

AI Explainability in the EU AI Act: A Case for an NLE Approach Towards Pragmatic Explanations

Anna Nicolis and Nigel Kingsman
Hughes Hall, University of Cambridge



© Anna Nicolis and Nigel Kingsman. This is an Open Access article distributed under the terms of the [Creative Commons Attribution Non-Commercial 4.0 License](https://creativecommons.org/licenses/by-nc/4.0/).

This paper navigates the implications of the emerging EU AI Act for artificial intelligence (AI) explainability, revealing challenges and opportunities. It reframes explainability from mere regulatory compliance with the Act to an organising principle that can drive user empowerment and compliance with broader EU regulations. The study's unique contribution lies in attempting to tackle the 'last mile' of AI explainability: conveying explanations from AI systems to users. Utilising explanatory pragmatism as the philosophical framework, it formulates pragmatic design principles for conveying "good explanations" through dialogue systems using natural language explanations. AI-powered robo-advising is used as a case study to assess the design principles, showcasing their potential benefits and limitations. The study acknowledges persisting challenges in the implementation of explainability standards and user trust, urging future researchers to empirically test the proposed principles.

Keywords: EU AI Act, Explainability, Explanatory Pragmatism, Natural Language Explanations, Robo-Advising

Introduction

This paper focuses on AI explainability under the EU AI Act, assessing its role in enhancing regulatory compliance, fostering a culture of ethical AI, and building trust with customers, thereby helping align AI practices with European societal values. Section 1 interrogates the concept of AI explainability under the EU AI Act; it is argued that the concept of explainability is tied to two key normative outcomes: user empowerment and regulatory compliance. Section 2 looks at the theoretical foundations of explanation and proposes the utility of the explanatory pragmatism framework based on a communicative view of explanation and an inferentialist view of understanding. Having laid the theoretical foundations, this section suggests five pragmatic principles of what a "good explanation" is. Section 3 introduces the concept of Natural Language Explanations (NLE) as a human-comprehensible verbalisation of the working of a model in natural language. Design principles for a dialogue system based on NLE are proposed. The final section applies the design principles for a good explanation and respective delivery mechanisms in a hypothetical Robo Advising Dialogue System (RADS) with model examples and test scenarios to evaluate the effectiveness of the design principles in practice. It is concluded that future work would comprise technical testing of

LLM/NLP to help drive application-building best practices.

1. EU AI Act: Meanings of Explainability

The EU AI Act is a central component of the EU's digital single market strategy, aiming to facilitate the efficient functioning of the internal market by establishing common regulations for the development, deployment and adoption of AI-powered products and services ([European Commission, 2021](https://ec.europa.eu/commission/presscorner/detail/en/ip21_1117)).

The cornerstone of the EU AI Act ("Act") is a risk-based taxonomy of AI systems; AI applications that are deemed to pose an unacceptable risk, such as state-driven social scoring, are prohibited outright. In contrast, AI systems with limited or minimal risk, such as spam filters, can enter the market with minimal new requirements. Title III constitutes the majority of the regulation, outlining prescriptive rules for so-called high-risk AI systems, which are defined as systems that pose significant threats to the health, safety or fundamental rights of natural persons.

This paper focuses on explainability, an aspect of the Act that involves not only a lack of prescriptive technical rules but also general definitional challenges, similar to its regulatory precedent constituted by the GDPR ([Casey et al., 2019](https://eur-lex.europa.eu/eli/reg/2016/679/oj); [GDPR, 2016](https://eur-lex.europa.eu/eli/reg/2016/679/oj)).

Although the Act does not contain an explicit mandate for “AI explainability” requirements, Recital 38 alludes to the concept of “explainable AI”, cautioning against potential negative impacts on individuals’ fundamental rights and power imbalances if AI systems lack sufficient transparency, explainability, and documentation.

The Act incorporates two essential articles that imply a degree of explanation for AI systems in use: Article 13, “Transparency and provision of information to users”, and Article 14, “Human oversight.”

Article 13 delineates three sub-articles with key requirements:

- Article 13(1) mandates that high-risk AI systems be designed with sufficient transparency to enable users to interpret the system’s output for appropriate use. The level of “transparency” functions as a condition to ensure compliance with other obligations specified in Chapter 3 of the regulation;
- Article 13(2) prescribes that high-risk AI systems must furnish users with relevant, accessible and comprehensible instructions;
- Article 13(3) further details the characteristics of information to be included in the user instructions outlined in Article 13(2), such as the system’s purpose, technical robustness, potential risks to health and safety or fundamental human rights, performance, training dataset, human oversight measures assisting users in interpreting the AI system’s output and lifetime and maintenance measures.

The notion of transparency is not value-neutral and has been subject to critical scrutiny (Ananny & Crawford, 2018). The term lacks a singular definition, and in the literature on explainability, “transparency” and “explainability” are often employed somewhat interchangeably (Grady, 2022).

