IA FIABLE
retos técnicos, éticos, legales, culturales y socio-económicos

MINISTERIO DE CIENCIA E INNOVACIÓN
AGENCIA ESTATAL DE INVESTIGACIÓN
Ayuda RED2022-134315-T financiada por MCIN/AEI /10.13039/501100011033

The CiTIUS-USC Z-Inspection® Lab

“How to Evaluate, Audit, Certify… AI systems?”

Jose María Alonso Moral
josemaria.alonso.moral@usc.es

Workshop on Social Impact of Artificial Intelligence
30 May 2024, Fundación Ramón Areces, Madrid

frA | FUNDACIÓN RAMÓN ARECES

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CiTIUS Centro Singular de Investigación en Tecnoloxías Intelixentes

CiTIUS

Centro Singular de Investigación
en **Tecnoloxías Intelixentes**

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

citius.gal

SCIENTIFIC AREAS

Green and smart computing and devices

- Electronic design of intelligent devices
- High Performance Computing

Support for machine intelligence

- Data and process science and engineering
- Automatic learning and reasoning

Intelligent technologies

- Virtual and augmented reality
- Language technologies
- Computer vision
- Robotics

Social, economic, ethical and political framework

- Trustworthy AI

R&D RESULTS 2021-23

ARTICLES

75% in Q1

44% in D1

27% in top 10% most cited

11,8M €

21% income from european projects

TEAM

+150 PEOPLE

31 SENIOR RESEARCHERS

SOME EUROPEAN PROJECTS

erc

NANOVR
ERC Consolidator Grant
Nanoscale Design using Virtual Reality

Marie Skłodowska-Curie Actions

NL4XAI
Interactive Natural Language for Explainable AI (Coordinator)

MENELAOS NT
Multimodal Fusion of Sensor Information (Coordinator)

hybrids
Hybrid intelligence to monitor, promote and analyse transformations in good democracy practices (Coordinator)

MISEL
Multispectral Intelligent Vision System with Embedded Low-Power Neural Computing

iR4S iRead4Skills
Intelligent Reading Improvement System for Fundamental and Transversal Skills Development

SOS food
Sustainability Optimization for Secure Food Systems

EDIH DATAlife
DIH for the deployment of AI and Data Analytics in SMEs in the primary, biotechnological and health sectors

SOME STRATEGIC PROJECTS

nós
Artificial Intelligence at the service of the Galician Language

CAMEIA
Plexus Tech - USC Chair in AI applied to personalised precision medicine

Televés - USC Chip Chair in microelectronic design

TECHNOLOGY TRANSFER

+30 contracts with industry

85% licences to international companies

FINSA soluciones en madera
Machine learning for digital twins and processes optimisation in industry

Avincis AVIATION CRITICAL SERVICES
Efficient processing of aerial images within the Civil UAVs Initiative

Lacer
Rapid analysis of the development of dental plaque

SPIN-OFFS

situm INDOOR POSITIONING

inverbis.

CONGRESS

ECAI 50th ANNIVERSARY

Our Trustworthy AI lab (ELSEC)

TELSEC4TAI

MINISTERIO DE CIENCIA E INNOVACIÓN

AGENCIA ESTATAL DE INVESTIGACIÓN

Episteme

GSi

MARTIN
PEREIRA FARIÑA

ALBERTO JOSE
BUGARIN DIZ

SENEN
BARRO AMENEIRO

MARCOS
GARCIA GONZALEZ

PABLO
GAMALLO OTERO

ALEJANDRO
CATALA BOLOS

JOSE MARIA
ALONSO MORAL

LComp

Z-Inspection®

IA FIABLE
retos técnicos, éticos, legales, culturales y socio-económicos

https://citius.gal/research/areas/ia-fiable/

Psychology

Marnuel
Gandoy

Law, Economics & Business Administration

LUIS
MIGUEZ MACHO

Marcos
Torres

Luis Otero

Isabel Neira

DIANA
SANTIAGO IGLESIAS

Liliana
Arroyo

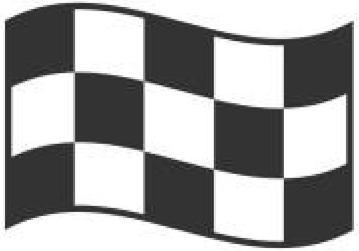
Manuela
Battaglini

Ulises
Cortes

Dagmar
Monett

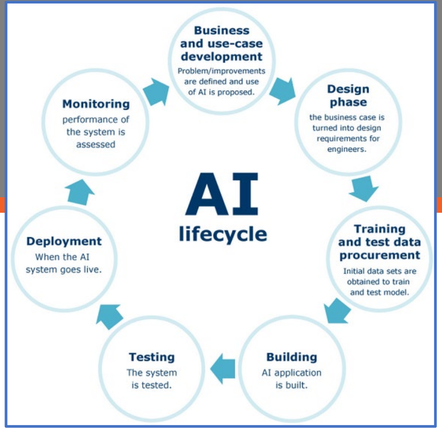
External Ethics Advisors

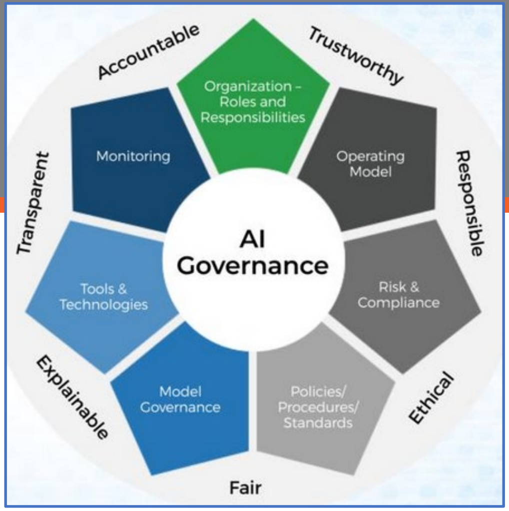
How to certify AI ?

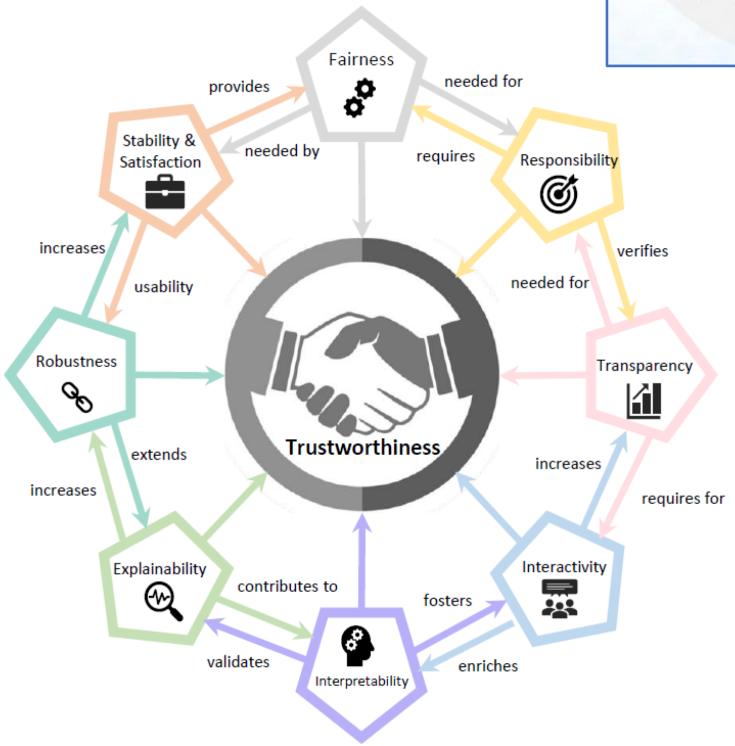


Certification

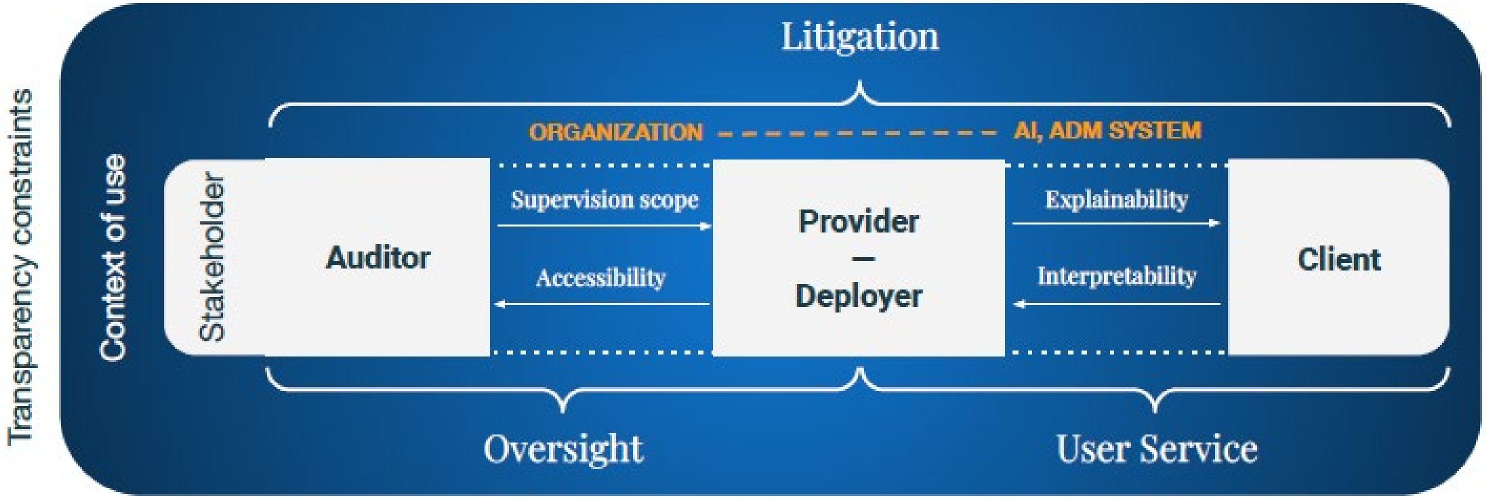
European AI Regulation
AI Act
European Commission


