Reinterpreting Vulnerability to Tackle Deception in Principles-Based XAI for Human-Computer Interaction

Conference Paper · July 2023

CITATIONS
0

READS
66

2 authors:

Rachele Carli
University of Bologna
9 PUBLICATIONS **15** CITATIONS
SEE PROFILE

Davide Calvaresi
HES-SO Valais-Wallis
113 PUBLICATIONS **1,679** CITATIONS
SEE PROFILE

Reinterpreting Vulnerability to Tackle Deception in Principles-Based XAI for Human-Computer Interaction.

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. Artificial intelligence (AI) systems have been increasingly adopted for decision support, behavioral change purposes, assistance, and aid in daily activities and decisions. Thus, focusing on design and interaction that, in addition to being functional, foster users' acceptance and trust is increasingly necessary. Human-computer interaction (HCI) and human-robot interaction (HRI) studies focused more and more on the exploitation of communication means and interfaces to possibly enact deception. Despite the literal meaning often attributed to the term, deception does not always denote a merely manipulative intent. The expression "banal deception" has been theorized to specifically refer to design strategies that aim to facilitate the interaction. Advances in explainable AI (XAI) could serve as technical means to minimize the risk of distortive effects on people's perceptions and will. However, this paper argues that how the provided explanations and their content can exacerbate the deceptive dynamics or even manipulate the end user. therefore, in order to avoid similar consequences, this analysis suggests legal principles to which the explanation must conform to mitigate the side effects of deception in HCI/HRI. Such principles will be made enforceable by assessing the impact of deception on the end users based on the concept of vulnerability – understood here as the rationalization of the inviolable right of human dignity – and control measures implemented in the given sytems.

Keywords: XAI · Deception · Vulnerability

1 Introduction

Interactive AI systems are now used for many purposes that require a constant exchange of information with the end user. Some of the main tasks performed by such applications include: e-health goals [16], decision-making activities, support and guide in behavioral changes [11], e-administration proceedings [31], assistance for e-services [19].

The quality and frequency of interaction in many of these cases are crucial for two fundamental reasons. First, this allows the application to refine its outcomes and, consequently, to pursue the purpose for which it was developed more effectively and efficiently. Second, an interaction that is not only technically satisfying but also pleasant, at times "familiar", enables users to be more consistent in their engagement, to adhere better to the recommendations provided, and to rely on them to develop trust [68].

In light of the above, there has been a growing interest of researchers in the fields of HCI, HRI, and XAI in those dynamics and elements that, if correctly implemented and elicited, could foster interaction by acting on the psychological, cognitive, and emotional mechanisms of the human interlocutors. This relied on research in neuroscience, behavioral psychology, cognitive science, communication science, and interdisciplinary working groups [33]. One of the main results achieved by scholars led back to an aspect already dear to computer science: the theme of deception. It has entered the context of human-AI interaction since the Turing Test, demonstrating how the very concept of AI is based on the ability a system has to emulate the capabilities of human beings, regardless of whether they are objectively present or not [66]. The credibility of this appearance has a far more impactful influence on the perceived quality of the interaction than pure technical efficiency. For this reason, efforts have been made to implement AI systems more and more with design characteristics and communication features capable of targeting the brain areas that are involved in the perception of positive emotions such as cuteness, trustworthiness, sympathy, tenderness, and empathy. It is worth emphasizing that the primary purpose of such implementations is to ensure the good functionality of the application and, as a direct consequence, the possibility for the user to benefit from the resulting beneficial and supportive effects.

Thus, the concept of "banal deception" has recently been theorized [53]. In particular, it aims to delimit the difference between deception, understood as a mechanism of mendacity, manipulation, and the phenomenon described above, which identifies an expedient that may contribute to the user's best interest.

According to such premises, the crucial role of explanation clearly emerges in this context. It may have the capacity to make many processes performed by the AI system clearer to the non-specialist interlocutor, redefining the boundaries between appearance and reality, technology and humanity, functionality and emotionality [27]. In all those cases in which – or within the limits in which – banal deception is essential for the success of the interaction, XAI can counterbalance the emotional and unconscious scope of the user's reactions with accurate, punctual explanations, which leverage the more logical-rational part of the brain in return.

This paper argues that the explanation risks becoming an element of deception due to how it is produced, communicated (e.g., styles and tones), and its content. In this sense, it could elicit deceptive effects, leading to real manipulative consequences.

Therefore, this paper suggests a framework – based on enforceable legal principles – paving the way for structuring two possible tools to mitigate potential adverse effects. The first proposal concerns structuring a Vulnerability Impact Assessment, aiming at establishing the different degrees of banal deception that can be considered admissible, depending on the level with which they impact, exasperate, or assist the unavoidable human vulnerability. The second proposal concerns the formalization of a knowledge graph that identifies features and relationships between elements that make up vulnerability and implement them in the XAI system.

The rest of this paper is organized as follows.

Section 2 presents the state of the art, focusing on the original concept of deception, presenting the new theorization of 'banal deception' and emphasizing the role the XAI could play in either facilitating manipulative drifts or – conversely – protecting the user from them. To pursue the second objective, Section 3 elaborates on the basic legal principles that should be the framework to which the design of explanations should adhere. Section 4.1 and Section 4.2 present the dual approach through which this framework should be concerted, namely a new Impact Assessment vulnerability-based and a knowledge graph implemented in the system itself.

2 Background & State of the Art

This section presents the background and state of the art of disciplines intersecting explainability in HCI, such as deception and XAI.

2.1 Deception in HCI

The subject of deception has a long history in computer science and, more specifically, HCI and – more recently – HRI. However, researchers in these disciplines are often skeptical about describing the outcome of their work or their approach in terms of deception due to the predominantly negative connotation that this concept has inherited from the humanities, especially the legal sciences. However, while the legal semantics is often connected to an act aimed at circumventing, misleading, and inducing disbelief for the benefit of the deceiver and necessarily to the detriment of the deceived, the same is not always true from a more strictly technological point of view.

Deception has become part of AI since the Touring Test. In what was essentially considered an HCI experiment, the computer program will only be able to win the game if it can "fool" the human interlocutor [41]. This means to assume a very human-centric perspective, where it is assessed whether the illusion – of intelligence in this case – has been programmed accurately and allegedly enough to not only be plausible but to convince the individual [60]. In other words, we start by analyzing human beings, their communicative strategies, and connotative semantics to understand how to program a successful interaction [6].