Hayes’s (2020) notion of transparency is applied to interpret its meaning in the Act. Transparency is a condition that is conducive to knowledge acquisition about X, a phenomenon, an object or, in our case, an AI algorithm. Such a condition is defined by several properties:

- Availability: the information about X is documented;
- Accessibility: the information about X can be accessed by the agent who is looking to find it;
- Understandability: the agent is able to understand the information about X that is made available.

It is argued that, in general terms, transparency serves as an enabling principle for explainability. Without availability and access to information, generating and sharing explanations about AI systems would not be possible.

1.1 Explainability: User Empowerment and Regulatory Compliance

The rules delineated in Articles 13 and 14 of the Act offer a perspective on explainability that enhances the user’s agency and establishes conditions that foster adherence to other requirements within the Act. Specifically, transparency and, subsequently, explainability function as conditions that permit the AI system user to employ the system for its intended purpose while concurrently ensuring compliance with the Act’s regulatory mandates. Below, it is explained how the notion of explainability in the Act can be understood through the lenses of user empowerment and regulatory compliance.

Explainability and User Empowerment

The concept of explainability as a user-empowering condition is articulated in Article 13(1) and Recital 47 of the Act. These provisions mandate that high-risk AI systems function in a manner comprehensible to their users, providing them with sufficient information to interpret the system’s operations and utilise it appropriately.

Explainability and Regulatory Compliance

The notion of explainability as a condition for compliance is articulated in Article 13(2), which prescribes information transparency requirements “with a view to achieving compliance with the relevant obligations of the user and the provider.” This compliance-oriented objective is also evident in Article 11, which outlines technical documentation requirements, and Article 40, which establishes compliance as a necessary safety condition

aligned with technical standards (Sovrano *et al.*, 2021).

2. Explanatory Pragmatism

This section transitions from examining the role of explainability in the EU AI Act and turns the discussion to an assessment of frameworks for implementing explainability requirements. Design principles for generating explanations and dialogue systems are proposed, drawing on the philosophical framework of “explanatory pragmatism” (Nyrup & Robinson, 2022).

Numerous theories of explanation have been developed in philosophy, which are often influenced by fields such as psychology and linguistics. Prominent theories in contemporary philosophy, including causal realism (Salmon, 1984), constructive empiricism (Van Fraassen, 1980), ordinary language philosophy (Achinstein, 1983), cognitive science (Holland *et al.*, 1989), naturalism and scientific realism (Sellars, 1962), offer distinct definitions of “explanation”, sometimes in complementary ways. All theories, except for causal realism, are pragmatic, as they aim to make explanations specific and tailored for the individual recipients.

Most definitions incorporate the process of question-answering as an element of the act of explaining. This pragmatism can also be seen in the objective to adapt explanations to suit individual users, ensuring that the same explainable information is presented and reformulated in a unique manner for each user.

The features of explanation under the explanatory pragmatism framework (Nyrup & Robinson, 2022) are based on the following key aspects:

Communicative View of Explanation

Explanations are communicative acts where an explainer shares certain information with an audience to help them achieve relevant comprehension. This definition contains two key concepts. First, explanations should be regarded as speech acts (Austin, 1962) and thus evaluated based on their effectiveness in fulfilling their communicative function. Second, the primary communicative function of explanations is to facilitate the audience's understanding of the information transferred by the explainer (Franco, 2019).

Inferentialist View of Understanding

Understanding is a context-dependent concept (Kelp, 2015; Wilkenfeld, 2017). Merely acknowledging someone's ability to draw inferences is not enough to claim that the person has “understood” something. Building on this idea proposed by Nyrup and Robinson (2022), the conversation plays a crucial role in determining the class of inferences that are relevant for achieving that purpose. The role of dialogue as a means for conveying explanations is further set out in section 3.

Drawing on these two dimensions of the explanatory pragmatism framework, explainability is defined as:

“Explainability: in the conversational context, C, a given phenomenon (model, system, prediction, ...), P, is explainable by an explainer, S, to an audience, A, to the extent S is able to convey information to A that enables A to draw inferences about P that are needed to achieve the purposes that are salient in C.” (Nyrup & Robinson, 2022)

After establishing a theoretical foundation for explanation and explainability based on explanatory pragmatism, a set of design principles for implementing explanations in a practical business product context is introduced. It is attempted to translate the philosophical structure presented in the previous section into terms that product managers and business analysts can readily understand and apply when designing explainability frameworks for AI products.

2.1 Design Principles for a “Good Explanation”

An Explanation Should Be Factually Correct

The explanation should follow a robust technical framework, where the information included addresses the object of the explanation. The accuracy of the explanation will depend on the level of explanation provided. A local explanation relates to a specific output prediction from the model. A global explanation provides information relating to the higher-level workings of the system. Both need to be factually correct with respect to the operations of the AI model.

