Risk and Exposure of XAI in Persuasion and Argumentation: The case of Manipulation

Rachele Carli[1,2][0000−00020−8689−285X], Amro Najjar[3][0000−0001−7784−6176], 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] University of Luxembourg, Luxembourg
[3] Luxembourg Institute of Science and Technology (LIST), Luxembourg amro.najjar@list.lu
[4] University of Applied Sciences Western Switzerland, Switzerland davide.calvaresi@hevs.ch

Abstract. In the last decades, Artificial intelligence (AI) systems have been increasingly adopted in assistive (possibly collaborative) decision-making tools. In particular, AI-based persuasive technologies are designed to steer/influence users' behaviour, habits, and choices to facilitate the achievement of their own – predetermined – goals. Nowadays, the inputs received by the assistive systems leverage heavily AI data-driven approaches. Thus, it is imperative to have transparent and understandable (to the user) both the process leading to the recommendations and the recommendations. The Explainable AI (XAI) community has progressively contributed to "opening the black box", ensuring the interaction's effectiveness, and pursuing the safety of the individuals involved. However, principles and methods ensuring the efficacy and information retain on the human have not been introduced yet. The risk is to underestimate the context dependency and subjectivity of the explanations' understanding, interpretation, and relevance. Moreover, even a plausible (and possibly expected) explanation can lead to an imprecise or incorrect outcome or its understanding. This can lead to unbalanced and unfair circumstances, such as giving a financial advantage to the system owner/provider and the detriment of the user.
This paper highlights that the sole explanations – especially in the context of persuasive technologies – are not self-sufficient to protect users' psychological and physical integrity. Conversely, explanations could be misused, becoming themselves a tool of manipulation. Therefore, we suggest characteristics safeguarding the explanation from being manipulative and legal principles to be used as criteria for evaluating the operation of XAI systems, both from an *ex-ante* and *ex-post* perspective.

Keywords: XAI · Manipulation · Persuasion

1 Introduction

Since the last decade, Artificial Intelligence (AI) systems have pervaded a wide range of daily-living applications. Smart homes and smart cities [58], AI-powered

job recruitment systems [2], and e-health applications [10] are examples of state of the art complex and personal AI systems. Those applications record remarkable results. However, the majority is powered by black-box machine learning (ML) and are trained on biased data and behave unintelligibly for human users [27]. Such a lack of understandability reduces the system's acceptability from the user perspective [49]. Furthermore, it has been shown that users tend to attribute a State of Mind (SoM) to AI systems to better process/make sense of their behavior. Therefore, if a user misunderstands the system's plans and intention, the resulting SoM is erroneous and leads to failures (or can even compromise the user safety) [29].

Explainable AI (XAI) strives to bridge this gap. The first contributions of the current wave of XAI date back to mid 2010s [26]. Since then, the XAI perspective has broadened, approaching virtual entities (so-called agents) and robotos [3], automated planning [23] and recommender systems [69].

Nevertheless, despite the recent advances in XAI, the latter still lacks solid principles and methods enforcing the efficacy of XAI. The main challenge is acknowledging that: (i) the user's understanding is context, domain, and user dependent, and (ii) a plausible explanation does not necessarily mirror precisely or faithfully the underlying decision-making mechanism. This can lead to imprecise or incorrect outcomes.

Therefore, the safety of the individuals involved in the interaction might be undermined. Indeed, besides the well-known threats concerning privacy and data protection [37,57,63], incorrect or unfair outcomes can interfere with the users' volitional and decisional processes. In such a case, liability's allocation problem – which is, anyway, already not easy to solve [5]– arises alongside the determination of the causal link [17], and the prompt identification and prevention of the possible damages [24].

This paper argues that explanation and XAI, especially in the contexts of persuasive technologies, are still prone to risks and cannot be considered sufficient to protect users' psychological and physical integrity. Conversely, it can itself be a tool of manipulation. Therefore, we investigate the characteristics necessary for an explanation or XAI system not to be manipulative. Moreover, we suggest desiderata that can be used as criteria for evaluating the operation of AI systems, both from an ex-ante and ex-post perspectives.

