The Wildcard XAI: from a Necessity, to a Resource, to a Dangerous Decoy.

Rachele Carli[1,2][0000−00020−8689−285X] and Davide Calvaresi[4][0000−0001−9816−7439]

[1] Alma Mater Research Institute for Human-Centered AI, University of Bologna, Italy rachele.carli2@unibo.it
[2] CLAIM Group and AI RoboLab University of Luxembourg, Luxembourg
[3] University of Applied Sciences Western Switzerland, Switzerland
davide.calvaresi@hevs.ch

Abstract. There has been a growing interest in Explainable Artificial Intelligence (henceforth XAI) models among researchers and AI programmers in recent years. Indeed, the development of highly interactive technologies that can collaborate closely with users has made explainability a necessity. This intends to reduce mistrust and the sense of unpredictability that AI can create, especially among non-experts. Moreover, the potential of XAI as a valuable resource has been recognized, considering that it can make intelligent systems more user-friendly and reduce the negative impact of black box systems. Building on such considerations, the paper discusses the potential dangers of large language models (LLMs) that generate explanations to support the outcomes produced. While these models may give users the illusion of control over the system's responses, they actually have persuasive and non-explanatory effects. Therefore, it is argued here that XAI, appropriately regulated, should be a resource to empower users of AI systems. Any other apparent explanations should be reported to avoid misleading and circumventing effects.

Keywords: XAI · Anthropomorphism · Dependency

1 Introduction

Rationality has long been considered a fundamental trait of human beings [58], not only distinguishing them from other forms of life but also serving as a basis for classifying individuals based on their ability to make conscious decisions, particularly in legal contexts [3]. However, numerous studies have shown that the concept of rationality is often a myth [70,29]. The mechanisms by which individuals acquire and analyze information, experience reality, and form opinions about events and phenomena are largely the result of subconscious processes in which emotions and sensations play a decisive role [43].

It is not surprising that these studies have become more prevalent or relevant, particularly in light of the ongoing discussion surrounding the relationship

between humans and AI systems. This emerging field of research has emphasized that despite being aware of the difference between human and artificial entities and the ability to differentiate between the two, individuals tend to engage emotional mechanisms and form attachments with AI that are more suited to living beings [43,63,7]. Consequently, additional mechanisms have been developed to facilitate the activation of rational processes of experience and information processing by subjects when interacting with new technologies.

With respect to that, XAI is a field of research that may help counterbalance the inherent lack of total rationality in individuals [23]. Its goal is to make the mechanisms underlying the functioning of data-driven predicting systems limit symbolic interpretations.

Therefore, it is possible to believe that XAI originated as a necessity to provide protection for non-experts interacting with opaque and complex technologies. To this end, XAI models also serve as a resource for opening the decision-making mechanisms in various sensitive fields of everyday life, such as health, finance, decision-making, and education [2]. Hence, thanks to the explanations, exploiting the positive aspects of technological innovation in these areas would have been possible, modulating exposure to risks.

However, XAI is still an evolving discipline with ambiguous boundaries and insufficient development. These hinder its ability to achieve its underlying goals, making it a potentially dangerous decoy. In fact, in some cases, complex explanations have been replaced with what hereafter is described as "persuasive justification" made by LLMs, which are easier to implement but may not be as effective.

This paper highlights how these models have undergone increasing development in recent years, supported primarily by what they are designed to appear like. Despite their advertised capabilities and design features, these models are not comparable to human intelligence, thought processes, or cognition [79]. Therefore, the analysis here developed argues that the lack of proper explanations produced by LLMs have the potential to increase the risk exposure for users. Thus, we claim a need for a renewed focus on responsible XAI development through the joint efforts of researchers, developers, and European regulation.

The rest of the paper is organized as follows.

Section 2 introduces the theme of anthropomorphism and its immanence in human nature. Thus, Section 3 outlines the perception of AI systems as social agents and their consequences. This introduces the centrality of the theme of XAI, which is illustrated here by highlighting its relevance and profiles of current ambiguity in Section 4. Section 5 introduces the recently developed alternative of using LLMs to perform the functions of XAI, highlighting its limitations and risks against the advertised advantages. Section 6 states the paper's claim to the responsible development of XAI, which contributes to user empowerment and limits possible harm to the integrity and fundamental rights. Section 7 concludes the paper.

2 Anthropomorphism and the Human Nature

Anthropomorphism is defined as the attribution of human characteristics to inanimate objects or phenomena [31].

While earlier theories considered it a bias common among fragile or still cognitively immature subjects [22], it is now recognized as an essential and intrinsic component of the human mind [41]. It has been argued that it cannot be completely eliminated [14] as it is reflected in many aspects of our psyche and nature. One example is humans' innate tendency to feel trust, regardless of personal individuality. Contemporary psychology posits that this may be an unconscious need to feel vulnerable to others and to rely on them rather than being entirely self-sufficient [19].

Such a perspective aligns with Aristotle's observation that humans are "social animals" driven to aggregation. Consequently, if this natural inclination were absent, no interaction would be possible, even those essential for a civilized society. Freud elaborated on the Aristotelian concept, defining human beings as "symbolic animals" who are led to conceptualize material reality through symbolic interpretation [17]. It implies that the way people experience reality and the data on which they base their emotional responses are never perfectly rational and objective. In fact, there is always a certain and variable degree of interpretation, which is not only due to the senses through which we know but also to the cognitive and psychic structures with which we are endowed [43].