Considering only this angle, it is still possible to interpret the deceptive dynamic as aimed at misleading the other party. Nevertheless, it should be noted that the Turing test was structured as a game in which individuals participated willingly, following the game's rules. The playful background conveys the deceptive interaction with an innocuous and ethically acceptable connotation [35]. This is precisely what modern voice assistants and much interaction software claim to have inherited from the test. However, even in that context, giving the application the ability to respond with jokes is a way of playfully engaging the interlocutors, pushing them to challenge the limits of the imitation of humanity [50].

In these contexts, deception is domesticated and does not carry the negative connotation that manipulative ends have in common with other misleading expedients. Therefore, the envisioned outcome, subsequent research, and experimentation were set to imagine a future in which deception, conceived in its functional and non-harmful form, would become a useful tool for developing successful interactions with new technologies used on a daily basis.

2.2 From Deception to the ELIZA effect

To make it possible efficient and effective interactions for the benefit of the users, HCI developed as a field of research aimed at adapting system interfaces, design features, and functionality to the perceptual and cognitive abilities of human beings. Thus, deception became a proper method to deflect any element that could uncover the artificial and aseptic nature of the AI system. Said otherwise, the standard approach in the design of applications deputed to interact continuously and closely with users became to exploit the fallibility, the unconscious psychological and cognitive mechanisms inherent in human nature [12].

An example of this evolution is represented by ELIZA, the software that pioneered new perspectives on chatbots [43]. It shows that the correctness and appropriateness of the outputs, given a certain input, are not the only crucial factors for a successful HCI. The so-called "social value" is also important [38]. That is to say that the fact that the application can play a certain role in the interaction, which remains consistent in itself throughout the whole exchange, has to be considered central [28]. This stems from the fact that individuals by nature attribute to their interlocutors – even humans – a specific role, which could be defined as a "social role" [61]. This has little to do with the actual identity of the other (e.g., if they are the professional to whom we turn for a consultation, a family member, or a stranger). What plays an essential role is the – social and not – value that people attribute to those with whom they interact. It consists of projections, past experiences, and emotional resistance. Referring more specifically to the HCI domain, the subconscious tendency of humans to believe that AI systems and software have their own behavior and that this is similar to that of peers has been termed the "ELIZA effect" [63].

Upon closer analysis, it demonstrated how even relatively unsophisticated programs can deceive the user through AI, creating an appearance of intelligence and agency [23].

This was instrumental in suggesting that humans are naturally inclined to attribute human appearance, faculties, and destinies to inanimate objects [29], and that this inescapable characteristic can be exploited to create efficient interactions.

Hence, the theorization of the ELIZA effect was conducive to bringing to light something already clear since the Imitation Game: AI is the result of projection mechanisms which ineradicably characterize individuals – despite their level of practical knowledge [59]. For "projection mechanisms" is meant that universal psychic modality by which people transfer subjective ideational content outwards – into other people, animals, even objects [45].

This led to an in-depth exploration of the unconscious mechanisms that lead human beings to prefigure a kind of "computer metaphor", according to which machines and software could be comparable to human beings. Such an examination was conducted mainly through disciplines like neuroscience, behavioral psychology, cognitive science, and communication science [48,20,24].

What mentioned so far has led to the creation of the CASA model: Computers Are Social Actors too [51]. According to such a paradigm, people applied to computers social rules and expectations similar to those they have towards humans. This is possible because each component in interface design conveys social meaning, even if this end is not pre-determined by programmers or designers. Concurrently, if it is somehow possible to anticipate this meaning, it is also possible to direct it with the result of programming for more efficient HCIs [65].

2.3 Banal Deception

The investigation conducted so far supports the idea, now widespread in the literature, that deception in HCI and HRI is often implemented and addressed as an essential element for the best functionality of AI systems and the increase of the user's comfort. Ultimately, it could even be described as a constitutive element of AI, without which it would not be possible to define artificial intelligence itself [53]. However, this is not to disregard its possible manipulative drifts. Especially from a legal standing point, the concrete outcome of a harmful event must often be considered more relevant, rather than the benevolent but unrealized intent with which the event was preordained. For this reason, research in the field of human-AI interaction has recently proposed a new terminology, namely *Banal Deception* [53]. It is adopted to frame that type of deception that does not arise with the direct intent to mislead but rather to facilitate the use of the application and the efficiency of achieving the intended purpose. Doing so would contribute to integrating AI technologies into everyday life for decision support, entertainment, and guidance in behavioral changes. Simone Natale [53] identifies five elements that can guide in profiling the phenomenon of banal deception:

Ordinary character: ELIZA may be a good example. It did not present anything particularly extraordinary, and the same can be said of Siri, Alexa, or

many other modern chatbots. Non-specialized users seem to focus on communicative and interactive aspects that make them often curious about other media. However, the AI technologies mentioned above induce them to believe that the appearance of personality and agency is actually real. This denotes the inherent vulnerability to deception, on the one hand. On the other hand, it also underlines that banal deception is probably imperceptible, but not without consequences. It may not be intent on manipulating, but it has the predetermined purpose of making AI systems enter the core of individuals' mental structure, and – to some extent – identity [54]. In fact, thanks to the mechanisms of trivial deception, in fact, such technologies can target specific areas of the human mind, elicit trust and emotional attachment, influence habits and tastes, shape the perception of reality.

Functionality: an application capable of eliciting positive emotions, trust, and reliability in the user will be used more often and with less skepticism. This allows a more intense flow of data, which is indispensable for improving performance.

Obliviousness: being extremely subtle, as well as being a decisive part of the design, this deceptive phenomenon is not perceived by the user and is often lowered to the rank of mere technical expedient without further investigation. Nonetheless, overcoming the barriers of consciousness and awareness is also effective in physiologically balanced, well-informed subjects [5]. They can recognize the artificial nature of the application rationally, without being able to "resist" the mechanisms of anthropomorphism and personification proper of their primordial cognitive structure.