The rest of this paper is organized as follows.

Section 2 presents the state of the art focusing on XAI and its milestones, the legal perspective on the topic, principles such as transparency, safety, and autonomy, and concepts such as persuasion and manipulation. Section 3 elaborates on the legal entanglements beyond XAI, explaining why an explanation alone cannot be considered sufficient to make the algorithm fully transparent, effectively safe, and, as a consequence, it cannot preserve the user's autonomy, and illustrating desiderata for a non-manipulative XAI. Finally, Section 4 concludes the paper.

2 Background & State of the Art

This section presents the background and state of the art of the disciplines intersecting explainability and persuasion such as XAI, AI & legal reasoning, and self-authorship.

2.1 Explainable AI

Between the 80s and 90s, XAI has been called by the widespread use of expert systems [67,28]. Since these relatively early days, several works have attempted to explain the decisions of expert systems, but also of neural network [64,15]. After a while, the interest in expert systems and XAI waned as AI entered one of its so called AI winters, which has seen the advent proliferation of white-box ML approaches (i.e., decision-trees) [12].

In the 2010s, the development of black-box ML and DL techniques achieved several breakthroughs, giving the sub-symbolic AI new momentum [39].

However, the intriguing results of some ML black-box have raised several concerns about the lack of transparency of these mechanisms [62,27] – thus, reviving the interest in "opening the black-box" [26]. Techniques such as LIME [55] and SHAP [42] have been proposed to interpret the cutting-edge ML, Deep Neural Networks (DNN), Reinforcement Learning, and Deep Reinforcement Learning (DRL) mechanisms [41]. Such initiative expanded beyond the pure domain of ML. Several works started to integrate XAI in the domains of automated planning [23], recommender systems [69], agents and multi-agent systems [3], and robots [29]. Moreover, this new advent of XAI has spurred several works aiming at defining explainability [30], getting inspiration from the way humans explain their behaviour [48], defining metrics for explainability [31], adopting a human-centric approach (where, unlike earlier works in XAI, the main determinant of how successful an explanation is the degree of understanding, trust and satisfaction it inspires in the human-user receiving their explanations) [49], and formalizing models for mixed human-agent-human explainable interactions [11]. This has resulted in a body of work aiming at exploring personalized and context-aware explanations, which improve the human understanding of AI systems and thereby increase their trustworthiness and their ability to influence human behavior [3]. Several initiatives from governmental and non-governmental, and international institutions [59] have supported this move for XAI and a broader view of ethical and trustworthy AI [4].

Nevertheless, it has been pointed out that explainability alone does not fully meet expectations and does not guarantee the achievement of the objectives for which it was theorized [20]. This is even clearer if we analyze the legal principles that algorithmic intelligibility would be required to pursue.

2.2 Explainable AI through the lens of legal reasoning

Recommender systems can provide explanations to the user to raise their awareness about the dynamics of the interaction, trust in the system, evaluation of

the quality of the interaction, and an idea about the willingness to follow the recommendation. Transposing these goals into the legal domain, we could say that XAI is functional in achieving (i) transparency, (ii) security of use, (iii) decision-making autonomy. However, given the nature of AI systems, and the psychological and cognitive mechanisms inherent in human beings, these purposes present some criticalities.

Transparency XAI is considered crucial to making the recommendation's process and the functionality of the system understandable by humans. That can be relevant to determine the quality of the output and to identify possible errors [25]. On the one hand, this perspective is grounded on the idea that people have the right to know why they were affected by the instructions or suggestions of a machine, but even how they could possibly be affected in the future [68]. On the other hand, transparency has often been conceived - even outside the context of new technologies - as an essential concept to allow an effective vindication of infringed rights and the consequent compensation for the damage caused to them [8].