An Explanation Should Be Useful

The explanation should not only be factually accurate but also useful to the recipient of the information. The concept of utility is defined as a function of the actionable insights that are meaningful within the specific context in which the recipient of the explanation operates.

An Explanation Should Be Context-Specific

The concept of utility cannot be defined in absolute terms. Utility is always bound by the meaning that information has in a specific context. The context is described by the normative principles of a particular setting, the goals of that system and the constraints within which the user operates.

An Explanation Should Be User-Specific

The degree of utility and the context-bound nature of an explanation are always experienced from the positionality of a defined

user. In the context of explainability, the main features that define user specificity are the user’s degree of technical knowledge and role in relation to the working of the model (e.g. creator, user, regulator, auditor).

An Explanation Should Provide Pluralism

While an explanation should be adapted for the context and the audience, that does not mean that it should be limited to what the model is programmed to interpret as the intentions and desires of the audience. The explanation framework should begin with the objectives that are important within the given context and offer inferences that agents can utilise to achieve their objectives (Nyrup & Robinson, 2022, p. 6). This implies that the explanation allows for a diversity of normative perspectives rather than imposing a single normative stance on the intended recipient.

Table 1: *User Empowerment and Regulatory Compliance Outcomes Mapped to “Good Explanation” Principles.*

Principles	User Empowerment	Regulatory Compliance
1. An explanation should be factually correct	The information provided in the explanation can be empowering to the user only if it is correct and relevant to the product or service concerned. Incorrect information will be misleading to the user and may lead to detrimental outcomes	Information about the workings of the system must be correct in order to meet external audit and record-keeping requirements.
2. An explanation should be useful; 3. An explanation should be context-specific; 4. An explanation should be user-specific. ¹	Providing explanations that are relevant to the user and the context of use will be useful to the recipient of the explanation, allowing the user to act on the information provided and make decisions in an empowering way.	The concept of utility in relation to regulatory compliance can be described as a meta-outcome. If the information is presented in a way that is not meaningful, for example, as a disorganised collection of code and training data, it will not be deemed suitable for the purpose it is meant to serve. This principle regarding information clarity and utility is already widely adopted in financial services regulation for retail customers (ESMA, 2014).

¹ The concept of utility (principle 2) is strictly linked to the context (principle 3) and user specificity (principle 4). The utility of something is measured as a function of the outcomes delivered to a particular user in a specific context. For example, a technical explanation using scientific language and formulae will be of little utility to a layperson. Similarly, an explanation using plain English with a simplified version of the information will not benefit a technical auditor or specialist, but it will be highly relevant to a layperson.

<p>5. An explanation should provide pluralism</p>	<p>By starting with the purposes that are important in the given context, the explanation can be tailored to the user’s needs and preferences, empowering the user to make better decisions. Allowing for a range of normative views can also help users understand different perspectives and make more informed decisions.</p>	<p>The diversity of explanations is not tied to a specific regulatory outcome. However, it aligns with the mandate that the information provided to the user should be clear and not misleading, allowing users to make decisions that best suit their individual circumstances rather than prioritising the interests of the business entities supporting the AI system.</p>
---	--	---

3. NLE and Dialogue Systems

After introducing the conceptual framework for a good explanation based on explanatory pragmatism, a pragmatic solution for delivering explanations from AI systems to users is set out.

The current literature on explainable AI (XAI) primarily focuses on technical solutions and challenges related to interpreting AI models and creating explanations (Cambria *et al.*, 2023). However, it is also important to consider how to deliver the explanation to non-technical users after it is generated by the model.

In particular, the paper focuses on the Natural Language Explanations (NLE) framework. This explanatory framework provides a human-comprehensible verbalisation of the working of the model in natural language. This type of explanation can be generated in different ways, ranging from deep generative models to simple template-based frameworks. It is argued that NLE has the potential to enhance the user experience and foster trust in AI systems by using familiar language and presenting information in a natural way (Paek & Horvitz, 2000).

By making information more accessible and user-friendly, NLE can improve user understanding and trust in the system. The role of NLE is examined due to the growing prominence of large language models (LLMs) as the default interface for users interacting with AI systems.

As academic resources and definitions of this approach are scarce, a generalised understanding of a “dialogue system” is used, encompassing conversational agents and chatbots (Lakkaraju *et al.*, 2022). The precedent for using dialogue systems for NLE can be traced back to expert systems, a category of symbolic

AI that emerged around the mid-1960s. These expert systems were based on the principle of transferring specific human expertise into a computer. This transferred knowledge enabled the computer to offer advice as needed, similar to a human advisor, and, if necessary, to clarify the reasoning behind its suggestions. However, the application of expert systems was restricted due to various overarching AI challenges, including issues related to knowledge representation, generalisation and learning (Liao, 2005).