AI lifecycle
Business and use-case development
Problem/improvements are defined and use of AI is proposed.
Design phase
the business case is turned into design requirements for engineers.
Training and test data procurement
Initial data sets are obtained to train and test model.
Building
AI application is built.
Testing
The system is tested.
Deployment
When the AI system goes live.
Monitoring
performance of the system is assessed





Exposure of third parties privacy; trade secret; intellectual property



Perceived regulatory pressure

Gaming Opportunities

Safe AI
Responsible AI
AIg
Trustworthy AI

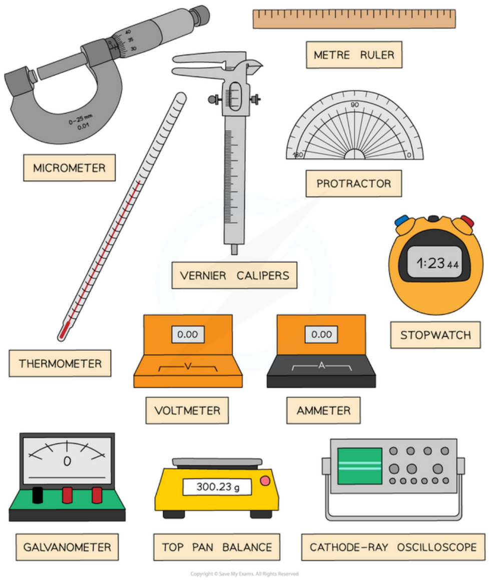
L. Nannini, J. M. Alonso-Moral, A. Catala, M. Lama, S. Barro, "**Operationalizing Explainable AI in the EU Regulatory Ecosystem**", IEEE Intelligent Systems, 2024, https://doi.org/10.1109/MIS.2024.3383155

S. Alia, T. Abuhmed, S. El-Sappagh, K. Muhammad, Jose M. Alonso-Moral, R. Confalonieri R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, "**Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence**", Information Fusion, 2023, https://doi.org/10.1016/j.inffus.2023.101805

How to audit and evaluate AI ?

Auditing and Evaluating AI

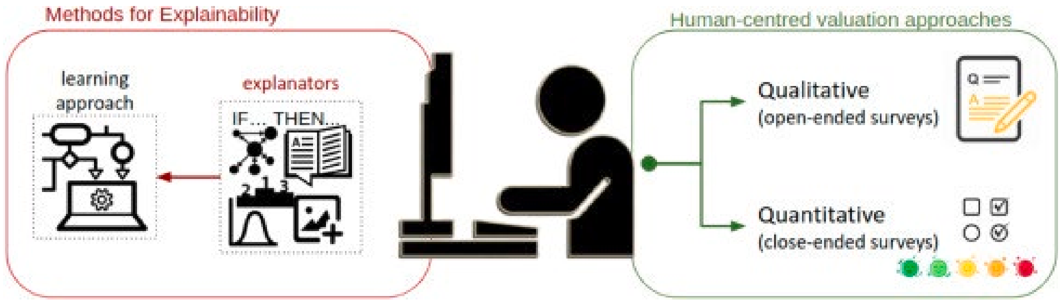
Some Preliminary Notions



- ☐ **Metrology is the Science of Measurement and its application**
  - ➢ To measure something, we need a Metric System (i.e., a standard reference and a related unit which is defined without any ambiguity)
- ☐ **SI: International System of Units** (accepted worldwide)
  - ✓Time is measured in seconds, minutes, hours…
  - ✓Length is measured in meters, inches…
  - ✓Mass is measured in Kg, Pounds…

Automatic Evaluation

- ✓ **Complexity Metrics**
- ✓ **Transparency Metrics**
- ✓ **Interpretability Metrics**


Methods for Explainability
learning approach
explanators
IF... THEN...
Human-centred valuation approaches
Qualitative (open-ended surveys)
Quantitative (close-ended surveys)

Human Evaluation

- ✓ **Interviews**
- ✓ **Focus Groups**
- ✓ **Free-text Questionnaires**

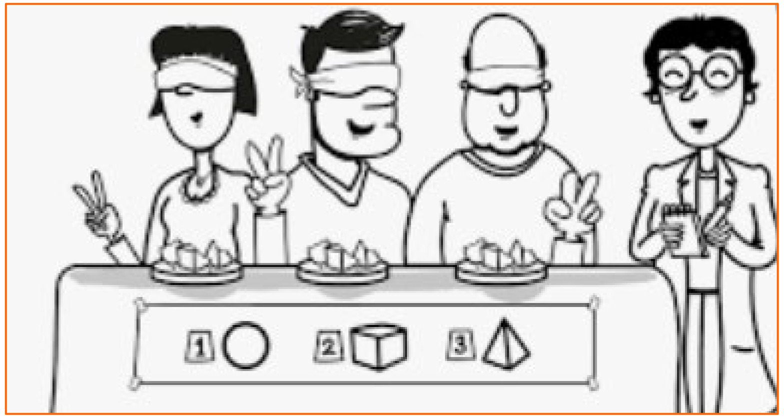
- ✓ **Panel of Experts**
- ✓ **Consumer Testing**
- ✓ **Randomized Control Tests**

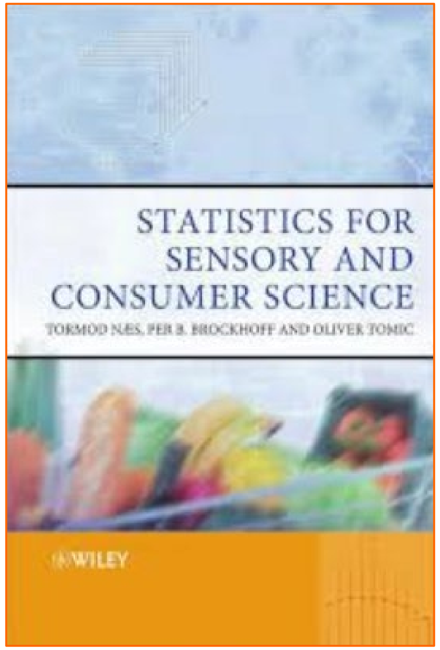
How to measure human-centric & subjective properties ?

☐ **What about subjective judgements (e.g., beauty, elegance, transparency,…) ?**

☐ **Which is the standard reference for sensory inputs (acid, sweet, salty, easy to read…) ?**