Furthermore, the human brain is naturally inclined to resist change and attempt to revert thoughts, behavior, or habits to those known or long-established [19]. This occurs because changing routines implies not only an expenditure of energy but also an increased exposure to risk. Therefore, the grey matter consolidates the propensity to maintain habits already acquired — both in terms of actions and thoughts — by releasing opioid substances that form a sort of addiction to those same habits [61,62]. Such a mechanism represents an archaic and inherent mode of the structure of the human brain for coping with uncertainty or unpredictability [19]. This is exemplified by the attribution of atmospheric storms to the wrath of the gods centuries ago due to a lack of scientific understanding. Similarly, in the present era, there is a tendency among the general public to ascribe intentionality, desires, and emotions to AI systems in order to cope with the uncertainty and non-predictability of their behaviour [64]. It implies that the actions, responses, or decisions of an artificial agent are interpreted as if they were based on and influenced by the same processes, including the irrational ones, that dominate human mechanisms. Such a perception makes these technologies appear less different and less distant from humankind, and therefore more easily integrated into users' everyday lives [15].

It follows that the phenomenon of anthropomorphism cannot be eradicated and, together with it, of the disposition to create with AI systems bonds that should be more appropriate among people only [12].

Thus, attempting to completely eliminate the empathic response to AI systems and determining its appropriateness is futile. Instead, a more practical approach would be to understand the psychological mechanisms that underlie

this reaction from users and the reasoning behind certain design choices. Such a human-centered development of new technologies would strike a balance between the interests at stake, leading to more effective and user-friendly systems.

3 AI systems are (perceived as) social actors too

In their work, entitled "Machine and Mindlessness: Social Responses to Computers", Clifford Nass and Youngme Moon emphasize how individuals attribute the same social norms to computers as they do to their human environment, regardless of the fact that they are fully aware of the artificial nature of the machine with which they are interacting [53]. The authors explain this attitude as a deliberate, albeit subconscious, denial of signals that indicate the artificial nature of the computer. Subsequent studies have shown that these same attributions can affect a wide range of other technologies, especially with the advent of highly interactive systems [81].

After studying the natural manifestations of the above described phenomenon, researchers investigated the features that elicit the most emotional and empathic responses in people. The objective was to incorporate such characteristics into the design of interactive technologies to attempt to anticipate and direct the empathic response of users to the greatest extent possible [77,27]. This has led to the development of technologies capable of effectively targeting the inherent human mechanisms of anthropomorphism, attachment, and empathy [14]. Hence, by predicting the subconscious response of users, it was possible to direct their perception towards goals pre-determined by programmers and designers.

The aim is to facilitate the creation of a bond of trust and acceptance that makes it more desirable for individuals to use the AI system. Such a goal can be achieved by making the system appear more familiar, user-friendly, and sometimes even reliable [21].

This is one way in which, by design, it is aimed to establish and maintain longer-term interactions, which are indispensable for system efficiency and the social roles that some applications are now programmed to fulfill.

3.1 Social actors and social roles

In the context of human-robot interaction, research has shown that the physical presence of a robot in the human space can make them appear as 'equal companions' without losing the awareness that they are inanimate objects [25]. Robots' ability to occupy a physical space and to move in the environment, as well as to act in it and potentially modify it, enhances their human-like perception in terms of the possession of free will, intentionality towards an end, and character aspects. If we consider that these robots are frequently tasked with performing duties typically carried out by humans, and are capable of emulating social skills, it becomes clear that they are closely associated with the role they are assigned [65].

However, this example does not exclude the possibility that even non-embodied systems can induce the same association of ideas.

An example is Replika, a chatbot that can be downloaded onto personal devices [56]. It is designed to emulate different types of relationships, as selected by the user. Therefore, depending on people's preferences, Replika can act as a friend, lover, or girlfriend. The fact that it is often given its own name facilitates the attribution of a single and distinct identity, rather than being perceived as an extension of programming techniques and user's will. This contributes to making it perceived as a separate and independent entity. Furthermore, Replika is designed to engage individuals in increasingly frequent and regular interactions through message exchanges [32,40]. Such interactions serve as a call for attention, analogous to that initiated by an individual who cares for and is attached to their beloved ones.

Similarly, also systems designed to assist users in behavioral changing paths can resemble a social role [74,67]. In that case, the purpose is not merely entertainment but health support — as in the case of applications that bring the user closer to a healthier eating style or to interrupt habits that are harmful to health. However, even in these scenarios, it is essential to maintain friendly and effective communication to encourage the person to trust the recommendations, continue using the application, and share the necessary information for its proper functioning [14].

The same happens in the case of systems with less typified interactions, such as ChatGPT. It is designed to answer various questions, from health-related queries to news about current affairs, daily life, or cultural issues [35,1]. Nevertheless, it is important to note that the user's perception is that of interacting with a personal "assistant" who is always informed, responsive, and reliable.

One of the most emblematic examples is that of Baby X, produced by the company Soul Machine which, for this discussion, is already a rather emblematic name. The application can reproduce an infant in two dimensions, simulating the development of real children's hearing, comprehension, and language capacities [66]. To appear sufficiently realistic, Baby X is programmed to display a range of behaviors, including blushing, communicating expressions through its gaze, crying, and calming itself through the receipt of human attention [69]. It is of interest to observe that the human being involved in the interaction tends to provide the requisite attention with a degree of diligence comparable to that which would be applied to a real infant and to address the application or refer to it using a language similar to that which would be used to discuss the needs and achievements of a real child [57].

As can be surmised, the ascription of what might be termed "social agency", as well as the designation of actual social roles to what are and remain, in the final analysis, artificial agents, is not without inherent risks. While this is indispensable if the technology in question is to provide support, help, companionship, and assistance, as intended, it can also bring non-negligible side effects.

