space E_{space} when being explained the given prediction.

For illustrative purposes, consider the classification task for a dataset of four classes: $Y = \{y_1, y_2, y_3, y_4\}$. Let some test instance x_{test} be predicted to be of class y_1. An explainer-preferred explanation would therefore include the most relevant piece of factual explanation for class y_1 as well as the most relevant explanations for all the other (CF) classes:

$$E_{\text{explainer}}(x_{\text{test}}, y_1) = e_{f1}(y_1) \cup e_{cf1}(y_1, y_2) \cup e_{cf1}(y_1, y_3) \cup e_{cf1}(y_1, y_4) \tag{5}$$

The explainee may consider (a part of) the offered explanation irrelevant, redundant, or poorly explanatory. Figure 2 illustrates a possible discrepancy between the automatically generated and some user-preferred explanations. Whereas the factual explanation may be satisfactory for him or her, the explainee may find optimal the third most relevant CF explanation (from the explainer's point of view) for class y_2 (if it were offered), the second most relevant CF explanation for class y_3, and not require any CF explanation for class y_4. In this case, the reconstructed user-preferred explanation could be formally represented as follows:

$$E_{\text{explainee}}(x_{\text{test}}, y_1) = e_{f1}(y_1) \cup e_{cf3}(y_1, y_2) \cup e_{cf2}(y_1, y_3) \tag{6}$$

As shown in the example above, there may exist only a slight overlap between the most relevant explainer-preferred explanation and that expected by the explainee. It therefore appears indispensable to provide end users with a means of interaction with the explanation generation module to enable them to interactively explore the explanation space and, subsequently, shape the explanation in accordance

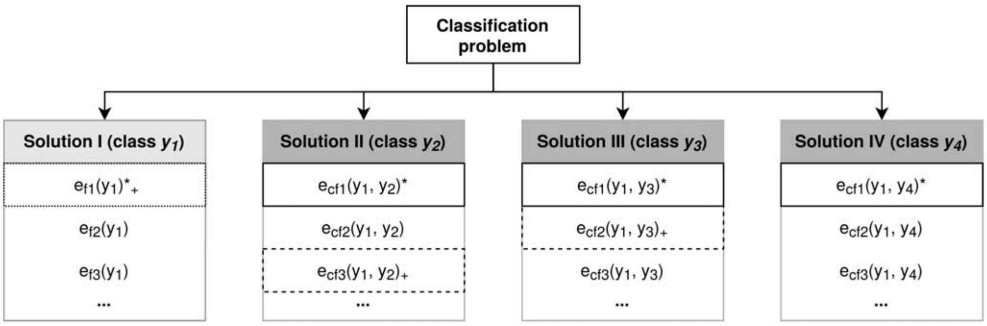


Fig. 2. A schema of a classification problem. Class y_1 is predicted by the classifier to be the solution to the problem. The other possible solutions (classes y_2, y_3, y_4) are considered hypothetical and form the set of CF classes. The corresponding explanations in solid rectangles (additionally marked with * as superscript) are those generated automatically. The explanations in dashed rectangles (additionally marked with + as subscript) are those preferred by the end user. Notably, the factual explanation in a double-dashed rectangle (that for class y_1) is both explainer-preferred and explainee-preferred.

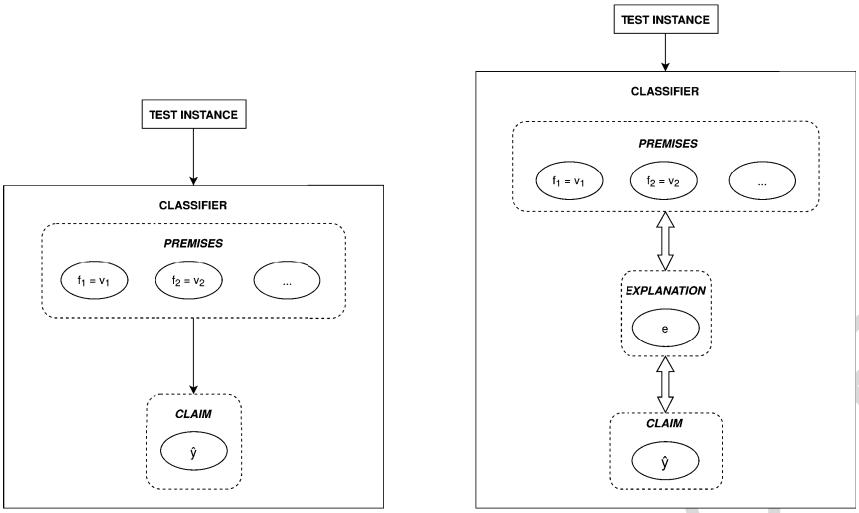
with their preferences. To do so, it is helpful to consider the classifier's reasoning from the argumentative point of view. Argumentation is regarded as an effective mechanism to communicate explanation in natural language [8]. Thus, various argumentation frameworks are shown to be particularly useful in the field of XAI for their ability to generate explanations of different modalities (e.g., textual, graphical, hybrid) [16]. Further, recent work on argumentation-based explanation generation shows that such frameworks provide efficient explanatory interfaces between AI-based systems and users of such systems, particularly, in the form of dialogue [77]. In addition, argumentation is shown to logically connect with, for example, abductive reasoning tools that are widely used for counterfactual reasoning [11].

In these settings, a prediction may be treated as a claim proposed by the classifier. Such a claim is then supported by the decisive feature value pairs (either specific values or intervals of values) that led the classifier to make the corresponding prediction (see Fig. 3(a)). However, ground-truth data-based premises cannot be attacked directly, as they can by no means be claimed invalid. Therefore, it appears necessary to introduce an intermediate explanation layer that approximates the premises and serves as an attackable natural language interface between the premise and the claim (see Fig. 3(b)).

Throughout this paper, we claim that rule-based explanations from interpretable classifiers serve this purpose well. First, they reflect the features retrieved from the data that the classifier is trained on. Second, their natural language representation allows the end user to construct a comprehensive mental representation of the underlying data. Following Hempel's definition of explanation [31], explanations themselves can be regarded as arguments. In the context of explanatory dialogue between the system and the user, explanations can then be attacked in the dialogic intercourse between the dialogue parties. In this manner, the end user is given the opportunity to interactively inspect explanations from the explanation space that do not make part of the explainer-preferred explanation by arguing over the initially (and, if necessary, also subsequently) offered pieces thereof.

3. Dialogue game for XAI

In this section, we formally define a dialogue game that serves to communicate explanation(-s) generated automatically by an explanation generation module (paired with the corresponding interpretable rule-based classifier) to its end user. Thus, Section 3.1 proposes formal components of explanatory di-



(a) A premise-claim representation of a classifier's prediction.

(b) A premise-explanation-claim representation of a classifier's prediction.

Fig. 3. Schematic representations of classifier's reasoning from the argumentative point of view.

alogue. Subsequently, Section 3.2 presents an example of an explanatory dialogue modelled in accordance with the principles outlined in Section 3.1. Finally, Section 3.3 generalises the proposed approach to explanatory information-seeking dialogue modelling in form of an explanatory context-free dialogue grammar.

3.1. *Formal description of explanatory information-seeking dialogue*

In order to construct a communication channel between the system and the end user, we propose that explanatory dialogue be modelled on the basis of the so-called "dialogue game" approach to argumentation [54]. Taking into consideration the aforementioned requirements to explanation, we formally define an explanatory dialogue between the explanation generation module and end user as a 10-tuple $D = \langle P, M, R, Pr\ , K, E, DET, CLAR, CFS, KB\ \rangle$ where

- P is the set of dialogue participants;
- M is the set of dialogue moves that the dialogue participants make in the course of a dialogue;
- R is the set of requests and responses that specify allowed utterances in the course of explanatory dialogue;
- Pr is the dialogue protocol governing the flow of the conversation in accordance with the set of predetermined locution rules specifying types of legitimate utterances;
- K is the knowledge store, i.e. the dynamically populated set of all the pieces of explanation that the user requests and receives during his or her interaction with the system;
- E is the explanation store, i.e. the dynamically updated set of the last offered pieces of explanation for each class under consideration;
- DET is the detailisation store, i.e. the set of features of the actually processed piece of high-level explanation whose values (i.e., linguistic terms) can be inspected for further details;
- $CLAR$ is the clarification store, i.e. the set of features of the actually processed piece of explanation whose definitions can be requested;

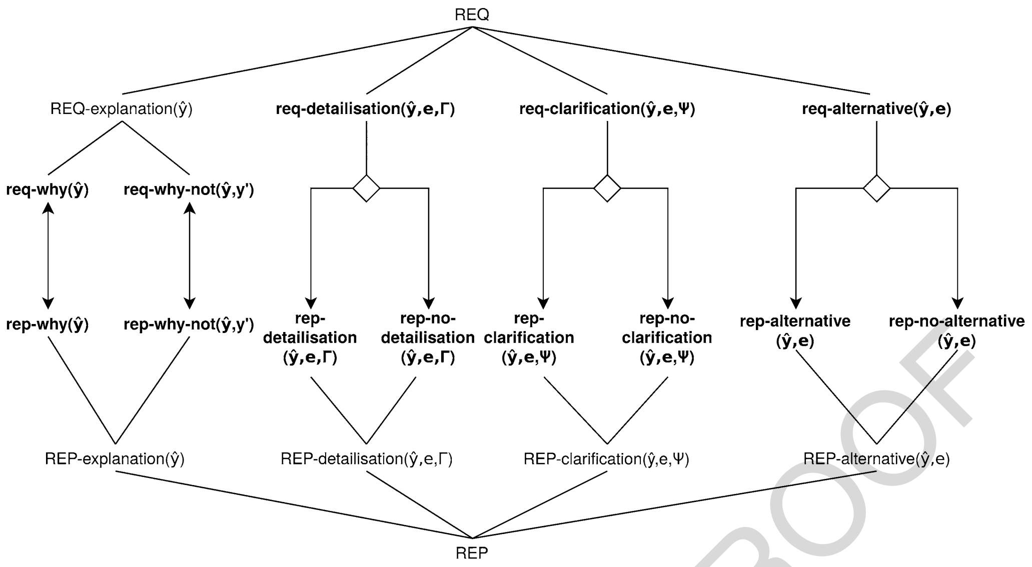


Fig. 4. A typology of requests and replies. Individual requests/responses are in bold. In addition, sets of request/responses are named with uppercase letters (i.e., REQ-/REP-).

- *CFS* is the CF class store, i.e. the set of CF classes whose explanations can be potentially offered to the end user;
- *KB* is a knowledge base containing the domain knowledge for the addressed problem.

Let us now define each component of the proposed explanatory dialogue model in detail.

1) Participants. An explanatory dialogue serves as an interface between two parties: the explainable classifier (or, in general, the system S) and the human agent interacting with the system (the user U). Therefore, the set of participants is defined to always consist of two items $P = \{S, U\}$ where the system S always plays the role of the explainer whereas the user U is always the explainee.

2) Moves. A single instance of a dialogue can be regarded as a sequence of finite legitimate moves $M = \langle m_0, m_1, \dots, m_n \rangle$, each of which is generated in accordance with the locution rules as well as those making part of the corresponding dialogue protocol.

3) Responses and requests. Our explanatory dialogue model presupposes that the explainer (i.e., the system) has the ability to present all the information available to it to the explainee (i.e., the user). The user is, in turn, capable of inquiring all such information. It is therefore crucially important to find a balance between the information that the user may require from the system and the information the system can provide the user with.

Driven by the assumption that high- and low-level explanations may accommodate both expert and lay users and inspired by previous work on formal explanatory dialogue modelling [9], we distinguish four types of user requests and responses that form the corresponding set $R = \{REQ, REP\}$. Namely, those are the requests for (either factual or CF) explanation, detailisation, clarification, and alternative explanation of either of the considered kinds. Figure 4 summarises all possible types of user's requests and the corresponding system's responses. All locutions generated by both parties fall into either of the two symmetric classes.

On the one hand, the set of requests from the user to the system *REQ*=\{*REQ-explanation(*$\hat{y}$*)*, *req-detailisation(*$\hat{y}$, e, Γ*)*, *req-clarification(*$\hat{y}$, e, Ψ*)*, *req-alternative(*$\hat{y}$, e*)*\} consists of the following items:[2]

- *REQ-explanation(*$\hat{y}$*)*: the set of user requests for explanation for system's prediction $\hat{y}$;
- *req-detailisation(*$\hat{y}$, e, Γ*)*: the user request for further details on feature Γ (i.e., the corresponding numerical intervals) that makes part of a high-level (either factual or CF) explanation e for prediction $\hat{y}$;
- *req-clarification(*$\hat{y}$, e, Ψ*)*: the user request for clarification of the meaning of a specific feature Ψ that makes part of (either factual of CF and either high-level or low-level) explanation e for prediction $\hat{y}$;
- *req-alternative(*$\hat{y}$, e*)*: the user request for an alternative (either factual or CF and either high-level or low-level) explanation provided that the user is not satisfied with the previously offered explanation e for system's prediction $\hat{y}$.

Further, the set of user explanation requests *REQ-explanation(*$\hat{y}$*)* consists of the following possible locutions:

- *req-why(*$\hat{y}$*)*: the user request for a factual explanation for the system's prediction $\hat{y}$;
- *req-why-not(*$\hat{y}$, y'*)*: the user request for a CF explanation concerning the CF class $y' \in Y \setminus \{\hat{y}\}$ for prediction $\hat{y}$ (i.e., to specify why some CF class y' was not predicted instead of $\hat{y}$).

On the other hand, the set of responses (replies) that the system sends back to the user *REP*=\{*REP-explanation(*$\hat{y}$*)*, *REP-detailisation(*$\hat{y}$, e, Γ*)*, *REP-clarification(*$\hat{y}$, e, Ψ*)*, *REP-alternative(*$\hat{y}$, e*)*\} mirrors the set of user requests:

- *REP-explanation(*$\hat{y}$*)*: the set of system responses in an attempt to explain prediction $\hat{y}$;
- *REP-detailisation(*$\hat{y}$, e, Γ*)*: the set of system responses in an attempt to provide details (i.e., numerical intervals) with respect to feature Γ of explanation e for system's prediction $\hat{y}$;
- *REP-clarification(*$\hat{y}$, e, Ψ*)*: the set of system responses in an attempt to clarify feature Ψ making part of (either factual or CF) explanation e for prediction $\hat{y}$;
- *REP-alternative(*$\hat{y}$, e*)*: the set of system responses in an attempt to provide the user with an explanation alternative to the previously offered (either factual or CF and either high-level or low-level) explanation e for prediction $\hat{y}$.

In addition, the set of replies to requests for (initial, non-alternative) explanation *REP-explanation(*$\hat{y}$*)* consists of the following items:

- *rep-why(*$\hat{y}$*)*: the system attempts to factually explain the prediction $\hat{y}$ on the basis of the known features that led to that decision and offers a factual explanation if it is able to, or refuses to offer it, otherwise;
- *rep-why-not(*$\hat{y}$, y'*)*: the system attempts to provide the user with a CF explanation for prediction $\hat{y}$ for the given CF class y' or refuses to offer it, otherwise.

The set of replies to detailisation requests *REP-detailisation(*$\hat{y}$, e, Γ*)* consists of the following items:

- *rep-detailisation(*$\hat{y}$, e, Γ*)*: the system provides the numerical intervals over which the corresponding linguistic term of the requested explanation feature Γ making part of explanation e is defined;

[2] Sets of requests are denoted using uppercase letters (as in, e.g., *REQ-explanation*) whereas single instances of requests are denoted using only lowercase letters (as in, e.g., *req-detailisation*).

- *rep-no-detailisation(*$\hat{y}$, e, Γ*)*: the system refuses to provide numerical intervals on the requested feature's linguistic term in explanation e, e.g. due to their unavailability.

The set of replies to clarification requests *REP-clarification(*$\hat{y}$, e, Ψ*)* consists of the following items:

- *rep-clarification(*$\hat{y}$, e, Ψ*)*: the system provides the user with a definition of the requested feature Ψ making part of explanation e for prediction $\hat{y}$ retrieving it from the knowledge base;
- *rep-no-clarification(*$\hat{y}$, e, Ψ*)*: the system refuses to clarify the requested feature Ψ making part of explanation e for prediction $\hat{y}$ due to, e.g., its absence in the knowledge base.

The set of replies to alternative explanation requests *REP-alternative(*$\hat{y}$, e*)* consists of the following items:

- *rep-alternative(*$\hat{y}$, e*)*: the system recognises the fact that the user is not satisfied with the offered (factual or CF) explanation e for prediction $\hat{y}$, seeks the most relevant alternative to it, generates and offers an alternative explanation to the user;
- *rep-no-alternative(*$\hat{y}$, e*)*: the system recognises the fact that the user is not satisfied with the offered (factual or CF) explanation e for prediction $\hat{y}$, seeks the most relevant alternative to it, but is unable to generate it.

4) Dialogue protocol. An explanatory dialogue between the system and the user is modelled following the rules specified in the dialogue protocol. The protocol determines turntaking rules, the rules governing user's and system's allowed moves at each stage of the explanatory dialogue, and the termination states of the dialogue. Thus, the locution types above are directly mapped to the speech acts produced by the system and the user as specified in the dialogue protocol. All of the aforementioned protocol rules are specified in Appendix B.

5) Knowledge store. Let K be the knowledge store which accumulates user's knowledge w.r.t. explanations requested during his or her interaction with the system. Knowledge store K is initialised to be an empty set: $K = \emptyset$. When the system generates a factual or CF explanation (locutions *explain-f*($\hat{y}$, E, e_f) and *explain-cf*($\hat{y}$, E, y', e_{cf}), as specified in the dialogue protocol), the corresponding piece of explanation is added to the knowledge store: $K = K \cup e_f(\hat{y})$ or $K = K \cup e_{cf}(\hat{y}, y')$, respectively. The same applies to alternative explanations of either kind (locutions *alter-f*($\hat{y}$, E, e_f, e'_f) and *alter-cf*($\hat{y}$, E, y', e_{cf}, e'_{cf})).

6) Explanation store. Let E be the explanation store which tracks the current state of the explainee-preferred explanation throughout the dialogue. Explanation store E is initialised to be an empty set: $E = \emptyset$. Similarly to the knowledge store, a factual or CF explanation is added to the explanation store once generated: $E = E \cup e_f(\hat{y})$ or $E = E \cup e_{cf}(\hat{y}, y')$, respectively. If the user finds the offered factual or CF explanation not satisfactory enough and asks for an alternative explanation (locutions *why-alternative*($\hat{y}$, E, e_f) and *why-not-alternative*($\hat{y}$, E, y', e_{cf}), respectively), the corresponding explanation is removed from the explanation store: $E = E \setminus e_f(\hat{y})$ or $E = E \setminus e_{cf}(\hat{y}, y')$, respectively. Noteworthy, the user cannot request an alternative explanation to any explanation non-offered previously. Further, the user can only submit explanation-related requests (detailisation, clarification, alternative) for the piece of explanation being processed. The resulting explainee-preferred explanation is the union of all the pieces of explanation found in the explanation store when a terminal dialogue state is reached.

7) Detailisation store. Let *DET* be the store that contains the features of the currently processed high-level explanation for which further details can be requested. *DET* is initialised to be empty, as the explanatory dialogue starts: $DET = \emptyset$. The user can submit a detailisation request to the system only if a high-level (either factual or CF) explanation $e = e^h_f | e^h_{cf}$ is being processed. Recall that for each feature Γ of the currently processed high-level explanation e, the feature is defined in terms of a linguistic variable mapped to the corresponding linguistic terms. When a new piece of high-level explanation is offered to the end user, *DET* is reinitialised with the set of features that the currently processed explanation contains: $DET = \{\Gamma\}, \forall\Gamma \in e$. The user can ask the system to provide him or her with the numerical intervals for the linguistic term of the given explanation feature only once during a sub-dialogue concerning a specific piece of explanation. Thus, the corresponding feature is eliminated from the detailisation store once the system has generated a response θ (locution *elaborate*$(\hat{y}, E[,y'], e, \Gamma, \theta)$): $DET = DET \setminus\{\Gamma\}$. If $DET = \emptyset$, it is prohibited for the user to submit a detailisation request (locution *what-details*$(\hat{y}, E[,y'], e, \Gamma)$). When the user makes the final decision w.r.t. the system's claim (i.e., either accepts or rejects it), the detailisation store is nullified: $DET = \emptyset$.

8) Clarification store. Let *CLAR* be the clarification store that contains the explanation features whose meaning can be clarified. Similarly to the detailisation store, *CLAR* is initialised to be empty: $CLAR = \emptyset$. When a new piece of explanation is offered, *CLAR* is populated with all the features that the explanation being processed $e = e^h_f | e^h_{cf} | e^l_f | e^l_{cf}$ contains: $CLAR = \{\Psi\}, \forall\Psi \in e$. Noteworthy, the definitions for all the features that the dataset contains are precollected, mapped to one another by an expert or retrieved from a dictionary, and stored in the knowledge base. The user can ask to clarify a specific feature from the clarification store only once during a sub-dialogue concerning a specific piece of explanation. Then, the corresponding feature is eliminated from the clarification store after the system's response υ (locution *clarify*$(\hat{y}, E[,y'], e, \Psi, \upsilon)$): $CLAR = CLAR \setminus\{\Psi\}$. If $CLAR = \emptyset$, it is prohibited for the end user to submit a clarification request (locution *what-is*$(\hat{y}, E[,y'], e, \Psi)$). When the user makes the final decision w.r.t. the system's claim (i.e., either accepts or rejects it), the clarification store is nullified: $CLAR = \emptyset$.