XAI represents just the first step towards accountability - in particular in the contest of digital data and algorithmic decision-making - and it cannot be considered a self-sufficient instrument [45]. Moreover, the explanation must be understandable to different stakeholders, who have different levels and types of knowledge. Therefore, an outcome might be transparent for a user, yet less effective for another individual (or group) [51].

As a consequence, in view of the recurrent reference - also from a regulatory point of view [18] [1] [47]- to the principle of transparency, some questions remain open. In particular, it would be appropriate to evaluate: (i) whether data-driven AI systems can be really transparent for non-specialized users and (ii) whether transparency is always possible and advantageous.

Safety Explanations aim to make the user able to develop an appropriate mental representation of the AI systems they are interacting with. In this way, it is possible to distinguish correct recommendations from incorrect ones, and so to offer the individuals a tool to mitigate *ex-post* errors that the machine can do, due to its design or as a result of the interaction itself [70]. This theory foresees the presence of the "human in the loop" as an expression of respect for human dignity [33]. However, it also assumes that users are cognitively engaged with the given explanation and that the provided information is useful to create an accurate mental representation of the actual characteristics of the system/device.

Since neuroscience demonstrates that such a mechanism cannot be taken for granted, it seems appropriate to further investigate: (i) whether the implementation of XAI models in AI systems has a real and direct consequence on the safety, (ii) whether a user is really able to foresee or prevent harmful consequences just relying on an explanation.

Autonomy The users ability to understand the outcome and the decision-making process of the technology with which they interact is considered a key element in encouraging people's autonomy [6]. This concept is underlined even by the High-Level Expert Group on AI (HLEG), which has clearly stated that individuals have to be put in condition to maintain the ability to self-determine themselves while interacting with an AI system [1].

In this view, the principle of autonomy implies the faculty to choose – and to live – by one's values [43]. Nonetheless, it should be noticed that being autonomous does not mean being entirely devoid of external influences and internal biases. Despite the abundant literature on the right of self-determination, in fact, human beings are just partially "their own making" [53].

Therefore, it should be further examined: (i) whether the explanation can make the users concretely aware of the dynamic of the interaction and, as a consequence, actually free to choose on their own, and (ii) whether XAI may be itself an instrument of manipulation.

In the light of the analysis carried out so far on the impact of the explanation on the principles of transparency, safety, and autonomy, it is clear that there are still some open questions. To address them, it could be advisable a multidisciplinary debate. Nevertheless, for what persuasive and argumentative systems are concerned, this must include a prior attempt to delineate the central differences between the concepts of manipulation and persuasion. To this end, some general distinctions can be drawn, on the basis of the studies carried out so far, especially concerning the studies on consumer protection regulation – as far as legal profiles are concerned –, and the dimension of manipulation of the will - as far as psychological profiles are concerned.

2.3 Persuasion and manipulation: the impact on self-authorship

Persuasion and manipulation are two important aspects of social sciences, yet the debate is stuck on a theoretical level. In particular, the determination of clear boundaries between these two concepts is still under discussion. Therefore, there is still no practical agreement on what unequivocally constitutes the extremes of manipulation. It follows that, from the point of view of enforcement practice, the law has encountered particular difficulties in effectively curbing this phenomenon. This is even more pronounced in the case of new technologies and their application as persuasive and argumentative systems.

Persuasion is linked to what has been defined as the concept of "resistible incentive" [56]. It means that when a system/component is implemented with persuasive techniques, it appeals to the user's ability to make accurate and informed decisions. This implies starting from the personal goal stated by the user and demonstrating that the given recommendation represents the best option to reach it through supporting arguments. Such arguments include evidence and example on which the user may build his own conviction. However, they are described as "resistible", for they do not compulsory determine the final choice, leaving open the possibility of ignoring the suggestion or even the one of acting

in disagreement with it [61]. Faced with each of these three scenarios - agreeing with the decision, denying it, or acting against it - the human being has kept track of their decision-making process and would be able to consciously reconstruct it if necessary or required.