3.2 The pitfalls of designing for anthropomorphism

Individuals tend to attribute value to others in terms of reliability and competence based on emotional impulses [39]. Therefore, when evaluating the truthfulness of a consultation, people unconsciously tend to favour those who have elicited empathy and positive feelings, rather than those who have been consistently correct [65]. Rationality and objective analysis only come into play later and to varying degrees depending on the individual.

So far, the analysis illustrates how the same mechanisms can be used to evaluate intelligent systems that are called upon to fulfill social roles otherwise performed by human beings. This is certainly due to the human inherent propensity to anthropomorphism and the pre-programmed and well-documented "design for anthropomorphism" process described earlier.

Furthermore, recent studies indicate that users who interacted with an application for an extended period, being able to report their thoughts, feelings, and concerns, preferred it to interact with friends or family members [47,9]. The experiment participants perceived the system as highly reliable, non-judgemental, capable of maintaining secrecy, and always willing to listen [10].

Then, it becomes clear how AI excels over other humans in the quality of the exchange insofar as it does exactly the only thing it is really capable of doing: it sticks precisely to a script that is pre-written in its code by the programmers, does not express (personal) opinions, and does not diverge from the expectations of the user it is called upon to please. All of this represents the very essence of it being *artificial* but is rather commonly interpreted as a sign of human-like reliability. As a result, users not only attribute to AI systems exactly the role they are meant to emulate but they are also inclined to appreciate them more than they would if a human being played the same social role.

In doing so, users may be more exposed to the system's technical limitations and biases.

The nature of the problem is directly proportional to the criticality of the role to be played by the system. In the case of applications whose outcomes may have repercussions or influence aspects related to both people's physical and psychological health, this dynamic certainly becomes more problematic. However, more generally speaking, such mechanisms may produce side-effects connected with attention-grabbing mechanisms [49,75], with possible repercussions on memory and attention [42], over-trust, and even effects comparable to those of a real addiction [51]. Similar consequences have already been demonstrated in the literature with regard to other technologies, including television, mobile telephones, and video games. Compared with AI systems, these technologies present a significantly lower level of targeting of emotional processes and human attachment.

In such a scenario, it is evident that a greater understanding of the technical features of the technologies with which one interacts can be advantageous. Nevertheless, the extent to which greater technical knowledge correlates with greater protection from the risks enumerated herein will be the subject of the following sections.

4 The phenomenon of explainability in AI

XAI refers to a set of techniques that aim to provide transparency to the user regarding the processes used by the AI system to produce a specific outcome [2,6]. It is often considered a tool for transparency, clarifying the process that occurred between the user's inputs and the machine's outcomes. Furthermore, it is crucial to ensure that the human involved in the interaction maintains an adequate level of supervision over the results of the application and identifies any errors that may have occurred [26,59]. The use of XAI enables humans to maintain awareness of their interaction with the system, thus ensuring the ability to change their mind and withdraw consent if necessary, thus maintaining a sufficient level of agency towards the operations supported, facilitated, or even replaced by the technology in use.

The development of recent European legislation has made responsible XAI development crucial, particularly in ensuring compliance with the AI Act, the first form of AI regulation in Europe [55]. The Act mandates that producers provide detailed and easily accessible information, especially for high-risk technologies, concerning the limitations and capabilities of the target system, even to individuals lacking technical expertise. Consequently, it is of paramount importance that the decision-making processes are traceable and transparent at every stage of operation [59]. Even in the case of technologies that fall into the limited-risk categories, for which there are no strict rules to adhere to, XAI is of crucial importance. Indeed, transparency and the ability to trace the manner in which systems select, utilize, and integrate the data at their disposal can facilitate compliance with data processing regulations and redefine liability profiles in the event of damage or malpractice [24].

Nevertheless, despite the popularity that the advent of AI explainability has had in the scientific literature, this practice is still far from having been perfected. The reasons are to be found in both theoretical-applicative aspects and inappropriate balance of interests at stake.

4.1 The ambiguities of XAI

To date, there are no standardized methods for developing and implementing XAI techniques, due to the lack of clear definitions and agreement over reasonable expectations [30]. Therefore, the integration of the various disciplines that are called upon to contribute to the development of explanations becomes even more complex. In fact, the range of users who should benefit from such accessible, clear, and detailed explanations is fairly diverse and not always precisely predictable.

Moreover, models that cannot perform complicated operations, such as those involving decision tree methods, are less likely to have a "black box effect". Conversely, the most advanced models can be indeed difficult to interpret. However, such models are — and need to be — also highly preformative, while the implementation of XAI systems, regardless of the complexity of the reference application, can result in a loss of performance. In this case, the term "efficiency"

refers purely to the possibility and accuracy of completing the task for which the systems were programmed, thus meeting user expectations. What is too often overlooked is that such an approach represents a complete reversal of priorities. Despite the economic demands of the market, which appear to suggest that profit should be the primary objective in transactions between private parties and in the market itself, the law imposes a precise order in the balancing of interests. In point of fact, the rules stipulate that fundamental human rights must always take precedence over the pursuit of profit or the acquisition of advantage. Another element that cannot be undervalued is the high costs associated with studying, implementing, and experimenting XAI models, which may discourage smaller or less established companies from doing so.

Furthermore, producing explanations that are truly useful for the abovementioned purposes and empowering the user may require additional data from the people involved. This could contradict one of the goals of XAI, which is to ensure effective data protection and more transparency. In some cases, the necessity for such a significant increase in data sharing may be difficult to justify, particularly in light of the current legislative framework, which places greater emphasis on the manner in which data is selected, used, and stored. Hence, contradiction emerges between the legal and ethical requirements for responsible XAI and the technical prerequisites for effective XAI.