A dialogue system can help implement the design guidelines articulated as part of the explanatory pragmatism framework by emulating the way humans typically convey explanations and share knowledge. Rather than offering a rigid, one-directional output that the user must merely accept, a dialogue system allows users to interact with the model using their own language. A dialogue system can improve contextual comprehension and promote user trust in the system.

3.1 Design Principles for Dialogue Systems

After establishing the design principles for a good explanation, the list below sets out the design principles and components for a dialogue system focussed on delivering user explanations in a business product context. (Lakkaraju *et al.*, 2022, p. 7).

Proposed design principles for a dialogue system

1. The user should be able to prompt the dialogue system using natural language, eliminating the requirement to write any code to interact with the system.
2. The system should be able to understand ongoing user requests and associate

them with appropriate explanations to be generated and presented to the user.

3. The system should understand the context of the questions and adapt the explanations accordingly.
4. The system should build on previous inputs and clarify/rephrase anything generated by a previous prompt or prompts.
5. The system should state when it does not have an answer due to a lack of information or because it does not understand the question instead of trying to persuade the user.
6. The system should provide a confidence level of its explanation.
7. The system should be near real-time to simulate a human verbal dialogue.

3.2 Key Benefits of Dialogue Systems

By enabling users to interact with the system using natural language (1), the dialogue system reduces the need for technical expertise and is accessible to a wider range of users, reducing audience comprehension issues. The ability of the system to understand ongoing user requests and associate them with appropriate explanations (2) aids in the delivery of relevant and coherent information, addressing the challenge of audience comprehension, particularly with complex models. By understanding the context of questions and adapting explanations accordingly (3), the dialogue system helps improve domain knowledge, allowing users to make meaningful inferences in their specific fields of interest. This implies that the user is being educated with new knowledge, or at least guided towards it, which adds an innovative dimension to XAI. Instead of treating the audience as a static receiver that the XAI system needs to accommodate, the system could actively improve their comprehension and use of explanations.

By building on previous inputs and offering clarification or rephrasing when needed (4), the system promotes better understanding, addressing both semantic mapping and audience comprehension issues. The system's capability to state when it lacks information or

does not understand a question (5) rather than misleading users addresses the truthfulness of information, thereby promoting trust between users and the system.

By providing a confidence level for its explanations (6), the dialogue system empowers users to assess the reliability of the information and make informed decisions, which further addresses the audience's ability to comprehend the information and its relevance to their domain knowledge. Lastly, a near-real-time dialogue system (7) simulates human verbal interactions, creating a seamless user experience.

Key conceptual components required to build the dialogue system are laid out below (Lakkaraju *et al.*, 2022, p. 2).

3.3 Core Components of a Dialogue System

1. **Graphical User Interface (GUI):** The GUI is the visual interface designed to facilitate user interaction with the dialogue system. It should be user-friendly and intuitive, making it easy for users to enter their input via text and receive the system's output in a clear and understandable format. The GUI may incorporate elements such as text boxes, buttons or visual cues to guide users through the conversation.
2. **Natural Language Understanding (NLU):** This component is responsible for processing and interpreting user input in natural language. NLU involves parsing the input text, identifying its structure and extracting relevant information, such as keywords, intents and entities. By doing so, the dialogue system can comprehend the user's query or request and determine the appropriate response or action.
3. **Explanation Mapping:** This component is responsible for connecting the user's input to the appropriate explanation. It may involve querying a knowledge base, database or another AI model to retrieve the necessary information to generate the explanation.
4. **Explanation Generation:** The explanation generation component is

responsible for creating human-readable explanations based on the retrieved information from the mapping component. Using NLE, it produces explanations that are both context-specific and comprehensible to the user based on the language of the initial query. The generated explanation is then returned to the user through the GUI.

In real-world applications, these capabilities can be implemented as separate components or as an end-to-end model. For instance, components 2 and 4 could be integrated within a single model, while component 3 might be part of the same model or a separate component. ChatGPT's plug-ins exemplify this approach, where a text model interacts with another system or model through model prompting (OpenAI, 2023).

4. Robo-Advising Case Study

This section evaluates the significance of explainability requirements by focussing on a specific case study within the financial services domain: robo-advising, which involves automated, algorithm-driven financial planning services in contrast to a human wealth manager. At the time of conducting this research project, robo-advising is not explicitly mentioned in Annex III of high-risk AI products, which outlines mandatory compliance with the Act's requirements for high-risk products, such as Article 13 and Article 14 (European Commission, 2021). A narrow interpretation of the regulation would imply that financial services firms offering robo-advising products are not required to address the explainability requirements described in Section 1.

However, the focus on robo-advising underscores the crux of the paper's argument: addressing explainability is not an end goal in itself tied to minimal regulatory compliance with the EU AI Act but rather an organising principle that can help achieve other normative outcomes.