9) CF class store. Let *CFS* be the CF class store that contains all CF classes. It is initialised upon the successful execution of the factual explanation request (locution *explain-f*$(\hat{y}, E, e_f)$) so that $CFS = Y \setminus \{\hat{y}\}$ for some prediction $\hat{y} \in Y$. The user is allowed to request a CF explanation for each class from *CFS* only once (locution *why-not-explain*$(\hat{y}, E, y')$). In addition, the user is allowed to ask for a (series of) alternative CF explanation(-s) for the same CF class (locution *why-not-alternative*$(\hat{y}, E, y', e_{cf})$ as many times as there are alternative CFs for that class. Once a CF explanation is requested for some CF class y', it is eliminated from the *CFS* store: $CFS = CFS \setminus\{y'\}$. When the user makes the final decision w.r.t. the system's claim (i.e., either accepts or rejects it), the CF class store is nullified: $CFS = \emptyset$.

10) Knowledge Base. The knowledge base contains the dataset-related domain knowledge including a specification of all the dataset features (e.g., linguistic terms, the corresponding intervals, and definitions of all the features that the dataset contains).

3.2. Illustrative example

Having introduced the proposed formalism for explanatory information-seeking dialogue modelling, let us now illustrate it taking the previously considered example for reference (see Table 1 for details). Thus, we are considering the beer style classification problem for the beer dataset that contains the following classes: $Y_{\text{beer}} = \{$*Blanche, Lager, Pilsner, IPA, Barleywine, Stout, Porter, Belgian strong ale*$\}$. Table 2 outlines the states of the detailisation, clarification, and CF class stores of the example explanatory dialogue after each dialogue move. Table 3 outlines the states of the knowledge and explanation stores for the same example dialogue.

Initially, the system claims that some instance of beer is of class *Blanche* (move m_1). All the stores that make part of the dialogue model (*K, E, DET, CLAR, CFS*) are initialised to be empty. At the next

Table 2

A move-by-move formal description of the stores governing the example of explanatory dialogue from Table 1

Move	*Locution*	*DET*	*CLAR*	*CFS*
m_1	*claim* ($\hat{y}$, E)	∅	∅	∅
m_2	*why-explain* ($\hat{y}$, E)	∅	∅	∅
m_3	*explain-f* ($\hat{y}$, E, e_f)	{*colour, bitterness*}	{*colour, bitterness*}	{*Lager, Pilsner, IPA, Barleywine, Stout, Porter, Belgian strong ale*}
m_4	*what-is*($\hat{y}$, E, e_f, Ψ)	{*colour, bitterness*}	{*colour, bitterness*}	{*Lager, Pilsner, IPA, Barleywine, Stout, Porter, Belgian strong ale*}
m_5	*clarify* ($\hat{y}$, E, e_f, Ψ, υ)	{*colour, bitterness*}	{*colour*}	{*Lager, Pilsner, IPA, Barleywine, Stout, Porter, Belgian strong ale*}
m_6	*why-not-explain* ($\hat{y}$, E, y')	{*colour, bitterness*}	{*colour*}	{*Lager, Pilsner, IPA, Barleywine, Stout, Porter, Belgian strong ale*}
m_7	*explain-cf* ($\hat{y}$, E, y', e_{cf})	{*colour, bitterness*}	{*colour, bitterness*}	{*Lager, Pilsner, IPA, Barleywine, Porter, Belgian strong ale*}
m_8	*what-details* ($\hat{y}$, E, e_{cf}, Γ)	{*colour, bitterness*}	{*colour, bitterness*}	{*Lager, Pilsner, IPA, Barleywine, Porter, Belgian strong ale*}
m_9	*elaborate* ($\hat{y}$, E, e_{cf}, Γ, θ)	{*colour*}	{*colour, bitterness*}	{*Lager, Pilsner, IPA, Barleywine, Porter, Belgian strong ale*}
m_{10}	*why-not-explain*($\hat{y}$, E, y'')	{*colour*}	{*colour, bitterness*}	{*Lager, Pilsner, IPA, Barleywine, Porter, Belgian strong ale*}
m_{11}	*explain-cf*($\hat{y}$, E, y'', e_{cf})	{*colour*}	{*colour*}	{*Lager, Pilsner, IPA, Barleywine, Belgian strong ale*}
m_{12}	*why-not-alternative*($\hat{y}$, E, y'', e_{cf})	{*colour*}	{*colour*}	{*Lager, Pilsner, IPA, Barleywine, Belgian strong ale*}
m_{13}	*alter-cf*($\hat{y}$, E, y'', e_{cf}, e'_{cf})	{*colour, strength*}	{*colour, strength*}	{*Lager, Pilsner, IPA, Barleywine, Belgian strong ale*}
m_{14}	*accept-u* ($\hat{y}$, E)	∅	∅	∅
m_{15}	*accept-s* ($\hat{y}$, E)	∅	∅	∅

step, the user requests a factual explanation for the given prediction (m_2). The system provides the user with a factual explanation (m_3). As the factual explanation is generated, both *DET* and *CLAR* stores are populated with the corresponding features (colour and bitterness). Further, the piece of factual explanation $e_f(\hat{y} = \text{Blanche})$ is placed to both the knowledge store and the explanation store. In addition, the CF store *CFS* is populated with all the CF classes. At the next stage, the user asks the system to clarify the notion of bitterness (m_4) and receives the corresponding definition from the system (m_5). As the clarification request for a given feature can only be submitted once while processing a specific piece of explanation, *bitterness* is then eliminated from the *CLAR* store.

Once the factual explanation is offered, the user may commit to the factual explanation offered and inquire a CF explanation for some CF class. In the present example, the user seeks, at this stage, to know why the classifier did not predict the given beer to be *Stout* (m_6). Then, the classifier presents the most relevant piece of CF explanation for this CF class in accordance with its ranking (m_7). The CF explanation $e_{cf}(y' = \text{Stout})$ is then added to both the knowledge and explanation stores, whereas the class *Stout* is removed from the *CFS* store. Then, the *DET* and *CLAR* stores are updated with the features that the newly offered CF explanation contains. As the user requires more detailed information on *bitterness* (m_8), the system retrieves the requested numerical interval over which the value of *bitterness* is defined to be high (m_9). The feature *bitterness* is then removed from the *DET* store. Then, the user proceeds to request a CF explanation for class *Porter* (m_{10}). Similarly to the previously offered explanations, *DET* and *CLAR* are updated accordingly, as the most relevant piece (from explainer's point of view) of CF

Table 3

An example explanatory dialogue schema

Block	*Move*	*Utterance*	*K*	*E*
C	m_1	**System:** The test instance is of class y.	$\emptyset$	$\emptyset$
E	m_2	**User:** Could you explain why you think so?	$\emptyset$	$\emptyset$
	m_3	**System:** It is of class y because $\langle$feature$_1\rangle$ is $\langle$term$_1\rangle$.	$\{e_f(\hat{y})\}$	$\{e_f(\hat{y})\}$
	m_4	**User:** What do you mean by $\langle$feature$_1\rangle$?	$\{e_f(\hat{y})\}$	$\{e_f(\hat{y})\}$
	m_5	**System:** $\langle$feature$_1\rangle$ is $\langle$definition for feature$_1\rangle$.	$\{e_f(\hat{y})\}$	$\{e_f(\hat{y})\}$
	m_6	**User:** But why is it not of class y'?	$\{e_f(\hat{y})\}$	$\{e_f(\hat{y})\}$
	m_7	**System:** It would be of class y' if $\langle$feature$_1\rangle$ were $\langle$term$_2\rangle$ and $\langle$feature$_2\rangle$ were $\langle$term$_3\rangle$.	$\{e_f(\hat{y}), e_{cf}(y')\}$	$\{e_f(\hat{y}), e_{cf}(y')\}$
	m_8	**User:** Could you specify how $\langle$feature$_1\rangle$ is defined?	$\{e_f(\hat{y}), e_{cf}(y')\}$	$\{e_f(\hat{y}), e_{cf}(y')\}$
	m_9	**System:** $\langle$feature$_1\rangle$ is defined to be $\langle$term$_2\rangle$ because it is found in the interval $\langle$[term$_{2\text{min}}$, term$_{2\text{max}}$]$\rangle$.	$\{e_f(\hat{y}), e_{cf}(y')\}$	$\{e_f(\hat{y}), e_{cf}(y')\}$
	m_{10}	**User:** But why is the test instance not of class y''?	$\{e_f(\hat{y}), e_{cf}(y')\}$	$\{e_f(\hat{y}), e_{cf}(y')\}$
	m_{11}	**System:** It would be of class y'' if $\langle$feature$_1\rangle$ were $\langle$term$_3\rangle$ and $\langle$feature$_3\rangle$ were $\langle$term$_3\rangle$.	$\{e_f(\hat{y}), e_{cf}(y'), e_{cf}(y'')\}$	$\{e_f(\hat{y}), e_{cf}(y'), e_{cf}(y'')\}$
	m_{12}	**User:** I am not quite satisfied with your explanation. Could you offer me another one?	$\{e_f(\hat{y}), e_{cf}(y'), e_{cf}(y'')\}$	$\{e_f(\hat{y}), e_{cf}(y')\}$
	m_{13}	**System:** Sure! It would be of class y'' if...	$\{e_f(\hat{y}), e_{cf}(y'), e_{cf}(y''), e'_{cf}(y'')\}$	$\{e_f(\hat{y}), e_{cf}(y'), e'_{cf}(y'')\}$
T	m_{14}	**User:** Okay, I trust your prediction.	$\{e_f(\hat{y}), e_{cf}(y'), e_{cf}(y''), e'_{cf}(y'')\}$	$\{e_f(\hat{y}), e_{cf}(y'), e'_{cf}(y'')\}$
	m_{15}	**System:** Thank you for your trust in me. Bye!	$\{e_f(\hat{y}), e_{cf}(y'), e_{cf}(y''), e'_{cf}(y'')\}$	$\{e_f(\hat{y}), e_{cf}(y'), e'_{cf}(y'')\}$

In the left-hand side column ("Block"), C stands for claim, E – for explanation, T – for termination).

explanation is generated and offered for the class *Porter* (m_{11}). Then, the class *Porter* is excluded from the *CFS* store whereas the newly offered CF explanation is added to the knowledge and explanation stores. However, as the user is left dissatisfied or not convinced enough with the offered explanation, he or she inquires an alternative explanation to the previously offered one (m_{12}). Then, the latest offered explanation is removed from the explanation store. Subsequently, if the next best ranked alternative can be offered, it is added to the explanation store (m_{13}). The *DET* and *CLAR* stores are then updated accordingly. Having processed the presented explanations in their entirety, the user makes an informed decision that the classifier's prediction can be accepted (m_{14}). The system terminates the dialogue outputting a farewell locution (m_{15}).

Table 3 generalises the presented example of explanatory dialogue for any dataset where features, linguistic terms, and classes serve as dataset-specific variables. It is possible to generalise any explanatory dialogue modelled in accordance with the proposed framework using the suggested template utterances. Noteworthy, three main building blocks of such explanatory dialogue (C – claim, E – explanation, and T – termination) can be distinguished. Figure 5 presents the corresponding (partial, for illustrative purposes) parse tree of such a generalised explanatory dialogue.

3.3. Explanatory dialogue grammar (EDG)

As follows from the example of dialogue presented in Section 3.2, the proposed dialogue model has a hierarchical structure with respect to its main building blocks. This observation allows us to reflect the modular composition of explanatory dialogue (following our model) in a context-free dialogue grammar. As the transitions between the states of the dialogue are finite and predefined, the use of the correspond-

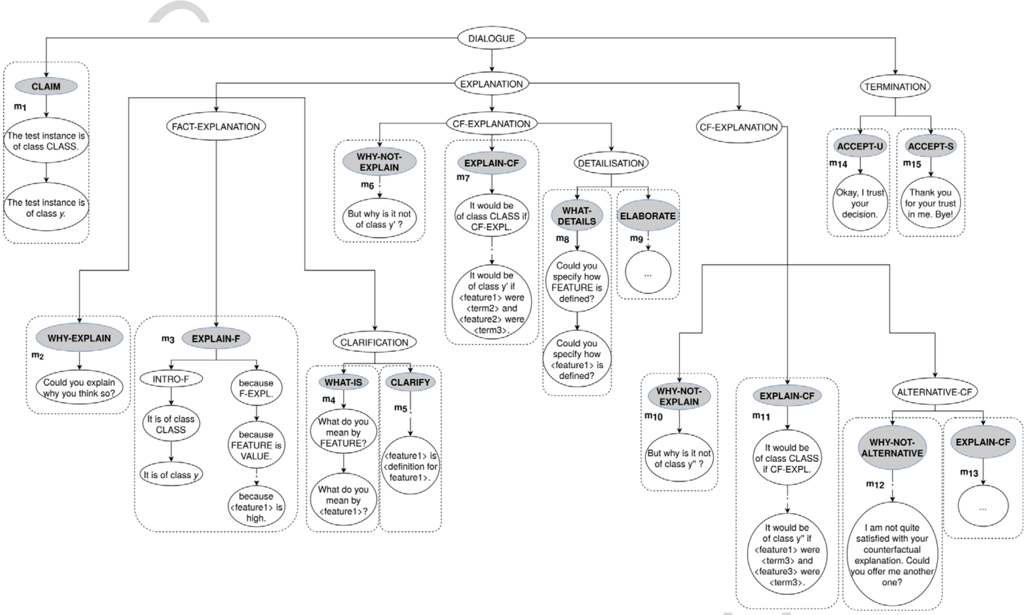

DIALOGUE
EXPLANATION
TERMINATION
CLAIM
m1
The test instance is of class CLASS.
The test instance is of class y.
FACT-EXPLANATION
CF-EXPLANATION
WHY-NOT-EXPLAIN
m6
But why is it not of class y' ?
EXPLAIN-CF
m7
It would be of class CLASS if CF-EXPL.
It would be of class y' if <feature1> were <term2> and <feature2> were <term3>.
DETAILISATION
WHAT-DETAILS
m8
Could you specify how FEATURE is defined?
Could you specify how <feature1> is defined?
ELABORATE
m9
...
CF-EXPLANATION
ACCEPT-U
m14
Okay, I trust your decision.
ACCEPT-S
m15
Thank you for your trust in me. Bye!
WHY-EXPLAIN
m2
Could you explain why you think so?
m3
EXPLAIN-F
INTRO-F
It is of class CLASS
It is of class y
because F-EXPL.
because FEATURE is VALUE.
because <feature1> is high.
CLARIFICATION
WHAT-IS
m4
What do you mean by FEATURE?
What do you mean by <feature1>?
CLARIFY
m5
<feature1> is <definition for feature1>.
WHY-NOT-EXPLAIN
m10
But why is it not of class y'' ?
EXPLAIN-CF
m11
It would be of class CLASS if CF-EXPL.
It would be of class y'' if <feature1> were <term3> and <feature3> were <term3>.
ALTERNATIVE-CF
WHY-NOT-ALTERNATIVE
m12
I am not quite satisfied with your counterfactual explanation. Could you offer me another one?
EXPLAIN-CF
m13
...

Fig. 5. A parse tree of the example of explanatory dialogue. Shaded nodes are non-terminals corresponding to specific speech acts. The subtrees in the dashed regions represent dialogue moves.

ing EDG allows us to (1) generate any explanatory dialogue that is valid in accordance with the dialogue protocol restrictions and (2) parse any actually valid explanatory dialogue or make a conclusion that the present explanatory dialogue is invalid with respect to the dialogue model constraints. Further, a grammar-based dialogue model can take into account modifications in the dialogue protocol if those are deemed necessary.

In light of the above, we define an EDG following Chomsky's definition of a context-free grammar as a tuple $G = \langle T, N, P, S \rangle$ where T is the set of terminals, N is the set of non-terminals, P is the set of production rules (productions), and S is the start token. In our model, T corresponds to a sentence actually uttered by each participant in the course of a dialogue. N encompasses the internal building blocks of the dialogue as well as the speech acts involved (see the shaded nodes in Fig. 5 for details). Thus, any explanatory dialogue is said to have three main building blocks (those corresponding to the non-terminals *CLAIM, EXPLANATION, TERMINATION*). In accordance with current legal requirements to explanation for AI, the block *EXPLANATION* enables the user to exercise the right to explanation and is made optional. All the non-terminals produced from the non-terminal *EXPLANATION* are designed in accordance with the predefined requests and responses (see Section 3.1 for details). In addition, P is composed in accordance with the dialogue protocol settings (see Appendix B for details). Note that productions can be subdivided in two groups, i.e., dataset-independent and dataset-specific productions. Dataset-independent production rules form the core of the proposed explanatory dialogue model and can be used in any application domain so long as it meets the settings of the classification problem as described in Section 2.1. The dataset-independent rules valid for the illustrative example of an explanatory dialogue are outlined in Appendix C. In turn, dataset-specific rules follow the structure of the given dataset and they are restricted by the information provided by the given interpretable rule-based classifier and the corresponding knowledge base. Finally, the start token S is known to always be the non-terminal *DIALOGUE* node, i.e., the root node in the tree depicted in Fig. 5.

4. Process mining for dialogue analytics

The proposed model of explanatory dialogue is designed in a top-down manner, which signals certain shortcomings. Thus, the dialogue protocol bases on the assumption that the taxonomy of requests and responses proposed in Section 3 inspired by findings from the literature exhaustively covers user's needs and system's abilities when engaged in an explanatory dialogue. However, in the absence of any empirical evaluation, such assumptions may result being purely speculative. For example, specific requests may be utilised to a very limited extent or even not utilised at all. Alternatively, there may exist requests that are not included in the original model, which may nevertheless be considered essential for human-machine interaction by the explainees. Either way, modifications to the model should be grounded on the data obtained from the end users. As such data-driven conclusions on the utility of the top-down dialogue model can only be made upon empirical evaluation, a user study is necessary to validate the proposed model.

In addition to analysis of free-form user feedback, evaluation of a dialogue model can be automated by inspecting dialogue patterns in the collected dialogue transcripts. In these settings, dialogues can be treated as iterative processes whose key patterns allow us to discern strengths and weaknesses of the dialogue model. To analyse dialogues as processes, we propose a use of process mining techniques.

Process mining is the subfield of data science that aims to provide tools for discovering insights into operational processes and thus supports process improvements [76]. Following the process mining terminology [50], an instance of a process (i.e., a specific explanatory dialogue) is denoted as a *trace* τ.

Table 4

An example of an event log (the activities in bold are those produced by the system; the user-produced activities are those in italics)

Case	*Activity*	*Start*	*End*
Dialogue$_1$	**claim**	2022-06-09 11:54:12	2022-06-09 11:54:12
Dialogue$_1$	*why-explain*	2022-06-09 11:54:12	2022-06-09 11:54:21
Dialogue$_1$	**explain-f**	2022-06-09 11:54:21	2022-06-09 11:54:22
Dialogue$_1$	*what-details*	2022-06-09 11:54:22	2022-06-09 11:54:42
Dialogue$_1$	**elaborate**	2022-06-09 11:54:42	2022-06-09 11:54:42
Dialogue$_1$	*why-not-explain*	2022-06-09 11:54:42	2022-06-09 11:55:58
Dialogue$_1$	**explain-cf**	2022-06-09 11:55:58	2022-06-09 11:56:00
Dialogue$_1$	*what-details*	2022-06-09 11:56:00	2022-06-09 11:56:32
Dialogue$_1$	**elaborate**	2022-06-09 11:56:32	2022-06-09 11:56:33
Dialogue$_1$	*accept-u*	2022-06-09 11:56:33	2022-06-09 11:57:28
Dialogue$_1$	**accept-s**	2022-06-09 11:57:28	2022-06-09 11:57:28
Dialogue$_2$	**claim**	2022-06-15 17:03:34	2022-06-15 17:03:34
Dialogue$_2$	*why-explain*	2022-06-15 17:03:34	2022-06-15 17:04:22
Dialogue$_2$	**explain-f**	2022-06-15 17:04:22	2022-06-15 17:04:23
Dialogue$_2$	*what-is*	2022-06-15 17:04:23	2022-06-15 17:04:50
Dialogue$_2$	**clarify**	2022-06-15 17:04:50	2022-06-15 17:04:50
Dialogue$_2$	*why-not-explain*	2022-06-15 17:04:50	2022-06-15 17:05:38
Dialogue$_2$	**explain-cf**	2022-06-15 17:05:38	2022-06-15 17:05:40
Dialogue$_2$	*why-not-alternative*	2022-06-15 17:05:40	2022-06-15 17:06:12
Dialogue$_2$	**alter-cf**	2022-06-15 17:06:12	2022-06-15 17:06:13
Dialogue$_2$	*what-details*	2022-06-15 17:06:13	2022-06-15 17:06:59
Dialogue$_2$	**elaborate**	2022-06-15 17:06:59	2022-06-15 17:07:00
Dialogue$_2$	*reject-u*	2022-06-15 17:07:00	2022-06-15 17:07:49
Dialogue$_2$	**reject-s**	2022-06-15 17:07:49	2022-06-15 17:07:49

Subsequently, each trace consists of the set of *activities* A (in this case, locutions). In turn, a specific instance (realisation) of an activity $\alpha \in A$ (i.e., a dialogue move) is referred to as an *event* ε. Altogether, a collection of explanatory dialogues makes up the so-called *event log*.