5 The advent of 'persuasive justifications' *versus* XAI explanations

Recently, GPTs have received plenty of attention from public opinion and the field literature due to what have been considered as their significant abilities in natural language processing [36,76].

Indeed, GPTs belong to the class of language models that generate LLMs generate verbal output that aims to be fluent and coherent, but their accuracy is just partial and situational [38,78].

Due to the inherent complexity of these systems, they require extensive training on a large amount of data to function at their full potential [72]. However, collecting such data is difficult, especially considering the limitations imposed by personal data protection and privacy regulations. Additionally, the outcomes of these systems are predictions that not only cannot be made on the amount of data that would be technically desirable — also for the security reasons above mentioned — but are also by definition never certain [68]. Although the produced text may be syntactically correct and the style may be appropriate for the context of use, this does not guarantee semantic correctness, which refers to the content component of the outcome [54].

It could be argued that these limitations in LLMs are simply due to an incomplete stage of development [80]. In other words, it could be assumed that the functionality and accuracy of these systems are still highly improvable. Once they have access to sufficient data and the sustainability of their training has been resolved, they could potentially access all available knowledge, which is

impossible for a single human being to possess, to advance human knowledge and awareness.

The vision presented has two main weaknesses. First, it appears unfeasible in the foreseeable future, and second, it loses sight of the technical nature of LLMs. LLMs are essentially structured as complex interface systems in information databases. Their complexity lies in their ability to compare and combine data, which is still obscure in many ways. This makes them efficient tools for facilitating research in many fields. However, it does not guarantee the production of new knowledge beyond what is already known, and even less so the correctness of such results.

Indeed, to be able to maintain a semantically appropriate verbal o text exchange, especially when dealing with open conversations whose content cannot be predicted — as in the case of ChatGPT — it is indispensable to possess what has been defined as 'Theory of Mind' [16]. This expression refers to the ability to foresee, admittedly partially and subjectively, what the state of mind of our interlocutor is. The result is not only to guide the style of the conversation but also to help modulate its tone and provide answers more in line with the interlocutor's real requests and intentions [18].

Recent studies have shown that the abilities deployed by GPT in emulating the possession of a user's Theory of Mind are not supported by actual capabilities [71,34,44]. In fact, LLMs cannot handle beliefs that the user considers axioms or takes for granted, including fallacies. This phenomenon has been addressed as "system's conservatism" [71], analogous to the corresponding philosophical concept [8]. The expression refers to the phenomenon where certain AI models tend to uphold the assumptions of the individual involved in the interaction rather than challenging them. This results in an inability to differentiate between true and false information in real-world scenarios and to identify any errors made by human in their epistemological or argumentative journey [44]. A similar mechanism is found in all those examples in which the system modifies a correct answer already given in the face of the skepticism of the user, who claims to own different notions about it.

The last aspect is particularly interesting in the context of the present analysis.

LLMs can solve complex problems and demonstrate abilities beyond natural language generation. However, upon critical evaluation of the results produced by these models about their ability to solve puzzles or complex problems, it is difficult to believe that the system is truly capable of arriving at a solution [45]. It may simply be reproducing an input-output scheme to which it was repeatedly exposed during its training. This suspicion seems well-founded, especially considering that the system radically alters its conclusion and even contradicts itself when faced with slight variations from standard examples [5,68]. The same applies to tasks where the LLM must make a choice, provide argumentation for a recommendation, or search for information on a given topic. The model exhibits challenges in making inferences based on causal relationships, particularly when completing these tasks [5,20].

Therefore, a risk arises of mistaking for intelligence what in practice is merely faithful adherence to a pre-established pattern that, being articulated in its composition and particularly complex in its concrete operation, is taken for "truly human-like" [4].

Such a risk materializes and is taken to its extreme consequences when these LLMs are used to replace XAI models [52]. An example of this is those cases in which LLM semantics produce explanations for the responses or general outcomes.

It has been suggested that such a solution would be desirable not only because the complexities inherent in XAI models could be simplified but also because more user-friendly explanations could be produced [48,46,82]. However, these models can only provide *justifications* for their previous answers, recommendations, or solutions. Moreover, such justifications will be as closely aligned as possible with the user's existing beliefs. If the last is not the case, the justification will be modified to align with the established patterns of the "system's conservatism". Consequently, the individuals involved in the interaction are not actually exposed to increased information about the system's doing, the data analyzed, and the manner of said analysis. Rather, a persuasion mechanism is initiated to convince people of the reasonableness of the outcome. This is achieved through an argumentative rather than an explanatory procedure.

Such a scenario represents a clear failure of one of the main purposes that the XAI field sets out to achieve, regardless of the current opportunities for success: empowering users to make autonomous and informed choices or evaluations.

On the contrary, the mechanisms mentioned in the current section reflect the features of design and programming mentioned earlier as described as "design for anthropomorphism" 2. These features contribute to the emphasis on the prevalence of 'appearance over reality' in technological innovation — both from a technical and developmental perspective. Therefore, this enhances reliability and increases the inclination to rely on AI systems despite the absence of tangible guarantees regarding the trustworthiness and respect for fundamental rights and physical and psychological integrity of human beings involved or otherwise impacted.

As a consequence, users are exposed to an increased risk of manipulation, distortion of their decision-making processes, reinforcement of false beliefs or biases, and the mere illusion of cognitive sovereignty over the application, which could lead to a real dependency dynamic.

5.1 From persuasion to a potential dependency loop

This analysis has demonstrated that anthropomorphism is an inherent aspect of the human mind and of human evolution 2. Then, it has been demonstrated that this characteristic inevitably exposes us to certain risks, including the risk of circumvention. However, it also represents an essential resource for coping with uncertainty and the unknown. Therefore, the phenomenon of anthropomorphism could be associated with the concept of inherent vulnerability, which is common to all human beings [13].