☐ **How to avoid confounding effects ?**

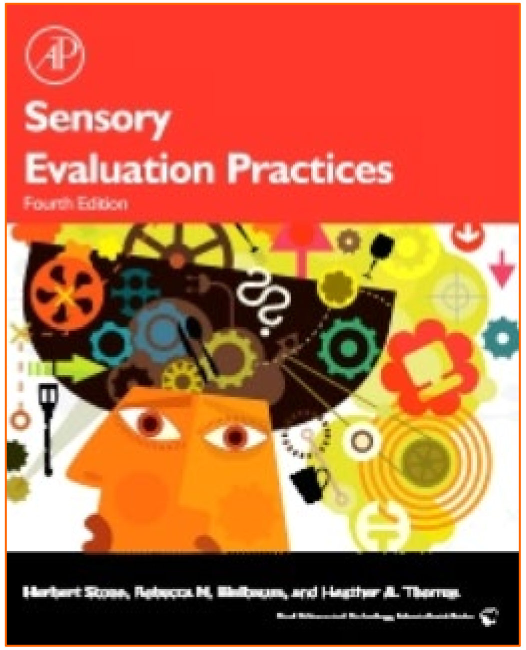




STATISTICS FOR
SENSORY AND
CONSUMER SCIENCE

TORMOD NÆS, PER B. BROCKHOFF AND OLIVER TOMIC

WILEY



Sensory
Evaluation Practices

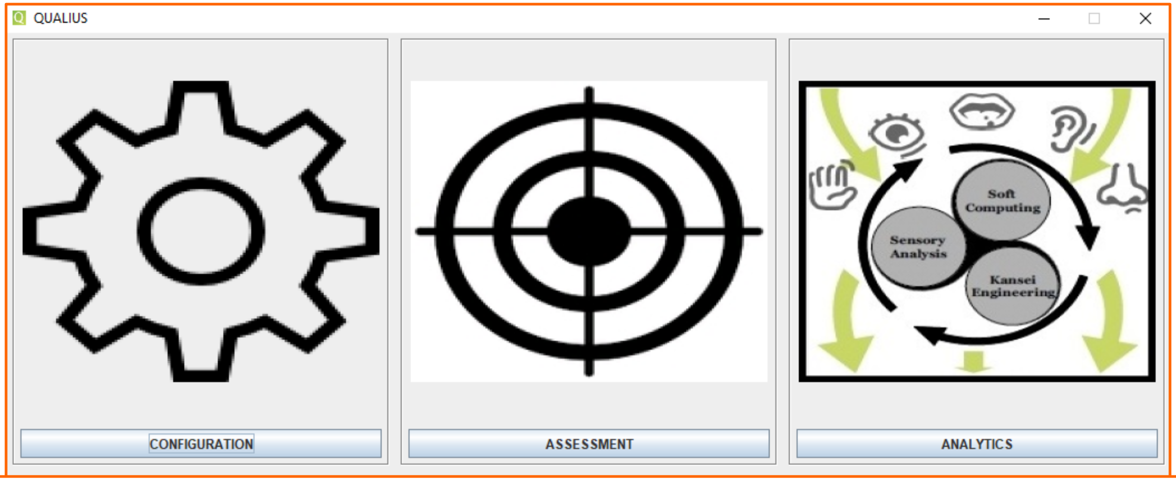
Fourth Edition

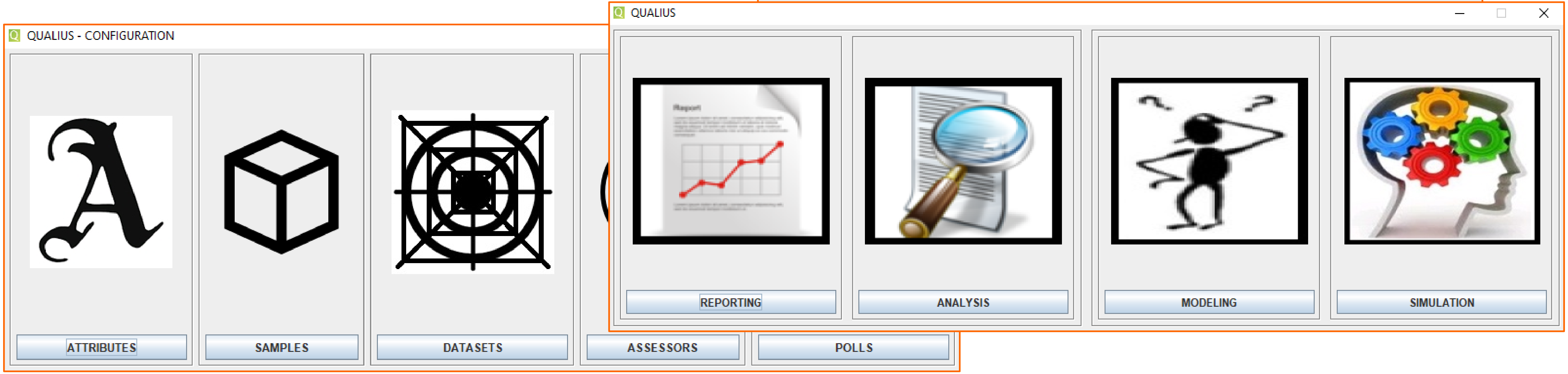
Herbert Stone, Rebecca N. Bleibaum, and Heather A. Thomas

QUALIUS

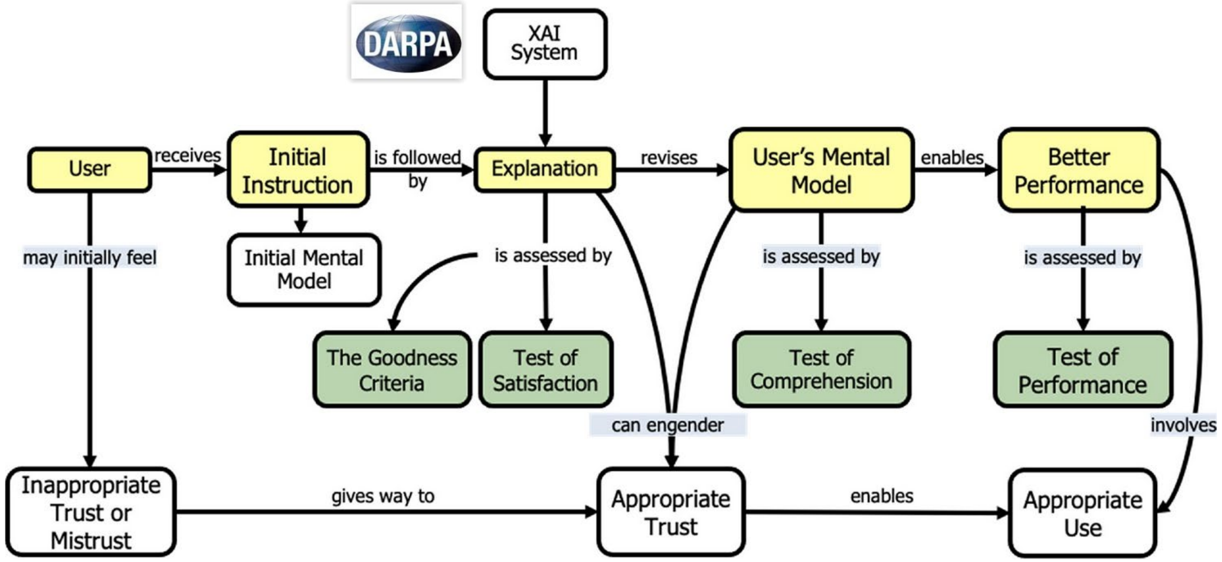
P. Quiros, J.M. Alonso, D.P. Pancho, "**Descriptive and comparative analysis of human perceptions expressed through fuzzy rating scale-based questionnaires**", International Journal of Computational Intelligence Systems, 9(3):450-467, 2016, https://doi.org/10.1080/18756891.2016.1175811

J.M. Alonso, D.P. Pancho, L. Magdalena, D.A. Nuñez, D.S. Sánchez, P.F. Suárez, J. Mingot, V. Iglesias, "**QUALE: A new Toolbox for Quantitative and Qualitative Analysis of Human Perceptions**", IFSA-EUSFLAT, 2015, https://doi.org/10.2991/ifsa-eusflat-15.2015.94



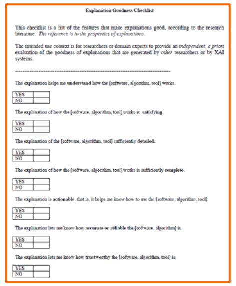


Guiding Human Evaluation in the context of Explainable AI



D. Gunning, E. Vorm, J. Y. Wang, M. Turek, "**DARPA's explainable AI (XAI) program: A retrospective**," Applied AI Letters, 2(4):e61, 2021 https://doi.org/10.1002/ail2.61

R.R. Hoffman, S.T. Mueller, G. Klein, J. Litman, "**Metrics for Explainable AI: Challenges and Prospects**", https://doi.org/10.48550/arXiv.1812.04608



Explanation Goodness Checklist

This checklist is a list of the features that make explanations good, according to the research literature. *The reference is to the properties of explanations.*

The intended use context is for researchers or domain experts to provide an *independent, a priori* evaluation of the goodness of explanations that are generated by *other* researchers or by XAI systems.

The explanation helps me **understand** how the [software, algorithm, tool] works.

YES	
NO	

The explanation of how the [software, algorithm, tool] works is **satisfying**.

YES	
NO	

The explanation of the [software, algorithm, tool] sufficiently **detailed.**

YES	
NO	

The explanation of how the [software, algorithm, tool] works is sufficiently **complete.**

YES	
NO	

The explanation is **actionable**, that is, it helps me know how to use the [software, algorithm, tool]

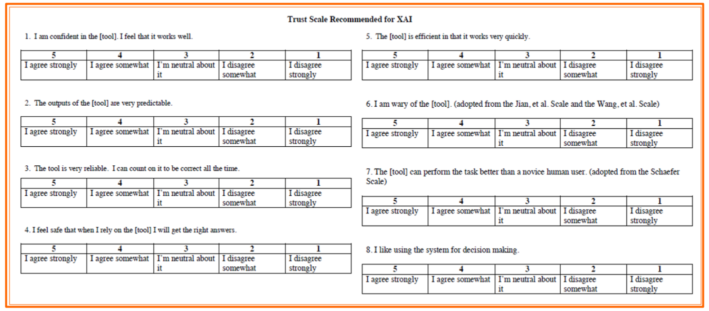
YES	
NO	

The explanation lets me know how **accurate or reliable** the [software, algorithm] is.

YES	
NO	

The explanation lets me know how **trustworthy** the [software, algorithm, tool] is.

YES	
NO	



Trust Scale Recommended for XAI

1. I am confident in the [tool]. I feel that it works well.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

2. The outputs of the [tool] are very predictable.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

3. The tool is very reliable. I can count on it to be correct all the time.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

4. I feel safe that when I rely on the [tool] I will get the right answers.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

5. The [tool] is efficient in that it works very quickly.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

6. I am wary of the [tool]. (adopted from the Jian, et al. Scale and the Wang, et al. Scale)

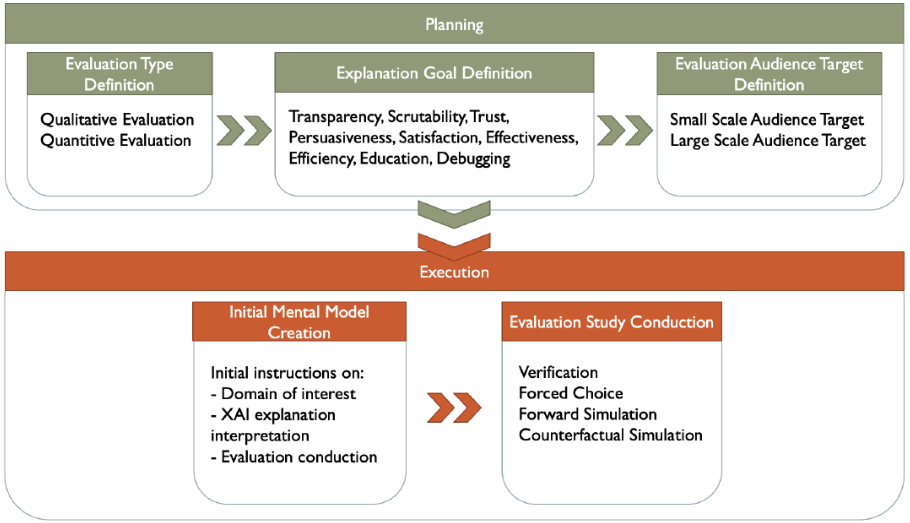
5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