Low definition: chatbots and other AI systems programmed according to the logic of the banal deception are neither necessarily very sophisticated from an aesthetic point of view nor particularly characterized in impersonating a single, fixed, communicative/social role. This is because human beings tend to attribute meaning to what they interact with in an intimate and/or continuous way [67]. Leaving (intentionally) the possibility to the user to exploit their imagination to fill the gaps left by programmers or designers allows customization that translates to an emotional level of familiarity, empathy, and attachment.

To be programmed: although banal deception relies on mechanisms inherent in human cognitive structures, it is voluntarily programmed by technical experts on the basis of studies aimed at investigating human perceptual mechanisms, with the precise purpose of targeting similar structures to pursue the "functionality" described above. In other words, banal deception needs the – unconscious – cooperation of the user to work, but it is *ex-ante* – consciously – pre-ordered by AI systems' developers.

2.4 Explainable AI

Interpretable and Explainable AI is a discipline that has not yet found a unanimous definition, as it lends itself to work among different disciplines [34]. Nevertheless, it can be detailed through its primary objective: to make data-driven recommendations, predictions/results, and data processing comprehensible to the final user [10,3].

This is necessary since human beings have a tendency to attribute mental states to artificial entities (a.k.a., agents) leveraging the evaluation of their objective behavior/outputs [42,7]. This in itself can lead to two possible inauspicious effects: (i) creating a false representation of the AI system and its capabilities and (ii) attributing an emotional-intentional valence to its answers/actions.

Starting from such assumptions, the explanation generated has been conceived by the scientific community as a valid aid so that the intentional stance [25] that the user will inevitably project onto the technology is as objective and realistic as possible. This should happen despite the prior knowledge possessed by the subject in question. Thus, according to XAI theorists, it would be possible to pursue a twofold result: to limit the negative effects of anthropomorphism and foster interaction.

Yet, this interpretation cannot in itself exhaust the complete analysis of a dynamic - that of HCI - which is multi-factorial.

This becomes clear considering the phenomenon called "Mindless behavior". Such an expression is commonly used to delineate the subconscious mechanism through which people apply social conventions to artificial agents. The reason why this evaluation seems to take place "mindlessly". Indeed, it stems from the fact that individuals reveal such a way of interpreting the interaction regardless of the level of awareness they have of the actual nature of the AI system [62].

This brings us back to the above discussion of the mechanisms of banal deception. In this context, making evident the mechanical and inanimate nature of the application, its lack of consciousness and intentionality, opening the black box by revealing the hidden mechanisms and rationals behind the processing of data would seem to have no bearing on the subconscious empathic dynamics that users are in any case naturally induced to enact.

2.5 XAI in the realm of Banal Deception

Explaining is considered critical in making the operation of the system/robot and the nature of the output as transparent as possible. This facilitates its use and (most importantly) its trustworthiness, desirability, and pleasant interaction. Ultimately, it is conceived as an essential tool to shorten the distance often perceived between the technicality of AI and the unskilled user.

Nevertheless, depending on the characteristics of the explanation and the manner in which it is given, it may itself represent an element of banal deception – as described above. Furthermore, in some circumstances, it may reinforce the deceptive mechanisms already inherent in the application, crossing the line between deception and manipulation.

This discussion will not delve into the conceptual and semantic analysis of these two themes, for which we refer to, among others [46,18]. For the purposes of the analysis conducted here, we point out that the circumscription of the manipulation concept is still much debated in the scientific community. This also applies to the legal sphere, where it is relevant to determine cases and means by which to intervene to protect people's will and psychological integrity. Hence, reporting at least an overall conceptualization of manipulation is deemed appropriate. It is conceived as a dynamic that can circumvent individuals' critical thinking and logic [44], making them do something different from what they would have done or justified if they had not been subjected to the same manipulative techniques for the benefit of the manipulator. Thus, targeting self-awareness more and before even affecting rationality, manipulation can have deception as one of the means through which the purpose is pursued [22].

Acknowledging this brief examination, the explanation might be structured according to the logic of the banal deception, going to strengthen confidence and trust in the outcome of the application, to the detriment of the real interest and goal set by the user. For instance, some explanations, or the methodology/expedient by which it is provided, could be aimed at (i) making the interlocutors dependent on the use or feedback of the AI system, (ii) inducing them to pursue ends that merely benefit the producer, (iii) generating behavioral change that is harmful to the user, but still useful for general profiling purposes, (iv) eliciting the loss of significant social contacts (including the 'second expert opinion' performed by a human specialist in the case of applications with potential impact on health).

To preserve the protective and positive purposes of XAI, to prevent it from becoming a tool of manipulation, and to make it a valuable aid in limiting the side effects of banal deception, it might be helpful to draw up a list of principles to which the explanation must conform. With this aim in mind, the principles suggested here are of a legal, rather than ethical, nature. This offers the benefit of making the framework below potentially enforceable with both *ex-ante* and *ex-post* logic. *Ex-ante*, ideally, it will have to be taken into account when programming and designing the AI system and its explainability. *Ex-post*, as it can be invoked in the event of a violation to require a forced adjustment or to correct any divergences that, through the interaction itself, the application will have developed.

3 Principles-based framework for explanations

The European approach to AI seems to be delineated around the recurring concept of human-centered AI (HCAI) [2]. This entails aiming to create AI systems that support human capabilities rather than replacing or impoverishing them. Therefore, technological development should be oriented toward the benefit of human beings. From a European perspective, it is possible through the protection and enhancement of fundamental rights referred to in the European Charter of Human Rights. They reflect the constitutive values of European policies, are

legally binding, and constitute the reference framework for the legal systems of the Member States - as well as often being used as a prototype for the legislation of other states at an international level.

Consequently, fundamental principles that must be considered essential for the design of human-centered explanations will be listed below. They have been identified starting with those most commonly referred to in the main regulations and guidelines issued by European Parliament and European Commission. A further skimming was carried out trying to identify those principles that were to be considered more directly involved in the dynamics analysed here – namely the possible manipulation of users' will, the potential distortion in the perception of reality, and the assessment of the risks attached to the interaction with AI systems. They could constitute the reference framework to make XAI a useful tool for mitigating the effects of banal deception on the end user, rather than exacerbating possible manipulative drifts.