An example of an event log basing on a collection of explanatory dialogues is depicted in Table 4. It contains two traces (i.e., Dialogue$_1$ and Dialogue$_2$) that represent instances of the recorded explanatory dialogues between (possibly, different) user(-s) and the given system (i.e., an interpretable rule-based classifier). In total, the process model contains 22 events each of which is essentially a specific dialogue move paired with the corresponding locution. Figure 6 illustrates the corresponding process model graph. The visual representation of the process model facilitates detection of the activity patterns (i.e., subprocesses characterising common parts of distinct dialogues) taking place in the collection of dialogues.

A dialogue protocol can be represented as a finite state machine whose nodes are the locutions modelled, edges being legitimate transitions between different states of the dialogue (e.g., from a request to all possible responses). In terms of process mining, one can represent the dialogue protocol as the so-called *process model* – a directed graph $M = \langle N, E\rangle$ where the set of nodes $N \subseteq A \cup \{\text{Start}, \text{End}\}$ is composed of the process activities and the set of edges $E \subseteq N \times N$ represents (possibly, causal) relations between pairs of activities where Start and End are, respectively, the start and end time of execution of the corresponding activity.

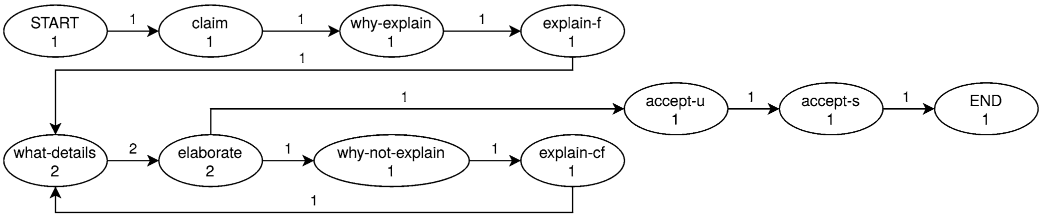


Fig. 6. The graphical view of the process model corresponding to the example Dialogue$_1$ in Table 4.

To analyse the actually recorded dialogues quantitatively, we suggest that the so-called *conformance checking* procedure be applied. In process mining, conformance checking is applied to relate the events in the actually registered processes and the process model in order to identify commonalities and discrepancies between the former and the latter. In the case of evaluating the proposed dialogue game, all the moves made by both dialogue game players follow the previously defined dialogue protocol. Hence, no deviation from the protocol can be observed. Instead, conformance checking allows us to highlight the most (and the least) frequent dialogue patterns in the event log and evaluate it against the process model (i.e., the dialogue protocol). Conformance checking can lead to obtaining data-driven knowledge of the least frequently submitted requests and/or dialogue state transitions, which can be used to modify the originally proposed dialogue protocol in order to increase its quality.

To sum it up, the proposed dialogue model can be evaluated in two complementary ways: qualitatively and quantitatively. On the one hand, qualitative free-form user feedback (e.g., in the form of a post-experiment survey) can point to missing requests or transitions between existing requests in the dialogue protocol. On the other hand, the least frequent dialogue patterns may signal their futility for explanatory purposes of the dialogue model. In process mining, a frequency threshold value can, for example, be set to subsequently optimise the process model by removing the least observed model patterns. Similarly, the least frequent requests or responses may be removed from the dialogue protocol if the empirically grounded threshold value is available and set prior to evaluation. As a result, process mining is shown to serve as a methodological basis for quantitative evaluation of the proposed dialogue model. In combination with free-form user feedback for qualitative evaluation of the dialogue protocol, process mining is able to provide us with further insights w.r.t. the quality of a dialogue model.

5. Experimental settings

In order to evaluate the proposed model of explanatory dialogue following the aforementioned evaluation framework, we carried out an exploratory user study. In the remainder of this section, we describe the setup of the human evaluation study. Thus, Section 5.1 describes the datasets used as the basis for training the classifiers for the study. Section 5.2 outlines technicalities of the explanation generation method used in the given experiment. Section 5.3 outlines the distinctive characteristics of the classifiers trained on the aforementioned datasets. Section 5.4 discusses the stimuli selection as well as the design of the dialogue system used in the experiment.

5.1. Datasets

In our study, we used the following three datasets: basketball player position [3], beer style [13], and thyroid disease diagnosis [19]. All three datasets serve to solve a multiclass classification problem in three different application domains. First, the basketball players position dataset presupposes five classes related to the following player positions: $Y_{\text{basketball}} = \{$*point-guard, shooting-guard, small-forward, power-forward, center*$\}$. Second, the beer style dataset (as was used in the illustrative example in Section 3.2) categorises instances of beer to belong to one of the following eight classes: $Y_{\text{beer}} = \{$*Blanche, Lager, Pilsner, IPA, Barleywine, Stout, Porter, Belgian strong ale*$\}$. Third, the thyroid disease dataset presupposes the following four potential labels: $Y_{\text{thyroid}} = \{$*no hypothyroid, primary hypothyroid, compensated hypothyroid, secondary hypothyroid*$\}$.

To guarantee consistent and comparable results, only numerical continuous features were used for training the corresponding classifiers. Further, all the features were mapped to linguistic terms as follows. The beer style dataset was annotated by an expert brewer, therefore it contains original feature-value partitions. The features from the other datasets were split in three uniform intervals of equal length, each of which was mapped to the following linguistic terms: $\langle$*low, medium, high*$\rangle$ (except for the feature *height*, which is described with 5 linguistic terms, in the basketball player position dataset). Table 5 summarises information on the features from all the datasets as well as the corresponding linguistic terms, with the numerical intervals attached.

5.2. *Explanation generation method*

To evaluate the dialogue game proposed in this paper as a communication interface between the system and the user, we generate multiple factual and CF explanations using the *XOR* method [72]. This explanation generation method operates on the rule base (i.e., a set of decision paths to each class) of a rule-based interpretable classifier (e.g., a fuzzy rule-based classification system or a decision tree DT where branches are first transformed into a list of rules). All automatic explanations follow the structure of the decision path (in the case of the factual explanation) or the minimally different decision path leading to the given CF class (in the case of the CF explanation). The following pipeline of four steps constitutes the explanation generation process:

(1) **Rule vectorisation.** Each rule found in the rule base is represented as a (binary, in the case of the XOR method) vector of all possible feature-value pairs. In the case of a DT, the values of the vector are all the unique conditions (e.g., "bitterness $\leqslant$ 10") found in the set of DT nodes.

(2) **Relevance estimation.** Once the rules are vectorised, a distance is calculated between vectors representing the decision path vector (responsible for the prediction) and each rule leading to the given (factual or CF) class. In the case of the XOR method, the exclusive-OR function calculates the distance between the vectors. The vectors are then ranked in accordance with the distances. The minimally distant rule is selected as a template for the output explanation following the conventional definition of a CF explanation.

(3) **Linguistic approximation.** Each interval found in the selected rule is mapped to the predefined linguistic terms by measuring the similarity between the set of numerical values corresponding to this interval and each set of numerical values for the corresponding feature. The most similar linguistic term is selected for the given feature.

(4) **Surface realisation.** The linguistically approximated rule is passed on to the surface realisation module that outputs a template-based grammatically correct high-level explanation. Similarly, the corresponding numerical intervals are used to generate a low-level explanation.

For DTs, factual explanations are essentially the feature-value intervals aggregated along the decision path. This explanation generation method presupposes that alternative factual explanations cannot be generated because alternative decision paths leading to the same predicted class would not adequately

Table 5
Numerical intervals of the features as well as the corresponding linguistic terms

Feature	*Linguistic term*	*Range of values*
Height	Short	[1.810, 1.888]
	Medium-height	[1.888, 1.966]
	Tall	[1.966, 2.044]
	Very tall	[2.044, 2.122]
	Extremely Tall	[2.122, 2.200]
Minutes	Low	[8.410, 14.290]
	Medium	[14.290, 20.160]
	High	[20.160, 26.040]
Points	Low	[2.800, 6.200]
	Medium	[6.200, 9.600]
	High	[9.600, 13.000]
2-points field points percentage	Low	[34.400, 45.500]
	Medium	[45.500, 56.600]
	High	[56.600, 67.700]
3-points field points percentage	Low	[0.000, 15.170]
	Medium	[15.170, 30.330]
	High	[30.330, 45.500]
Free throws	Low	[43.900, 59.300]
	Medium	[59.300, 74.700]
	High	[74.700, 90.100]
Rebounds	Low	[1.600, 3.330]
	Medium	[3.330, 5.070]
	High	[5.070, 6.800]
Assists	Low	[0.200, 1.930]
	Medium	[1.930, 3.670]
	High	[3.670, 5.400]
Blocks	Low	[0.000, 0.570]
	Medium	[0.570, 1.130]
	High	[1.130, 1.700]
Turnovers	Low	[0.200, 0.630]
	Medium	[0.630, 1.070]
	High	[1.070, 1.500]
Global assessment	Low	[4.000, 8.370]
	Medium	[8.370, 12.730]
	High	[12.730, 17.100]

(a) Basketball player position

Feature	*Linguistic term*	*Range of values*
Colour	Pale	[0.000, 3.000]
	Straw	[3.000, 7.500]
	Amber	[7.500, 19.000]
	Brown	[19.000, 29.000]
	Black	[29.000, 45.000]
Bitterness	Low	[7.000, 21.000]
	Low-medium	[21.000, 32.500]
	Medium-high	[32.500, 47.500]
	High	[47.500, 250.000]
Strength	Session	[0.035, 0.052]
	Standard	[0.052, 0.067]
	High	[0.067, 0.090]
	Very high	[0.090, 0.136]

(b) Beer style

Feature	*Linguistic term*	*Range of values*
Thyroid-stimulating hormone (TSH)	Low	[0.000, 3.333]
	Medium	[3.333, 6.666]
	High	[6.666, 10]
Triiodothyronine (T3)	Low	[0.050, 3.560]
	Medium	[3.560, 7.080]
	High	[7.080, 10.060]
Total thyroxine (TT4)	Low	[2.000, 94.660]
	Medium	[94.660, 187.330]
	High	[187.330, 280.000]
Thyroxine utilization (T4U)	Low	[0.250, 7.900]
	Medium	[7.900, 15.550]
	High	[15.550, 23.200]
Free thyroxine (FTI)	Low	[2.000, 84.660]
	Medium	[84.660, 167.330]
	High	[167.330, 250.000]

(c) Thyroid disease

explain the exact reasoning of the DT for the given test instance. On the contrary, alternative CF explanations are considered for explaining hypothethical, non-predicted outcomes. Once the explainer generates an explanation, it is then passed on to dialogue system upon request.

Table 6
Main characteristics of the datasets and the corresponding classifiers used in the experiments

Dataset	*# of instances*	*Accuracy*	*Precision*	*Recall*	*F-score*
Basketball	50	54.000%	0.535	0.540	0.529
Beer	400	93.500%	0.936	0.935	0.935
Thyroid	3772	95.334%	0.947	0.953	0.948

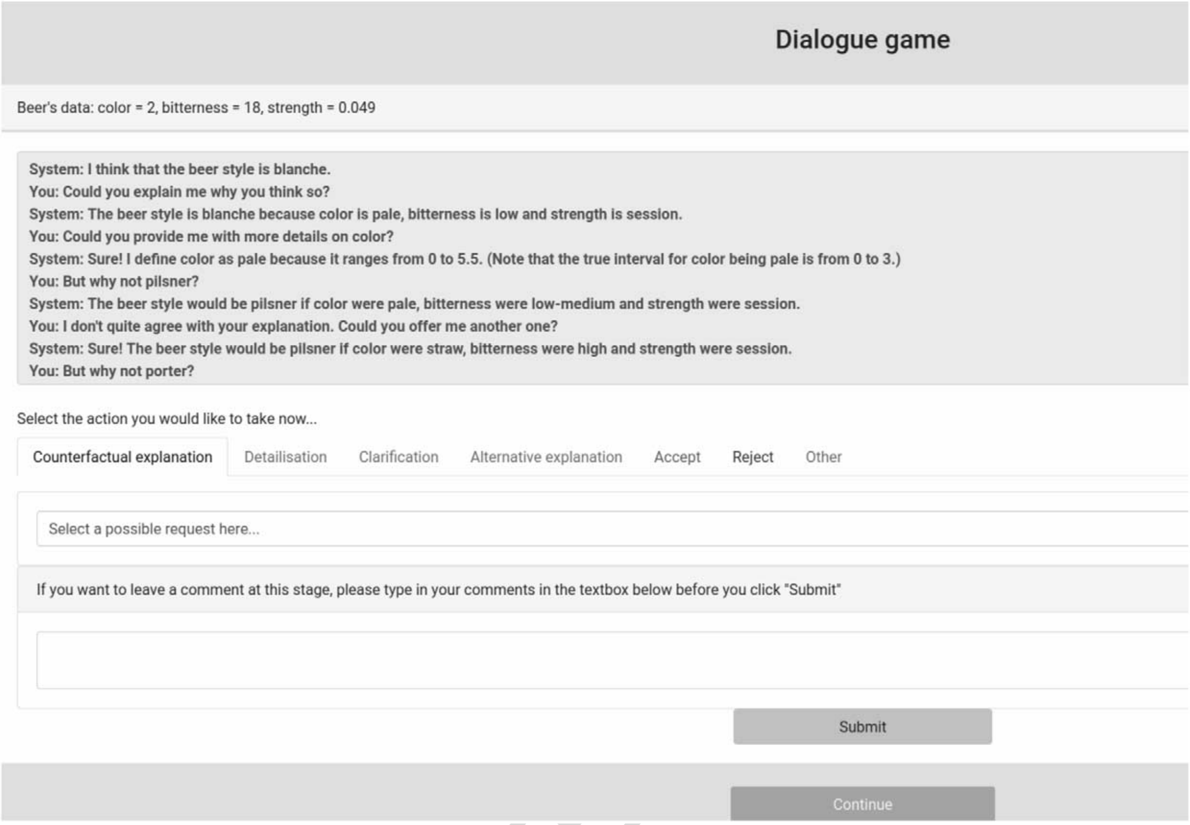
Table 7
Number of decision paths and CF classes for each dataset under consideration

Dataset	*Class*	*# of decision paths*	*# of alternative CF explanations*
Basketball	Point-guard	2	1
	Shooting-guard	2	1
	Small-forward	3	2
	Power-forward	3	2
	Center	4	3
Beer	Blanche	1	–
	Lager	2	1
	Pilsner	6	5
	IPA	8	7
	Barleywine	4	3
	Stout	2	1
	Porter	4	3
	Belgian strong ale	1	–
Thyroid	No hypothyroid	220	219
	Primary hypothyroid	49	48
	Secondary hypothyroid	2	1
	Compensated hypothyroid	186	185

5.3. *Classifiers*

In our human evaluation study, we use DTs as classifiers. Notably, DTs offer interpretable rule-based explanations that can be retrieved from their readily available internal structure. Three variants of DTs (*J48, RandomTree, REPTree*) were generated using the data mining tool Weka [30] and inspected for all the considered datasets. All the DTs were trained using 10-fold cross-validation.

It turns out that only the *RandomTree* algorithm generates at least two decision paths to all the classes in all the datasets under consideration (except for classes *Blanche* and *Belgian Strong Ale* in the beer style dataset). First, this guarantees the existence of at least one CF explanation for any class in each dataset for any test instance selected. Subsequently, it provides at least one alternative explanation for the given CF class. Since the other inspected DT algorithms did not provide at least one alternative CF explanation for the considered datasets, the *RandomTree*-based DTs were selected for all the use cases as classifiers whose predictions were to be explained in the study. Table 6 summarises main characteristics of the DTs used in the human evaluation study. Table 7 indicates numbers of decision paths for each CF class for each dataset.



Beer's data: color = 2, bitterness = 18, strength = 0.049

System: I think that the beer style is blanche.
You: Could you explain me why you think so?
System: The beer style is blanche because color is pale, bitterness is low and strength is session.
You: Could you provide me with more details on color?
System: Sure! I define color as pale because it ranges from 0 to 5.5. (Note that the true interval for color being pale is from 0 to 3.)
You: But why not pilsner?
System: The beer style would be pilsner if color were pale, bitterness were low-medium and strength were session.
You: I don't quite agree with your explanation. Could you offer me another one?
System: Sure! The beer style would be pilsner if color were straw, bitterness were high and strength were session.
You: But why not porter?

Select the action you would like to take now...

If you want to leave a comment at this stage, please type in your comments in the textbox below before you click "Submit"

Fig. 7. An example of a dialogue game human evaluation survey (the beer style dataset scenario).

5.4. Online evaluation settings

In order to execute human-machine interaction governed by means of the dialogue game proposed, we designed and implemented an online evaluation system. The corresponding ethical considerations are outlined in Appendix A. Figure 7 presents an example screen of the implemented software tool.[3] Further, the source code of the dialogue game survey, the DTs used in the experiments, and the collected experimental data are made publicly available.[4]

In the course of the study, the participants were presented the characteristics of a test instance following the chosen scenario (dataset). The participants did not have any prior knowledge about the dataset. They were asked to interact with the system until they could make an *informed* decision on acceptance or rejection of the system's claim. The participants determined the flow of the dialogue, as they requested necessary information to make a final decision.

Three test instances (one per dataset) were selected so that they would represent correctly predicted real data. Table 8 outlines the characteristics of the test instances used in the study. The following factual explanations were generated for the considered test instances:

[3] https://tec.citius.usc.es/dialgame
[4] https://gitlab.citius.usc.es/ilia.stepin/fcfexpgen (branch "dialgame").

Table 8

Test instance characteristics

Height	*Minutes*	*Points*	*2-points field goals percentage*	*3-points field goals percentage*	*Free throws*	*Rebounds*	*Assists*	*Blocks*	*Turnovers*	*Global assessment*	*Class*
1.85	21.19	9.2	43.1	40.0	81.9	1.9	3.8	0.0	0.7	8.8	Point-guard

(a) Basketball player position

Colour	*Bitterness*	*Strength*	*Class*
2	18	0.049	Blanche

(b) Beer style

Thyroid-stimulating hormone (TSH)	*Triiodothyronine (T3)*	*Total thyroxine (TT4)*	*Thyroxine utilization rate (T4U)*	*Free thyroxine index (FTI)*	*Class*
4.6	1.2	48	0.89	54	Secondary hypothyroid

(c) Thyroid diagnosis

- **Basketball:** "The player's position is point-guard because the number of rebounds is low and the number of assists is high."
- **Beer:** "The beer style is Blanche because colour is pale, bitterness is low and strength is session."
- **Thyroid:** "The patient has secondary hypothyroid because thyroid-stimulating hormone is medium, triiodothyronine is medium and total thyroxine is low."

Similarly, all the high-level automatically generated CF explanations contained only textual descriptions of the features involved. As all the features are numerical (either integer or real-valued), responses to detailisation requests would provide subjects with intervals to which the linguistic terms are mapped. Further, the users were then informed about the classifier's numerical intervals found for the given feature along the given decision path. These details were assumed to facilitate matching the system's claim with the feature-value pairs of the test instance.

Noteworthy, the same study participants could select multiple datasets to play the dialogue game. Therefore, the numbers of records for each dataset do not represent unique users. For this reason, whenever we hereinafter mention the study participants (subjects), we refer to the actually collected transcripts of explanatory dialogues.

Upon completion of the experiment, the study participants were asked to optionally provide their demographic data and leave free-text responses to the following questions and/or suggestions:

Q1 "If you could add other types of requests to the system, what would those be?";
Q2 "Did the interaction with the system change your initial (dis-)belief in the system's prediction? Why (not)?";
Q3 "If you have any other comments for us, please leave them in the textbox below."

Last but not least, all the collected dialogue transcripts were transformed into event logs. On the basis of the event logs, process models were then constructed for each use case. In addition, a global process model of all the event logs was calculated.

Table 9

General properties of the collected dialogues

Property	Dataset			All datasets
	Basketball	Beer	Thyroid	
		Number of dialogue moves		
Mean	12.57	15.76	10.11	14.17
Median	12.00	15.00	9.00	13.00
St.dev.	6.98	7.00	3.18	6.83
		Time taken (min)		
Mean	04 m 09 s	08 m 47 s	05 m 17 s	07 m 10 s
Median	04 m 01 s	05 m 42 s	04 m 54 s	04 m 35 s
St.dev.	01 m 39 s	09 m 39 s	02 m 31 s	07 m 55 s

6. Experimental results

In this section, we report the collected human evaluation results. Section 6.1 presents the quantitative results of the study (i.e., descriptive analytics of the collected dialogues and insights from the process models). Section 6.2 reports the qualitative results of the evaluation study (i.e., the free-form feedback that the study participants left optionally after their interaction with the dialogue system).

6.1. Dialogue analytics

A total of 60 dialogue transcripts have been collected in the course of the empirical study. In particular, 14 (23.33%) of the records relate to the basketball player position dataset. In turn, 37 (61.67%) transcripts are composed as the result of interaction with the classifier trained on the beer style dataset. In addition, 9 (15.00%) records reflect user interaction with the thyroid dataset-based classifier. All the collected dialogue transcripts were converted into event logs. The event logs were subsequently used to generate two process models: (1) the one related to the main building blocks of the modelled explanatory dialogue (i.e., claim, explanation, and termination) and (2) the one covering all the locutions produced by the study participants. Process model (1) gives a high-level overview of the user behaviour whereas process model (2) provides insights w.r.t. specific moves made by the study participants.

On average, it took the dialogue game participants around 14 moves for the users to make their final decision with respect to the system's claim. As for the time taken to complete the dialogue game, the study participants spent about 7 minutes to either accept or reject the claim. Table 9 reports average numbers of dialogue moves and the time taken to complete the dialogue for each dataset under consideration.