Despite a boom in Fintech services, robo-advising accounts for less than 1% of the overall assets under management in the EU (ESMA, 2023). The ESMA TRV risk analysis (2023) highlights that the lack of explainability provisions offered as part of the robo-advising

products impacts consumer trust. Furthermore, the report suggests that fund management companies may be discouraged from advertising their businesses' use of AI in the investment process to avoid attracting regulatory attention and potential reputational risk. Firms may also be reluctant due to their use of AI in more limited contexts (e.g. not directly affecting investment strategy) and the murky definitions surrounding it.

By going beyond minimal regulatory compliance, robo-advising firms can develop tools that future-proof their products and services as regulations continue to evolve.

Furthermore, there are systemic benefits to consider. Without consumer trust enabled by improved explainability, financial services providers may hesitate to invest in innovation related to robo-advising capabilities. Consequently, the democratising potential of such products, which aim to make financial advice more accessible and affordable, may not be fully realised (Bianchi & Briere, 2021, p. 20).

In the final section it is sought to set the foundations for a plan to empirically test AI explainability via dialogue systems using LLMs and an explanation generation mode within robo-advising.

Hypothetical dialogues were created between users and the Robo Advising Dialogue System (RADS), each demonstrating how the system could fulfil the design principles for effective explanations. These can be found in the appendix to this paper.

4.1 RADS User Vignettes

Model user vignettes have been crafted into hypothetical dialogues, assessing user empowerment and regulatory compliance to define what constitutes a good explanation.

Three distinct user profiles—a retail consumer, a data scientist, and an ESMA regulator—represent stakeholders typically interfacing with AI explanation systems. Their diverse expertise ranges from non-technical to highly specialised, and their expectations from AI systems differ correspondingly, from simple usability to stringent compliance with technical and regulatory benchmarks.

The dialogues exemplify how RADS might handle inquiries, varying from basic consumer questions about credit scores to a data scientist's technical scrutiny of prediction models and an ESMA regulator's audit on privacy metrics.

The retail consumer dialogue illustrates how RADS provides factually correct, context-specific, and user-tailored explanations without requiring technical knowledge from the user. This approach not only empowers users by demystifying financial recommendations but also aligns with regulatory standards by ensuring clarity and utility of information as mandated in financial regulations.

For the data scientist user vignette, RADS offers technical data on forecast errors, demonstrating its capacity to communicate complex information clearly and transparently, thus facilitating expert users to assess the model's performance critically. Such interactions are key to ensuring that AI systems remain under

informed human supervision, a core requirement of regulatory compliance.

The dialogue with the ESMA regulator presents a contrasting scenario where RADS fails to meet the design principles. When asked to translate technical details into plain English, RADS responds with an oversimplified answer that neither empowers the regulator with actionable information nor complies with the necessary regulatory transparency and insight.

4.2 Analysis of RADS Vignettes

Each dialogue scenario is assessed against the dual criteria of user empowerment and regulatory compliance, with a scoring mechanism evaluating the “goodness” of the explanation provided. The assessment gauges whether the explanations enable users to make informed decisions (user empowerment) and whether they adhere to the transparency and information quality required by regulators (regulatory compliance).

Table 2: Model User Vignettes Mapped Against Good Explanation Principles, User Empowerment and Regulatory Compliance Outcomes.

Vignette	Adherence to Good Explanation Principles 1-5	Adherence to User Empowerment	Adherence to Regulatory Compliance
1. Retail consumer	<ul style="list-style-type: none"> RADS converses in natural language (Principle 1) Provides targeted responses to each user's context (Principles 2 and 3) Ensures a coherent, ongoing dialogue (Principle 4) Presents pluralistic, open-ended explanations, recognising its limitations and fostering user autonomy (Principle 5) 	<ul style="list-style-type: none"> RADS enhances user experience with natural language responses and personalised insights, like advising on investment changes after a pay rise It is transparent about its limits, encouraging informed user decisions Unlike static Terms and Conditions, RADS supports financial literacy with dynamic, contextual dialogue, 	<ul style="list-style-type: none"> RADS meets Article 13(1) and 13(2) of the Act by providing transparent, tailored, and easily understandable information, enabling effective use of system outputs It also aligns with the ESMA Final Report guidelines by clarifying portfolio allocation processes, exemplifying how enhanced explainability can exceed basic regulatory expectations