Such vulnerability has been recognized in legal theory as a result of individuals being embodied — i.e., subject to the needs and frailties of their bodies — and embedded in a system of social relations [28]. This *embeddedness* also means that individuals are subject to variable balances of power in relation to the State, laws, institutions, and policy choices. The material manifestation of inherent vulnerability is our dependence on others and on political superstructures [37]. Greater dependency increases vulnerability's exposure, hence increasing the exposure to damage to psychological integrity and fundamental rights [13,11].

In the context of human-AI interaction, and specifically in the absence of regulation or technical containment measures, the development of dependency dynamics towards LLMs is facilitated. Thus, the risk of harm to psychological and physical integrity is increased.

The dependence on frequently used AI systems can be attributed to the programming of their interaction patterns, which are intertwined with the innate propensity for anthropomorphism and the consequent attribution of social roles to artificial agents.

Due to the fact that these applications are characterized by the phenomenon of "system's conservatism", the apparent possession of a Theory of Mind, and being constantly available, they may induce compulsive use and overtrust [50]. Indeed, LLMs are programmed to always provide an answer, conform to the user's strong beliefs, never offer a judgment on the nature of the requests, make their data processing non-transparent, and present themselves as companions ready to provide support. A similar dynamic may cause individuals who are fragile, insecure, or overly reliant on others to expect that all requests will be granted and all questions will be answered. Additionally, it has been illustrated that intelligent systems are often perceived as more accessible and reliable than human professionals 3. Such a combination increases the potential risks significantly.

This is particularly pertinent when one considers that the systems here under analysis are also capable of providing what might be termed 'persuasive justifications' rather than explanations. mechanism would serve to reinforce not only the subconscious belief that technology is immune to error but also that it is necessarily led to pursue the interests of the user, even to the extent of pursuing what might be considered the 'good' for the user, without providing any valid evidence. This would serve to enhance the effectiveness of manipulative and exploitative mechanisms that take advantage of the inherent vulnerability of human beings while also rendering them more difficult to identify and interrupt.

It follows that the solution that is most in line with technological demands for efficiency and accuracy, as well as regulatory demands for protecting individuals and society as a whole, must be sought elsewhere. More specifically, it could be important to start conceiving human-AI interaction as a system itself, which is powered by the distinctive characteristics of its two pillars: technology, with its pre-programmed features and design, and the human being, with their often subconscious and inherent traits.

6 The claim for a responsible XAI

The discussion surrounding the development of new technologies in Europe in recent years has largely been dominated by the concept of responsible innovation [73,33]. This term summarises the need to produce AI systems that not only comply with current regulations but also with those that have been progressively implemented. Furthermore, producers and the scientific community were urged to focus on researching and experimenting with technology that is human-centred [60]. Thus the idea of 'progress at any cost' is opposed and instead technologies that can truly support the flourishing of every individual are promoted.

In this context, the term "responsible" emphasizes the importance of collaboration between legislators and the scientific community to act in science and conscience, taking responsibility for the choices made, to which it must be possible to be accountable in case of harm or malpractice.

Therefore, the paper proposes the development of a responsible XAI as a necessary step towards achieving a truly responsible AI. To achieve this goal, research should focus on developing uniform procedures for the development of XAI systems that can overcome or mitigate technical obstacles. The potential costs in terms of performance should not be a hindrance. European regulations on AI all invoke the legal principle of balancing interests. Following prevailing norms, fundamental rights — including the right to integrity, privacy, and dignity — are considered an unbreakable core that takes precedence over all other rights, especially those of an economic nature. Responsible XAI must prioritize safety in use over technical efficiency. This requires a balancing act between the two, but any potential loss in efficiency should be outweighed by the gain in safety.

In order to guarantee that assessments are impartial and to prioritise consumer interests and rights, it is necessary for the European legislature to intervene more assertively. In the absence of laws that define the characteristics, scope and modalities of XAI, as well as its non-optional nature, it is not possible to enforce and make it binding on companies. It is not being asserted here that XAI represents the solution to the risk of manipulation, distortion of decision-making, and injury to the integrity of users. As discussed elsewhere, the human tendency towards anthropomorphism and vulnerability makes it impossible to eliminate some of the suggestions and empathic responses described above [15,11]. However, the analysis presented here aims to emphasize the potential for XAI to hold producers accountable and provide individuals with a source of resilience when interacting with AI systems.

It is important to note that these goals have not yet been fully achieved. Investigating only the technical profiles of explanations, ignoring the nature of human cognition and psyche, or using solely LLMs as substitutes for explanation generators will not suffice.

On the contrary, having an integrated approach to XAI development, as suggested in this paper, will prove to be effective at a double level. Analyzing the mechanisms underlying the outcomes of AI systems by experts is crucial in order to fully understand how they work, to anticipate the risks that could arise from

direct contact with the user, and to intervene — already in the experimentation or testing phase — for their mitigation or solution. Hence, it would not only be possible to demonstrate commitment to legislation that is certainly complex to implement. Such anticipatory measures would ensure a higher quality of the products placed on the market, facilitate positive reception by buyers, and result in savings in costs that may be necessary in the event of forced product withdrawals from the market and subsequent compliance with current legislation. Furthermore, the implementation of appropriate XAI systems also addresses the call for responsible development of AI, as advocated by both the scientific community and the European legislator. This concept emphasizes the importance of designers and programmers prioritising human needs and rights in technological innovation, while taking into account their ineradicable nature as members of the humankind. In fact, intelligent systems should not aim to induce behavior or solicit choices that support interests other than those expressed by the user or intended by the application. Such systems lay the foundations for manipulative mechanisms that can be detrimental to personal integrity.