Figure 8 illustrates the process model corresponding to the three main building blocks of the proposed dialogue game (i.e., claim, explanation, and termination). Thus, all but three participants required (at least, factual) explanation for the given prediction. Almost all of them eventually accepted the system's claim. In the remainder of this section, we are analysing only those transcripts where explanations were requested.

Figure 9 depicts the process model of the collection of explanatory dialogues that displays all the locutions produced. Thus, 331 explanation-related requests (all those covered by the EXPLANATION non-terminal in EDG) have been registered from the 57 participants who required explanation for the system's claim. The edge labels for the explanation-related requests in Fig. 9 show that the study participants actively exploited all the explanation-related requests that were designed in the original protocol.

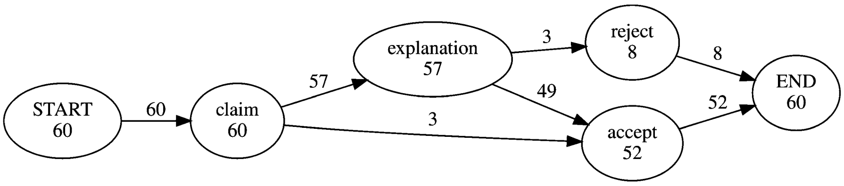


Fig. 8. The process model of all the collected explanatory dialogues based on the main EDG building blocks. The block “termination” is split into “accept” and “reject”.

On the one hand, a majority of the participants submitted further explanation-related requests (in this case, detailisation or clarification) upon receiving the factual explanation. On the other hand, a quarter of all the study participants considered the factual explanation sufficiently comprehensive to immediately request a (set of) CF explanation(-s).

The locution-level process model (see Fig. 9 for details) allows us to observe the answers to which requests were the most decisive for the participants to make their final decisions. Thus, the system’s claim was mainly accepted immediately after CF explanations (including those alternative) were presented whereas only one participant accepted the system’s claim did so as soon as the factual explanation was offered. The other explanation-related requests (i.e., detailisation and clarification) are found to have contributed less to immediate acceptance of the system’s claim. As for claim rejections, alternative CF explanations happen to most frequently trigger negative user decisions. Notably, alternative CF explanations were requested for nearly a half of all 76 CF explanations offered. In most cases, study participants stopped exploring the explanation space for the given CF class after the second-best ranked CF explanation was offered. However, third-best ranked CFs were requested to a limited extent.

It is worth noting that further insights into the quantitative results for individual use cases can be found in Appendix D.

6.2. *User feedback*

In this section we present all the free-form comments that the study participants left upon finishing their interaction with the system and summarise the most informative of them. Recall that study participants were encouraged to leave answers to two questions (Q1 and Q2) and/or indicate their free-form suggestions (Q3) unrelated to Q1 or Q2 after their interaction with the implemented dialogue system. The collected responses to Q1–Q3 are presented in Tables 10–12. As all the comments shown are original, some may contain grammatical, lexical, and/or orthographic errors. All the users’ statements are codified as follows: “$Cx.y$” where C stands for “comment”, x is the corresponding question number and y is the answer number.

Table 10 presents all the answers to Q1 (“If you could add other types of requests to the system, what would those be?”) that we collected throughout the study. Two comments (C1.1 and C1.2) are related to the basketball player position. Six statements (C1.3–C1.8) were made as a result of interaction with the system in the beer style case settings. One study participant left his or her comment (C1.9) after playing with the thyroid disease diagnosis scenario.

Regarding Q1, the study participants would like to extend the actual dialogue model so that it could inform them about the second most probable decision, or the technicalities of the decision-making system (e.g., the accuracy of the system). In addition, further definitions of notions related to the domain knowledge (see Comment C1.6, Table 10) were desired. Notably, concerns were raised about the in-

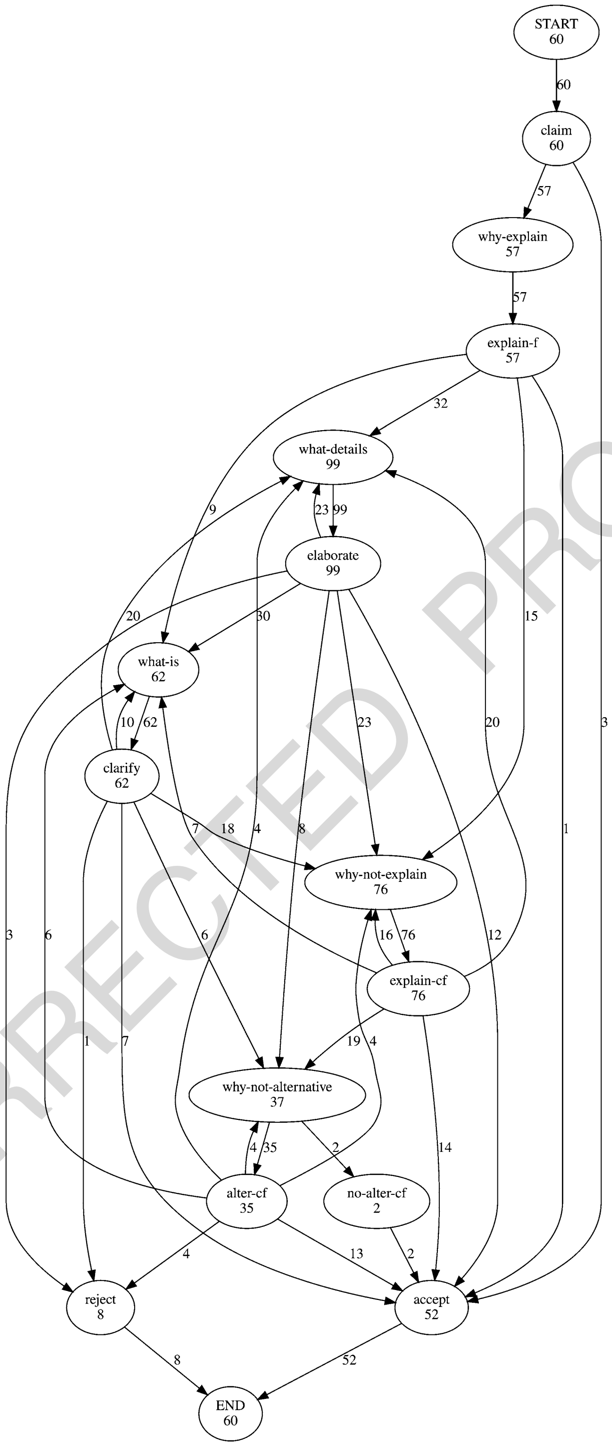


Fig. 9. The full process model of all the collected explanatory dialogues. For illustrative purposes, pairs of termination nodes, i.e. *{accept-u, accept-s}* and *{reject-u and reject-s}*, are merged into *accept* and *reject*, respectively.

Table 10

Study participants' answers to Q1 ("If you could add other types of requests to the system, what would those be?")

ID	User's statement
C1.1	"I'm unsure"
C1.2	"explain what is your primary goal for the predictions you are making"
C1.3	"Summarisation"
C1.4	"In clarifications, I'd like to not only get the definition of the strength but also the types of strength that exist. For example, Blanche's strength is session but I have no idea what session means."
C1.5	"It would be good to have some clarification of different terms than fixed one like color"
C1.6	"I would add more elaborated set of definitions, i.e. definitions of technical terms which are used for definitions."
C1.7	"how did you measure the (.); what is the accuracy of this measurement tool? What is the probability of your prediction?; how did you calculate this probability?"
C1.8	"I would like the possibility of going back to previous points. It seems to me that after the counterfactual explanation I was stuck on it, and going back to the original prediction was at least not intuitive. A graph of the history of dialogue that would allow me to travel through explanations would be great. Predefined options were not very clear to me I think a better explanation with examples would be beneficial. There might be corner cases on different topics that would make differentiating those options even harder."
C1.9	"Second most probable choice (differential diagnoses in the case of the thyroid case)"

ability to post-process the pieces of explanation that had already been discussed (see Comment C1.8, Table 10).

Table 11 shows all the collected answers to Q2 ("Did the interaction with the system change your initial (dis-)belief in the system's prediction? Why (not)?"). Five study participants (C2.1–C2.5) answered Q2 after making their decision on the automatic basketball player position classification. Ten statements (C2.6–C2.15) were made as a result of interaction with the system in the beer style case settings. Two study participants (C2.16–C2.17) commented on their interaction with the system, as the thyroid disease classification scenario was executed.

Regarding Q2, a fair number of commentators found the offered automated explanations convincing and satisfactory. Comment C2.5 (Table 11) illustrates that this was, in part, achieved due to the possibility to opt for factual explanations. In addition, some study participants positively assessed the ability to query the system for CF explanations (see Comment C2.8, Table 11) and further details and clarifications (see Comment C2.3, Table 11). Some of the commentators whose initial (dis-)belief in the system's claim did not change in the course of their interaction with the system remarked that the explanations offered were nevertheless satisfying (see Comment C2.2, Table 11) and supportive enough w.r.t. the system's claim (see Comment C2.11, Table 11).

Table 12 presents all users' free-form suggestions (Q3: "If you have any other comments for us, please leave them in the textbox below."). One comment (C3.1) was left after a dialogue with system w.r.t. the basketball player position classification whereas two statements (C3.2–C3.3) were made as a result of interaction with the system in the beer style case settings.

Regarding Q3, one study participant commented that the system's responses were too fast (see Comment C3.1, Table 12). In addition, another participant pointed out the need for supportive visualisation tools, a clearer distinction between detailisation and clarification requests, and different structures for alternative explanations for the same CFs (see Comment C3.2, Table 12). Finally, predictions for other data instances are found desired to be inspected to develop big picture thinking about the reasoning of the system (see Comment C3.3, Table 12).

Table 11

Study participants' answers to Q2 ("Did the interaction with the system change your initial (dis-)belief in the system's prediction? Why (not)?")

ID	User's statement
C2.1	"Yes. It provided a counter argument of why they had provided that prediction specifically and not another that I suggested."
C2.2	"No because the system had the numbers, so I believed it from start to finish."
C2.3	"I have no knowledge of basketball but the explanations were convincing so I was happy to accept the prediction after asking further questions"
C2.4	"It made me feel that the system has a certain etos but did not teach me about how these predictions are actually computed"
C2.5	"The system was able to successfully convince me of the prediction based on the factual information it provided."
C2.6	"No"
C2.7	"It didn't describe the details of the low bitterness when I asked about bitterness following a discussion about ipa. It provided me with details about high bitterness and outlined that ipa has high bitterness. I could not clarify the bitterness low level range that was the suggested prediction of Blanche."
C2.8	"Yes, seeing the classifications of the other types that is suspected made me accept that this prediction must be correct"
C2.9	"Yes, it gave me a deeper understanding of beer classification. It is a nice way to learn and to gain trust in AI system."
C2.10	"The system responses were good and straight to the point so it was quite convincing."
C2.11	"It did not. I thought it was pretty accurate from the start and given the example before the experimental item I could already gather a good idea of what was expected."
C2.12	"yes, in the beginning I didn't understand one of the words and my first thought was that the word, which was awkward to me, was an effect of system's malfunctioning."
C2.13	"I did not have a strong initial belief about the system prediction. However, it was convincing enough for me."
C2.14	"No – I had no experience or grounds on which to doubt what I was being told. The questions and answers seemed a matter of technical specification and not a matter of beliefs."
C2.15	"Not really, I know it is difficult for an AI system to have long dialogues as it needs to take account with everything that has been said before."
C2.16	"Not really, because I didn't have any expectations"
C2.17	"Clarification of the prediction terms as well as the features would be useful. For example, what hypothyroid means etc"

7. Discussion

The findings reported in the previous section enable us to outline several remarkable observations. As expected, high numbers of detailisation and clarification requests have been registered from the users interacting with a classifier in the settings where they did not have any prior knowledge of the dataset that the classifier had been trained on. As the users started their interaction with the system only having feature-value pairs of the test instance at their disposal, they oftentimes required not only an explanation to the system's claim but, perhaps, more importantly, definitions of the features that made part of the explanation or the numerical ranges over which the features were defined. The fact that a high number of requests for alternative explanations have been registered across all the use cases confirms that the most relevant explanation from the system's point of view may be far from the most relevant (or satisfactory) from the user's point of view.

As the same prediction can be explained in different ways, it turns out to be particularly important to extend the protocol so that it does not only offer the opportunity to rephrase the initially offered explanation but also enables the system to send requests to the user. For instance, if two pieces of

Table 12

Study participants' suggestions w.r.t. to Q3 ("If you have any other comments for us, please leave them in the textbox below")

Comment ID	User's statement
C3.1	"The responses were very fast, a slight delay after receiving a request would improve how the answer appears"
C3.2	"In the beginning, it'd be nice to have some kind of photo prompt together with the beer data to help vizualise what we are talking about. It's a bit hard to distinguish between detalisation and clarification. I didn't see the difference in the structures of counterfactual explanation and alternative explanation. In my case, for the counterfactual explanation, I asked about pilsner and when giving me an alternative explanation the system also used pilsner so I didn't get new information from the last request."
C3.3	"I would be curious to learn more about other topics and other predictions on the subject I took (in this case, beer)."

explanation are deemed equally relevant by the explanation generation module, requiring additional information from the user about his or her preferences may be crucially important for successful fine-tuning of the explanation being processed. On the one hand, both such explanations can be presented simultaneously. Then, the user is to decide the format and/or ordering of the output explanations. On the other hand, the system can submit a request to the user to infer the actual user's needs taking into consideration the known differences between two explanations.

The qualitative results of the human evaluation study allow us to suggest a number of empirically-driven critical questions (CQ) to the system's prediction. Recall that our factual and CF textual explanations (in the simplest form) follow the templates "The test instance is [CLASS] because [FEATURE] is [VALUE]" and "The test instance would be [CLASS] if [FEATURE] were [VALUE]", respectively. We can therefore address CQs both to the prediction (variable CLASS in the example above) and to (components of) the explanation (the variables FEATURE and VALUE in the example above). Driven by the registered user feedback, the prediction-related CQs (CQ1, CQ2, and CQ3) can be exemplified as follows:

CQ1 Is the system's prediction correct?

CQ2 What is/are the accuracy/precision/recall/F-score of the system that predicted [CLASS]? (following C1.7 from Table 10);

CQ3 How were the accuracy/precision/recall/F-score calculated? (following C1.7 from Table 10).

In turn, the features and values of the given explanation may give rise to explanation-related CQs. For example, the feature values may be subject to explanation-related CQs that may occur when processing responses to detailisation requests (CQ4 and CQ5) while the definitions of the features themselves may be questioned upon performing clarification requests (CQ6):

CQ4 What data justify [VALUE] for [FEATURE]? (in the case of high-level explanations);

CQ5 Is [VALUE] consistently defined for [FEATURE] in [INTERVAL]? (where [VALUE] is the linguistic term of some high-level explanation's feature and [INTERVAL] is the corresponding numerical interval of the low-level explanation);

CQ6 Is the source of information of the definition of [FEATURE] credible?

The proposed dialogue model has a number of limitations. As it can be applied directly only to interpretable rule-based classifiers enhanced with explainers providing textual explanations, the communication between the system and the user may appear overly restricted. In light of the assumptions made in Section 2, parts of the protocol may have to be adjusted when dealing with, for example, categorical variables or a poorly interpretable feature space. In addition, the structure of the protocol may have to be made more flexible, as handling the previously processed explanations (for example, those for other CF classes) is not permitted.

Remarkably, the set of locutions included in the presented protocol is by no means exhaustive. The qualitative results of the human evaluation study signal a number of desired extensions to the proposed dialogue model. The users would, for example, appreciate to know more about the definitions of the linguistic terms. The modular architecture of the EDG production rules allows for adapting the dialogue game for developer's as well as user's needs. In this regard, the clarification requests can be made applicable not only to the features themselves but also to the values of the linguistic variable that appear in high-level explanations as well as domain knowledge-related terms. In addition, the proposed dialogue protocol might as well incorporate visual information (e.g., pictures of the domain knowledge available upon request) for detailisation requests.

8. Related work

A variety of computational argumentation models have proven to be efficient tools for explanatory dialogue modelling in the context of XAI. For instance, Arioua et al. [4] propose a formal model of argumentative explanatory dialogue to acquire new knowledge in inconsistent knowledge bases. Calegari et al. [10] implement a mechanism of reasoning over defeasible preferences using elements of abstract and structured argumentation. Groza et al. [26] model explanatory dialogues combining rule-based arguments extracted from both ML classifiers and expert knowledge in favour or against a given classification of retinal disorder. Subsequently, the arguments are used to persuade the other parties in multi-agent system settings.

Argumentative explanatory dialogues are of particular interest among XAI researchers, as they provide means for customisation of automated CF explanations in light of the collected user feedback [70]. There exist a large number of distinct techniques that allow for integrating user feedback to personalise initially generated CF explanations. For example, Suffian et al. [73] operate on user's preferred features and the corresponding ranges of values to fine-tune the originally generated explanation. Their FCE method first generates synthetically a set of CF data points where the preferred features range in the selected intervals. Then, the model aims to detect the most relevant (yet personalised) CF by searching for the minimally different (in terms of distance) CF data point from the generated synthetic data. Behrens et al. [6] propose a dynamically updated framework for user-specific explanation generation for knowledge graphs. More precisely, the user expresses his or her preferences by selecting two desired sets of graph nodes and, subsequently, ordering the selected generated meta-paths (i.e., sequences of alternated nodes and edges). Ghazimatin et al. [24] collect user feedback on explanations themselves for a recommender system to improve its performance. In this case, the user feedback is essentially a binary value signalling the similarity of an explanation to the recommendation. De Toni et al. [18] consider the problem of causal CF explanation generation as algorithmic recourse (i.e., overturning unfavourable ML-based model's prediction). In their reinforcement learning-based model, the user is asked to choose the best subsequent action from the so-called "choice set". The user's responses are then used to optimise the model's weights via Bayesian estimation and update the user's state.

Early computational models of explanatory dialogue stress that the context of explanation should depend on user's familiarity with the concepts presented to him or her [14]. Further, the end user is argued to necessarily build a sound mental model of the system to successfully interact with it [84]. However, only a few of argumentative explanatory dialogue implementations allow for direct dialogic interaction between an AI-based system and a given user for explanation customisation. Despite little evidence, human evaluation of the automatically generated explanations may lead to groundbreaking conclusions. For instance, Rago et al. [56] emphasise the need for multi-modality of the generated argumentative explanations, as users are found to generally prefer tabular explanations over textual ones but also textual over conversational. In addition, explanations containing a greater number of features (aspects) are, in general, found to be preferred.

Formal dialogue games provide an intuitive transparent tool of information exchange between the agents involved [54]. They have been extensively used in a wide range of AI applications, such as multi-agent systems [44] and recommendation systems [42]. Dialogue games have shown to have great potential for explanatory dialogue modelling [36]. The first dialogue games for (computational) explanatory dialogue modelling trace back to works by Walton [81] and Modgil and Caminada [46]. Arioua and Croitoru [5] propose a dialogue game to formalise Walton's dialectical system of explanatory dialogues. However, their formalism does not take into account some key properties of explanation (contrastive, selected, and social) as well as user-specific needs addressed in the field of XAI. On the other hand, Shao et al. [67] explain a neural network's classification output enabling the user to adjust the classifier's prediction by enabling the user to prove feedback on the arguments correcting the prediction. Shaheen et al. [65] design two dialogue game-based protocols for generating and communicating explanations for satisfiability modulo theory (SMT) solvers. Thus, their approach distinguishes between a passive explanatory dialogue game where the explainee only inquires explanation and an active game where the user is explicitly asked to confirm or refute the system's assertion. Unfortunately, both protocols lack any empirical evaluation. Alternatively, Sklar and Azhar [69] perform a user study to evaluate a dialogue game-based framework for making cooperative actions in the treasure hunt game. They show that explanations communicated using a dialogue game-based communication protocol lead to above-average user satisfaction. Shams et al. [66] design a dialogue game to explain and justify the best agent's plan in normative practical reasoning settings. Finally, argumentative dialogue game-based models have been proposed for generating model-agnostic local explanations to justify given predictions [55]. To the best of our knowledge, no other dialogue games (including those aforementioned) have ever been evaluated (quantitatively) using process mining techniques like those introduced in this paper.

The previously mentioned protocols were mainly proposed for modelling information-seeking or inquiry explanatory dialogues. However, the formalism of dialogue games is also suitable for (and extensively applied to) modelling persuasive explanatory dialogues. Thus, Sassoon et al. [61] center explanatory dialogue around instances of a domain-specific argumentation scheme guided with the corresponding critical questions. Depending on the degree of agreement between the agents, the explanatory dialogue is then modelled in one of the three following modalities: information-seeking, deliberation, or persuasion. Morveli-Espinoza et al. [48] propose a protocol for persuasive negotiation dialogues where agents exchange explanatory and rhetorical arguments. Similarly to our approach, they consider alternative responses to be, in part, attacks to the previously uttered arguments. However, their protocol does not tackle CF explanations.

Last but not least, a large body of research has attempted to formalise dialogue by means of dialogue grammars [32,58]. Thus, they have been regarded as a natural interface between the underlying speech acts and actually produced utterances [64]. Dialogue grammars have been shown to disambiguate between distinct dialogue flow patterns (e.g., elaboration, digression, problem resolution, to name a few) [33]. In addition, dialogue grammars facilitate induction of task-based dialogue systems [22]. Beneficially, such grammars can be learned from dialogic data in an unsupervised manner [23]. Further, dialogue grammars are scalable yet universally induced from any domain [38]. Subsequently, the grammar-based approach to dialogue modelling has been enhanced with methods of corpus-based query generation for natural language understanding [34].