		fostering trust through transparency and avoiding false certainty	
2. Data scientist dialogue	<ul style="list-style-type: none"> • Data scientist converses with RADS in technical language • RADS effectively addresses inquiries about error forecasting (Principles 2 and 3), and • Offers options like MSE, MAE, and MAPE metrics (Principle 1) • Once a metric is chosen, it details the forecast error percentage (Principle 4), and • Provides thorough global and local explanations for troubleshooting (Principle 5) 	<ul style="list-style-type: none"> • RADS facilitates a transparent QA process by providing the data scientist with pertinent information about forecasting error rates • It adapts to the technical language used by the user • Empowers the user to supervise the model's performance and address any shortcomings 	<ul style="list-style-type: none"> • Dialogue demonstrates how the system can help meet the requirements under Article 14 of the Act concerning human oversight • For instance, if the average deviation rate of 12.75% from the actual value, as determined by the MAPE for the specified period, exceeds the targeted deviation rate, the user may further investigate the relevant information and the model components that may impact this calculation
3. ESMA regulatory dialogue	<ul style="list-style-type: none"> • An ESMA regulator audits model compliance, but RADS inadequately addresses their plain English request for privacy and fairness metrics, violating principles • Oversimplifies, lacking detail and context (violations of Principles 2 and 3) • Fails to build on previous non-technical queries (Principle 4) • Provides general explanations without meaningful insights, not admitting its own limitations (Principle 5) 	<ul style="list-style-type: none"> • RADS dialogue lacks adaptability and detail, failing to provide information tailored to the regulator's needs • The inefficiency hampers the audit process, forcing the representative to seek further clarification elsewhere • RADS does not meet the requirements for providing accessible and specific information necessary for effective regulatory oversight. 	<ul style="list-style-type: none"> • RADS aims to provide detailed explanations of its operations and metrics to regulators, aligning with regulatory compliance goals • However, its failure to understand the regulator's knowledge level hinders this aim • RADS does not offer the necessary access and transparency, impairing effective oversight as mandated by Article 14 of the Act

For practical application, these hypothetical dialogues serve as benchmarks for developing

and refining dialogue systems. They help identify where the system performs well and

where improvements are needed to align with the stated principles of effective explanation. By including these vignettes as part of a training dataset, researchers and developers can enhance the dialogue system's ability to provide more nuanced and contextually appropriate explanations.

5. Limitations

Inherent risks and limitations of the proposed approach are acknowledged. Technical challenges include LLMs' handling of complex queries, generating coherent and relevant explanations, discerning context, retaining information, ensuring truthfulness, assessing confidence levels, and delivering timely responses.

To address these, multi-disciplinary approaches should be considered. Confidence scoring mechanisms can help monitor performance; reinforcement learning from human feedback (RLHF) can adjust model accuracy; active learning could refine data handling. Regular expert reviews are necessary to evaluate and improve the system.

Beyond technical challenges, dialogue systems risk fostering overreliance and a false sense of assurance in users. There's a delicate balance between building trust in AI and preventing uncritical acceptance of AI decisions, especially in sensitive fields like financial services. Privacy remains a critical ethical concern, with the need for stringent protocols for data handling.

6. Conclusion and Future Work

This research evaluates the EU AI Act's implications for AI explainability, underscoring explainability as not only a regulatory need but as an ethical AI guide fostering user empowerment and broad regulatory compliance outcomes. This study particularly focuses on the 'last mile' problem of explainability, presenting design principles derived from explanatory pragmatism for effectively communicating AI explanations to users. These principles aim to translate academic insights into actionable strategies for businesses and product developers.

A case study on AI-driven robo-advisory services exemplifies applying these principles,

suggesting benefits beyond compliance, like competitive edge and consumer trust.

Acknowledging the challenges in implementing explainability through dialogue systems and potential user overreliance, the paper calls for empirical testing of the proposed principles in practice.

Whilst the research presented in this paper proposed the seven design principles to create an NLE Dialogue System and showcase the conformance of case study examples against those design principles, it stopped short of investigating the feasibility of applying the principles in a scalable, practicable manner. Future research would comprise building on top of the rich research in the Large Language Model and Natural Language Processing space to understand whether such technologies could deliver robust real-time scoring of dialogue systems' conformance to the principles, which in turn can inform the real-time assessment of dialogue systems' ability to suitably empower their users and to adhere to regulations. Such research, if successful, could help drive best practices across a range of applications, including, but not limited to, robo-advisory dialogue systems.

Appendix

Vignette 1: Retail Customer to RADS – User Clarifications about Credit Score, Weightings and Portfolio Allocations

Retail Customer: *I have a credit score of 350, which I know is considered a high-risk profile. Does this impact my portfolio allocation?*

RADS: *Yes, your credit score has an impact on your portfolio allocation. A low credit score, such as 350, may indicate higher financial risk, and my recommendations would likely be adjusted to account for this factor. I would typically recommend a more conservative investment strategy, focussing on lower-risk assets and diversification to help address the financial risks associated with a lower credit score. It is important to note that other factors, like your financial goals, risk tolerance and investment horizon, will also play a role in creating your portfolio allocation.*

Retail Customer: *What weighting does my credit score have in relation to my annual income, £87k per year, in my portfolio allocation?*

RADS: *In determining your portfolio allocation, the system accounts for multiple factors, including your credit score, annual income, risk tolerance, financial goals and investment horizon. While I cannot provide an exact weighting for each factor, as the model dynamically adjusts based on the information provided, I can assure you that both your credit score and annual income are considered when generating recommendations.*

Your credit score may influence the risk profile of your portfolio, while your annual income can help determine your investment capacity and how aggressively or conservatively you might want to invest. The combination of these factors, along with other personal details, enables me to create a personalised and well-rounded investment strategy tailored to your specific needs and financial situation.