On the contrary, Sunstein, one of the main theorists of manipulation, described it as "an influence that does not sufficiently engage or appeal to the target's capacity for reflection and deliberation" [40]. Said otherwise, manipulating a person means pulling on the same strings of seduction, which means subverting their capacity for critical reflection and reasoned action, even before altering them [44].

Therefore, its target is not rationality but (self-)awareness. It can undermine people's decision-making process and make them decide something different from what they should have decided if they were lucid [14]. However, as mentioned above, it is not just a matter of rationality but awareness of the process that led to the choice. The decisions are not all rational. Some individuals are often influenced by their emotional states, their beliefs – not necessarily based on factual evidence –, or the mechanisms through which they interpret reality. At the same time, they can be convinced to rationally choose something as a result of a suffered manipulation. In the latter case, though, they could not be able to reconstruct the reason that led them to that conclusion and to recognize it as their own [7]. People are not fully self-transparent, and they have no constant consciousness about what drives their choices and thoughts. However, they can make reliable assumptions, which can become the basis for present or future actions or evaluations. Said otherwise, to be aware of a choice may mean to be conscious of the fact that any human being is conditioned by external elements, but to be anyway able to find a personal, own reason, to act in accordance with that conditioning [38].

In turn, although the possibility/risk of a manipulative result is declared – or if the manipulation is not otherwise hidden – it is just as likely to occur and lead to harmful consequences for those involved [40].

In the context of persuasive technologies, it is possible not only for a device to manipulate the user through its output but also through the explanation it provides to support it. Therefore, the need to analyze possible manipulative issues of XAI in persuasion and argumentation is impelling.

Conversely, due to the approach that considers explanations as security guarantees and an incentive to autonomy, this aspect is still largely underestimated in research — both from a legal and a technical point of view. It is common to focus only on the liability and personal data protection profiles – which certainly still require attention and in-depth research – while ignoring implications on the manipulative level for what explainable AI is concerned. Nonetheless, they are equally able to harm individual rights, being also more difficult to intercept and contain. Moreover, considering the ability of manipulation techniques to curb individual self-governance, the current regulatory approach based on the principle of informed consent and enforceable transparency is inadequate for the purpose [9].

In this paper, by manipulation, we intend those situations in which: (i) the explanations supporting a recommendation to achieve an agreed-upon goal are "incorrect'/biased", (ii) the explanations are correct, but the goal diverges from the previously agreed-upon, and (iii) both explanations and overall goal present "incorrect'/biased" elements.

Therefore, it is essential and timely to evaluate how to address these challenges to prevent the explanations from becoming an instrument of manipulation themselves.

3 Legal entanglements beyond XAI

As highlighted in the previous section, just focusing on the indiscriminate implementation of "optimal" XAI models (focusing solely on effectiveness and efficiency) is insufficient to produce explanations holding against the risk of being manipulative — unacceptable in persuasive and behavioral change scenarios.

To this end, it should be acknowledged that XAI has inherent drawbacks due to both its models and human nature. Such limits affect the principles of transparency, security, and autonomy. Addressing them requires a multi-disciplinary approach that would allow answering crucial questions.

3.1 Can data-driven AI systems be actually transparent for non-expert users?

Transparency has often been referred to in sectoral legislation – as in the case of the financial or insurance field [19,65] – and in documents dealing more generally with the regulation of AI systems [1,66]. The main claim is that it should be able (i) to provide the user with awareness about the interaction with the system or its components and (ii) to guarantee an effective liability regime for damages.

The nature of the stakeholders is the first element to take into account. They are usually – but not exclusively – non-specialized individuals, with the most diverse expertise, neither known nor identifiable *ab-origine*. Thus, it is nearly impossible to develop a prototype of a "transparent explanation" that can be entirely understandable and effective to meet everyone's needs.