Dialogue grammars are found to model human-human dialogue [68] as well as human-machine dialogue [39]. Thus, dialogue grammars appear particularly useful for multimodal human-machine interaction. For instance, hybrid multiset grammars are proposed to govern speech and textual input jointly [20]. On a similar note, Kottur et al. [40] propose a dialogue grammar for visual co-reference resolution. In contrast to the aforementioned approaches where the explanatory dialogue is formalised by means of dialogue grammars, our EDG allows for producing natural language output only. However, a high degree of modularity that dialogue grammars offer makes it possible to extend the dialogue model so that it also outputs visual data (e.g., saliency maps) if such visual explanations are included in the set of terminals of the grammar.

9. Conclusions and future work

In this paper, we presented a new approach for explanatory dialogue modelling. Namely, we designed a dialogue game for the task of communicating explanations for predictions of interpretable rule-based classifiers. Unlike previous approaches, the dialogue protocol proposed in this work allows for effective communication of both factual and CF explanations for expert and lay users. The protocol offers a transparent means of conveying personalised textual rule-based explanations. Its use can be extended to other interpretable rule-based classifiers (e.g., other DT algorithms or fuzzy rule-based classification systems).

Subsequently, we validated the dialogue protocol by carrying out a human evaluation study. The quantitative results (i.e., the reconstructed process models) confirm the necessity in all the proposed requests for explanatory dialogue between the classifier and its user and therefore proves them indispensable for explanatory dialogue modelling. Thus, detailisation and clarification requests are found particularly useful when natural language explanations are presented in the settings where users have no prior knowledge of the dataset. In addition, end users show a high degree of interest in CF explanations in addition to their factual counterparts. Further, they appear to appreciate the possibility to question the initially offered CF explanations across different application domains. Provided that such CF explanations are generated automatically and presented to the user in accordance with their relevance to the test instance (e.g., the distance from the test instance), the proposed protocol allows the explainer to communicate multiple explanations. Hence, it favours diversity of the offered explanations, which is shown to increase their explanatory power. Moreover, the qualitative results show that the proposed dialogue game appears to be an effective tool to convey appealing explanations which were convincing enough for a good number of users. In this sense, the set of the proposed requests and replies turns out to be a potentially effective tool for measuring the effectiveness of (counter-)factual explanation generation frameworks outputting textual explanations in the course of interaction with end users. Finally, the protocol is flexible enough to be adapted in the near future for estimating the trustworthiness, satisfaction, or persuasive capability of automatically generated explanations while preserving the original structure of the given explanatory dialogue modelled. Nevertheless, the proposed protocol may be found somewhat overrestrictive, as it does not enable end users to submit explanation-related requests for the pieces of explanation whose processing is considered finalised.

The present piece of research opens the door for several lines of future work. Importantly, the proposed dialogue model should be adapted to handle other types of classifiers including those that do not reveal any interpretable information about their internals. In many settings, knowledge of the feature space is unavailable or hard to interpret. Then, the detailisation requests may result being of little utility unless additionally adapted to the functionality of the given classifier. In addition, we intend to enlarge the argumentative potential of the proposed dialogue model by developing further methods of capturing user's preferences. Further work is also necessary to incorporate explanations of other modalities (e.g., visual) for dialogic communication. Whereas the concept of explanation space may be directly applicable to other settings (e.g., a prediction can be explained by means of different pieces of visual information), this may require redefinition of sub-components of the explanation space.

Another important line of future work consists in extending the actual protocol to incorporate explanations of different content and tasks. For instance, it is of peculiar interest to test the applicability of the dialogue protocol in the settings of regression, recommendation, or planning tasks. Finally, we aim to design and carry out further human evaluation experiments on the trade-off between the limitations of the protocol (e.g., underrepresentated locution types) and the persuasive power of explanations that it communicates. Such experiments (e.g., disabling users to perform specific acts) would allow us to estimate the impact of specific requests and further shape the protocol.

Appendix A. Ethical considerations

All the information collected from the human evaluation study participants was in agreement with the European Union's General Data Protection Regulation (GDPR). In addition, this piece of research has been approved by the Ethics Committee of the University of Santiago de Compostela (Spain). Human evaluation was based solely on non-personal or anonymous data. Further, all the participants gave informed consent confirming the following:

- the participant reached the age of majority;
- participation in the study was completely voluntary;
- participation in the study could be terminated at any time;
- participant's anonymous responses would be used for research purposes in accordance with GDPR.

Appendix B. Dialogue protocol

In our model, any explanatory dialogue is modelled in accordance with the protocol outlined below. Thus, the protocol presupposes the following rules:

(1) **Turntaking.** The system initiates the dialogue, i.e. it makes the move m_1 by claiming the prediction from the domain-specific finite set of all possible predictions $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$ corresponding to the dataset classes. Every subsequent even ($m_2, m_4, \ldots$) and odd ($m_3, m_5, \ldots$) moves are made by the user and the system, respectively. Each participant is allowed to produce only one locution at a time.

(2) **User's U allowed moves.**

(a) *why-explain*($\hat{y}$, E): U requests to factually explain $\hat{y}$. The explanation store E remains empty. The system is allowed to respond in either of the following ways:

- *explain-f*($\hat{y}$, E, e_f) iff S is able to produce a factual explanation;
- *no-explain-f*($\hat{y}$, E) otherwise.

(b) *why-not-explain*($\hat{y}$, E, y'): U requests to counterfactually explain why $\hat{y}$ and not y'. E must contain a factual explanation for $\hat{y}$: $E = \{e_f(\hat{y})\}$. The system is allowed to produce either of the following locutions:

- *explain-cf*($\hat{y}$, E, y', e_{cf}) if S is able to produce a CF explanation;
- *no-explain-cf*($\hat{y}$, E, y') otherwise.

(c) *what-details*($\hat{y}$, E[,y'], e, Γ) where $e = e^h_f | e^h_{cf}$: U requests details on a feature Γ used in a previously uttered (factual or CF) high-level explanation e ($\Gamma \in e$). In response, S generates one of the locutions below:

- *elaborate*($\hat{y}$, E[,y'], e, Γ, θ) if S is capable of providing U with details on feature Γ;
- *no-elaborate*($\hat{y}$, E[,y'], e, Γ) otherwise. Note that the parameter y' is optional and passed on iff $e = e^h_{cf}$.

(d) *what-is*($\hat{y}$, E[,y'], e, Ψ) where $e = e^h_f | e^h_{cf} | e^l_f | e^l_{cf}$: U requests a definition of a specific feature Ψ being part of (factual or CF, high- or low-level) explanation e ($\Psi \in e$). The system is allowed to respond using one of the following locutions:

- *clarify*($\hat{y}$, E[,y'], e, Ψ, υ) if S can provide U with such a definition;
- *no-clarify*($\hat{y}$, E[,y'], e, Ψ) otherwise. Note that the parameter y' is optional and passed on iff $e = e^h_{cf} | e^l_{cf}$.

(e) *why-alternative*($\hat{y}$, E, e_f): U disagrees (or is not satisfied) with the offered factual explanation e_f and requires an alternative factual explanation. The system responds producing one of the following locutions:

- *alter-f*($\hat{y}$, E, e_f, e'_f) if S is capable of providing U with a different factual explanation;
- *no-alter-f*($\hat{y}$, E, e_f) otherwise.

(f) *why-not-alternative*($\hat{y}$, E, y', e_{cf}): U disagrees with the offered CF explanation e_{cf} and requires that S provide an alternative CF explanation. The system replies using one of the following locutions:

- *alter-cf*($\hat{y}$, E, y', e_{cf}, e'_{cf}) provided that an alternative CF explanation e'_{cf} can be offered;
- *no-alter-cf*($\hat{y}$, E, y', e_{cf}) otherwise.

(g) *accept-u*($\hat{y}$, E): U accepts the prediction $\hat{y}$. In response, the system generates the fairwell locution *accept-s*($\hat{y}$, E).

(h) *reject-u*($\hat{y}$, E): U rejects the prediction $\hat{y}$. In response, the system generates the fairwell locution *reject-s*($\hat{y}$, E).

(3) **System's S allowed moves.**

(a) *claim*($\hat{y}$, E): S claims prediction $\hat{y}$. The knowledge store K and the explanation store E are initialised to be empty. U is allowed to:

- require a factual explanation (locution *why-explain*($\hat{y}$, E));
- accept prediction $\hat{y}$ without any subsequent explanation (locution *accept-u*($\hat{y}$, E));
- reject prediction $\hat{y}$ without any subsequent explanation (locution *reject-u*($\hat{y}$, E)).

(b) *explain-f*$(\hat{y}, E, e_f)$: S factually explains $\hat{y}$ with e_f and provides U with its high-level component (recall that $e_f(\hat{y}) = \langle e^h_f(\hat{y}), e^l_f(\hat{y})\rangle$). The factual explanation is added to the knowledge store $K = K \cup e_f(\hat{y})$ and the explanation store $E = E \cup e_f(\hat{y})$. The detailisation and clarification stores are populated with the features making part of the explanation e_f. User U is then allowed to:

- require a CF explanation for some CF class $y' \in CFS$ (locution *why-not-explain*($\hat{y}$, E, y'));
- ask for details on a feature $\Gamma \in DET$ of the factual explanation (locution *what-details*($\hat{y}$, E, e_f, Γ));
- demand a definition of some feature $\Psi \in CLAR$ making part of the factual explanation e_f (locution *what-is*($\hat{y}$, E, e_f, Ψ));
- disagree with the factual explanation e_f for prediction $\hat{y}$ and require an alternative factual explanation (locution *why-alternative*($\hat{y}$, E, e_f));
- accept prediction $\hat{y}$ (locution *accept-u*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject-u*($\hat{y}$, E)).

(c) *no-explain-f*$(\hat{y}, E)$: S is unable to factually explain $\hat{y}$. U may nevertheless:

- require a CF explanation for some CF class y' (locution *why-not-explain*($\hat{y}$, E, y'));
- accept prediction $\hat{y}$ (locution *accept*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject*($\hat{y}$, E)).

(d) *explain-cf*$(\hat{y}, E, y', e_{cf})$: S counterfactually explains why $\hat{y}$ and not y' with e_{cf} and provides U with its high-level component (recall that $e_{cf}(\hat{y}, y') = \langle e^h_{cf}(\hat{y}, y'), e^l_{cf}(\hat{y}, y')\rangle$). The CF explanation is added to the knowledge store $K = K \cup e_{cf}(\hat{y}, y')$ and the explanation store: $E = E \cup e_{cf}(\hat{y}, y')$. The CF class y' is then eliminated from the CF class store: $CFS = CFS \setminus \{y'\}$. In response, U is allowed to:

- require details on some feature $\Gamma \in DET$ for the given CF explanation (locution *what-details*($\hat{y}$, E, y', e_{cf}, Γ));
- request a definition of a feature making part of the CF explanation e_{cf} (locution *what-is*($\hat{y}$, E, y', e_{cf}, Ψ));
- disagree with the offered CF explanation and ask for an alternative one for the same CF class (locution *why-not-alternative*($\hat{y}$, E, y', e_{cf}));
- require a CF explanation for another CF class from the CF class store $y'' \in CFS$ (locution *why-not-explain*($\hat{y}$, E, y''));
- accept prediction $\hat{y}$ (locution *accept*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject*($\hat{y}$, E)).

(e) *no-explain-cf*$(\hat{y}, E, y')$: S is unable to counterfactually explain why $\hat{y}$ and not y'. The CF class y' is eliminated from the CF class store: $CFS = CFS \setminus \{y'\}$. User U is allowed to:

- require a CF explanation for another CF class $y'' \in CFS$ (locution *why-not-explain*($\hat{y}$, E, y''));

- disagree with the factual explanation and require an alternative to it provided that it is the only explanation that the explanation store E contains (locution *why-alternative*($\hat{y}$, E, e_{cf})) iff $\nexists\, e_{cf} \in E$;
- accept prediction $\hat{y}$ (locution *accept*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject*($\hat{y}$, E)).

(f) *elaborate* ($\hat{y}$, E [,y'], e, Γ, θ) where $e = e^h_f | e^h_{cf}$: S provides required details θ on feature Γ of a high-level (factual or CF) explanation e. The feature Γ is therefore excluded from the detailisation store: $DET = DET \setminus \{\Gamma\}$. U is allowed to:

- require further details on another feature of the same explanation remaining in the detailisation store ($\Gamma' \in DET$) (locution *what-details*($\hat{y}$, E[,y'], e, Γ') where $\Gamma' \neq \Gamma$;
- require a CF explanation for an arbitrary CF class $y' \in CFS$ if a factual explanation is being processed (locution *why-not-explain*($\hat{y}$, E, y')) or require a CF explanation for another CF class if a CF explanation is being processed (locution *why-not-explain*($\hat{y}$, E, y''));
- specify how a feature Ψ of the explanation e is defined (locution *what-is*($\hat{y}$, E[,y'], e, Ψ) where $\Psi \in CLAR$;
- disagree with the factual explanation and require an alternative to it, provided that it is the only explanation that E contains (locution *why-alternative*($\hat{y}$, E, e_f)) iff the currently processed explanation is factual (i.e., $e = e^h_f$ and $\nexists\, e_{cf} \in E$);
- require another CF explanation for the same CF class (locution *why-not-alternative*($\hat{y}$, E, y', e_{cf}) iff the explanation currently processed is counterfactual (i.e., $e = e^h_{cf}$);
- accept prediction $\hat{y}$ (locution *accept-u*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject-u*($\hat{y}$, E)).

(g) *no-elaborate*($\hat{y}$, E[,y'], e, Γ) where $e = e^h_f | e^h_{cf}$: S is unable to provide details on feature $\Gamma \in e$ because either all the available details have already been provided or the details required are not found in the knowledge base of the system. U can respond using one of the following locutions:

- require further details on another feature of the same explanation remaining in the detailisation store ($\Gamma' \in DET$) (locution *what-details*($\hat{y}$, E[,y'], e, Γ') where $\Gamma' \neq \Gamma$;
- require a CF explanation for an arbitrary CF class $y' \in CFS$ if a factual explanation is being processed (locution *why-not-explain*($\hat{y}$, E, y')) or require a CF explanation for another CF class if a CF explanation is being processed (locution *why-not-explain*($\hat{y}$, E, y''));
- specify how a feature Ψ of the explanation e is defined (locution *what-is*($\hat{y}$, E[,y'], e, Ψ) where $\Psi \in CLAR$;
- disagree with the factual explanation and require an alternative to it, provided that it is the only explanation that E contains (locution *why-alternative*($\hat{y}$, E, e_f)) iff the currently processed explanation is factual (i.e., $e = e^h_f$ and $\nexists\, e_{cf} \in E$);
- require another CF explanation for the same CF class (locution *why-not-alternative*($\hat{y}$, E, y', e_{cf}) iff the explanation currently processed is counterfactual (i.e., $e = e^h_{cf}$);
- accept prediction $\hat{y}$ (locution *accept-u*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject-u*($\hat{y}$, E)).

(h) *clarify*($\hat{y}$, E[,y'], e, Ψ, υ): S provides a definition υ for feature Ψ of the currently processed explanation e. The explanation e can be of any modality: factual or CF, high-level or low-level. The corresponding feature is then eliminated from the clarification store: $CLAR = CLAR \setminus \{\Psi\}$. U uses one of the following locutions to respond:

- require details on a feature $\Gamma \in e$ remaining in the detailisation store ($\Gamma \in DET$) (locution *what-details*($\hat{y}$, $E[,y']$, e, Γ));
- require a CF explanation for an arbitrary CF class $y' \in CFS$ if a factual explanation is being processed (locution *why-not-explain*($\hat{y}$, E, y')) or require a CF explanation for another CF class if a CF explanation is being processed (locution *why-not-explain*($\hat{y}$, E, y''));
- specify how another feature Ψ of the explanation e is defined (locution *what-is*($\hat{y}$, $E[,y']$, e, Ψ') where $\Psi' \in CLAR$;
- require an alternative factual explanation (locution *why-alternative*($\hat{y}$, E, e)) iff $e = e^h_f | e^l_f$;
- require an alternative CF explanation for the same CF class (locution *why-not-alternative* ($\hat{y}$, E, e)) iff $e = e^h_{cf} | e^l_{cf}$;
- accept prediction $\hat{y}$ (locution *accept-u*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject-u*($\hat{y}$, E)).

(i) *no-clarify*($\hat{y}$, $E[,y']$, e, Ψ): S is unable to provide a definition of the feature $\Psi \in e$ because the feature is specified incorrectly, or the definition is not found in the system's knowledge base, or the definition has already been provided. U is allowed to respond using one of the following locutions:

- require details on a feature $\Gamma \in e$ remaining in the detailisation store ($\Gamma \in DET$) (locution *what-details*($\hat{y}$, $E[,y']$, e, Γ));
- require a CF explanation for an arbitrary CF class $y' \in CFS$ if a factual explanation is being processed (locution *why-not-explain*($\hat{y}$, E, y')) or require a CF explanation for another CF class if a CF explanation is being processed (locution *why-not-explain*($\hat{y}$, E, y''));
- specify how another feature Ψ of the explanation e is defined (locution *what-is*($\hat{y}$, $E[,y']$, e, Ψ') where $\Psi' \in CLAR$;
- require an alternative factual explanation (locution *why-alternative*($\hat{y}$, E, e)) iff $e = e^h_f | e^l_f$;
- require an alternative CF explanation for the same CF class (locution *why-not-alternative*($\hat{y}$, E, e)) iff $e = e^h_{cf} | e^l_{cf}$;
- accept prediction $\hat{y}$ (locution *accept-u*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject-u*($\hat{y}$, E)).

(j) *alter-f*($\hat{y}$, E, e_f, e'_f): S provides U with a factual explanation e'_f alternative to e_f. The previous (possibly also alternative to the original) piece of factual explanation is removed from the explanation store. The newly generated alternative factual explanation is added to the knowledge store $K = K \cup e'_f$ and the explanation store $E = E \cup e'_f$. The detailisation and clarification stores are populated with the features of the newly generated alternative factual explanation. U responds using one of the following locutions:

- require details on a feature Γ of the offered alternative factual explanation e'_f (locution *what-details($\hat{y}$, E, e'_f, Γ)*);
- require a CF explanation for some CF class y' (locution *why-not-explain*($\hat{y}$, E, y'));
- specify how a feature Ψ of the offered alternative factual explanation e'_f is defined (locution *what-is*($\hat{y}$, E, e'_f, Ψ));
- require another alternative factual explanation (locution *why-alternative*($\hat{y}$, E, e'_f));
- accept prediction $\hat{y}$ (locution *accept-u*($\hat{y}$, E));

- reject prediction $\hat{y}$ (locution *reject-u*($\hat{y}$, E)).

(k) *no-alter-f*($\hat{y}$, E, e_f): S is unable to offer a factual explanation alternative to e_f because there exists no explanation alternative to the factual or all the alternatives have already been offered. U responds using one of the following locutions:

- require details on a feature Γ of the latest offered (either original or alternative) factual explanation (locution *what-details*($\hat{y}$, E, e_f, Γ));
- require a CF explanation for some CF class y' (locution *why-not-explain*($\hat{y}$, E, y'));
- specify how a feature Ψ of the latest offered (either original or alternative) factual explanation is defined (locution *what-is*($\hat{y}$, E, e_f, Ψ));
- accept prediction $\hat{y}$ (locution *accept-u*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject-u*($\hat{y}$, E)).

(l) *alter-cf*($\hat{y}$, E, y', e_{cf}, e'_{cf}): S provides U with a CF explanation e'_{cf} alternative to e_{cf}. The previous (possibly also alternative to the original) piece of CF explanation is removed from the explanation store. The newly generated alternative CF explanation is added to the knowledge store $K = K \cup e'_{cf}$ and the explanation store $E = E \cup e'_{cf}$. The detailisation and clarification stores are populated with the features of the newly generated alternative CF explanation. U is allowed to respond using one of the following locutions:

- require details on a feature Γ of the offered alternative CF explanation e'_{cf} (locution *what-details*($\hat{y}$, E, y', e'_{cf}, Γ));
- require a CF explanation for some other CF class y'' (locution *why-not-explain*($\hat{y}$, E, y'') iff $y'' \neq y'$);
- specify how a feature Ψ of the offered alternative CF explanation e'_{cf} is defined (locution *what-is*($\hat{y}$, E, y', e'_{cf}, Ψ));
- accept prediction $\hat{y}$ (locution *accept-u*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject-u*($\hat{y}$, E)).

(m) *no-alter-cf*($\hat{y}$, E, y', e_{cf}): S is unable to offer a CF explanation alternative to e_{cf}. U is allowed to make one of the following actions:

- require details on a feature Γ of the latest offered alternative CF explanation e'_{cf} (locution *what-details*($\hat{y}$, E, y', e'_{cf}, Γ));
- require a CF explanation for some other CF class y'' (locution *why-not-explain*($\hat{y}$, E, y'') iff $y'' \neq y'$);
- specify how a feature Ψ of the latest offered alternative CF explanation e'_{cf} is defined (locution *what-is*($\hat{y}$, E, y', e'_{cf}, Ψ));
- accept prediction $\hat{y}$ (locution *accept-u*($\hat{y}$, E));
- reject prediction $\hat{y}$ (locution *reject-u*($\hat{y}$, E)).