7 Conclusions and Future Works

This study examines the dynamics of human-AI interaction and the tools required to empower non-expert users.

It illustrates and dissects the subconscious, cognitive, and psychological mechanisms underlying the integration of intelligent systems into everyday life. The study establishes the ineradicability of anthropomorphism and the propensity to regard technology as a reliable and infallible ally. It is also stated the preordination with which designers and programmers implement specific features and functionality, with the aim of facilitating acceptance and usage, while also opening the way to potential risks to the integrity, freedom of decision-making, and psychological independence of those involved.

Hereof, XAI emerges as a necessary tool and valuable resource for making the functionality of AI systems more knowable and accessible to its users. Despite its potential, this discipline still faces challenges that hinder its theoretical development and practical application. These challenges stem from a lack of consensus on theoretical frameworks and empirical models, as well as a decrease in system efficiency when implemented with XAI tools. It is important to address these issues in order to fully realize the benefits of the field iyself.

Therefore, the recent rapid development of LLMs has been introduced, underlining how its interaction interfaces tend to support — or even elicit — the human tendency towards anthropomorphism. Thus, the belief that the system is equipped with a Theory of Mind, analysis and processing capabilities, and knowledge that makes it seem an indispensable resource for the user is reinforced, due to the fact that it appears not only extremely efficient but also reliable. If these models are given the ability to provide explanations to humans, it lays the foundation for a potentially risky and dependency-oriented bond between AI and individuals. This paper argues that such explanations are very different in na-

ture from those envisaged in XAI and only aim to justify the outcome already produced by the system, seeking to align it as closely as possible with the user's preconceptions rather than with the technological reality behind such outcomes.

Therefore, the paper advocates for a shift in the scientific field towards the study and implementation of responsible XAI. This involves balancing the economic and performance interests of IT companies with user advocacy and protection interests, prioritizing the latter. In fact, research should aim to implement systems that prioritize explainability and human-centeredness. In doing so, it is suggested that legislation is needed to provide certain and uniform guidelines regarding XAI features, characteristics, scopes, and modalities of application.

Future work will focus on creating guidelines to better define what is meant by responsible XAI and what concrete measures should be taken to guide XAI research in that direction.

Acknowledgments

This work is partially supported by the Joint Doctorate grant agreement No 814177 LAST-JD-Rights of Internet of Everything, and the Chist-Era grant CHIST-ERA19-XAI-005 — *(i)* the Swiss National Science Foundation (G.A. 20CH21_195530), *(iii)* the Luxembourg National Research Fund (G.A. INTER/CHIST/19/14589586).

References

1. Aljanabi, M., Ghazi, M., Ali, A.H., Abed, S.A., et al.: Chatgpt: open possibilities. Iraqi Journal For Computer Science and Mathematics **4**(1), 62–64 (2023)
2. Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K.: Explainable agents and robots: Results from a systematic literature review. In: 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019. pp. 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems (2019)
3. Balakrishnan, P., Nataraajan, R., Desai, A.: Consumer rationality and economic efficiency: Is the assumed link justified? Marketing Management Journal **10**(1) (2000)
4. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. pp. 610–623 (2021)
5. Binz, M., Schulz, E.: Using cognitive psychology to understand gpt-3. Proceedings of the National Academy of Sciences **120**(6), e2218523120 (2023)
6. Biran, O., Cotton, C.: Explanation and justification in machine learning: A survey. In: IJCAI-17 workshop on explainable AI (XAI). vol. 8, pp. 8–13 (2017)
7. Block-Lieb, S., Janger, E.J.: The myth of the rational borrower: Rationality, behavioralism, and the misguided reform of bankruptcy law. Tex. L. Rev. **84**, 1481 (2005)
8. Bourke, R.: What is conservatism? history, ideology and party. European Journal of Political Theory **17**(4), 449–475 (2018)
9. Brandeis, L.D.: Other people's money and how the bankers use it, 1914. Boston, MA and New York, NY (Bedford/St. Martin's) (1995)