3.1 Right to the integrity of the person.

Article 3 of the European Chart of Human Rights protects individual physical and also psychological integrity [21]. In addition, the article refers to the value of free and informed consent in healthcare treatment. However, it is a commonly accepted interpretation that consent is conceived as the pivotal instrument of any act affecting a person or one of their available rights – as also demonstrated in the GDPR.

The reference to informed consent is certainly fundamental at a conceptual level. Indeed, it may be considered one of the reasons why a branch of legal experts sees the explanation as a valid tool for shortening – even removing – the information gap that recognizes the non-specialized user of AI as disadvantaged by default. In the scope of this study, on the contrary, this issue does not seem to be decisively relevant. Informing an individual of what is happening, why a given recommendation is being made, or how their data will be processed and stored is certainly essential. At the same time, if the primary purpose of the framework in question is to prevent explanation from becoming an instrument of manipulation, informing is a practice that is neither sufficient nor goal-oriented. This is mainly due to the phenomenon of mindless behavior described above and to the subliminal nature of banal deception. Furthermore, even if the user accepted the dynamic of banal deception, if the result implies the possible infringement of fundamental rights, this presumed acceptance would be considered null and void. This is because fundamental human rights are considered by law to be "unavailable", namely, not subjected to renunciation or negotiation by the holder.

It follows that, in pursuing the scope of the principle of individual integrity, an explanation should also guarantee the respect of the subsequent principles:

Physical Health: This is mainly affected by those AI systems that involve medical aspects or habits that impact health (e.g., quitting smoking or adopting

a different diet). Here it is important to ensure that the interaction and the justifications provided for the recommendations offered follow the standards of care, medical guidelines, and principles of good medical practice that are also followed by human specialists [49,40]. More specifically, it will be important to give the user the most truthful and up-to-date view of their progress and of the appropriateness of their goal. In no way, for the pure purpose of incentivizing and increasing interaction, should the individual be induced to persist in use once the limits set by health standards have been reached (e.g., to continue to lose weight or to increase muscle mass once beyond good medical practice). The system must also be able to interrupt the flow of recommendations/explanations if the user gives signals that they want to use the service outside of safe standards (e.g., setting weight loss standards too low, unbalancing nutrient intake in an unhealthy way, persisting in refusing explanations to bring their goals closer to those set by medical standards). Moreover, according to *ex* Article 3, it is impossible to impose any treatment that the patient does not understand and accept. Likewise, the system cannot use an explanation that exploits means of subliminal and subtle persuasion such as those of banal deception, push to accept recommendations or outcomes that induce potentially health-impacting actions.

Physiological Health: Although the subject of mental and psychological health is becoming increasingly pervasive in the law, there is still no objective and uniquely accepted definition of it in doctrine. Among the first steps taken by jurisprudence was to decouple this concept from the occurrence of mental disorders in the clinical-pathological sense [57]. By interpretation, it could be useful to start from the very concept of integrity, which is brought back – by analogy with other areas – to the preservation of unity, of the compactness of the subject of analysis. Anything that interferes with this idea of integrity, causing a split in an individual's coherence with themselves, their beliefs, and their feelings, alters the integrity thus understood [13]. Therefore, the explanation must aim at mitigating those aspects of banal deception that may manipulate users' perception and will, thus leading them to prefigure a conception of reality, of themselves, and of their own needs. If, as briefly investigated above, manipulation is that phenomenon that goes beyond reasoning, the explanation must be structured in such a way as to counterbalance the functional effects of deception without leading to the distortion of individuals' self-awareness.

3.2 Respect for private and family life

Article 7 of the European Chart of Human Rights protects the respect for private and family life. In this concept, the security and confidentiality of the home environment and correspondence are explicitly mentioned [21].

This article is often considered to have a rather broad semantic and applicative scope that is not easy to substantiate. Indeed, the concept of private life includes instances belonging to the aforementioned Article 3, encompassing aspects of physical and psychological integrity. However, its primary focus should be on aspects of identity and autonomy [1].

Identity: Personal identity from the perspective of pure private law is understood as the unique, personal recognition of an individual. However, nowadays, the law has also opened up in interpreting identity as the set not only of objective and verifiable data attributable to an individual. It also included the specialty of people's cognitive-psychological dimension, the way they perceive themselves, their beliefs, and their will. Thus, personal identity can be harmed by expedients aimed at inducing changes, habits, and desires that are not consistent with the idea that an individual has of themself and with the lifestyle and beliefs they have chosen for themselves. Manipulation is precisely able to target self-awareness and induce attitudes that are lucidly not justifiable or recognizable as proper by those who perform them. Therefore, in this context, the explanation must have the primary role of allowing the users at each stage of the interaction to realign with themselves, accepting only the recommendations they consider in line with their own convictions and goals. It must also always put the user in the position to question and challenge a recommendation/ motivation received with an exchange that includes acceptance and/ or rejection and a more active and argumentative understanding.

Inviolability of private space: Following the examination of the concept of identity above, arguing that the domestic environment – or more generally private – should be protected from external inferences means also to include elements that go beyond the mere concept of home or private property. In this *space*, it is also necessary to bring back habits and the deeper aspects of daily life [53]. For this reason, while banal deception has the primary purpose of facilitating the inclusion of AI systems in the most intimate contexts – both physically and cognitively – through the explanation, we should aim to ensure that this happens only to the extent that it is essential for more effective interaction.

Autonomy of private choice: The two satellite principles analyzed so far lead us to an incontrovertible conclusion: the protecting the individual autonomy, here to be understood as "freedom to choose for oneself" [47]. This implies the negative freedom to reject what one does not want or is not willing to accept. The explanation, this view, must be designed in such a way as to always ensure the possibility of rejecting both a given justification and a recommendation, as well as to go back on decisions previously taken in order to modify them potentially. XAI must become the main tool for the user to always keep in mind their ability to release themselves from the application and perceive that they can always choose for themselves in each phase of the interaction.

3.3 Human Dignity

The right to human dignity is presented last, but certainly not in importance. In fact, it is expressed in Article 1 of the European Charter of Human Rights [21], just as it is often the first right to be enunciated and guaranteed in most Constitutional Charters and international treaties [8]. The reason why it is the last

to be analyzed here is due to the twofold approach with which it is addressed by the doctrine.