(4) **Termination states.** The dialogue ends when the system generates a concluding locution (either *accept-s*($\hat{y}$, E) or *reject-s*($\hat{y}$, E)) immediately after the end user accepts or rejects the system's prediction, respectively.

An explanatory dialogue is governed in accordance with the aforementioned rules. Table 13 summarises and exemplifies the dialogue protocol outlined above.

Table 13

The set of allowed moves for the participants of an explanatory dialogue game

Locution	Interpretation	Utterance template	Possible response(-s)
		System (S):	
claim($\hat{y}$, E)	S claims prediction $\hat{y}$ to be true	The test instance is of class $\hat{y}$.	• *why-explain*($\hat{y}$, E) • *accept-u*($\hat{y}$, E) • *reject-u*($\hat{y}$, E)
explain-f($\hat{y}$, E, e_f)	S factually explains $\hat{y}$ with e_f	The test instance is of class $\hat{y}$ because ⟨feature$_1$⟩ is ⟨term$_1$⟩ [and ⟨feature$_2$⟩ is ⟨term$_2$⟩,. . .].	• *why-not-explain*($\hat{y}$, E, y') • *what-details*($\hat{y}$, E, e_f, Γ) • *what-is*($\hat{y}$, E, e_f, Ψ) • *why-alternative*($\hat{y}$, E, e_f) • *accept-u*($\hat{y}$, E) • *reject-u*($\hat{y}$, E)
no-explain-f($\hat{y}$, E)	S is unable to factually explain $\hat{y}$	Sorry, I don't have a factual explanation for you.	• *why-not-explain*($\hat{y}$, E, y') • *accept-u*($\hat{y}$, E) • *reject-u*($\hat{y}$, E)
explain-cf($\hat{y}$, E, y', e_{cf})	S counterfactually explains why $\hat{y}$ and not y' with e_{cf}	The test instance would be of class y' if ⟨feature$_1$⟩ were ⟨term$_2$⟩ [and ⟨feature$_2$⟩ were ⟨term$_1$⟩,. . .].	• *what-details*($\hat{y}$, E, y', e_{cf}, Γ) • *why-not-explain*($\hat{y}$, E, y'') iff $y'' \neq y'$ • *what-is*($\hat{y}$, E, y', e_{cf}, Ψ) • *why-not-alternative*($\hat{y}$, E, y', e_{cf}) • *accept-u*($\hat{y}$, E) • *reject-u*($\hat{y}$, E)
no-explain-cf($\hat{y}$, E, y')	S is unable to counterfactually explain why $\hat{y}$ and not y'	Sorry, I don't have a CF explanation for you.	• *why-not-explain*($\hat{y}$, E, y'') iff $y'' \neq y'$ • *why-alternative*($\hat{y}$, E, e_f) iff $\nexists e_{cf} \in E$ • *accept-u*($\hat{y}$, E) • *reject-u*($\hat{y}$, E)
elaborate($\hat{y}$, E[,y'], e, Γ, θ) where $e = e^h_f \| e^h_{cf}$	S provides requested details θ on feature Γ of a high-level explanation e	I define Γ to be ⟨term⟩, as it ranges from ⟨min$_{\text{term}}$⟩ to ⟨max$_{\text{term}}$⟩.	• *what-details*($\hat{y}$, E[,y'], e, Γ') iff $\Gamma' \neq \Gamma$ • *why-not-explain*($\hat{y}$, E, y') iff $e = e^h_f$ or *why-not-explain*($\hat{y}$, E, y'') iff $e = e^h_{cf}$ and $y'' \neq y'$ • *what-is*($\hat{y}$, E[,y'], e, Ψ) • *why-alternative*($\hat{y}$, E,e) iff $e = e^h_f$ • *why-not-alternative*($\hat{y}$, E, y', e) iff $e = e^h_{cf}$ • *accept-u*($\hat{y}$, E) • *reject-u*($\hat{y}$, E)
no-elaborate($\hat{y}$, E[,y'], e, Γ) where $e = e^h_f \| e^h_{cf}$	S is unable to provide details on feature Γ of a high-level explanation e (e.g., because all the required details have already been provided)	Sorry, I don't have details on Γ.	• *what-details*($\hat{y}$, E[,y'], e, Γ') iff $\Gamma' \neq \Gamma$ • *why-not-explain*($\hat{y}$, E, y') iff $e = e^h_f$ or *why-not-explain*($\hat{y}$, E, y'') iff $e = e^h_{cf}$ and $y'' \neq y'$ • *what-is*($\hat{y}$, E[,y'], e, Ψ) • *why-alternative*($\hat{y}$, E,e) if $e = e^h_f$ • *why-not-alternative*($\hat{y}$, E,e) if $e = e^h_{cf}$ • *accept-u*($\hat{y}$, E) • *reject-u*($\hat{y}$, E)

Table 13

(Continued)

Locution	Interpretation	Utterance template	Possible response(-s)
clarify$(\hat{y}, E[,y'], e, \Psi, \upsilon)$ where $e = e^h_f \vert e^h_{cf} \vert e^l_f \vert e^l_{cf}$	S provides definition υ for feature Ψ making part of explanation e	Ψ is υ.	• *what-details*$(\hat{y}, E[,y'], e, \Gamma)$ • *why-not-explain*$(\hat{y}, E, y')$ iff $e = e^h_f \vert e^l_f$ or *why-not-explain*$(\hat{y}, E, y'')$ iff $e = e^h_{cf} \vert e^l_{cf}$ and $y'' \neq y'$ • *what-is*$(\hat{y}, E[,y'], e, \Psi')$ iff $\Psi' \neq \Psi$ • *why-alternative*$(\hat{y}, E,e)$ iff $e = e^h_f \vert e^l_f$ • *why-not-alternative*$(\hat{y}, E, y', e)$ iff $e = e^h_{cf} \vert e^l_{cf}$ • *accept-u*$(\hat{y}, E)$ • *reject-u*$(\hat{y}, E)$
no-clarify$(\hat{y}, E[,y'], e, \Psi)$ where $e = e^h_f \vert e^h_{cf} \vert e^l_f \vert e^l_{cf}$	S is unable to provide a definition for feature $\Psi \in e$ (e.g., because it is absent in the knowledge base or the inquired term is not found in the set of features)	Sorry, I cannot clarify what Ψ is.	• *what-details*$(\hat{y}, E[,y'], e, \Gamma)$ • *why-not-explain*$(\hat{y}, E, y')$ iff $e = e^h_f \vert e^l_f$ or *why-not-explain*$(\hat{y}, E, y'')$ iff $e = e^h_{cf} \vert e^l_{cf}$ and $y'' \neq y'$ • *what-is*$(\hat{y}, E[,y'], e, \Psi')$ where $\Psi' \neq \Psi$ • *why-alternative*$(\hat{y}, E, e)$ iff $e = e^h_f \vert e^l_f$ • *why-not-alternative*$(\hat{y}, E, e)$ iff $e = e^h_{cf} \vert e^l_{cf}$ • *accept-u*$(\hat{y}, E)$ • *reject-u*$(\hat{y}, E)$
alter-f$(\hat{y}, E, e_f, e'_f)$	S provides a factual explanation e'_f alternative to e_f	The test instance is of class $\hat{y}$ because ⟨feature$_1$⟩ is ⟨term$_3$⟩ [and ⟨feature$_2$⟩ is ⟨term$_4$⟩,...].	• *what-details*$(\hat{y}, E, e'_f, \Gamma)$ • *why-not-explain*$(\hat{y}, E, y')$ • *what-is*$(\hat{y}, E, e'_f, \Psi)$ • *why-alternative*$(\hat{y}, E, e'_f)$ • *accept-u*$(\hat{y}, E)$ • *reject-u*$(\hat{y}, E)$
no-alter-f$(\hat{y}, E, e_f)$	S is unable to provide a factual explanation alternative to e_f	Sorry, I don't have an alternative factual explanation for you.	• *what-details*$(\hat{y}, E, e_f, \Gamma)$ • *why-not-explain*$(\hat{y}, E, y')$ • *what-is*$(\hat{y}, E, e_f, \Psi)$ • *accept-u*$(\hat{y}, E)$ • *reject-u*$(\hat{y}, E)$
alter-cf$(\hat{y}, E, y', e_{cf}, e'_{cf})$	S provides a CF explanation e'_{cf} alternative to e_{cf} for some CF class y'	The test instance would be of class y' if ⟨feature$_1$⟩ were ⟨term$_4$⟩ [and ⟨feature$_2$⟩ were ⟨term$_3$⟩,...].	• *what-details*$(\hat{y}, E, y', e'_{cf}, \Gamma)$ • *why-not-explain*$(\hat{y}, E, y'')$ iff $y'' \neq y'$ • *what-is*$(\hat{y}, E, y', e'_{cf}, \Psi)$ • *accept-u*$(\hat{y}, E)$ • *reject-u*$(\hat{y}, E)$
no-alter-cf$(\hat{y}, E, y', e_{cf})$	S is unable to provide a CF explanation alternative to e_{cf}	Sorry, I don't have an alternative CF explanation for you.	• *what-details*$(\hat{y}, E, y', e_{cf}, \Gamma)$ • *why-not-explain*$(\hat{y}, E, y'')$ iff $y'' \neq y'$ • *what-is*$(\hat{y}, E, y', e_{cf}, \Psi)$ • *accept-u*$(\hat{y}, E)$ • *reject-u*$(\hat{y}, E)$
accept-s$(\hat{y}, E)$	S utters the farewell locution, as the user accepted the system's claim	Ok, thank you for your trust in me. Bye!	–

Table 13
(Continued)

Locution	Interpretation	Utterance template	Possible response(-s)
reject-s$(\hat{y}, E)$	S utters the farewell locution, as the user rejected the system's claim	Sorry about my poor explanatory capacities. Bye!	–
User (U):			
why-explain$(\hat{y}, E)$	U requests to factually explain prediction $\hat{y}$	Could you explain why you think so?	• *explain-f*$(\hat{y}, E, e_f)$ • *no-explain-f*$(\hat{y}, E)$
why-not-explain$(\hat{y}, E, y')$	U requests to counterfactually explain why $\hat{y}$ and not y'	But why not y'?	• *explain-cf*$(\hat{y}, E, y', e_{cf})$ • *no-explain-cf*$(\hat{y}, E, y')$
what-details$(\hat{y}, E[,y'], e, \Gamma)$ where $e = e^h_f \| e^h_{cf}$	U requests details on a specific feature Γ of a (factual or counterfactual) high-level explanation e $(\Gamma \in e)$	Could you provide me with details on Γ?	• *elaborate*$(\hat{y}, E[,y'], e, \Gamma, \theta)$ • *no-elaborate*$(\hat{y}, E[,y'], e, \Gamma)$
what-is$(\hat{y}, E[,y'], e, \Psi)$ where $e = e^h_f \| e^h_{cf} \| e^l_f \| e^l_{cf}$	U requests a definition for a specific feature Ψ making part of (factual or counterfactual, high- or low-level) explanation e $(\Psi \in e)$	What do you mean by Ψ?	• *clarify*$(\hat{y}, E[,y'], e, \Psi, \upsilon)$ • *no-clarify*$(\hat{y}, E[,y'], e, \Psi)$
why-alternative$(\hat{y}, E, e_f)$	U disagrees with the offered factual explanation e_f and requires an alternative factual explanation	I do not agree (or, I am not satisfied/convinced) with your (factual) explanation. Could you offer me another one?	• *alter-f*$(\hat{y}, E, e_f, e'_f)$ • *no-alter-f*$(\hat{y}, E, e_f)$
why-not-alternative$(\hat{y}, E, y', e_{cf})$	U disagrees with the offered CF explanation e_{cf} and requires an alternative CF explanation for some CF class y'	I do not agree (or, I am not satisfied/convinced) with your (CF) explanation. Could you offer me another one?	• *alter-cf* $(\hat{y}, E, y', e_{cf}, e'_{cf})$ • *no-alter-cf*$(\hat{y}, E, y', e_{cf})$
accept-u$(\hat{y}, E)$	U accepts all pieces of explanation contained in explanation store E and therefore definitely accepts prediction $\hat{y}$	Ok, I trust (or agree/am satisfied/am convinced) with your prediction.	• *accept-s* $(\hat{y}, E)$
reject-u$(\hat{y}, E)$	U rejects (a) piece(-s) of explanation contained in explanation store E and therefore definitely rejects prediction $\hat{y}$	I don't really trust (or am not satisfied/am not convinced/agree with) your prediction and you won't be able to convince me.	• *reject-s* $(\hat{y}, E)$

Appendix C. Explanatory dialogue grammar productions

Recall that an EDG can be formalised by means of a context-free grammar $G = \langle N, T, R, S\rangle$ (see Section 3.3 for details). Outlined below is the set of the generalised dataset-independent production rules (R):

(1) DIALOGUE → CLAIM EXPLANATION TERMINATION
(2) CLAIM → The test instance is of class CLASS.
(3) EXPLANATION → FACT-EXPLANATION (CF-EXPLANATION)* | ϵ

(4) TERMINATION → ACCEPT-U ACCEPT-S | REJECT-U REJECT-S

(5) ACCEPT-U → Okay, I trust your prediction.

(6) ACCEPT-S → Thank you for your trust in me. Bye!

(7) REJECT-U → I don't trust your prediction and you won't convince me.

(8) REJECT-S → Sorry for my poor explanatory capacities. Bye!

(9) FACT-EXPLANATION → WHY-EXPLAIN [EXPLAIN-F | NO-EXPLAIN-F]

(10) WHY-EXPLAIN → Could you explain why you think so?

(11) EXPLAIN-F → SURE INTRO-F [B|b]ecause F-EXPL (and F-EXPL)*. [DETAILISATION | CLARIFICATION | ALTERNATIVE-F | ϵ]

(12) INTRO-F → It is of class CLASS | ϵ

(13) F-EXPL → FEATURE is VALUE

(14) NO-EXPLAIN-F → Sorry, I don't have a factual explanation for you.

(15) SURE → Sure! | ϵ

(16) CF-EXPLANATION → WHY-NOT-EXPLAIN [EXPLAIN-CF | NO-EXPLAIN-CF]

(17) WHY-NOT-EXPLAIN → But why is it not of class CLASS?

(18) EXPLAIN-CF → SURE It would be of class CLASS if CF-EXPL (and CF-EXPL)*. [DETAILISATION | CLARIFICATION | ALTERNATIVE-CF | ϵ]

(19) CF-EXPL → FEATURE were VALUE

(20) NO-EXPLAIN-CF → I don't have an explanation for why it is not of class CLASS.

(21) DETAILISATION → WHAT-DETAILS [ELABORATE | NO-ELABORATE] [DETAILISATION | CLARIFICATION | ALTERNATIVE-F | ALTERNATIVE-CF | ϵ]

(22) WHAT-DETAILS → Could you FURTHER specify how TERM FEATURE is defined?

(23) ELABORATE → Sure! FEATURE is defined to be TERM because it lies in the range RANGE.

(24) NO-ELABORATE → Sorry, I don't any FURTHER details on the requested term. [CLARIFICATION | ALTERNATIVE-F | ALTERNATIVE-CF | ϵ]

(25) FURTHER → further | ϵ

(26) CLARIFICATION → WHAT-IS [CLARIFY | NO-CLARIFY]

(27) WHAT-IS → What do you mean by FEATURE?

(28) CLARIFY → FEATURE is DEFINITION. [DETAILISATION | CLARIFICATION | ALTERNATIVE-F | ALTERNATIVE-CF | ϵ]

(29) NO-CLARIFY → Sorry, I cannot clarify the term FEATURE. [DETAILISATION | ALTERNATIVE-F | ALTERNATIVE-CF | ϵ]

(30) ALTERNATIVE-F → WHY-ALTERNATIVE [EXPLAIN-F | NO-EXPLAIN-F]

(31) ALTERNATIVE-CF → WHY-NOT-ALTERNATIVE [EXPLAIN-CF | NO-EXPLAIN-CF]

(32) WHY-ALTERNATIVE → REQ-ALTERNATIVE-BEG EXPL-TYPE-F REQ-ALTERNATIVE-END

(33) WHY-NOT-ALTERNATIVE → REQ-ALTERNATIVE-BEG EXPL-TYPE-CF REQ-ALTERNATIVE-END

(34) REQ-ALTERNATIVE-BEG → I am not quite satisfied with your

(35) REQ-ALTERNATIVE-END → explanation. Could you offer me another one?

(36) EXPL-TYPE-F → factual | ϵ

(37) EXPL-TYPE-CF → counterfactual | ϵ

Table 14

Aggregated self-reported demographic user data for all the use cases

18–25	14 (26.92%)
26–35	24 (46.16%)
36–45	10 (19.23%)
46–55	3 (5.77%)
56–65	1 (1.92%)

(a) Age

Male	28 (53.85%)
Female	24 (46.15%)

(b) Gender

Doctorate (Ph.D)	20 (38.46%)
Master's (M.A./M.Sc.)	25 (48.08%)
Bachelor's (B.A./B.Sc.)	6 (11.54%)
Prefer not to say	1 (1.92%)

(c) Education

Native speaker	20 (38.46%)
Proficient (C2)	17 (32.69%)
Advanced (C1)	10 (19.23%)
Upper intermediate (B2)	5 (9.62%)

(d) English proficiency

Student	30 (57.69%)
Non-student	22 (42.31%)

(e) Occupation

Appendix D. Further details on human evaluation use cases

This appendix outlines the quantitative results of the human evaluation study. First, we report the demographic data of all the study participants who decided to disclose it. Recall that 60 people participated in the evaluation of the proposed dialogue game. All in all, 52 out of all the 60 (86.67%) study participants disclosed their demographic data. In summary, the overall collection of dialogue transcripts is gender-balanced. In addition, the participants who reported their education level had at least a Bachelor degree. Further, all the subjects had at least the B2 level of English proficiency. Table 14 summarises all the self-reported demographic data collected from all the participants.

Subsequently, we provide the reader with the demographic data of the study participants and the process models grouped by use case. Thus, Section D.1 presents the results for the collection of the basketball dataset-related data. Section D.2 displays the results for the beer style classification explanatory dialogues. Section D.3 highlights the results collected for the thyroid disease classification scenario.

D.1. Basketball player position classification

Fourteen (23.33%) of the 60 collected dialogue transcripts relate to the basketball player position dataset. 12 out of the 14 (85.71%) participants who selected the basketball player position scenario attached their demographic data. In summary, 7 (58.33%) participants who chose this scenario and disclosed the demographic data were males, 5 (41.67%) were females. In addition, all the participants who disclosed their demographic data reported to have at least a Bachelor degree and the C1 level of English proficiency. Table 15 summarises all the self-reported demographic data collected from the participants who selected the basketball player position scenario.

Fig. 10 depicts the process model based on the main building blocks (i.e., claim, explanation, and termination) within the collected explanatory dialogues (see Rule 1 of the EDG, Appendix C, for reference). Thus, 12 out of 14 (85.71%) participants required (at least, factual) explanation(-s) for the given prediction. Further, 11 out of 12 (91.67%) such participants accepted the system's prediction after processing the explanation offered. On the contrary, only one out of the 12 (8.33%) participants rejected

Table 15

Self-reported demographic data of the users who interaction with the classifier trained on the basketball player position dataset

18–25	5 (41.67%)
26–35	6 (50.00%)
36–45	1 (8.33%)

(a) Age

Male	7 (58.33%)
Female	5 (41.67%)

(b) Gender

Doctorate (Ph.D)	2 (16.67%)
Master's (M.A./M.Sc.)	8 (66.66%)
Bachelor's (B.A./B.Sc.)	2 (16.67%)

(c) Education

Native speaker	8 (66.66%)
Proficient (C2)	2 (16.67%)
Advanced (C1)	2 (16.67%)

(d) English proficiency

Student	10 (83.33%)
Non-student	2 (16.67%)

(e) Occupation

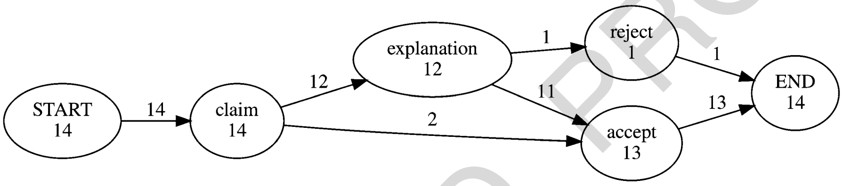


Fig. 10. The process model of the collected basketball player position classification explanatory dialogues based on the main EDG building blocks.

the claim after the explanation was presented. Alternatively, 2 out of 14 (14.29%) participants did not require any explanation for the system's claim. Both of them eventually accepted the system's claim.

As for all the 12 participants who required explanation for the system's claim, 67 explanation-related requests (i.e., those for factual or (alternative) CF explanation, detailisation, and clarification) have been registered. Figure 11 depicts the locution-level process model for the collected explanatory dialogues. Thus, 12 out of the 67 requests (17.91%) were those for factual explanation. In addition, 18 out of the 67 (26.87%) explanation-related requests were those for CF explanation. Further, alternative CF explanations were requested 9 times (13.43%). In addition, 15 out of 67 (22.39%) requests addressed numerical details for the offered linguistic terms whereas only 13 out of 67 requests (19.40%) were clarification requests.