Retail Customer: *Next year I am expecting a 15% raise in my base salary. Can you show me how my current portfolio allocation would change on that basis?*

RADS: *Congratulations on your expected raise! I can certainly help you understand how a 15% increase in your base salary might impact your portfolio allocation.*

With a higher annual income, you may have an increased capacity for investment, which could lead to several adjustments in your portfolio allocation:

- *Increased diversification: With more funds available for investment, you might consider diversifying your portfolio across more asset classes, sectors or geographic regions.*
- *Risk tolerance adjustment: Depending on your financial goals and risk tolerance, a higher salary might enable you to take on more risk in your investments, potentially leading to a higher allocation in equities and other higher-risk assets.*
- *Rebalance based on goals: If your financial goals have changed due to your increased income, your portfolio allocation might need to be adjusted accordingly. This could involve revisiting your target allocations*

and ensuring they align with your updated financial objectives.

Vignette 2: Data Scientist to RADS Error Metrics and Technical Robustness Check

Data Scientist: *Can you provide information on the forecast errors for the past seven days? I want to know how well your predictions have performed.*

RADS: *I can provide you with forecast error metrics for the past seven days. However, I need to know which specific metric you're interested in, such as mean squared error (MSE), mean absolute error (MAE) or mean absolute percentage error (MAPE).*

Data Scientist: *I want to start with MAPE for the past seven days. Can you also explain how this metric helps evaluate forecast accuracy?*

RADS: *MAPE is an accuracy metric that measures the average percentage error between the actual values and the forecasted values. It is calculated by taking the average of the absolute percentage errors across all predictions. Lower MAPE values indicate better forecasting accuracy.*

For the past seven days, the MAPE of our predictions is 12.75%. This means that, on average, our forecasts deviated by 12.75% from the actual values.

Data Scientist: *How reliable are these forecast error metrics in determining the overall accuracy and robustness of your system?*

RADS: *Forecast error metrics like MAPE provide valuable insights into the accuracy of predictions. However, it's important to acknowledge their limitations. For example, MAPE may be less informative when dealing with small values, as the percentage error can be disproportionately large. It's best to consider multiple error metrics to gain a comprehensive understanding of the system's overall accuracy and robustness.*

Data Scientist: *Which prediction was the largest contributor to this error?*

RADS: *The largest residual error occurred with security X. The model predicted a 5% increase, but in reality, security X saw an 11% decrease.*

Data Scientist: *Interesting. Could you provide some context as to why the model forecasted this increase for security X?*

RADS: *The model's forecast for security X was largely influenced by the lagged behaviour of security Y. Historically, there has been a tendency for increases in security X to follow increases in security Y with a one-day lag. The feature 'security_y_lag_1_day' had the highest SHAP value for this prediction, indicating its significant contribution to the model's forecast.*

Vignette 3: ESMA Official to RADS – Audit on Privacy Metrics

ESMA Representative: *How do you ensure that your AI model does not inadvertently learn any patterns or biases that could lead to privacy breaches or unfair treatment of certain clients?*

RADS: *The model is designed with fairness and privacy as key considerations. We use techniques such as differential privacy, which introduces a controlled amount of noise into the training data, to prevent the model from overfitting or memorising specific client details. Additionally, we perform regular audits of the model's outputs to identify and mitigate any potential biases or privacy concerns.*

ESMA Representative: *Can you provide some specific metric examples or parameters that you use for differential privacy to ensure the protection of client data during the model training phase?*

RADS:

```
import numpy as np
```

```
from diffprivlib.models import LogisticRegression
```

```
# Set privacy budget
```

```
epsilon = np.log(2)
```

```
# Initialise a differentially private Logistic Regression model
```

```
model = LogisticRegression(epsilon=epsilon, data_norm=1)
```

ESMA Representative: *I am not a technical user. Can you explain this in plain English?*

RADS: *We use complex mathematical algorithms to balance privacy and model effectiveness.*

References

Achinstein, P. (1983). *The Nature of Explanation* (1st ed.). Oxford University Press.