7. The [tool] can perform the task better than a novice human user. (adopted from the Schaefer Scale)

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly

8. I like using the system for decision making.

5	4	3	2	1
I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly



Planning

Evaluation Type Definition
Qualitative Evaluation
Quantitive Evaluation

Explanation Goal Definition
Transparency, Scrutability, Trust, Persuasiveness, Satisfaction, Effectiveness, Efficiency, Education, Debugging

Evaluation Audience Target Definition
Small Scale Audience Target
Large Scale Audience Target

Execution

Initial Mental Model Creation
Initial instructions on:
- Domain of interest
- XAI explanation interpretation
- Evaluation conduction

Evaluation Study Conduction
Verification
Forced Choice
Forward Simulation
Counterfactual Simulation

R. Confalonieri, J.M. Alonso-Moral, "**An operational framework for guiding human evaluation in Explainable and Trustworthy AI**", IEEE Intelligent Systems, 2023, https://doi.org/10.1109/MIS.2023.3334639

The ReproHum Project

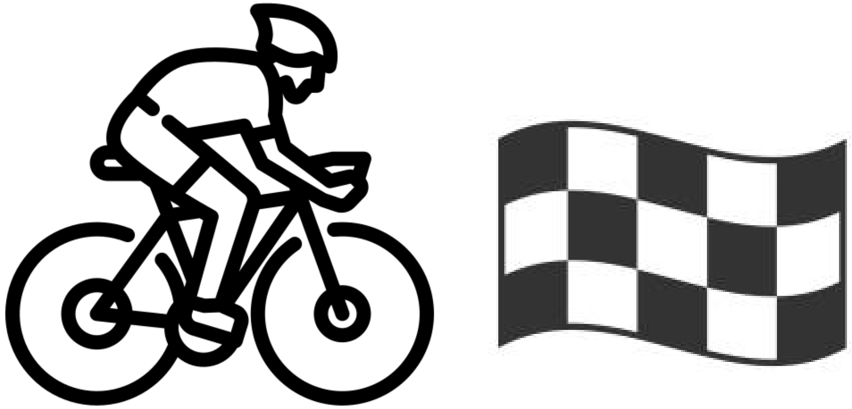
EPSRC
Pioneering research and skills

Investigating Reproducibility of Human Evaluations in NLP

https://reprohum.github.io/

A. Belz et al., "**Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP**", Proceedings of the 4th ACL Workshop on Insights from Negative Results in NLP, 2023, https://aclanthology.org/2023.insights-1

Use Case

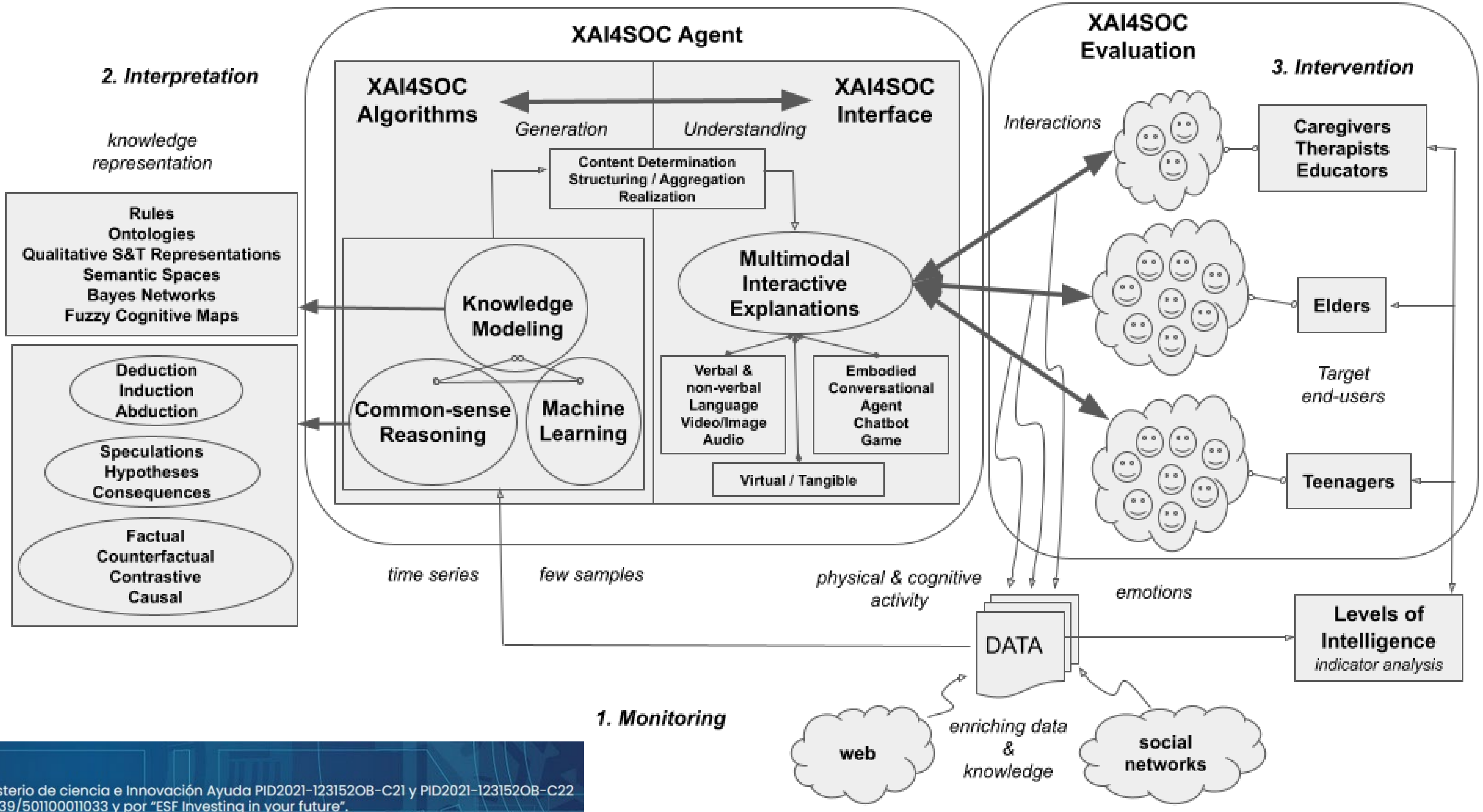


Centro Singular de Investigación
en **Tecnoloxías Intelixentes**

XAI4SOC will create an open-source framework for the design, development, validation, evaluation and deployment of a new generation of ubiquitous XAI agents (algorithms and interfaces) for enhancing the prominence of their target users, with the support of caregivers, therapists and educators, in pursuing a healthy ageing and wellbeing development

UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

UNIVERSITAT
JAUME I



2. Interpretation

knowledge representation

Rules
Ontologies
Qualitative S&T Representations
Semantic Spaces
Bayes Networks
Fuzzy Cognitive Maps

Deduction
Induction
Abduction

Speculations
Hypotheses
Consequences

Factual
Counterfactual
Contrastive
Causal

XAI4SOC Agent

XAI4SOC Algorithms

Generation

Understanding

XAI4SOC Interface

Content Determination
Structuring / Aggregation
Realization

Knowledge Modeling

Common-sense Reasoning

Machine Learning

Multimodal Interactive Explanations

Verbal & non-verbal Language
Video/Image
Audio

Embodied Conversational Agent
Chatbot
Game

Virtual / Tangible

XAI4SOC Evaluation

3. Intervention

Interactions

Caregivers
Therapists
Educators

Elders

Target end-users

Teenagers

time series

few samples

physical & cognitive activity

emotions

DATA

Levels of Intelligence
indicator analysis

1. Monitoring

web

enriching data & knowledge

social networks

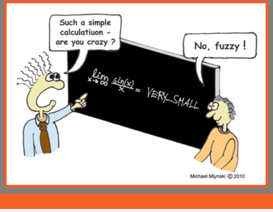
https://xai4soc.es/

MINISTERIO
DE CIENCIA
E INNOVACIÓN

AGENCIA
ESTATAL DE
INVESTIGACIÓN

Proyecto financiado por el Ministerio de ciencia e Innovación Ayuda PID2021-123152OB-C21 y PID2021-123152OB-C22
financiada por MCIN/AEI/ 10.13039/501100011033 y por "ESF Investing in your future".

Evaluating the consistency of explanations by LLMs


Such a simple calculatiuon - are you crazy ?
No, fuzzy !
$\lim_{x \to \infty} \frac{\sin(x)}{x}$ = VERY_SMALL
Michael Mlynski © 2010

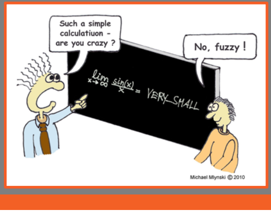
Fragment of a real conversation between TimeVersa and a user. Underlined, temporal expressions relevant for the conversation.