The factual explanation seemed clear and explanatory enough to a half of the participants. Thus, 6 out of 12 (50.00%) study participants who requested a factual explanation did not inquire any further details or clarifications before requesting their first CF explanation. As for the other 6 participants, detailisation requests have been more frequently registered for the factual explanation offered: 7 out of 15 times (46.67%) – 5 (33.33%) times immediately after the factual explanation was offered, 2 (13.33%) times subsequently to the first detailisation request related to the factual explanation. Also, clarification requests are found when processing 6 out of 13 factual explanations (46.15%): once – immediately after it was generated, five times – following detailisation requests. On the other hand, 5 of the 12 (41.67%) participants who requested explanation in the first place were interested in obtaining CF explanations (recall that the 5 participants submitted 18 CF explanation requests altogether). Further, numerical intervals specifying linguistic terms of the corresponding CF explanations were inquired 8 out of the overall

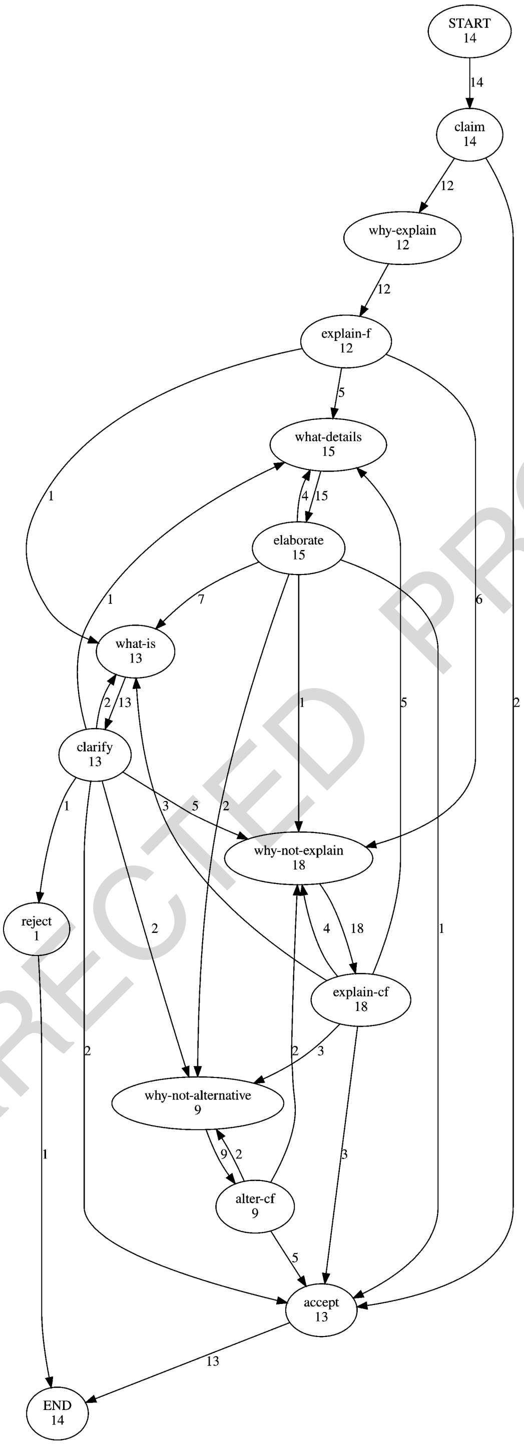


Fig. 11. The full process model of the basketball player position classification explanatory dialogues. For illustrative purposes, pairs of termination nodes, i.e. *{accept-u, accept-s}* and *{reject-u and reject-s}*, are merged into *accept* and *reject*, respectively.

Table 16

Number of times the CF explanations (sorted by rank) were requested by participants (basketball player position classification)

	CF class			
CF rank	*Shooting guard*	*Small forward*	*Power forward*	*Center*
#1	5	6	5	2
#2	3	2	1	1
#3	–	2	–	–

15 times (53.33%), 5 of them submitted as soon as the corresponding CF was offered. In addition, 7 (53.85%) out of all the 13 clarification requests were registered when processing CF explanations, 3 of them – submitted immediately upon receiving the corresponding CF explanation.

Importantly, the locution-level process model (see Fig. 11 for details) shows us the responses to which requests were the most decisive for the study participants to make their final decisions. Recall that 11 out of the 12 participants who required explanation accepted the system's claim. Thus, 3 (27.27%) of the 11 participants accepted the system's claim immediately after a CF explanation had been presented. Further, 5 out of 11 participants (45.46%) found themselves in the position to make the final decision after an alternative CF explanation was displayed. For 2 out of the 11 (18.18%) participants who accepted the claim, the response to their clarification requests triggered their final decision. In addition, 1 out of 11 (9.09%) such participants accepted the system's claim after (s-)he was provided with the details on the inquired feature. Recall that only one subject rejected the claim after having been provided with the explanation. In this case, a response to a clarification request motivated that decision.

Recall that 18 CF explanation requests were registered in the basketball player position classification dialogues. All such CF explanations are those deemed most relevant to the test instance by the system. However, there have as well been registered 9 requests for alternative CF explanations, 7 of them being an alternative to the best ranked CFs.

Table 16 presents numbers of CF explanation requests for each CF class (row "#1") as well those related to second and third best-ranked alternative CF explanations (rows "#2" and "#3", accordingly). Thus, in 7 out of the 18 (38.89%) cases where (the best-ranked) CF explanations were offered, the users did not find them satisfactory. Further, when exposed to 2 out of the 7 (28.57%) second-best ranked CF explanations were offered, the participants required third-best ranked CFs. In particular, both such cases occur when CF explanations were asked for the CF class "Small forward". Importantly, 5 out of all the 9 (55.56%) alternative CF explanations turned out to be crucially decisive from the end user's point of view (i.e., they led to making an immediate decision – in this case, acceptance of the system's claim).

D.2. Beer style classification

Thirty-seven (61.67%) of all the collected dialogue transcripts relate to the beer style classification scenario. All in all, 31 out of the 37 (83.78%) participants who played the beer scenario disclosed their demographic data. In summary, 17 (54.84%) of all the participants who chose this scenario and left their demographic data were males, 14 (45.16%) – females. In addition, all the participants who reported their education level had at least a Bachelor degree and the B2 level of English proficiency. Table 17 summarises all the self-reported demographic data collected from the participants who selected the beer style dataset as the basis of the dialogue game.

Fig. 12 illustrates the process model corresponding to the three main building blocks of the proposed dialogue game. Thus, 36 out of 37 (97.30%) participants required (at least, factual) explanation for the given prediction. Further, 33 out of the 36 (91.67%) participants accepted the system's prediction after

Table 17

Self-reported demographic user data (the beer style classification dataset)

18–25	6 (19.35%)
26–35	14 (45.16%)
36–45	8 (25.81%)
46–55	2 (6.45%)
56–65	1 (3.23%)

(a) Age

Male	17 (54.84%)
Female	14 (45.16%)

(b) Gender

Doctorate (Ph.D)	16 (51.61%)
Master's (M.A./M.Sc.)	12 (38.71%)
Bachelor's (B.A./B.Sc.)	2 (6.45%)
Prefer not to say	1 (3.23%)

(c) Education

Native speaker	9 (29.03%)
Proficient (C2)	11 (35.48%)
Advanced (C1)	8 (25.81%)
Upper intermediate (B2)	3 (9.68%)

(d) English proficiency

Student	14 (45.16%)
Non-student	17 (54.84%)

(e) Occupation

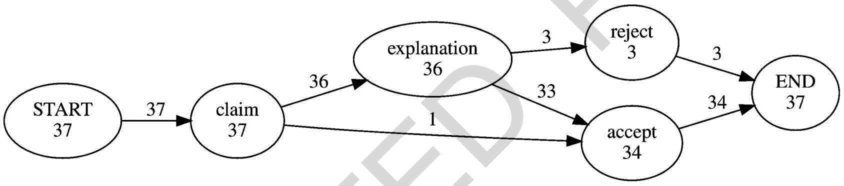


Fig. 12. The process model of the collected beer style classification explanatory dialogues based on the main EDG building blocks.

processing the explanation offered whereas only 3 (8.33%) rejected the system's prediction. In addition, only 1 out of 37 (2.70%) participants did not require any explanation for the system's claim. Eventually, that participant accepted the system's claim.

Figure 13 depicts the locution-level process model for the collected explanatory dialogues. Thus, 235 explanation-related requests (all those covered by the EXPLANATION non-terminal in EDG) were registered from the 36 participants who required explanation for the system's claim. More precisely, 36 out of the 235 (15.32%) requests were those for factual explanation. In addition, 50 out of the 235 (21.28%) explanation-related requests were those for CF explanation. Further, alternative CF explanations were requested 25 times (10.64%). Moreover, 78 out of 235 (33.19%) requests addressed numerical details for the offered linguistic terms whereas 46 out of 235 (19.57%) requests were clarification requests.

It is worth noting that the factual explanation seemed rather unclear to most of the participants. Thus, 31 out of the 36 (86.11%) study participants who requested a factual explanation inquired either further details or clarifications before requesting their first CF explanation. Thus, 52 out of all the 78 detailisation requested registered were concerned with the factual explanation. In 24 (46.15%) cases, numerical intervals for specific features were requested as soon as the factual explanation was presented whereas the other 28 (53.85%) cases of detailisation requests were follow-ups to other (including detailisation) requests. Also, 32 out of 46 (69.57%) clarification requests were found when processing the factual explanation: 7 times (21.88%) – immediately after it was generated, 25 times (78.12%) – following

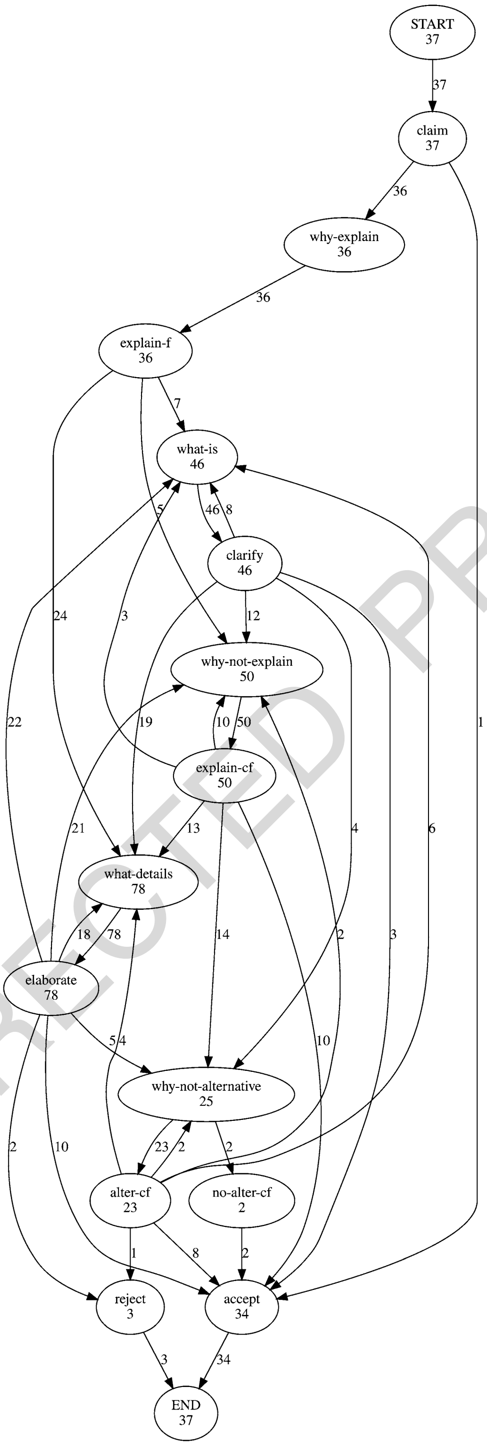


Fig. 13. The full process model of the collected beer style classification explanatory dialogues. For illustrative purposes, pairs of termination nodes, i.e. *{accept-u, accept-s}* and *{reject-u and reject-s}*, are merged into *accept* and *reject*, respectively.

Table 18

Number of times the CF explanations (sorted by rank) were requested by participants (beer style classification)

	CF class						
CF rank	*Lager*	*Pilsner*	*IPA*	*Barleywine*	*Stout*	*Porter*	*BSA*
# 1	12	9	10	5	5	3	6
# 2	7	6	3	1	3	2	1
# 3	1	1	–	–	–	–	–

detailisation or other clarification requests. On the other hand, 29 of the 36 (80.56%) participants who requested explanation in the first place were interested in obtaining CF explanations. Further, numerical intervals specifying linguistic terms of the corresponding CF explanations were inquired 26 out of the overall 78 times (33.33%), a half of them submitted as soon as the corresponding CF was offered. In addition, 14 (30.43%) out of all the 46 clarification requests were registered when processing CF explanations, 3 of them – immediately after the CF explanation was presented. Last but not least, out of the 25 alternative CF explanations requested, 14 (56.00%) were requested immediately after the questioned CF was presented whereas 11 (44.00%) – after detailisation or clarification requests concerning the CF explanation in question or subsequent to other alternative CF requests.

The locution-level process model (see Fig. 13 for details) also shows the responses to which requests were the most decisive for the study participants to make their final decisions in the beer style classification scenario. Thus, 10 out of the 33 participants (30.30%) who inquired an explanation and accepted the system's claim did so immediately after a CF explanation was presented. Further, 8 out of 33 participants (24.24%) found themselves in the position to make the final decision after an alternative CF explanation was displayed. In addition, 2 out of 33 participants (6.06%) accepted the system's prediction despite the fact that the system could not offer the participant an alternative CF upon request. In addition, for 3 out of the 33 (9.09%) participants who accepted the claim, the response to their clarification requests triggered their final decision. Finally, 10 out of 33 (30.30%) such participants accepted the system's claim after (s-)he was provided with the details on the inquired feature. On the other hand, 2 out of 3 (66.67%) participants rejected the claim when offered details on a specific explanation feature whereas 1 out of 3 (33.33%) did so upon receiving an alternative CF explanation.

Recall that 50 CF explanation requests were registered in the beer style classification scenario dialogues. The best ranked CFs (from the system's points of view) were questioned in 23 out of 50 (46.00%) cases, as the participants asked for an alternative CF explanation. Further, in 2 of the 23 (8.70%) such cases, third-best ranked CFs were requested. Table 18 shows the distribution of requests for CF explanation as well as their alternative variants by CF class. Remarkably, 8 out of the 25 (32.00%) alternative CFs turned out to be decisive (led to immediate acceptance of the system's prediction) whereas 1 alternative CF (4.00%) motivated immediate rejection of the system's claim. Finally, 2 out of all the 33 (6.06%) positive decisions made by the participants who requested an explanation were made after the system did not manage to offer an alternative CF explanation.

D.3. *Thyroid diagnosis classification*

Nine (15.00%) of all the 60 collected dialogue transcripts relate to the thyroid disease dataset. In summary, 4 (44.44%) participants who chose this scenario were males, 5 (55.56%) were females. Similarly to the other classification scenarios, all the participants reported to, at least, have a Bachelor degree and the B2 level of English proficiency. Table 19 summarises all the self-reported demographic data collected from the participants who selected the thyroid disease classification scenario.

Table 19

Self-reported demographic data of the users who interacted with the classifier trained on the thyroid disease dataset

18–25	3 (33.33%)
26–35	4 (44.45%)
36–45	1 (11.11%)
46–55	1 (11.11%)

(a) Age

Male	4 (44.44%)
Female	5 (55.56%)

(b) Gender

Doctorate (Ph.D)	2 (22.22%)
Master's (M.A./M.Sc.)	5 (55.56%)
Bachelor's (B.A./B.Sc.)	2 (22.22%)

(c) Education

Native speaker	3 (33.33%)
Proficient (C2)	4 (44.45%)
Upper intermediate (B2)	2 (22.22%)

(d) English proficiency

Student	6 (66.67%)
Non-student	3 (33.33%)

(e) Occupation

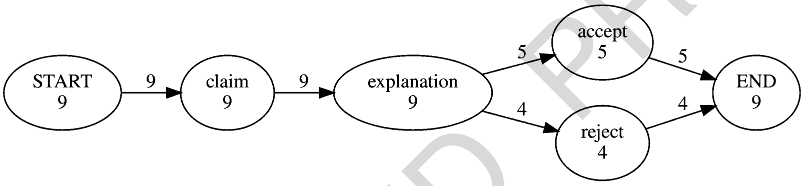


Fig. 14. The process model of the collected thyroid diagnosis classification explanatory dialogues based on the main EDG building blocks.

Fig. 14 illustrates the process model related to the three main building blocks of the proposed model of explanatory dialogue. Thus, all the 9 out of 9 (100.00%) participants required (at least, factual) explanation for the given prediction. Eventually, 5 out of 9 (55.56%) participants accepted the system's claim. On the contrary, 4 out of 9 (44.44%) study participants rejected the system's claim.

As for all the 9 participants who required explanation for the system's claim, 29 explanation-related requests have been registered. Figure 15 depicts the locution-level process model for the corresponding collection of explanatory dialogues. Due to the design of the protocol, 9 out of the 29 requests (31.04%) were those for factual explanation. In addition, 8 out of the 29 (27.59%) explanation-related requests were those for CF explanation. Further, 3 alternative CFs were inquired (10.34% of the explanation-related requests). In addition, 6 out of the 29 (20.69%) requests addressed numerical details for the offered linguistic terms whereas only 3 out of 29 requests (10.34%) were clarification requests.

Out of the nine participants who required (factual) explanation for the system's claim, three (33.33%) requested details for one of the corresponding features that the factual explanation contained. In addition, one participant (11.11%) requested to clarify a term that the factual explanation contained. Besides, one person (11.11%) concluded the dialogue by accepting the system's claim immediately after the factual explanation was displayed whereas four (44.44%) study participants inquired a CF explanation right after processing the factual explanation. Most of the detailisation (4 out of 6, 66.67%) and clarification requests (2 out of 3, 66.67%) addressed the factual explanation. All but one detailisation requests were submitted to the system as soon as the factual explanation was processed whereas one detailisation request followed one of the previously sent detailisation requests. One of the clarification requests was

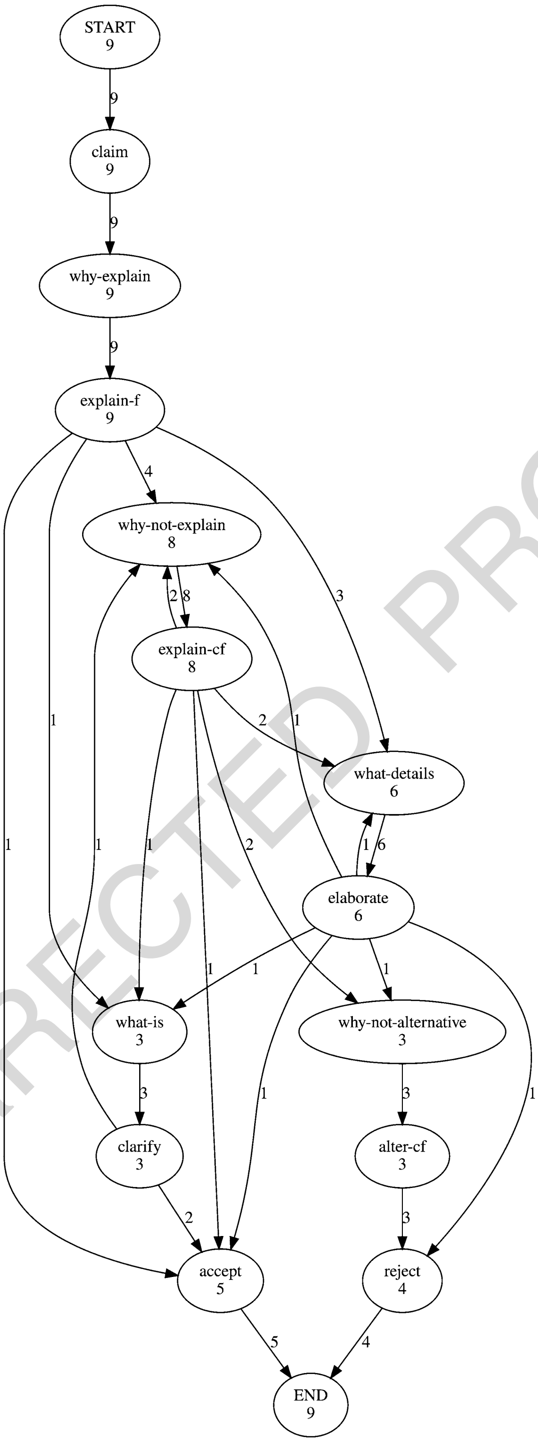


Fig. 15. The full process model of the collected thyroid disease classification explanatory dialogues. For illustrative purposes, pairs of termination nodes, i.e. *{accept-u, accept-s}* and *{reject-u and reject-s}*, are merged into *accept* and *reject*, respectively.

Table 20

Number of times the CF explanations (sorted by rank) were requested by participants (thyroid disease classification)

	CF class		
CF rank	*No hypothyroid*	*Primary hypothyroid*	*Compensatory hypothyroid*
# 1	4	2	2
# 2	2	1	–

sent to the system immediately after the factual explanation was generated whereas the other clarification request followed a detailisation request. Conversely, two (33.33%) detailisation requests were registered for all the CF explanations generated.

The locution-level process model (see Fig. 15 for details) shows the responses to which requests were the most decisive for the study participants to make their final decisions. Out of the 5 participants who accepted the system's claim, one (20.00%) did so immediately after a factual explanation was presented. Similarly, 1 out of 5 (20.00%) accepted the claim in response to a CF explanation offered and a detailisation request each. In addition, 2 (40.00%) participants accepted the system's claim after having received a response to their clarification requests. Out of the 4 participants who rejected the claim, three (75.00%) did so after an alternative CF explanation was offered whereas one (25.00%) was driven by a response to his or her detailisation request.