This would require the creation of many protocols to ensure transparency tailored to the user. Such a solution is complex both technically and conceptually. Indeed, all individuals belonging to the same category, age group, and professional background should be assumed to share the same knowledge. This represents a simplification that not only – however necessary – would end up not solving the problem, but which also would show not to consider the incidence of internal subjective and experiential dynamics on cognitive phenomena. Notably, XAI aims to make the AI system both understandable for the final user and more easily interpretable by experts, who can thus identify potential bugs or criticalities to be corrected. However, the explanations' nature, level of details, and quality of the required language are difficult to homogenize.

Moreover, transparency is not always a guarantee of substantial justice [16]. In the context of applying new technologies to government practice, for instance, it has been proven that data disclosure does not lead to a higher level of security or a better governance [46]. In some cases, then, transparency can also be irrelevant [36]. This is another sign that it is neither essential nor sufficient for accountability. Ordinary individuals regularly use tools ignoring their technical specifications or process leading to a given output (i.e., personal computers or very sophisticated industrial machines). However, their use is possible based on a – probably subconscious – trust in the competence of the technician who developed the device and the possibility to ask experts in case of malfunctioning. This does not deny the specificity of modern AI systems compared to earlier machines, mainly because they are increasingly implemented with ML techniques and NN. Nevertheless, it should be emphasized that, for this very reason, it is not possible to guarantee yet that the utterly independent use of these technologies by non-specialized people can be considered safe solely on the basis of the principle of transparency. Indeed, recent studies demonstrate that, sometimes, systems that lack transparency seem to be very efficient in practice. In contrast, highly transparent systems seem to perform less accurately and to be more exposed, as a consequence, to harmful effects [32].

With regards to the theme of accountability, then, it should be noticed that the legal setting is used to deal with partial, incomplete or inaccessible evidence. Therefore, a lack or even the absence of transparency, even if not advisable, may not represent a detrimental obstacle for the judicial system.

This does not mean that the research for transparency in AI is a useless exercise. Conversely, it must be balanced with other aspects concerning the technical functionality of the system. Balancing the interests – and rights – at stake, we realize that between safety and explainability, it would be more appropriate to invest in resources that make the technology secure, even if they are not able to turn it entirely in a "glass box" [52].

3.2 Can the mere fact of giving an explanation make the system safer?

XAI is often conceived as an element allowing the user to foresee or prevent harmful consequences. For this reason, it is assumed able to foster the system's safety.

However, such a dynamic should not be taken for granted. The desired result is quite explanation-dependant. Suppose the explanation is too complex and might need an unreasonable effort to be understood. In that case, people are induced to pay less attention and to draw – paradoxically - much more hasty and uninformed conclusions [34]. Moreover, even if the information provided is appropriate and well received by the user, this is not in itself sufficient to ensure that it is correctly understood [13]. Consequently, the mere act of providing an explanation cannot be deemed sufficient – though necessary – to ensure that individuals are equipped with the appropriate tools to ensure a conscious and safe use of the system [35].

Another critical aspect – often overlooked – is that including the explanations or making them increasingly rigorous can lower the system's accuracy. At that point, the system might even be safer if the quality of the XAI model is lower or if it is not present at all.

This is not to say that explainability cannot play a relevant role in promoting "a safe" human-machine interaction. However, we should try to act at the source of the problem. For example, we should not limit to provide information that justifies a certain level of trust (and therefore usability) of the system without interfering with individual integrity. Instead, we should focus on the system's technical characteristics to minimize the effects of a possible explanation's failure. XAI should be seen as a control mechanism, a tool provided to the users, through which they can verify that the machine is functioning properly, not as a solution to the malfunction itself.