Human dignity is considered a "constellation principle", around which all others orbit and by reason of which all others find their justification and their – possible – balance [69]. This is why it is considered the founding element of freedom, justice, and peace [4], enforced as such by the United Nations General Assembly. Nonetheless, the difficulty in providing an objective and universally accepted empirical demonstration, its imperative character, and the lack of definition [30] have meant that – without discussing its legal value – its direct concrete application has been questioned.

Consequently, it might be useful to identify a still legally relevant concept that serves as an element of operability in the practice of the higher principle of human dignity – at least with regard to the human-AI interaction context.

Vulnerability: The principle of human dignity protects the intrinsic value that each individual possesses only as a human being [26] and, consequently, protects the individual's autonomy against forms of constraint. Both such aspects are the foundation of any comprehensive discussion of vulnerability [32]. The main difference is that any reference to human dignity often lends a universalist approach that has not always been easy to apply to concrete cases and disciplines. In other words, the reference to human dignity denotes a reference to a pivotal foundation of modern Constitutions, to a fundamental and inalienable human right that, as such, are easier to include in a principle-oriented argumentation that enshrines the theoretical frame of reference rather than an instrument directly applicable, without being translated into concepts of more immediate practical implementation [9]. On the contrary, vulnerability has already been used by the European Court of Human Rights as an indirect tool to evaluate the impact that some phenomena have on human dignity [64].

It follows that the concept of vulnerability can be used to substantiate the influence of banal deception at many levels. Depending on the result of such an investigation, it could become possible to draw up a range of possible repercussions. Depending on the range taken as a reference, it may be determined how to react – whether to correct, reevaluate, or stop the practice.

In particular, the objective and factual analysis using vulnerability as the materializing principle of the universal right of human dignity will allow the following satellite principles to be monitored and ensured:

Inclusion: An AI system capable of engaging the users at a psychological level, often going beyond their cognitive structures, is also able to reduce the level of socially relevant interactions, including those with healthcare specialists or psychotherapists (e.g., in the case of virtual nutritional coaches or behavioral changes applications [15,14]). The explanation can act as a pivotal element so that the phenomenon of banal deception does not result in induced or encouraged addiction and that the user always has the opportunity to interface with domain experts when the system recognizes the establishment of dynamics of dependence

on interaction and loss of connection with reality (including the reality of one's physical or mental condition)

Humanisation of the interaction: The above-mentioned dynamics could also lead to a phenomenon of dehumanization of individuals [56], conceived by the system as an aggregate of data and inputs, more than as human beings with their own weaknesses, their doubts, their inherent biases. The privileged, continuous, and often unique interaction with a responsive AI system that appears reliable and friendly can lead to conceiving that mode of interaction as the benchmark for evaluating all the others. This means creating – and consolidating – expectations of readiness in output, systematic argumentation, and acclimatization to errors that can generate two possible situations. On the one hand, the phenomenon of mechanomorphism [17], according to which the user becomes accustomed to the methods of communication, the timing, and the content provided by a given application, reshaping on it the expectations that are created on interactions with other human beings. In other words, technology becomes the model through which to navigate and act in the real world rather than the opposite. On the other hand, people can lower their expectations, their communicative level, and their complexity of thoughts to facilitate understanding of the system and its work. In this case, humans put themselves at the service of technology, anthropomorphizing its technical shortcomings instead of being its owners and users. In such a context, explanations can modulate interactive dynamics, ensure individuals always maintain control, specify technical dysfunctions, and modulate interaction times and pause from usage.

4 Principles in action

The principles listed above represent the starting framework into which the concrete approach to deception in human-AI interaction can be inserted. We have realized that it is impossible to completely remove how banal deception impacts human experience with new technologies, both technically and functionally (i.e., from programming and psychological-perceptual reasons). Therefore, an approach that aims at realizing a true human-centered AI will have to address the issue, trying to modulate its impact, maximizing the benefits, and reducing the potential harmful effects.

This may be made possible through a two-phase approach: (i) a new method of assessing the impact of new technologies and their design and (ii) a control system to be implemented in the explanation and communications stages themselves.

For the sake of completeness, both levels will be presented. However, for the more specific purpose of this discussion, only the latter will be explored in depth, leaving a more detailed analysis of the former for future work.

4.1 Vulnerability Impact Assessment

The Vulnerability Impact Assessment (VuAI) aims to systematically identify, predict, and respond to the potential impacts of the technology used on human vulnerability. Moreover, in a broader sense, it could become crucial in assessing government policies at both European and Member States levels. It would be framed by international legal and ethical principles and fundamental human rights.

This could be an important instrument for mitigating possible harms occurred in, or because of, the interaction with AI systems designed in accordance with banal deception dynamics, while ensuring accountability. To this end, it could be relevant to make VuAI mandatory human rights due diligence for providers. Such an essential step can foster the achievement of the EU goals for the development and deployment of human-centered AI. It is also central for understanding and determining the levels of risk of AI systems, even when it is not immediate or objectively identifiable *ex-ante* the impact of the AI system on human rights, and even when there is little evidence and knowledge for detecting the risk level.

Once the reference structure is concretely developed, it will make it possible to divide the interaction models into classes, which consider (i) the interactive mode, (ii) the nature of the expected average user, and (iii) the ultimate goal of the interaction itself. Each of them will be linked to a range of impacts estimated on the profiles of human vulnerability, investigating whether it is respected and supported, exalted, or exploited for purposes not aligned with the right of human dignity (which we said underlies vulnerability and legitimizes enforceability). Each estimate of the impact on human vulnerability must correspond to a range of deception to be considered, not further reducible for reasons related to the functionality and acceptability of the AI system.

To this end, the Impact Assessment may consist of two elements: an assessment tool (e.g., a questionnaire or a semi-automatized feedback analysis tool) and an expert committee. The first is useful to define the features which may elicit or directly exploit vulnerability, thus inducing over-dependency and manipulating people's will. What would make hypothetical harms to vulnerability legally enforceable is the connection it has with human dignity. More precisely, the relation of direct derivation vulnerability has with this foundational right, as previously addressed.

The Expert Committee would analyze these aspects in the specific context of usage or with regard to the given technology under evaluation.