10. Brandtzaeg, P.B., Skjuve, M., Følstad, A.: My ai friend: How users of a social chatbot understand their human–ai friendship. Human Communication Research **48**(3), 404–429 (2022)
11. Carli, R.: Deception in Social Robotics: Problematic Profiles of Human Robot Interaction and the Universality of Human Vulnerability. Forthcoming (Forthcoming)
12. Carli, R., Najjar, A.: Rethinking trust in social robotics. arXiv preprint arXiv:2109.06800 (2021)
13. Carli, R., Najjar, A.: A vulnerability-oriented impact assessment for the development of human-centred and fundamental rights-empowering social robots. In: Proceedings of the 11th International Conference on Human-Agent Interaction. pp. 404–406 (2023)
14. Carli, R., Najjar, A., Calvaresi, D.: Human-social robots interaction: The blurred line between necessary anthropomorphization and manipulation. In: Proceedings of the 10th International Conference on Human-Agent Interaction. pp. 321–323 (2022)
15. Carli, R., Najjar, A., Calvaresi, D.: Risk and exposure of xai in persuasion and argumentation: The case of manipulation. In: Explainable and Transparent AI and Multi-Agent Systems: 4th International Workshop, EXTRAAMAS 2022, Virtual Event, May 9–10, 2022, Revised Selected Papers. pp. 204–220. Springer (2022)
16. Carlson, S.M., Koenig, M.A., Harms, M.B.: Theory of mind. Wiley Interdisciplinary Reviews: Cognitive Science **4**(4), 391–402 (2013)
17. Cassirer, E.: Filosofia delle forme simboliche: Il linguaggio/[trad. di Eraldo Arnaud]. Nuova Italia (1996)
18. Castano, E., Martingano, A.J., Basile, G., Bergen, E., Jeong, E.H.K.: Listening in to a conversation enhances theory of mind. Current Research in Ecological and Social Psychology **4**, 100108 (2023)
19. Chan, A.A.Y.H.: Anthropomorphism as a conservation tool. Biodiversity and Conservation **21**, 1889–1892 (2012)
20. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al.: A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology (2023)
21. Crevier, D.: AI: the tumultuous history of the search for artificial intelligence. Basic Books, Inc. (1993)
22. Dacey, M.: Anthropomorphism as cognitive bias. Philosophy of Science **84**(5), 1152–1164 (2017)
23. Das, A., Rad, P.: Opportunities and challenges in explainable artificial intelligence (xai): A survey. arXiv preprint arXiv:2006.11371 (2020)
24. Directive, T.: Directive 2004/109/ec of the european parliament and of the council of 15 december 2004 on the harmonisation of transparency requirements in relation to information about issuers whose securities are admitted to trading on a regulated market and amending directive 2001/34/ec. OJ L **390**(15.12) (2004)
25. Dumouchel, P., Damiano, L.: Living with robots. Harvard University Press (2017)
26. Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., et al.: Explainable ai (xai): Core ideas, techniques, and solutions. ACM Computing Surveys **55**(9), 1–33 (2023)
27. Epley, N., Waytz, A., Cacioppo, J.T.: On seeing human: a three-factor theory of anthropomorphism. Psychological review **114**(4), 864 (2007)
28. Fineman, M.A.: Vulnerability: reflections on a new ethical foundation for law and politics. Ashgate Publishing, Ltd. (2013)
29. Gatt, L., Caggiano, I.A.: Consumers and digital environments as a structural vulnerability relationship (2022)

30. Graziani, M., Dutkiewicz, L., Calvaresi, D., Amorim, J.P., Yordanova, K., Vered, M., Nair, R., Abreu, P.H., Blanke, T., Pulignano, V., et al.: A global taxonomy of interpretable ai: unifying the terminology for the technical and social sciences. Artificial intelligence review pp. 1–32 (2022)
31. Guthrie, S.E.: Anthropomorphism: A definition and a theory. (1997)
32. Hakim, F.Z.M., Indrayani, L.M., Amalia, R.M.: A dialogic analysis of compliment strategies employed by replika chatbot. In: Third International conference of arts, language and culture (ICALC 2018). pp. 266–271. Atlantis Press (2019)
33. Herrmann, H.: What's next for responsible artificial intelligence: a way forward through responsible innovation. Heliyon (2023)
34. Holterman, B., van Deemter, K.: Does chatgpt have theory of mind? arXiv preprint arXiv:2305.14020 (2023)
35. Iftikhar, L., Iftikhar, M.F., Hanif, M.I.: Docgpt: Impact of chatgpt-3 on health services as a virtual doctor. EC Paediatrics **12**(1), 45–55 (2023)
36. Imamguluyev, R.: The rise of gpt-3: Implications for natural language processing and beyond. Journal homepage:www.ijrpr.com ISSN **2582**, 7421 (2023)
37. Ippolito, F.: La vulnerabilità quale principio emergente nel diritto internazionale dei diritti umani? Ars interpretandi **24**(2), 63–93 (2019)
38. Johnson, D., Goodman, R., Patrinely, J., Stone, C., Zimmerman, E., Donald, R., Chang, S., Berkowitz, S., Finn, A., Jahangir, E., et al.: Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the chat-gpt model. Research square (2023)
39. Kim, J., Park, K., Ryu, H.: Social values of care robots. International journal of environmental research and public health **19**(24), 16657 (2022)
40. Köbis, N., Bonnefon, J.F., Rahwan, I.: Bad machines corrupt good morals. Nature Human Behaviour **5**(6), 679–685 (2021)
41. Levillain, F., Zibetti, E.: Behavioral objects: The rise of the evocative machines. Journal of Human-Robot Interaction **6**(1), 4–24 (2017)
42. Lin, H.F.: Examination of cognitive absorption influencing the intention to use a virtual community. Behaviour & Information Technology **28**(5), 421–431 (2009)
43. Lotto, B.: Percezioni: come il cervello costruisce il mondo. Bollati Boringhieri (2022)
44. Ma, Z., Sansom, J., Peng, R., Chai, J.: Towards a holistic landscape of situated theory of mind in large language models. arXiv preprint arXiv:2310.19619 (2023)
45. Marcus, G., Davis, E.: Gpt-3, bloviator: Openai's language generator has no idea what it's talking about. Technology Review p. 294 (2020)
46. Mariotti, E., Alonso, J.M., Gatt, A.: Towards harnessing natural language generation to explain black-box models. In: 2nd Workshop on interactive natural language technology for explainable artificial intelligence. pp. 22–27 (2020)
47. Marriott, H.R., Pitardi, V.: One is the loneliest number... two can be as bad as one. the influence of ai friendship apps on users' well-being and addiction. Psychology & marketing **41**(1), 86–101 (2024)
48. Mavrepis, P., Makridis, G., Fatouros, G., Koukos, V., Separdani, M.M., Kyriazis, D.: Xai for all: Can large language models simplify explainable ai? arXiv preprint arXiv:2401.13110 (2024)
49. Misztal, A.: From ticks to tricks of time: narrative and temporal configuration of experience. Phenomenology and the Cognitive Sciences **19**(1), 59–78 (2020)
50. Mohseni, S., Yang, F., Pentyala, S., Du, M., Liu, Y., Lupfer, N., Hu, X., Ji, S., Ragan, E.: Machine learning explanations to prevent overtrust in fake news detection. In: Proceedings of the international AAAI conference on web and social media. vol. 15, pp. 421–431 (2021)