Move	Participant	Utterance
m_1	User	Hello! I would like to be in Cíes Islands around 12 PM. What time should I take the boat from Vigo?
m_2	TimeVersa	There is a boat at 11.15 AM that is suitable for you.
m_3	User	Why do you think this boat is suitable for me?
m_4	TimeVersa	The boat trip from Vigo to Cíes Islands lasts about 35 minutes, so if you take the boat at 11.15 AM you can be there around 12 PM.
m_5	User	Could you specify what “around 12 PM” means for you?
m_6	TimeVersa	It means between 11.50 AM and 12.10 PM.
m_7	User	OK. Book the boat for me, please.
m_8	TimeVersa	Could I help you with anything else?
m_9	User	Yes. Once we have reached Cíes Islands around 12 PM, we had thought of doing the hiking route “Faro de Cíes”. Do you think we can be back in Vigo before 2 PM?
m_{10}	TimeVersa	No, this is not possible.
m_{11}	User	Why not?
m_{12}	TimeVersa	Because the route will take you about 3 hours. If you arrive the Islands around 12 PM, you can’t be back in Vigo before 2 PM.
m_{13}	User	OK, I understand. So, what time can we be back?
m_{14}	TimeVersa	You can be back in Vigo around 4.35 PM if you take the boat at 4 PM.

M. Canabal-Juanatey, J.M. Alonso-Moral, A. Catala, A. Bugarín-Diz, "**Enriching Interactive Explanations with Fuzzy Temporal Constraint Networks**", International Journal of Approximate Reasoning, 109128, 2024, https://doi.org/10.1016/j.ijar.2024.109128

Centro Singular de Investigación
en **Tecnoloxías Intelixentes**

Evaluating the consistency of explanations by LLMs

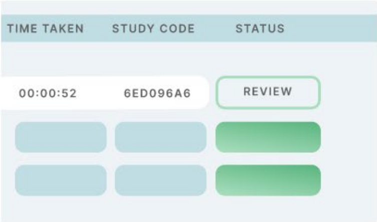

Such a simple calculatiuon - are you crazy ?
No, fuzzy !
$\lim_{x \to \infty} \frac{\sin(x)}{x}$ = VERY_SMALL
Michael Mlynski © 2010

Prolific

https://www.prolific.com/

Insert your study link

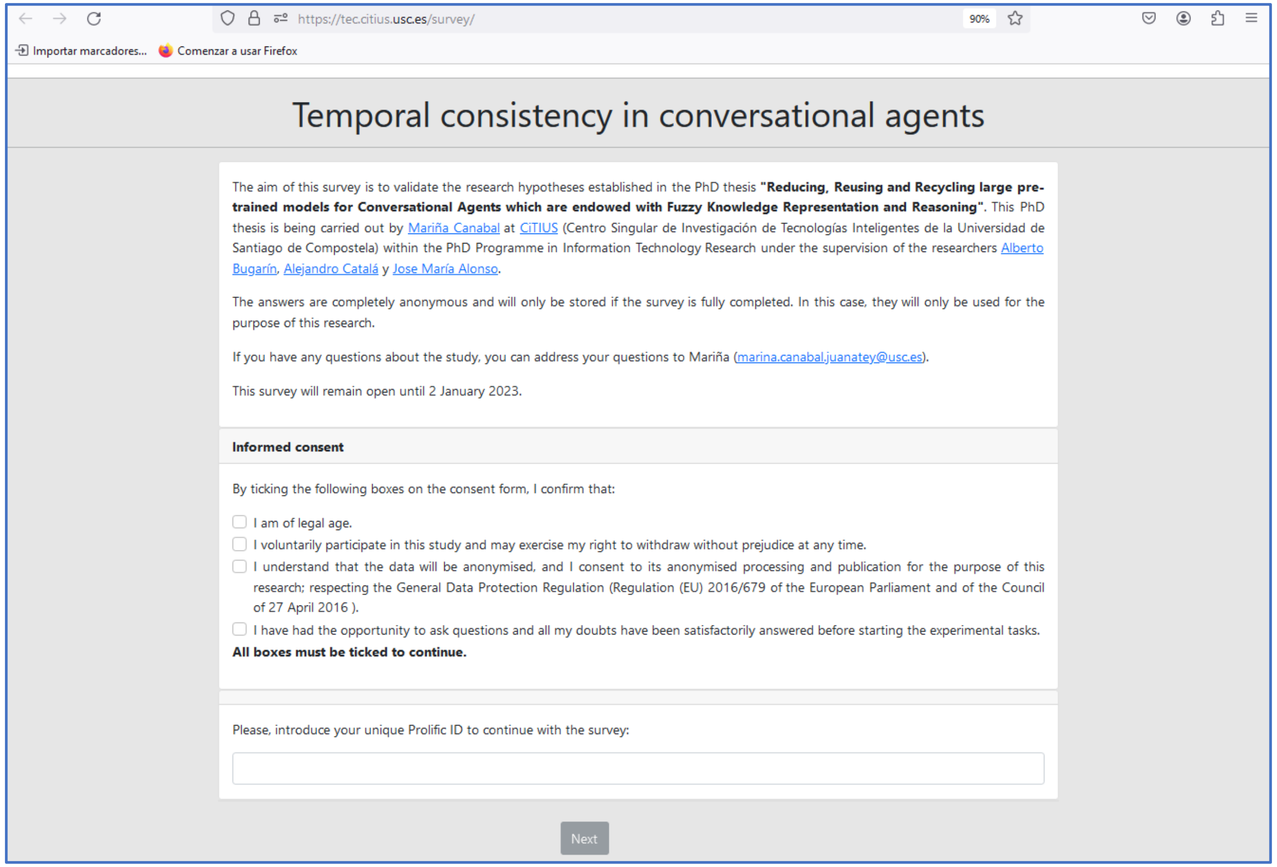
Find your target audience



Collect your responses

UNIVERSITY OF OXFORD · European Commission · Google · nugget.ai

https://tec.citius.usc.es/survey/



Temporal consistency in conversational agents

The aim of this survey is to validate the research hypotheses established in the PhD thesis **"Reducing, Reusing and Recycling large pre-trained models for Conversational Agents which are endowed with Fuzzy Knowledge Representation and Reasoning"**. This PhD thesis is being carried out by Mariña Canabal at CiTIUS (Centro Singular de Investigación de Tecnologías Inteligentes de la Universidad de Santiago de Compostela) within the PhD Programme in Information Technology Research under the supervision of the researchers Alberto Bugarín, Alejandro Catalá y Jose María Alonso.

The answers are completely anonymous and will only be stored if the survey is fully completed. In this case, they will only be used for the purpose of this research.

If you have any questions about the study, you can address your questions to Mariña (marina.canabal.juanatey@usc.es).

This survey will remain open until 2 January 2023.

Informed consent

By ticking the following boxes on the consent form, I confirm that:

☐ I am of legal age.
☐ I voluntarily participate in this study and may exercise my right to withdraw without prejudice at any time.
☐ I understand that the data will be anonymised, and I consent to its anonymised processing and publication for the purpose of this research; respecting the General Data Protection Regulation (Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016).
☐ I have had the opportunity to ask questions and all my doubts have been satisfactorily answered before starting the experimental tasks.

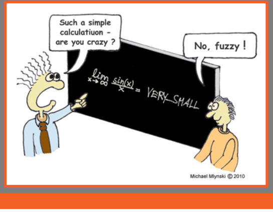
All boxes must be ticked to continue.

Please, introduce your unique Prolific ID to continue with the survey:

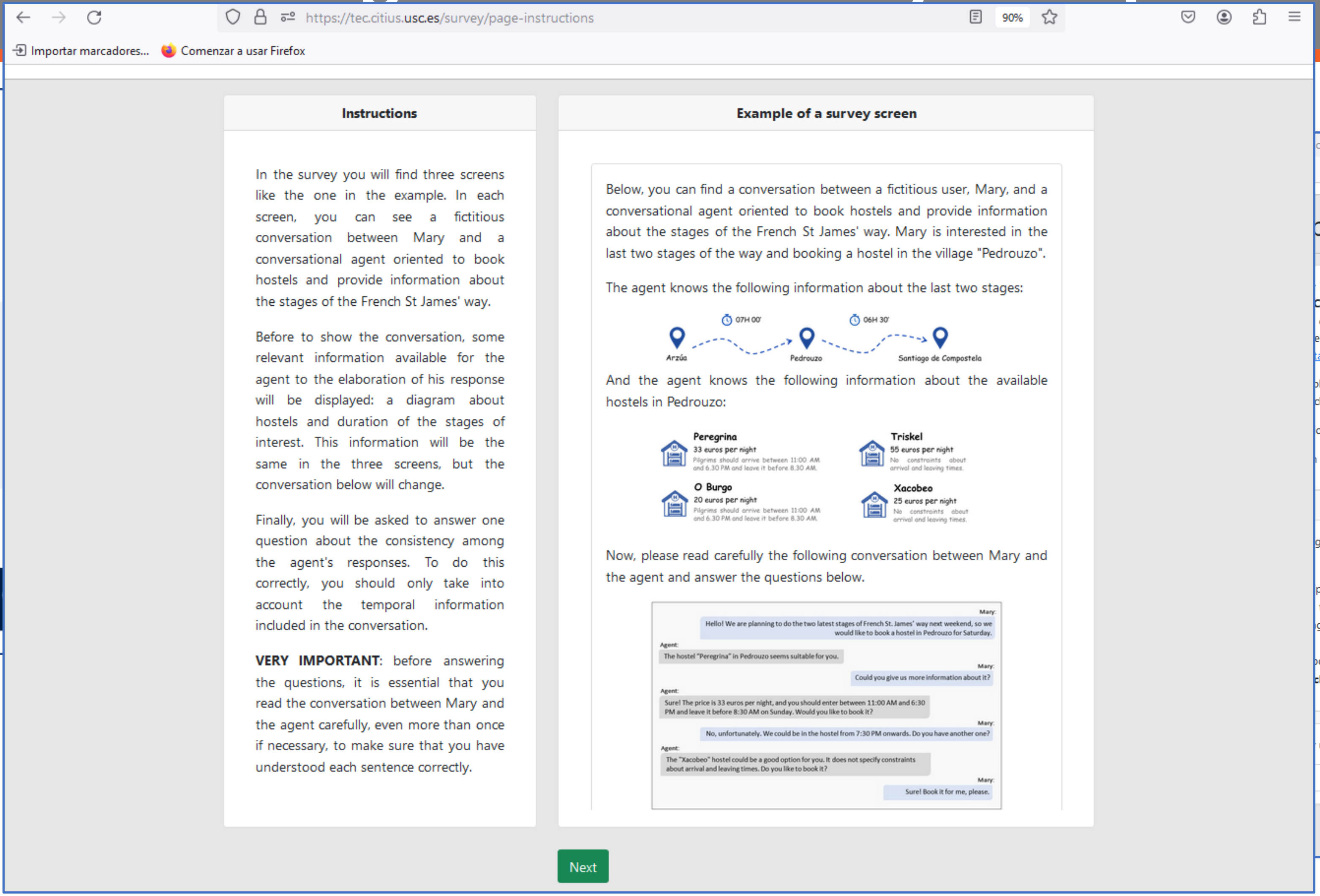
Next

M. Canabal-Juanatey, J.M. Alonso-Moral, A. Catala, A. Bugarín-Diz, "**Enriching Interactive Explanations with Fuzzy Temporal Constraint Networks**", International Journal of Approximate Reasoning, 109128, 2024, https://doi.org/10.1016/j.ijar.2024.109128

Evaluating the consistency of explanations by LLMs


Such a simple calculatiuon - are you crazy ?
No, fuzzy !
$\lim_{x \to \infty} \frac{\sin(x)}{x}$ = VERY_SMALL
Michael Mlynski © 2010

https://tec.citius.usc.es/survey/



Instructions

In the survey you will find three screens like the one in the example. In each screen, you can see a fictitious conversation between Mary and a conversational agent oriented to book hostels and provide information about the stages of the French St James' way.

Before to show the conversation, some relevant information available for the agent to the elaboration of his response will be displayed: a diagram about hostels and duration of the stages of interest. This information will be the same in the three screens, but the conversation below will change.

Finally, you will be asked to answer one question about the consistency among the agent's responses. To do this correctly, you should only take into account the temporal information included in the conversation.

VERY IMPORTANT: before answering the questions, it is essential that you read the conversation between Mary and the agent carefully, even more than once if necessary, to make sure that you have understood each sentence correctly.

Example of a survey screen

Below, you can find a conversation between a fictitious user, Mary, and a conversational agent oriented to book hostels and provide information about the stages of the French St James' way. Mary is interested in the last two stages of the way and booking a hostel in the village "Pedrouzo".

The agent knows the following information about the last two stages:

And the agent knows the following information about the available hostels in Pedrouzo:

Peregrina
33 euros per night
Pilgrims should arrive between 11:00 AM and 6.30 PM and leave it before 8.30 AM.

Triskel
55 euros per night
No constraints about arrival and leaving times.

O Burgo
20 euros per night
Pilgrims should arrive between 11:00 AM and 6.30 PM and leave it before 8.30 AM.

Xacobeo
25 euros per night
No constraints about arrival and leaving times.

Now, please read carefully the following conversation between Mary and the agent and answer the questions below.

Mary: Hello! We are planning to do the two latest stages of French St. James' way next weekend, so we would like to book a hostel in Pedrouzo for Saturday.

Agent: The hostel "Peregrina" in Pedrouzo seems suitable for you.

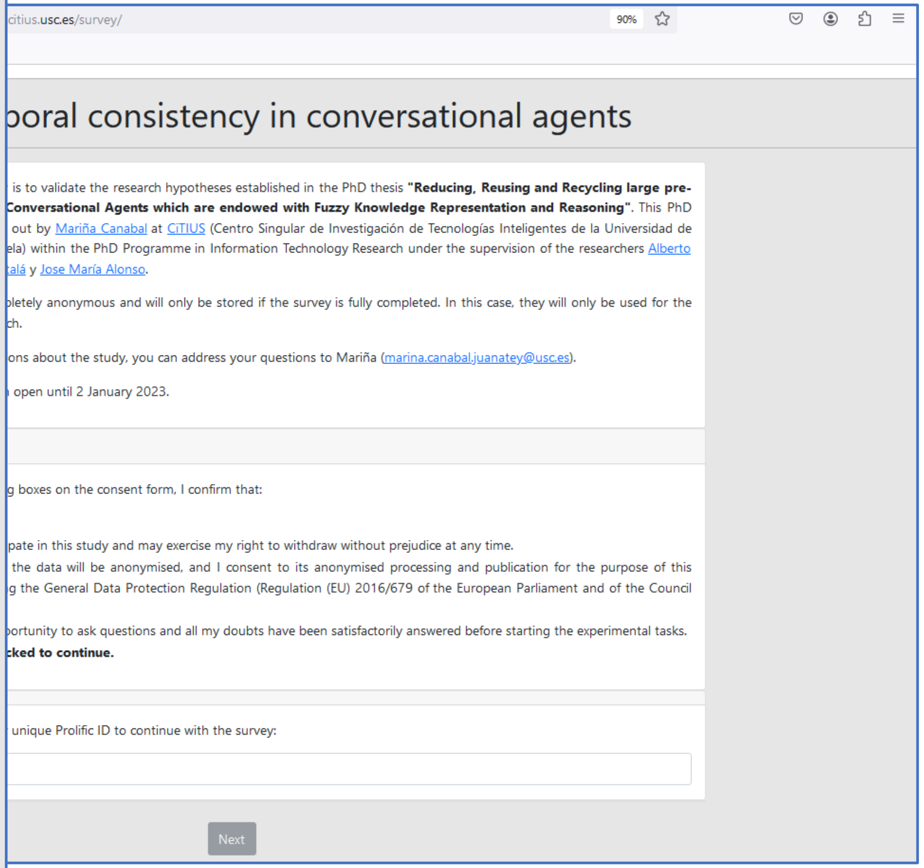
Mary: Could you give us more information about it?

Agent: Sure! The price is 33 euros per night, and you should enter between 11:00 AM and 6:30 PM and leave it before 8:30 AM on Sunday. Would you like to book it?

Mary: No, unfortunately. We could be in the hostel from 7:30 PM onwards. Do you have another one?

Agent: The "Xacobeo" hostel could be a good option for you. It does not specify constraints about arrival and leaving times. Do you like to book it?

Mary: Sure! Book it for me, please.



poral consistency in conversational agents

is to validate the research hypotheses established in the PhD thesis **"Reducing, Reusing and Recycling large pre-Conversational Agents which are endowed with Fuzzy Knowledge Representation and Reasoning"**. This PhD out by Mariña Canabal at CiTIUS (Centro Singular de Investigación de Tecnologías Inteligentes de la Universidad de ela) within the PhD Programme in Information Technology Research under the supervision of the researchers Alberto talá y Jose María Alonso.

pletely anonymous and will only be stored if the survey is fully completed. In this case, they will only be used for the ch.

ons about the study, you can address your questions to Mariña (marina.canabal.juanatey@usc.es).

open until 2 January 2023.

g boxes on the consent form, I confirm that:

pate in this study and may exercise my right to withdraw without prejudice at any time.
the data will be anonymised, and I consent to its anonymised processing and publication for the purpose of this g the General Data Protection Regulation (Regulation (EU) 2016/679 of the European Parliament and of the Council

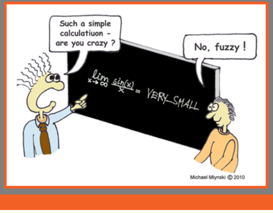
portunity to ask questions and all my doubts have been satisfactorily answered before starting the experimental tasks.
cked to continue.

unique Prolific ID to continue with the survey:

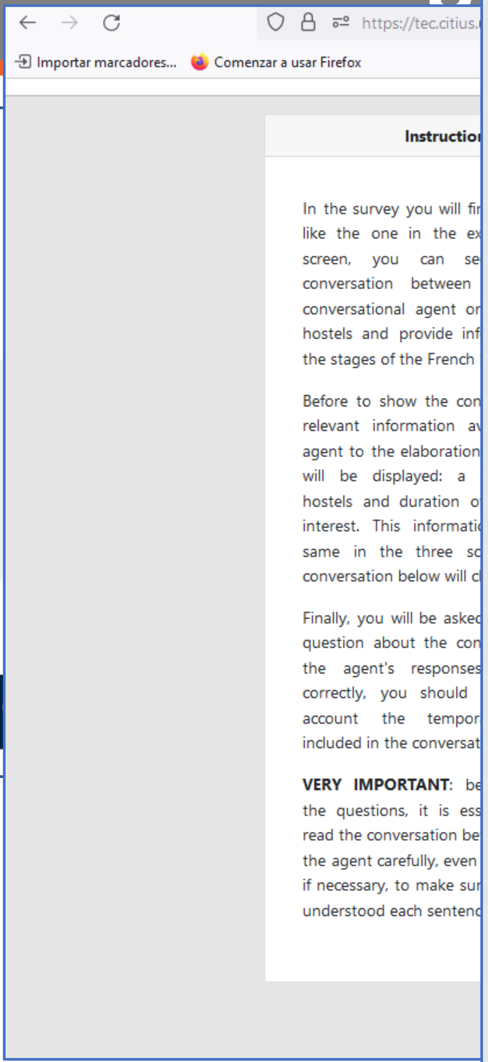
M. Canabal-Juanatey, J.M. Alonso-Moral, A. Catala, A. Bugarín-Diz, "**Enriching Interactive Explanations with Fuzzy Temporal Constraint Networks**", International Journal of Approximate Reasoning, 109128, 2024, https://doi.org/10.1016/j.ijar.2024.109128