A more detailed and systematic definition of this new impact assessment and its scope will be further discussed in future works.

4.2 Principles-based XAI

The framework described above represents the theoretical basis and justification element in a legal perspective of the modulation of deception in HCI.

To give substance to this new perspective, centered on a reassessment of the concept of vulnerability, it would be useful to formalize a knowledge graph. It would consist of a way to represent and structure the contextual (i.e. domain/application-related) concepts and information.

Its purpose would be to identify the characteristics of vulnerability, the nature of the relationships existing between its elements, and the influence of the context of use. In this way, systems leveraging XAI techniques would be able to identify and parse, with a good degree of approximation, any risk elements that might arise, even at run time, due to extensive interaction with the user. Once these "warnings" have been identified, the system will have to determine, select, and execute the ideal countermeasures.

The first approach would be to analyze/revise the explanations themselves, to counterbalance the otherwise exaggerated effects of banal deception. The main rebalancing effects of the interaction would consist in trying to engage the user as much as possible on a logical-rational level – both in the way the explanation is given and in its content. One way is to exploit design features that target areas of the brain antagonistic to those affected by the phenomenon of banal deception. Those same studies that have guided researchers in structuring the interaction "for deception" may provide insights into how to structure it to mitigate the same phenomenon (e.g., imposing semantic and thematic boundaries structured as logic rule sets). Moreover, contact with a second human opinion should be reiterated and encouraged, especially in the case of e-health applications or decision-making procedures. In doing so, it is important to reaffirm the individual's right to make autonomous choices and to ask for all necessary confirmation or information to form as critical and autonomous a thought as possible. In cases of intense risk to psychological integrity, mainly concerning aspects of presumed addiction, excessive dependence, isolation, and distortion of one's own initial will/goal, XAI-powered systems must suggest, even enforcing, a suspension of use and/or regular interaction (even periodically), until a decrease in the risk factors is registered, based on specific requests addressed to the user by the system itself. A possible strategy to enact such an intuition can be to periodically question the user's understanding and alignment with the necessary knowledge to safely use a given system and the "integrity" of their judgment/standing point.

If such an intervention might not be sufficient – or if the risk is high or difficult to assess by the application alone – the case should be handed over to a human domain expert that, for health/safety-critical applications, must always have the means to assess and intervene if necessary (e.g., a psychologist in stress-relieve personal assistants or a nutritionist/medical doctor in nutrition assistant scenarios).

5 Vulnerability as a guiding tool: scepticism and potentials

The application of the framework here proposed and the future theoretical investigations to be developed in this regard imply that vulnerability assumes a central role.

Especially in the European tradition, it might be natural to ask whether it is necessary to refer to a concept that is not purely legal in the strict sense. Indeed, it could be objected that a multiplication of legal principles does not benefit a clear, coherent, and streamlined application of the law, with the additional risk to become a mere exercise in style. This concern is certainly to be welcomed. However, the choice made here is intended to respond to an issue – that of the protection of the user and their psychological and physical integrity – which is particularly challenged by AI technologies and XAI systems designed according to the logic of banal deception. From this point of view, ignoring the central concept of human vulnerability, or relegating it to a particular condition, which does not change the application or formulation of the law in general, appears short-sighted and not resolving.

Moreover, embracing a universalistic conception of vulnerability means aligning with the framework outlined by Martha Fineman in her Vulnerability Theory (herenforth VT) [32]. It suggests a theoretical framework of redistribution of responsibility, burdens, support tools, and resilience, starting from the assumption that these measures are functional to the well-being of society, overcoming individual particularisms. This approach seems well suited to the analysis of the interaction between non-specialized users and AI, even when mediated by XAI. In fact, as demonstrated above, the dynamics of banal deception target cognitive and emotional constructs common to humans, not specific categories.

Despite what is sometimes disputed, VT is not a mere argumentative exercise, devoid of practical evidence. It is true that it has never been used holistically as a means of reforming, drafting, or adapting legislation yet. Nevertheless, some of its instances – universalistic interpretation of vulnerability, dependence as transposition of its concept, resilience as its opposite – have been cited or applied without explicit reference in the rulings of the European Court of Justice [39] and the European Court of Human Rights [36], and to address international human rights issues [37]. Moreover, it should not be considered that if vulnerability is no longer diversified in degrees, this implies not having regard to situations of particular fragility . It only means changing the perspective of the investigation – excluding that there may be users completely immune to the possible negative effects of deceptive mechanisms – and to encourage the research and formulation of legal and technological interventions. The latter should aim at creating resources and resilience tools for all, without excluding the possibility that they may be more decisive in some circumstances than in others.

6 Conclusions and Future Works

This study has focused on deception in human-AI interaction, arguing that *banal* deception plays a central role in enhancing the effectiveness of the system's functionalities, raising confidence and appreciation from the users in the given technology. However, the very concept of deception often has a negative connotation, being conceived as a tool of manipulation. Thus, we have pointed out that banal deception is intrinsic in human-AI interaction (HCI, HRI), and it consists of five fundamental elements: ordinary character, functional means, people's obliviousness-centered, low definition, and pre-definition.

Nevertheless, the fact that the banal deception has arisen with the precise intention of encouraging the most efficient interaction possible does not exclude the possibility of harmful side effects — above all, manipulative drifts.

In such a scenario, although XAI can be relevant to counteract such negative effects, the design and content of explanations can also exacerbate the phenomenon described above. This is because individuals are naturally led to attribute human qualities to inanimate objects, even more in the case of AI systems. Such a statement has already been proved by the Media Equation Theory [55], the CASA model [52], and the mechanism of Mindless Behaviour [58].

It has been emphasized here that this tendency is an integral part of an irreducible profile of vulnerability that characterizes human beings as such.

For this reason, this study claims the need to identify a framework to which XAI must refer (or embed) in the design of explanations — to counterbalance the possible harmful effects of banal deception and enhance its benefits. The principles here identified are: (i) the right to the integrity of a person — which consists of the right to both physical and psychological health; (ii) the respect for private and family life — which also includes the protection of personal identity, the inviolability of private space, and the autonomy of one's own choices –; (iii) right of human dignity – from which the right to inclusion, and the need to enforce a humanisation of the interaction derive. From this perspective, we suggested a new interpretation of the concept of vulnerability as an indirect instrument to evaluate the impact of the phenomenon of banal deception on human dignity.