51. Morava, M., Andrew, S.: Loneliness won't end when the pandemic ends. (2021)
52. Nam, D., Macvean, A., Hellendoorn, V., Vasilescu, B., Myers, B.: Using an llm to help with code understanding. In: 2024 IEEE/ACM 46th International Conference on Software Engineering (ICSE). pp. 881–881. IEEE Computer Society (2024)
53. Nass, C., Moon, Y.: Machines and mindlessness: Social responses to computers. Journal of social issues **56**(1), 81–103 (2000)
54. Oviedo-Trespalacios, O., Peden, A.E., Cole-Hunter, T., Costantini, A., Haghani, M., Rod, J., Kelly, S., Torkamaan, H., Tariq, A., Newton, J.D.A., et al.: The risks of using chatgpt to obtain common safety-related information and advice. Safety science **167**, 106244 (2023)
55. Parliament, E., the Council: Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828, (artificial intelligence act) (Emendaments, 6 March 2024)
56. Pentina, I., Hancock, T., Xie, T.: Exploring relationship development with social chatbots: A mixed-method study of replika. Computers in Human Behavior **140**, 107600 (2023)
57. Pitetti-Heil, J.: Artificial intelligence from science fiction to soul machines:(re-) configuring empathy between bodies, knowledge, and power. Artificial Intelligence and Human Enhancement: Affirmative and Critical Approaches in the Humanities **21**, 287 (2022)
58. Quinn, W.: Rationality and the human good. Social Philosophy and Policy **9**(2), 81–95 (1992)
59. Rai, A.: Explainable ai: From black box to glass box. Journal of the Academy of Marketing Science **48**(1), 137–141 (2020)
60. Regulation, P.: Regulation (eu) 2016/679 of the european parliament and of the council. Regulation (eu) **679**, 2016 (2016)
61. Roth, G.: Why long-lasting therapeutic changes in the brain need time. Psychotherapeut **61**, 455–461 (2016)
62. Roth, G., Strüber, N.: Emotion, motivation, personality and their neurobiological foundations. In: Psychoneuroscience, pp. 143–174. Springer (2023)
63. Sahlin, N.E., Brännmark, J.: How can we be moral when we are so irrational? Logique et Analyse pp. 101–126 (2013)
64. Salles, A., Evers, K., Farisco, M.: Anthropomorphism in ai. AJOB neuroscience **11**(2), 88–95 (2020)
65. Schreiber, D.: On social attribution: implications of recent cognitive neuroscience research for race, law, and politics. Science and engineering ethics **18**, 557–566 (2012)
66. Seymour, M., Riemer, K., Kay, J.: Actors, avatars and agents: Potentials and implications of natural face technology for the creation of realistic visual presence. Journal of the association for Information Systems **19**(10), 4 (2018)
67. Smestad, T.L.: Personality Matters! Improving The User Experience of Chatbot Interfaces-Personality provides a stable pattern to guide the design and behaviour of conversational agents. Master's thesis, NTNU (2018)
68. Sobieszek, A., Price, T.: Playing games with ais: the limits of gpt-3 and similar large language models. Minds and Machines **32**(2), 341–364 (2022)
69. Sookkaew, J., Saephoo, P.: "digital influencer": development and coexistence with digital social groups. International Journal of Advanced Computer Science and Applications **12**(12) (2021)

70. Stanovich, K.E.: Why humans are (sometimes) less rational than other animals: Cognitive complexity and the axioms of rational choice. Thinking & Reasoning **19**(1), 1–26 (2013)
71. Strachan, J., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Rufo, A., Manzi, G., Graziano, M., Becchio, C.: Testing theory of mind in gpt models and humans (2023)
72. Stringhi, E.: Hallucinating (or poorly fed) llms? the problem of data accuracy. i-lex **16**(2), 54–63 (2023)
73. Taddeo, M., Floridi, L.: How ai can be a force for good. Science **361**(6404), 751–752 (2018)
74. Tian, X., Risha, Z., Ahmed, I., Lekshmi Narayanan, A.B., Biehl, J.: Let's talk it out: A chatbot for effective study habit behavioral change. Proceedings of the ACM on Human-Computer Interaction **5**(CSCW1), 1–32 (2021)
75. Tourinho, A., de Oliveira, B.M.K.: Time flies when you are having fun: Cognitive absorption and beliefs about social media usage. AIS Transactions on Replication Research **5**(1), 4 (2019)
76. Trajtenberg, M.: Artificial intelligence as the next gpt. The economics of artificial intelligence: An agenda **175** (2019)
77. Trower, T.: Bob and beyond: A microsoft insider remembers (2010)
78. Wang, W., Shi, J., Tu, Z., Yuan, Y., Huang, J.t., Jiao, W., Lyu, M.R.: The earth is flat? unveiling factual errors in large language models. arXiv preprint arXiv:2401.00761 (2024)
79. Wu, J., Yang, S., Zhan, R., Yuan, Y., Wong, D.F., Chao, L.S.: A survey on llm-gernerated text detection: Necessity, methods, and future directions. arXiv preprint arXiv:2310.14724 (2023)
80. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al.: The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864 (2023)
81. Xu, K., Lombard, M.: Media are social actors: expanding the casa paradigm in the 21st century. In: Annual Conference of the International Communication Association. Fukuoka, Japan (2016)
82. Zhang, X., Guo, Y., Stepputtis, S., Sycara, K., Campbell, J.: Explaining agent behavior with large language models. arXiv preprint arXiv:2309.10346 (2023)