3.3 Can the explanation make the user "really" aware of the dynamic of the interaction?

The fact that a system is able to "explain itself" is seen as a way of making the user aware of how the technology works and the nature of the produced output. Thus, it is claimed that the human user can be the undisputed active subject of the interaction rather than its passive object. However, this perspective seems to neglect that XAI may itself be distorting the user's freedom of choice, although not necessarily in bad faith. The analysis of risk factors could be difficult for two main reasons:

First, humans can simplify complex information or groups of information. It means that if there are too many possible scenarios to take into account (even if some are not immediately intelligible), all of them can be brought under the general category of perils. The users may make a decision that does not reflect an actual level of danger. Thus, the resulting elaboration – favorable or not – cannot be described either as "informed" or as "conscious", in a concrete and literal sense. Second, even in light of a proper understanding of risks, people are often not able to accurately assess the likely consequences of their actions, tending to underestimate them [21].

Therefore, explanations may, in turn, lead an individual to create a misrepresentation of (i) the real capabilities and functionality of the device, (ii) the extremes of the interaction, and (iii) the reasons behind the recommendation. In other words, XAI might give rise to manipulative dynamics or conceal or facilitate manipulative recommendations provided by AI systems. This can happen on the basis of the way the device has been programmed, because of the dynamics developed during the interaction, or even simply because of specific conditions inherent to the psychological and cognitive dimension of human beings (e.g., biases, tendency to assimilate only information that supports one's theory, tendency to value the goal one intends to achieve more than possible and future risks, etc.).

All this requires a careful reflection on the boundaries that delimit the concept of persuasion and differentiate it from the one of manipulation. On the

basis of this analysis, it is, therefore, necessary to identify which characteristics persuasive systems must have in order not to be manipulative.

3.4 Desiderata for a not manipulative XAI

It has been pointed out that it is impossible (i) to have unequivocal certainty that an explanation is fully understood, (ii) that the understanding is sufficient to limit damages, and (iii) that an effective physical safety can exclude more subtle harm - namely, the distortion of the user's volitional process.

For these reasons, our analysis emphasizes that XAI should be considered one of the instruments we have to reach an end, not an end itself. The latter should be represented by the integrity of the human person as a whole, understood as the union of thoughts, will, and actions. Otherwise, we could risk excessive focus in making explanations as accurate as possible without investing enough in working on the technical safety of new technological approaches. Thus, we would potentially leave room for new critical profiles, such as those related to manipulation and possible coercion of end-users.

Before equipping people with tools to assess for themselves whether the recommendation provided by an AI system is reliable, we should worry about ensuring that the system is as technically safe and dependable as possible. Only after, we should focus on equipping it with as sophisticated as necessary explanation systems. This can produce three benefits: (i) it helps to efficiently manifest properties – safety and reliability – that the device already has and that the users might doubt because they are not experts; (ii) intervening to prevent or mitigate issues that might arise later through use and that could not be predicted with certainty; (iii) ensuring that the final decision remains concretely in the hands of the human user.

For its part, the European Union has recently expressed the attempt to address the issue of users' manipulation by AI systems preferring a legal approach to one focused purely on ethical guidelines. This intention is proven by the AI Act proposal, which is not yet an effective regulation, but offers interesting insights [66].

Starting from this recent act promulgated by the European authorities and basing ourselves on principles applied by legal experts in cases with similar critical profiles, possibly applicable principles and criteria will be suggested below.
Adaptivity. It is still related to the concept of personalized interaction. However, in the context here analyzed, personalization could represent another manipulation instrument. Indeed, targeting people's needs and preferences could make distinguishing good from mischievous intents difficult. An alternative could be to interpret this concept as "adaptivity to the interaction, not to the user". Otherwise, the system should be able to provide an argumentation that is built on the basis of the user's counter-arguments. In this way, the explanation is not a mere transmission of predetermined information, possibly modeled to the producer's interests. It would be shaped by the approach, doubts, and needs that the specific user involved perceives as necessary at that specific moment and for that specific purpose. In doing so, there would be no exploitation of people's

personal data for profiling purposes, for instance, which represents one of the main issues related to the theme of target advertising. The explanation would be generated on the basis of requests and considerations made extemporaneously. Moreover, these characteristics could serve as a further, implicit guarantee that the motivations provided are genuine and not polluted by interests other than those of the end-user.