Such a theoretical framework could represent the essential mean towards a: Vulnerability Impact Assessment and the implementation of related knowledge graphs enabling a semi-automated pre-check and possible handover to humans domain expert if necessary. The first measure could serve as a tool to assess how/how much banal deception impact humans' inherent vulnerability, taking into account the application under analysis and the nature of both the users and the interaction. Thus, it could be possible to address and legally enforce the level of deception to be considered admissible case by case. The second measure would act at a system level, placing a continuous run-time assessment of possible manipulative drift "warnings". Hence, these dynamics could be mitigated and/or limited promptly through the timely intervention of the systems themselves and the human specialist (triggered) intervention if necessary.

Future works will focus on in-depth analysis of the vulnerability concepts and their formalization (from a schematic/systemic perspective) to then enable the design and implementation of semi-automated reasoners bridging data-driven (run-time) generated explanations, legally relevant vulnerability concepts, and the underneath rule-based system vehiculating the overall system dynamics.

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, and by *(i)* the Swiss National Science Foundation (G.A. 20CH21_195530), *(ii)* the Italian Ministry for Universities and Research, *(iii)* the Luxembourg National Research Fund (G.A. INTER/CHIST/19/14589586), *(iv)* the Scientific and Research Council of Turkey (TÜBİTAK, G.A. 120N680).

References

1. Adrienne, K.: Effective enforcement of human rights: The tysiac v. poland case. Studia Iuridica Auctoritate Universitatis Pecs Publicata **143**, 186 (2009)
2. AI, H.: High-level expert group on artificial intelligence (2019)
3. 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)
4. Assembly, U.G., et al.: Universal declaration of human rights. UN General Assembly **302**(2), 14–25 (1948)
5. Astromskė, K., Peičius, E., Astromskis, P.: Ethical and legal challenges of informed consent applying artificial intelligence in medical diagnostic consultations. AI & SOCIETY **36**, 509–520 (2021)
6. Baker, R.S., De Carvalho, A., Raspat, J., Aleven, V., Corbett, A.T., Koedinger, K.R.: Educational software features that encourage and discourage "gaming the system". In: Proceedings of the 14th international conference on artificial intelligence in education. pp. 475–482 (2009)
7. Banks, J.: Theory of mind in social robots: Replication of five established human tests. International Journal of Social Robotics **12**(2), 403–414 (2020)
8. Barroso, L.R.: Here, there, and everywhere: human dignity in contemporary law and in the transnational discourse. BC Int'l & Comp. L. Rev. **35**, 331 (2012)
9. Beyleveld, D., Brownsword, R.: Human dignity in bioethics and biolaw (2001)
10. 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)
11. Bissoli, L., Bonacina, D., Dalla Riva, N., Demrozi, F., Jereghi, M., Marchiotto, N., Perbellini, G., Pernice, B., Pizzocaro, E., Pravadelli, G., et al.: A virtual coaching platform to support therapy compliance in obesity. In: 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC). pp. 694–699. IEEE (2022)

12. Bradeško, L., Mladenić, D.: A survey of chatbot systems through a loebner prize competition. In: Proceedings of Slovenian language technologies society eighth conference of language technologies. vol. 2, pp. 34–37. sn (2012)
13. Bublitz, J.C.: The nascent right to psychological integrity and mental self-determination. The Cambridge handbook of new human rights: Recognition, novelty, rhetoric pp. 387–403 (2020)
14. Calvaresi, D., Calbimonte, J.P., Siboni, E., Eggenschwiler, S., Manzo, G., Hilfiker, R., Schumacher, M.: Erebots: Privacy-compliant agent-based platform for multi-scenario personalized health-assistant chatbots. Electronics **10**(6), 666 (2021)
15. Calvaresi, D., Carli, R., Piguet, J.G., Contreras, V.H., Luzzani, G., Najjar, A., Calbimonte, J.P., Schumacher, M.: Ethical and legal considerations for nutrition virtual coaches. AI and Ethics pp. 1–28 (2022)
16. Calvaresi, D., Cesarini, D., Sernani, P., Marinoni, M., Dragoni, A.F., Sturm, A.: Exploring the ambient assisted living domain: a systematic review. Journal of Ambient Intelligence and Humanized Computing **8**(2), 239–257 (2017)
17. Caporael, L.R.: Anthropomorphism and mechanomorphism: Two faces of the human machine. Computers in human behavior **2**(3), 215–234 (1986)
18. 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)
19. Ch'ng, S.I., Yeong, L.S., Ang, X.Y.: Preliminary findings of using chat-bots as a course faq tool. In: 2019 IEEE Conference on e-Learning, e-Management & e-Services (IC3e). pp. 1–5. IEEE (2019)
20. Cisek, P.: Beyond the computer metaphor: Behaviour as interaction. Journal of Consciousness Studies **6**(11-12), 125–142 (1999)
21. Commission, E.: Charter of fundamental rights of the european union, 2012/c 326/02. Official Journal of the European Union (2012)
22. Coons, C., Weber, M.: Manipulation: theory and practice. Oxford University Press (2014)
23. Crevier, D.: AI: the tumultuous history of the search for artificial intelligence. Basic Books, Inc. (1993)
24. Crowther-Heyck, H.: George a. miller, language, and the computer metaphor and mind. History of Psychology **2**(1), 37 (1999)
25. Dennett, D.C.: The intentional stance. MIT press (1987)
26. Dicke, K.: The founding function of human dignity in the universal declaration of human rights. In: The concept of human dignity in human rights discourse, pp. 111–120. Brill Nijhoff (2001)
27. Druce, J., Niehaus, J., Moody, V., Jensen, D., Littman, M.L.: Brittle ai, causal confusion, and bad mental models: Challenges and successes in the xai program. arXiv preprint arXiv:2106.05506 (2021)
28. Edmonds, B.: The constructibility of artificial intelligence (as defined by the turing test). The Turing test: the elusive standard of artificial intelligence pp. 145–150 (2003)
29. Epley, N., Waytz, A., Cacioppo, J.T.: On seeing human: a three-factor theory of anthropomorphism. Psychological review **114**(4), 864 (2007)
30. Fabre-Magnan, M.: La dignité en droit: un axiome. Revue interdisciplinaire d'études juridiques **58**(1), 1–30 (2007)
31. Fejes, E., Futó, I.: Artificial intelligence in public administration–supporting administrative decisions. PÉNZÜGYI SZEMLE/PUBLIC FINANCE QUARTERLY **66**(SE/1), 23–51 (2021)