Austin, J. L. (1962). *How to do things with words*. Clarendon Press.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), pp. 973–989. <https://journals.sagepub.com/doi/10.1177/1461444816676645>

Bianchi, M., & Briere, M. (2021). *Augmenting investment decisions with robo-advice* (Report No. 3751620). Université Paris-Dauphine. <https://doi.org/10.2139/ssrn.3751620>

Cambria, E., Malandri, L., Mercurio, F., Mezzanzanica, M., & Nobani, N. (2023). A survey on XAI and natural language explanations. *Information Processing & Management*, 60(1), 103111. <https://www.sciencedirect.com/science/article/pii/S0306457322002126?via%3Dihub>

Casey, B., Farhangi, A., & Vogl, R. (2019). Rethinking explainable machines: The GDPR's right to explanation debate and the rise of algorithmic audits in enterprise. *Berkeley Technology Law Journal*, 34(1), p. 143. <https://doi.org/10.15779/Z38M32N986>

Council of the EU. (2022, December 6). Artificial Intelligence Act: Council calls for promoting safe AI that respects fundamental rights. <https://www.consilium.europa.eu/en/press/press-releases/2022/12/06/artificial-intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental-rights/>

European Commission. (2021). Proposal for a regulation of the European Parliament and the Council laying down harmonised rules on artificial intelligence and amending certain

Union legislative acts. <https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=celex%3A52021PC0206>

European Parliament. (2023). AI Act: A step closer to the first rules on Artificial Intelligence [Press release]. <https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence#:~:text=The%20new%20law%20promotes%20regulatory,that%20significantly%20impact%20their%20rights.>

European Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 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 and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L 119/1. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>

European Parliament and Council of the European Union. (2022). 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 (recast). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32014L0065>

European Securities and Markets Authority (ESMA). (2014). Enhanced protection for retail investors: MiFID II and MiFIR. https://www.esma.europa.eu/sites/default/files/library/2015/11/2014-726_enhanced_protection_for_retail_investors_-_mifid_ii_and_mifir.pdf

European Securities and Markets Authority (ESMA). (2022). Final report. https://www.esma.europa.eu/sites/default/files/library/esma35-43-3172_final_report_on_mifid_ii_guidelines_on_suitability.pdf

European Securities and Markets Authority (ESMA). (2023). TRV risk analysis 1 February 2023: Financial innovation – Artificial intelligence in EU securities markets. <https://www.esma.europa.eu/sites/default/fil>

[es/library/ESMA50-164-6247-AI_in_securities_markets.pdf](https://www.esma.europa.eu/sites/default/files/library/ESMA50-164-6247-AI_in_securities_markets.pdf)

Franco, P. L. (2019). Speech act theory and the multiple aims of science. *Philosophy of Science*, 86, pp. 1005–1015. <https://doi.org/10.1086/705452>

Grady, P. (2022). The EU should clarify the distinction between explainability and interpretability in the AI Act. <https://datainnovation.org/2022/08/the-eu-should-clarify-the-distinction-between-explainability-and-interpretability-in-the-ai-act/>

Hayes, P. (2020). An ethical intuitionist account of transparency of algorithms and its gradations. *Business Research*, 13(3), pp. 849–874. https://ideas.repec.org/a/spr/busres/v13y2020i3d10.1007_s40685-020-00138-6.html#refs

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P.R. (1989). *Induction: Processes of Inference, Learning, and Discovery* (1st ed.). MIT Press.

Kelp, C. (2015). Understanding phenomena. *Synthese*, 192, pp. 3799–3816. <https://doi.org/10.1007/s11229-014-0616-x>

Lakkaraju, H., Slack, D., Chen, Y., Tan, C., & Singh, S. (2022). Rethinking explainability as a dialogue: A practitioner’s perspective. <https://arxiv.org/abs/2202.01875>

Liao, S. (2005). Expert system methodologies and applications – A decade review from 1995 to 2004. *Expert Systems with Applications*, 28, pp. 93–103. <https://www.sciencedirect.com/science/article/abs/pii/S0957417404000934>

Nyrup, R., & Robinson, D. (2022). Explanatory pragmatism: A context-sensitive framework for explainable medical AI. *Ethics and Information Technology*, 24(13). <https://doi.org/10.1007/s10676-022-09632-3>

OpenAI. (2023). Language models can explain neurons in language models. <https://openai.com/research/language-models-can-explain-neurons-in-language-models>

Paek, T., & Horvitz, E. (2000). Conversation as action under uncertainty. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI'00)*, pp. 455- 464. San Francisco, CA.

Saeed W., & Omlin C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263, 110273 <https://www.sciencedirect.com/science/article/pii/S0950705123000230#section-cited-by>

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World* (1st ed.). Princeton University Press.

Sellars, W. S. (1962). Philosophy and the scientific image of man. In R. Colodny (Ed.),

Science, Perception, and Reality, pp. 35–78. Humanities Press/Ridgeview.

Sovrano, F., Vitali, F., & Palmirani, M. (2021). Making things explainable vs explaining: Requirements and challenges under the GDPR. <https://arxiv.org/abs/2110.00758>

Van Fraassen, B. C. (1980). *The Scientific Image* (1st ed.). Oxford University Press.

Wilkenfeld, D. (2017). MUDdy understanding. *Synthese*, 194, pp. 1273–1293. <https://doi.org/10.1007/s11229-015-0992-x>