Finally, recall that 8 CF explanation requests were registered in the thyroid disease classification dialogues. Table 20 presents occurrences of CF explanation requests for each CF class as well those related to alternative CF explanations. Thus, three participants requested an alternative CF explanation for the CF class (two for the class "No hypothyroid" and one – for the class "Primary hypothyroid". Hence, almost a half of the CF explanation requests (3 out of 8, 37.50%) left end users with unsatisfactory responses. Further, all the three such alternative CF explanations turned out to be the final users' dialogue moves before they made their decision (in all the cases the system's claim was eventually rejected).

Acknowledgements

Ilia Stepin is an *FPI* researcher (PRE2019-090153). This work was supported by the Spanish Ministry of Science, Innovation and Universities (grants PID2021-123152OB-C21, TED2021-130295B-C33 and RED2022-134315-T) and the Galician Ministry of Culture, Education, Professional Training and University (grants ED431G2019/04 and ED431C2022/19). These grants were co-funded by the European Regional Development Fund (ERDF/FEDER program). Katarzyna Budzynska was funded in part by POB CyberDS of Warsaw University of Technology within the *Excellence Initiative: Research University* (IDUB) programme under grant 1820/1/Z01/POB3/2021 and in part by VW foundation (VolkswagenStiftung) under grant 98 542. In addition, the authors would like to express their gratitude to Dr. Elena Musi for her invaluable support in gathering experimental data.

References

[1] A. Adadi and M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), *IEEE Access* **6** (2018), 52138–52160. doi:10.1109/ACCESS.2018.2870052.

[2] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J.M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. del Ser, N. Díaz-Rodríguez and F. Herrera, Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence, *Information Fusion* (2023). doi:10.1016/j.inffus.2023.101805.

[3] J.M. Alonso, Teaching explainable artificial intelligence to high school students, *International Journal of Computational Intelligence Systems* **13**(1) (2020), 974–987. doi:10.2991/ijcis.d.200715.003.

[4] A. Arioua, P. Buche and M. Croitoru, Explanatory dialogues with argumentative faculties over inconsistent knowledge bases, *Expert Systems with Applications* **80** (2017), 244–262. doi:10.1016/j.eswa.2017.03.009.

[5] A. Arioua and M. Croitoru, Formalizing explanatory dialogues, in: *International Conference on Scalable Uncertainty Management*, Springer, 2015, pp. 282–297. doi:10.1007/978-3-319-23540-0_19.

[6] F. Behrens, S. Bischoff, P. Ladenburger, J. Rückin, L. Seidel, F. Stolp, M. Vaichenker, A. Ziegler, D. Mottin, F. Aghaei et al., MetaExp: Interactive explanation and exploration of large knowledge graphs, in: *Companion Proceedings of the Web Conference 2018*, 2018, pp. 199–202. doi:10.1145/3184558.3186978.

[7] T.J.M. Bench-Capon, D. Lowes and A.M. McEnery, Argument-based explanation of logic programs, *Knowledge-Based Systems* **4**(3) (1991), 177–183, ISSN 0950-7051. doi:10.1016/0950-7051(91)90007-O.

[8] F. Bex and D. Walton, Combining explanation and argumentation in dialogue, *Argument & Computation* **7**(1) (2016), 55–68. doi:10.3233/AAC-160001.

[9] K. Budzynska, A. Rocci and O. Yaskorska, Financial dialogue games: A protocol for earnings conference calls, in: *Computational Models of Argument (COMMA)*, IOS Press, 2014, pp. 19–30. doi:10.3233/978-1-61499-436-7-19.

[10] R. Calegari, A. Omicini, G. Pisano and G. Sartor, Arg2P: An argumentation framework for explainable intelligent systems, *Journal of Logic and Computation* **32** (2022), 0955. doi:10.1093/logcom/exab089.

[11] R. Calegari, A. Omicini and G. Sartor, Argumentation and logic programming for explainable and ethical AI, in: *Proceedings of the Italian Workshop on Explainable Artificial Intelligence Co-Located with 19th International Conference of the Italian Association for Artificial Intelligence (XAI.it@AIxIA)*, 2020, pp. 55–68.

[12] E. Cambria, L. Malandri, F. Mercorio, M. Mezzanzanica and N. Nobani, A survey on XAI and natural language explanations, *Information Processing & Management* **60**(1) (2023), 103111–103116. doi:10.1016/j.ipm.2022.103111.

[13] G. Castellano, C. Castiello and A.M. Fanelli, The FISDeT software: Application to beer style classification, in: *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2017, pp. 1–6. doi:10.1109/FUZZ-IEEE.2017.8015503.

[14] A. Cawsey, *Explanation and Interaction: The Computer Generation of Explanatory Dialogues*, MIT Press, 1992. doi:10.1007/BF00387398.

[15] F. Cheng, Y. Ming and H. Qu, Dece: Decision explorer with counterfactual explanations for machine learning models, *IEEE Transactions on Visualization and Computer Graphics* **27**(2) (2020), 1438–1447. doi:10.1109/TVCG.2020.3030342.

[16] K. Čyras, A. Rago, E. Albini, P. Baroni and F. Toni, Argumentative XAI: A survey, in: *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI)*, 2021, pp. 4392–4399, Survey Track. doi:10.24963/ijcai.2021/600.

[17] S. Dandl, C. Molnar, M. Binder and B. Bischl, Multi-objective counterfactual explanations, in: *International Conference on Parallel Problem Solving from Nature*, Springer, 2020, pp. 448–469. doi:10.1007/978-3-030-58112-1_31.

[18] G. De Toni, P. Viappiani, B. Lepri and A. Passerini, Generating personalized counterfactual interventions for algorithmic recourse by eliciting user preferences, 2022, arXiv preprint arXiv:2205.13743.

[19] D. Dua and C. Graff, *UCI Machine Learning Repository*, 2017, http://archive.ics.uci.edu/ml.

[20] A. D'Ulizia, F. Ferri and P. Grifoni, A hybrid grammar-based approach to multimodal languages specification, in: *OTM Confederated International Conferences on the Move to Meaningful Internet Systems*, Springer, 2007, pp. 367–376. doi:10.1007/978-3-540-76888-3_59.

[21] D. Engelmann, J. Damasio, A.R. Panisson, V. Mascardi and R.H. Bordini, Argumentation as a method for explainable AI: A systematic literature review, in: *Proceedings of the 17th Iberian Conference on Information Systems and Technologies (CISTI)*, IEEE, 2022, pp. 1–6. doi:10.23919/CISTI54924.2022.9820411.

[22] A. Eshghi, I. Shalyminov and O. Lemon, Bootstrapping incremental dialogue systems from minimal data: The generalisation power of dialogue grammars, in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2220–2230. doi:10.18653/v1/D17-1236.

[23] J. Geertzen, Dialogue act prediction using stochastic context-free grammar induction, in: *Proceedings of the EACL Workshop on Computational Linguistic Aspects of Grammatical Inference*, 2009, pp. 7–15.

[24] A. Ghazimatin, S. Pramanik, R. Saha Roy and G. Weikum, ELIXIR: Learning from user feedback on explanations to improve recommender models, in: *Proceedings of the Web Conference*, 2021, pp. 3850–3860. doi:10.1145/3442381.3449848.

[25] H.P. Grice, Logic and conversation, in: *Syntax and Semantics: Speech Acts*, P. Cole and J.L. Morgan, eds, Academic Press, 1975, pp. 41–58. doi:10.1163/9789004368811_003.

[26] A. Groza, L. Toderean, G.A. Muntean and S.D. Nicoara, Agents that argue and explain classifications of retinal conditions, *Journal of Medical and Biological Engineering* **41**(5) (2021), 730–741. doi:10.1007/s40846-021-00647-7.

[27] R. Guidotti, Counterfactual explanations and how to find them: Literature review and benchmarking, *Data Mining and Knowledge Discovery* (2022), 1–55. doi:10.1007/s10618-022-00831-6.

[28] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri and F. Turini, Factual and counterfactual explanations for black box decision making, *IEEE Intelligent Systems* **34**(6) (2019), 14–23. doi:10.1109/MIS.2019.2957223.

[29] D. Gunning and D. Aha, DARPA's explainable artificial intelligence (XAI) program, *AI magazine* **40**(2) (2019), 44–58. doi:10.1609/aimag.v40i2.2850.

[30] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, The WEKA data mining software: An update, *SIGKDD Explor. Newsl.* **11**(1) (2009), 10–18, ISSN 1931-0145. doi:10.1145/1656274.1656278.

[31] C.G. Hempel, *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, The Free Press, New York, 1965.

[32] G. Hindelang, Dialogue grammar. A linguistic approach to the analysis of dialogue, *Concepts of Dialogue. Considered from the Perspective of Different Disciplines* (1994), 37–48. doi:10.1515/9783111332062-004.

[33] M. Hirzel, L. Mandel, A. Shinnar, J. Siméon and M. Vaziri, I can parse you: Grammars for dialogs, in: *2nd Summit on Advances in Programming Languages (SNAPL)*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017. doi:10.4230/LIPIcs.SNAPL.2017.6.

[34] E. Iosif, I. Klasinas, G. Athanasopoulou, E. Palogiannidi, S. Georgiladakis, K. Louka and A. Potamianos, Speech understanding for spoken dialogue systems: From corpus harvesting to grammar rule induction, *Computer Speech & Language* **47** (2018), 272–297. doi:10.1016/j.csl.2017.08.002.

[35] A.-H. Karimi, G. Barthe, B. Balle and I. Valera, Model-agnostic counterfactual explanations for consequential decisions, in: *International Conference on Artificial Intelligence and Statistics, PMLR*, 2020, pp. 895–905.

[36] N.C. Karunatillake, N.R. Jennings, I. Rahwan and P. McBurney, Dialogue games that agents play within a society, *Artificial intelligence* **173** (2009), 935–981.

[37] M.T. Keane, E.M. Kenny, E. Delaney and B. Smyth, If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques, in: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2021, pp. 4466–4474. doi:10.24963/ijcai.2021/609.

[38] I. Klasinas, A. Potamianos, E. Iosif, S. Georgiladakis and G. Mameli, Web data harvesting for speech understanding grammar induction, in: *Interspeech*, 2013, pp. 2733–2737. doi:10.21437/Interspeech.2013-627.

[39] M. Koit, O. Gerassimenko, R. Kasterpalu, A. Raabis and K. Strandson, Towards computer-human interaction in natural language, *International journal of computer applications in technology* **34**(4) (2009), 291–297. doi:10.1504/IJCAT.2009.024082.

[40] S. Kottur, J.M.F. Moura, D. Parikh, D. Batra and M. Rohrbach, CLEVR-dialog: A diagnostic dataset for multi-round reasoning in visual dialog, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 582–595. doi:10.18653/v1/N19-1058.

[41] M. Kuźba and P. Biecek, What would you ask the machine learning model? Identification of user needs for model explanations based on human-model conversations, in: *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2020, pp. 447–459. doi:10.1007/978-3-030-65965-3_30.

[42] C. Labreuche, N. Maudet, W. Ouerdane and S. Parsons, A dialogue game for recommendation with adaptive preference models, in: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems*, 2015. doi:10.5555/2772879.2773275.

[43] J. Lawrence, M. Snaith, B. Konat, K. Budzynska and C. Reed, Debating technology for dialogical argument: Sensemaking, engagement, and analytics, *ACM Trans. Internet Technol.* **17**(3) (2017). ISSN 1533-5399. doi:10.1145/3007210.

[44] P. McBurney and S. Parsons, Dialogue games for agent argumentation, in: *Argumentation in Artificial Intelligence*, Springer, 2009, pp. 261–280. doi:10.1007/978-0-387-98197-0_13.

[45] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence* **267** (2019), 1–38. doi:10.1016/j.artint.2018.07.007.

[46] S. Modgil and M. Caminada, Proof theories and algorithms for abstract argumentation frameworks, in: *Argumentation in Artificial Intelligence*, G. Simari and I. Rahwan, eds, Springer US, Boston, MA, 2009, pp. 105–129. doi:10.1007/978-0-387-98197-0_6.

[47] C. Molnar, in: *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*, 2nd edn, Leanpub, 2022, https://christophm.github.io/interpretable-ml-book.

[48] M. Morveli-Espinoza, A. Possebom and C.A. Tacla, A protocol for argumentation-based persuasive negotiation dialogues, in: *Brazilian Conference on Intelligent Systems*, Springer, 2021, pp. 18–32. doi:10.1007/978-3-030-91702-9_2.

[49] R.K. Mothilal, A. Sharma and C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 607–617. doi:10.1145/3351095.3372850.

[50] M. Ocana, D. Chapela-Campa, P. Alvarez, N. Hernández, M. Mucientes, J. Fabra, Á. Llamazares, M. Lama, P.A. Revenga, A. Bugarín, M. García-Garrido and J.M. Alonso, Automatic linguistic reporting of customer activity patterns in open malls, *Multimedia Tools and Applications* (2021), 1–27. doi:10.1007/s11042-021-11186-3.

[51] Official Journal of the European Union L119, Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA, Vol. 59, 2016, pp. 89–131, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv.

[52] Official Journal of the European Union L173, Directive 2014/65/EU of the European Parliament and of the Council of 15 May 2014 on markets in financial instruments and amending Directive 2002/92/EC and Directive 2011/61/EU, Vol. 57, 2014, pp. 349–485, https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:32014L0065.

[53] Parliament and Council of the European Union, Proposal for laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts, 2021, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN.

[54] H. Prakken, Coherence and flexibility in dialogue games for argumentation, *Journal of logic and computation* **15**(6) (2005), 1009–1040. doi:10.1093/logcom/exi046.

[55] H. Prakken and R. Ratsma, A top-level model of case-based argumentation for explanation: Formalisation and experiments, *Argument & Computation* **7**(1) (2021), 1–36. doi:10.3233/AAC-210009.

[56] A. Rago, O. Cocarascu, C. Bechlivanidis, D. Lagnado and F. Toni, Argumentative explanations for interactive recommendations, *Artificial Intelligence* **296** (2021), 103506. doi:10.1016/j.artint.2021.103506.

[57] M. Ravi, A. Negi and S. Chitnis, in: *A Comparative Review of Expert Systems, Recommender Systems, and Explainable AI, in: Proceedings of the IEEE 7th International Conference for Convergence in Technology (I2CT)*, 2022, pp. 1–8. doi:10.1109/I2CT54291.2022.9824265.

[58] J.J. Robinson, Diagram: A grammar for dialogues, *Communications of the Association for Computing Machinery* **25**(1) (1982), 27–47. doi:10.1145/358315.358387.

[59] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature Machine Intelligence* **1**(5) (2019), 206–215. doi:10.1038/s42256-019-0048-x.

[60] C. Russell, Efficient search for diverse coherent explanations, in: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 20–28. doi:10.1145/3287560.3287569.

[61] I. Sassoon, N. Kökciyan, E. Sklar and S. Parsons, Explainable argumentation for wellness consultation, in: *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, Springer, 2019, pp. 186–202. doi:10.1007/978-3-030-30391-4_11.

[62] M. Schleich, Z. Geng, Y. Zhang and D. Suciu, GeCo: Quality counterfactual explanations in real time, *Proc. VLDB Endow.* **14**(9) (2021), 1681–1693. doi:10.14778/3461535.3461555.

[63] U. Schmid and B. Wrede, What is missing in XAI so far?, *KI-Künstliche Intelligenz* **36** (2022), 303–315. doi:10.1007/s13218-022-00786-2.

[64] F. Sebastian, From Speech Act Theory to Dialog: Dialog Grammar, *The Routledge Handbook of Language and Dialogue* (2017), 162–173. doi:10.4324/9781315750583.

[65] A.T.Q. Shaheen and J. Bowles, Dialogue games for explaining medication choices, in: *Procedings of the 4th International Joint Conference on Rules and Reasoning*, Vol. 97, Springer Nature, 2020. doi:10.1007/978-3-030-57977-7_7.

[66] Z. Shams, D.V. Marina, O. Nir and P. Julian, Normative practical reasoning via argumentation and dialogue, in: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, 2016, pp. 1244–1250. doi:10.17863/CAM.58863.

[67] X. Shao, T. Rienstra, M. Thimm and K. Kersting, Towards understanding and arguing with classifiers: Recent progress, *Datenbank-Spektrum* **20**(2) (2020), 171–180. doi:10.1007/s13222-020-00351-x.

[68] H. Shi, R.J. Ross, T. Tenbrink and J. Bateman, Modelling illocutionary structure: Combining empirical studies with formal model analysis, in: *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, 2010, pp. 340–353. doi:10.1007/978-3-642-12116-6_28.

[69] E.I. Sklar and M.Q. Azhar, Explanation through argumentation, in: *Proceedings of the 6th International Conference on Human-Agent Interaction*, 2018, pp. 277–285. doi:10.1145/3284432.3284470.

[70] K. Sokol and P. Flach, One explanation does not fit all: The promise of interactive explanations for machine learning transparency, KI – Künstliche Intelligenz, 2020. ISSN 1610-1987. doi:10.1007/s13218-020-00637-y.

[71] I. Stepin, J.M. Alonso, A. Catala and M. Pereira-Fariña, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence, *IEEE Access* **9** (2021), 11974–12001, ISSN 2169-3536. doi:10.1109/ACCESS.2021.3051315.

[72] I. Stepin, J.M. Alonso-Moral, A. Catala and M. Pereira-Fariña, An empirical study on how humans appreciate automated counterfactual explanations which embrace imprecise information, *Information Sciences* **618** (2022), 379–399, ISSN 0020-0255. doi:10.1016/j.ins.2022.10.098.

[73] M. Suffian, P. Graziani, J.M. Alonso and A. Bogliolo, FCE: Feedback based counterfactual explanations for explainable AI, *IEEE Access* **10** (2022), 72363–72372. doi:10.1109/ACCESS.2022.3189432.

[74] W.R. Swartout and S.W. Smoliar, On making expert systems more like experts, *Expert Systems* **4**(3) (1987), 196–208. doi:10.1111/j.1468-0394.1987.tb00143.x.

[75] B. Ustun, A. Spangher and Y. Liu, Actionable recourse in linear classification, in: *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT)*, 2019, pp. 10–19. doi:10.1145/3287560.3287566.

[76] W. Van Der Aalst, *Process Mining: Data Science in Action*, Vol. 2, Springer, 2016.

[77] A. Vassiliades, N. Bassiliades and T. Patkos, Argumentation and explainable artificial intelligence: A survey, *The Knowledge Engineering Review* **36** (2021). doi:10.1017/S0269888921000011.

[78] S. Verma, J. Dickerson and K. Hines, Counterfactual explanations for machine learning: A review, in: *Proceedings of the Machine Learning Retrospectives, Surveys & Meta-Analyses (ML-RSA) Workshop at the Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[79] J. Waa, M. Robeer, J. Diggelen, M. Brinkhuis and M. Neerincx, Contrastive explanations with local foil trees, in: *Proceedings of the International Conference on Machine Learning (ICML) Workshop on Human Interpretability (WHI) in Machine Learning*, 2018.

[80] S. Wachter, B. Mittelstadt and C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, *Harvard Journal of Law & Technology* **31** (2018), 841–887.

[81] D. Walton, Dialogical models of explanation, in: *Proceedings of the Conference on Explanation-Aware Computing (ExaCt) Workshop*, 2007, pp. 1–9.

[82] D. Walton and E.C. Krabbe, *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*, SUNY Press, 1995.

[83] D. Wang, Q. Yang, A. Abdul and B.Y. Lim, Designing theory-driven user-centric explainable AI, in: *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI'19*, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–15. doi:10.1145/3290605.3300831.

[84] K. Weitz, L. Vanderlyn, N.T. Vu and E. André, "It's our fault!": Insights into users' understanding and interaction with an explanatory collaborative dialog system, in: *Proceedings of the 25th Conference on Computational Natural Language Learning, Association for Computational Linguistics*, 2021, pp. 1–16. doi:10.18653/v1/2021.conll-1.1.

[85] T. Wu, M.T. Ribeiro, J. Heer and D. Weld, Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models, in: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, Online, 2021, pp. 6707–6723. doi:10.18653/v1/2021.acl-long.523.

[86] L.A. Zadeh, Linguistic variables, approximate reasoning and dispositions, *Medical Informatics* **8**(3) (1983), 173–186. doi:10.3109/14639238309016081.

[87] D. Zhang, S. Mishra, E. Brynjolfsson, J. Etchemendy, D. Ganguli, B. Grosz, T. Lyons, J. Manyika, J.C. Niebles, M. Sellitto et al., *The AI Index 2022 Annual Report, AI Index Steering Committee, Stanford Institute for Human-Centered AI*, Stanford University, 2022.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

CAMPUS VIDA
CAMPUS DE EXCELENCIA INTERNACIONAL

Recent years have witnessed a striking rise of artificial intelligence algorithms that are able to show outstanding performance. However, such good performance is oftentimes achieved at the expense of explainability. Not only can the lack of algorithmic explainability undermine the user's trust in the algorithmic output, but it can also cause adverse consequences. In this thesis, we advocate the use of interpretable rule-based models that can serve both as stand-alone applications and proxies for black-box models. More specifically, we design an explanation generation framework that outputs contrastive, selected, and social explanations for interpretable (decision trees and rule-based) classifiers. We show that the resulting explanations enhance the effectiveness of AI algorithms while preserving their transparent structure.