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
CiTIUS Centro Singular de Investigación en Tecnoloxías Intelixentes

Evaluating LMs



https://tec.citius.usc.es/survey/



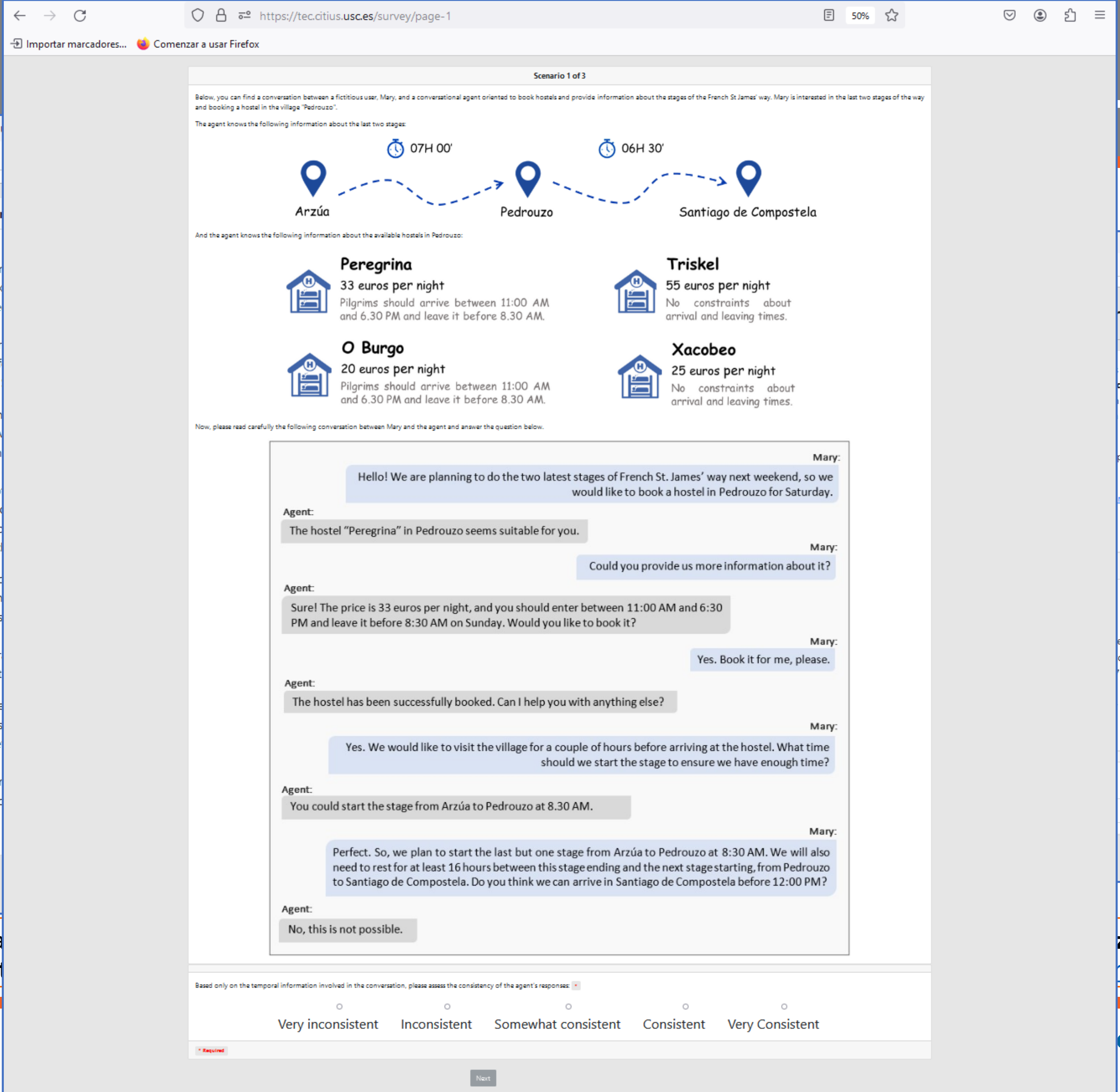
Instructio

In the survey you will fi
like the one in the ex
screen, you can se
conversation between
conversational agent or
hostels and provide inf
the stages of the French

Before to show the con
relevant information av
agent to the elaboration
will be displayed: a
hostels and duration o
interest. This informati
same in the three sc
conversation below will c

Finally, you will be asked
question about the con
the agent's responses
correctly, you should
account the tempor
included in the conversat

VERY IMPORTANT: be
the questions, it is ess
read the conversation be
the agent carefully, even
if necessary, to make su
understood each senten



Scenario 1 of 3

Below, you can find a conversation between a fictitious user, Mary, and a conversational agent oriented to book hostels and provide information about the stages of the French St James' way. Mary is interested in the last two stages of the way and booking a hostel in the village "Pedrouzo".

The agent knows the following information about the last two stages:

And the agent knows the following information about the available hostels in Pedrouzo:

Peregrina
33 euros per night
Pilgrims should arrive between 11:00 AM and 6.30 PM and leave it before 8.30 AM.

Triskel
55 euros per night
No constraints about arrival and leaving times.

O Burgo
20 euros per night
Pilgrims should arrive between 11:00 AM and 6.30 PM and leave it before 8.30 AM.

Xacobeo
25 euros per night
No constraints about arrival and leaving times.

Now, please read carefully the following conversation between Mary and the agent and answer the question below.

Mary:
Hello! We are planning to do the two latest stages of French St. James' way next weekend, so we would like to book a hostel in Pedrouzo for Saturday.

Agent:
The hostel "Peregrina" in Pedrouzo seems suitable for you.

Mary:
Could you provide us more information about it?

Agent:
Sure! The price is 33 euros per night, and you should enter between 11:00 AM and 6:30 PM and leave it before 8:30 AM on Sunday. Would you like to book it?

Mary:
Yes. Book it for me, please.

Agent:
The hostel has been successfully booked. Can I help you with anything else?

Mary:
Yes. We would like to visit the village for a couple of hours before arriving at the hostel. What time should we start the stage to ensure we have enough time?

Agent:
You could start the stage from Arzúa to Pedrouzo at 8.30 AM.

Mary:
Perfect. So, we plan to start the last but one stage from Arzúa to Pedrouzo at 8:30 AM. We will also need to rest for at least 16 hours between this stage ending and the next stage starting, from Pedrouzo to Santiago de Compostela. Do you think we can arrive in Santiago de Compostela before 12:00 PM?

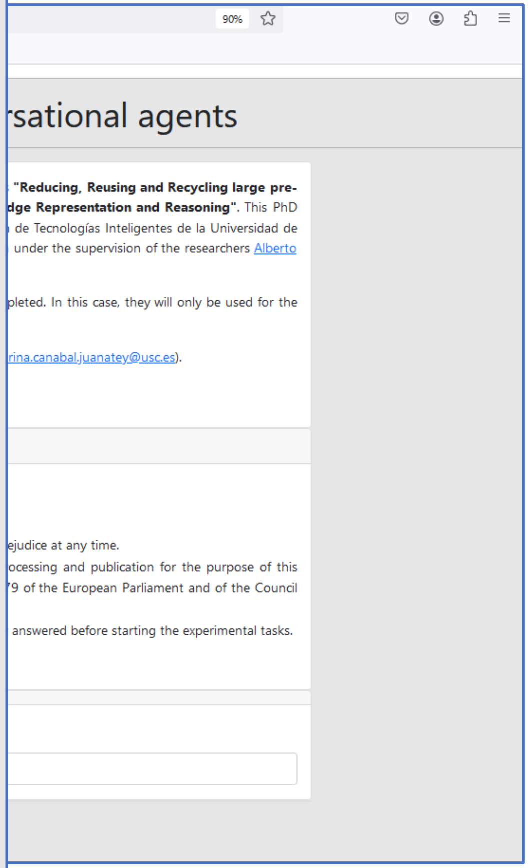
Agent:
No, this is not possible.

Based only on the temporal information involved in the conversation, please assess the consistency of the agent's responses: *

○ Very inconsistent ○ Inconsistent ○ Somewhat consistent ○ Consistent ○ Very Consistent

* Required

Next



rsational agents

"**Reducing, Reusing and Recycling large pre-**
dge Representation and Reasoning". This PhD
de Tecnologías Inteligentes de la Universidad de
under the supervision of the researchers Alberto

pleted. In this case, they will only be used for the

rina.canabal.juanatey@usc.es).

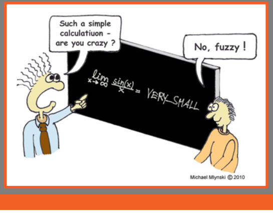
ejudice at any time.
ocessing and publication for the purpose of this
79 of the European Parliament and of the Council

answered before starting the experimental tasks.