32. Fineman, M.A.: Vulnerability: reflections on a new ethical foundation for law and politics. Ashgate Publishing, Ltd. (2013)
33. Glocker, M.L., Langleben, D.D., Ruparel, K., Loughead, J.W., Gur, R.C., Sachser, N.: Baby schema in infant faces induces cuteness perception and motivation for caretaking in adults. Ethology **115**(3), 257–263 (2009)
34. 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)
35. Guzman, A.L.: Making ai safe for humans: A conversation with siri. In: Socialbots and their friends, pp. 85–101. Routledge (2016)
36. Heri, C.: Responsive Human Rights: Vulnerability, Ill-treatment and the ECtHR. Bloomsbury Academic (2021)
37. Ippolito, F.: La vulnerabilità quale principio emergente nel diritto internazionale dei diritti umani? Ars interpretandi **24**(2), 63–93 (2019)
38. Kim, J., Park, K., Ryu, H.: Social values of care robots. International journal of environmental research and public health **19**(24), 16657 (2022)
39. Knijn, T., Lepianka, D.: Justice and vulnerability in Europe: An interdisciplinary approach. Edward Elgar Publishing (2020)
40. Kopelman, L.M.: The best interests standard for incompetent or incapacitated persons of all ages. Journal of Law, Medicine & Ethics **35**(1), 187–196 (2007)
41. Korn, J.H.: Illusions of reality: A history of deception in social psychology. SUNY Press (1997)
42. Lee, S.l., Lau, I.Y.m., Kiesler, S., Chiu, C.Y.: Human mental models of humanoid robots. In: Proceedings of the 2005 IEEE international conference on robotics and automation. pp. 2767–2772. IEEE (2005)
43. Leonard, A.: Bots: The Origin of the New Species. Wired Books, Incorporated (1997)
44. Leonard, T.C.: Richard h. thaler, cass r. sunstein, nudge: Improving decisions about health, wealth, and happiness (2008)
45. Magid, B.: The meaning of projection in self psychology. Journal of the American Academy of Psychoanalysis **14**(4), 473–483 (1986)
46. Margalit, A.: Autonomy: Errors and manipulation. Jerusalem Review of Legal Studies **14**(1), 102–112 (2016)
47. Marshall, J.: Personal freedom through human rights law?: Autonomy, identity and integrity under the European convention on human rights. Brill (2008)
48. Massaro, D.W.: The computer as a metaphor for psychological inquiry: Considerations and recommendations. Behavior Research Methods, Instruments, & Computers **18**, 73–92 (1986)
49. for the Study of Ethical Problems in Medicine, P.C., Biomedical, (US)., B.R.: Making health care decisions v. 1, vol. 1. President's Commission for the Study of Ethical Problems in Medicine and ... (1982)
50. Mitnick, K.D., Simon, W.L.: The art of deception: Controlling the human element of security. John Wiley & Sons (2003)
51. Nass, C., Moon, Y.: Machines and mindlessness: Social responses to computers. Journal of social issues **56**(1), 81–103 (2000)
52. Nass, C., Steuer, J., Tauber, E.R.: Computers are social actors. In: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 72–78 (1994)
53. Natale, S.: Deceitful media: Artificial intelligence and social life after the Turing test. Oxford University Press, USA (2021)

54. Papacharissi, Z.: A networked self and human augmentics, artificial intelligence, sentience. Routledge UK (2018)
55. Reeves, B., Nass, C.: Media equation theory. Retrieved March **5**, 2009 (1996)
56. Roberts, T., Zheng, Y.: Datafication, dehumanisation and participatory development. In: Freedom and Social Inclusion in a Connected World: 17th IFIP WG 9.4 International Conference on Implications of Information and Digital Technologies for Development, ICT4D 2022, Lima, Peru, May 25–27, 2022, Proceedings. pp. 377–396. Springer (2022)
57. Sabatello, M.: Children with disabilities: A critical appraisal. The International Journal of Children's Rights **21**(3), 464–487 (2013)
58. Sætra, H.S.: The parasitic nature of social ai: Sharing minds with the mindless. Integrative Psychological and Behavioral Science **54**, 308–326 (2020)
59. Sarrafzadeh, A., Alexander, S., Dadgostar, F., Fan, C., Bigdeli, A.: "how do you know that i don't understand?" a look at the future of intelligent tutoring systems. Computers in Human Behavior **24**(4), 1342–1363 (2008)
60. Schneider, B.: You are not a gadget: A manifesto. Journal of Technology Education **23**(2) (2012)
61. Schreiber, D.: On social attribution: implications of recent cognitive neuroscience research for race, law, and politics. Science and engineering ethics **18**, 557–566 (2012)
62. Seymour, W., Van Kleek, M.: Exploring interactions between trust, anthropomorphism, and relationship development in voice assistants. Proceedings of the ACM on Human-Computer Interaction **5**(CSCW2), 1–16 (2021)
63. Switzky, L.: Eliza effects: Pygmalion and the early development of artificial intelligence. Shaw **40**(1), 50–68 (2020)
64. Timmer, A.: A quiet revolution: vulnerability in the european court of human rights. In: Vulnerability, pp. 147–170. Routledge (2016)
65. Trower, T.: Bob and beyond: A microsoft insider remembers (2010)
66. Turing, A.M.: Computing machinery and intelligence. Springer (2009)
67. White, L.A.: The symbol: The origin and basis of human behavior. Philosophy of Science **7**(4), 451–463 (1940)
68. Yang, Y., Liu, Y., Lv, X., Ai, J., Li, Y.: Anthropomorphism and customers' willingness to use artificial intelligence service agents. Journal of Hospitality Marketing & Management **31**(1), 1–23 (2022)
69. Zatti, P.: Note sulla semantica della dignità. Maschere del diritto volti della vita pp. 24–49 (2009)