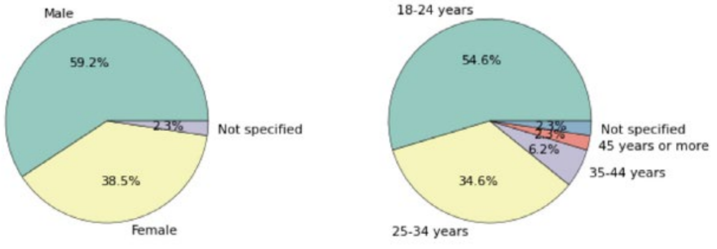
M. Canaba zzy Temporal
Constraint .2024.109128

orthy AI

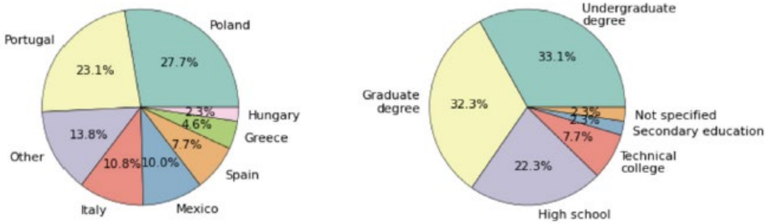
Evaluating the consistency of explanations by LLMs


Such a simple calculatiuon - are you crazy ?
No, fuzzy !
lim x→∞ sin(x)/x = VERY_SMALL
Michael Mlynski © 2010

130 subjects

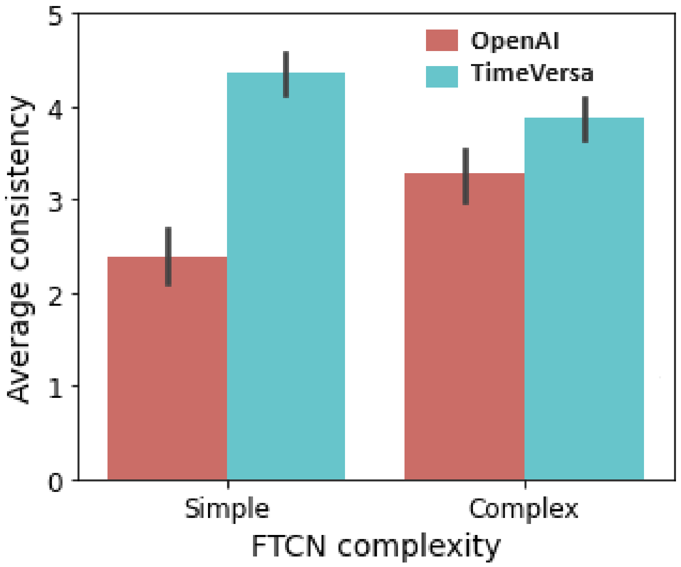


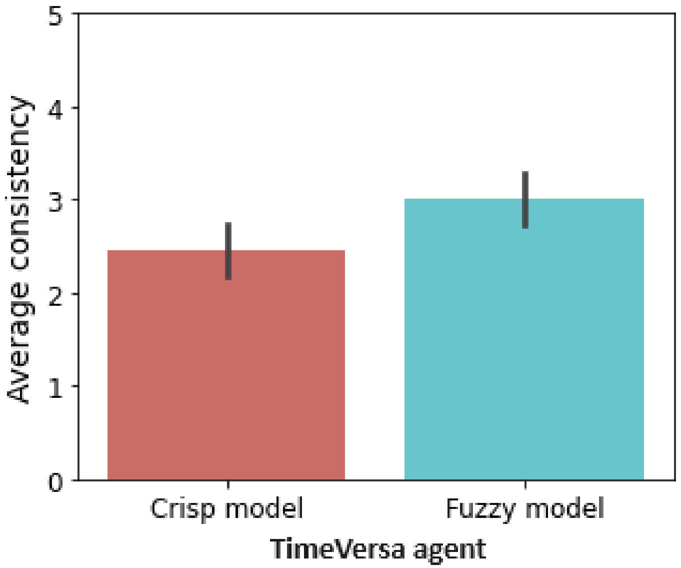
(a) Gender. (b) Age.



(c) Country of residence. (d) The highest education level completed.

Experimental design: two factors as independent variables at two levels each (2x2)

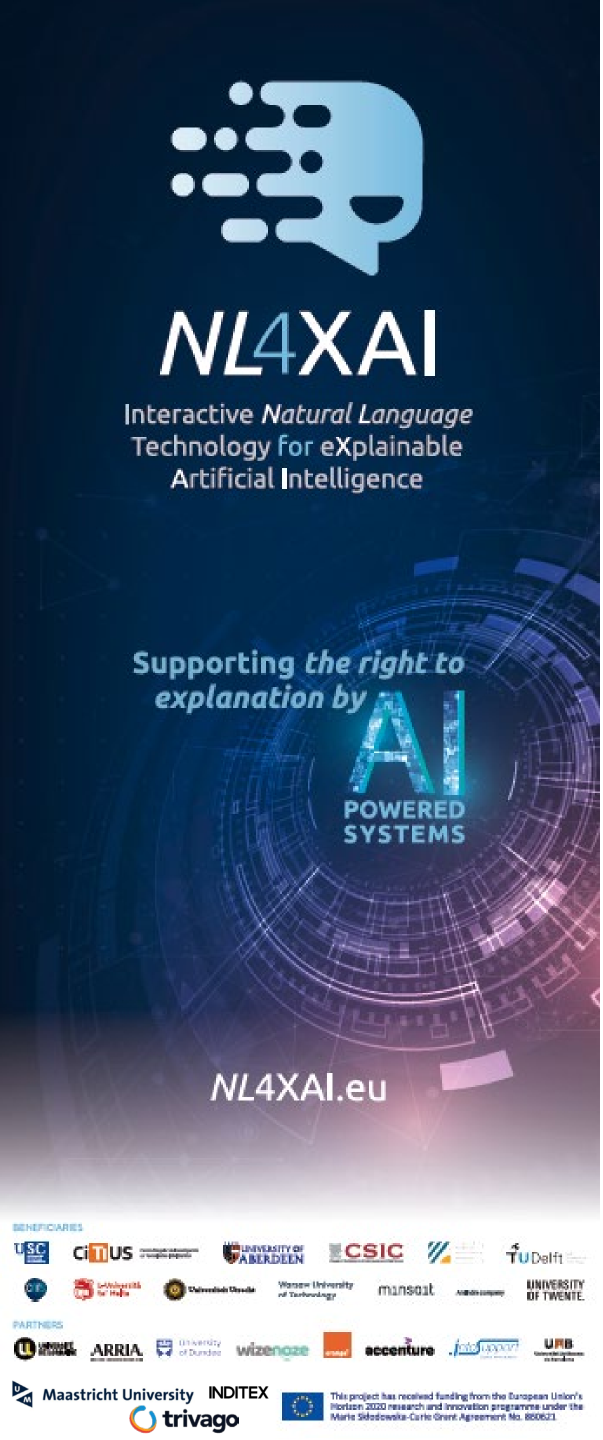




***H*1:** “In case of involving vague temporal information, the interaction with TimeVersa is perceived as more consistent than the interaction with the OpenAI API.”

***H*2:** “In case of involving vague temporal information, the interaction with TimeVersa incorporating the temporal reasoning model is perceived as more consistent than the interaction with TimeVersa incorporating the crisp version of the temporal reasoning model.”

M. Canabal-Juanatey, J.M. Alonso-Moral, A. Catala, A. Bugarín-Diz, "**Enriching Interactive Explanations with Fuzzy Temporal Constraint Networks**", International Journal of Approximate Reasoning, 109128, 2024, https://doi.org/10.1016/j.ijar.2024.109128


NL4XAI
Interactive Natural Language Technology for eXplainable Artificial Intelligence
Supporting the right to explanation by AI POWERED SYSTEMS
NL4XAI.eu
BENEFICIARIES
PARTNERS
Maastricht University
INDITEX
trivago

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 860621.

Jose Maria Alonso-Moral

https://orcid.org/
0000-0003-3673-421X

Profesor Titular de Universidad

GSi USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA CAMPUS VIDA CiTIUS

josemaria.alonso.moral@usc.es
https://citius.gal/team/jose-maria-alonso-moral
https://gitlab.citius.usc.es/jose.alonso/xai

International Ambassador

FM square

https://fmsquare.org/

IEEE Computational Intelligence Society

Task Force on Explainable Fuzzy Systems

https://sites.google.com/view/tf-explainable-fuzzy-systems/

Z-Inspection®

https://z-inspection.org/

XAI for Healthy Aging and Social Wellbeing

XAI 4 SOC

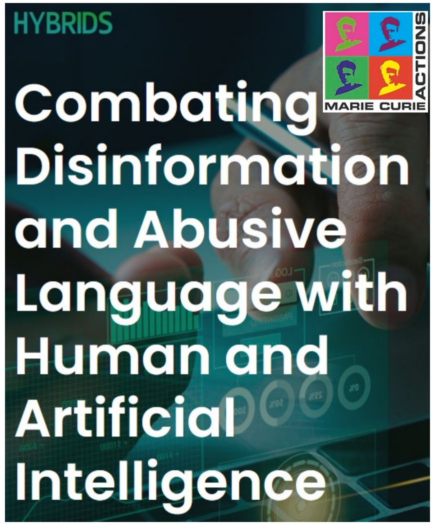
https://xai4soc.es/

TELSEC4TAI https://iafiable.es/

R&D Network on Trustworthy Artificial Intelligence: Technical, Ethical, Legal, Cultural and Socio-economic Challenges

DeepR3: Reducing, Reusing, and Recycling large models for developing Responsible and Green Language Technologies

http://ixa2.si.ehu.eus/deepr3/



Combating Disinformation and Abusive Language with Human and Artificial Intelligence

https://hybridsproject.eu/