Granularity of explanations. On the one hand, each recommendation should be accompanied by an explanation that connects it to the specific purpose for which it was provided. On the other hand, aggregating recommendations henceforth their explanations might be more appealing and sophisticated. However, this could conceal their implications/risks and, as a consequence, the traceability of the causal link with possible harmful/manipulative effects.

Another possible scenario is that a single recommendation that may induce several outcomes may be justified only once or with regard to one of them. In both those cases, the user could be induced to have a partial or misled representation of the role and impact of that recommendation on the final goal. That would have repercussions on the individual ability to assess the appropriateness of the outcome provided and whether it actually meets one's own interests.

Paternity of the choice over autonomy. The traditional conception of the principle of autonomy is grounded on an ideal model of judicious decision-makers, capable of pursuing their own interest optimally, through the instrument of legally manifested consent [54]. However, it was also proved that personal biases, level of education, cultural background, and inner motivations can alter the way people perceive and evaluate the information on the basis of which they decide to consent [50]. Therefore, as economics and the consumer protection law field can demonstrate, the idea of a perfectly rational (so properly autonomous) average individual is just a myth [60]. Then, an explanation must aim to preserve or re-establish the user's "Paternity of choice". This concept should be intended as the faculty a person has to recognize the authorship of a life-impacting decision, based on reasons – whether rational, sub-rational, or merely emotional – which make sense for that specific individual, according to internal – psychological, cultural, experiential – characteristics.

Order matters. People have the tendency to gradually decrease their level of attention, interest, and tolerance towards explanations [22]. Consequently, the most relevant information – with potential repercussions on safety, the exercise of freedom of choice, economic management, and, in general, is to be considered essential for an informed and responsible use of the device – should be provided at the outset. To decide which explanations fall within the list of priorities and in which order they should be placed, it may be helpful to adopt the principle of balancing, which is implemented in doctrine and case law with regards to fundamental rights. Thus, a human-right impact assessment would be helpful, able to link to each potentially affected right a potentially useful explanation, and able to structure the form and sequence of these explanations on the basis of the rights that should be considered primary to be protected.

From manipulation to manipulative techniques. The idea that manipulation does not have to be covert to be defined as such and to be effective is becoming increasingly widespread. However, this makes it even more difficult to determine if a person acted under the effect of a manipulative influence or in accordance with a deliberate choice. For this reason, the focus should be moved from the effect to the mean. It is necessary to distinguish manipulation from manipulative practices and act on the second ones to remove the first one.

If the practices that lead to the choice have been directed to the exploitation of the vulnerabilities of a particular person or group chosen as a designed target, those practices should be considered manipulative and regulated as such. In such a view, the result is almost superfluous. If an individual is subjected to manipulative techniques and (i) they are able to recognize them and so to act in accordance to their own deliberation, or (ii) they do not realize the manipulative procedures, yet do not act in accordance with them, this does not change the responsibility of those who tried to manipulate.

This perspective could be useful to settle a certainty in the regulation and identify ex-ante risky situations.

Concerning persuasive technologies, it should be guaranteed that the recommendation maximizes the users' utility while respecting their predetermined goals. Therefore, the explanation should focus on how the suggested behavior can fulfill this requirement rather than encouraging the act itself. Thus, individuals will preserve their own faculty to analyze whether the recommendation is feasible and whether it really allows them to get closer to their ends. This would also include the possibility to disagree with the given outcome and the following explanation.

From the discussion so far, it emerges that the analysis of XAI in persuasive technologies should be conducted in the light of Article 5[5] of the AIA – focused on the issue of user manipulation through AI system - instead of Article 13[6] – which addresses the issue of transparency and explainable AI, more generally. This, however, would require further research and an interpretative and adaptive breakthrough of the aforementioned norm. Indeed, we do not intend that the XAI-powered persuasive techniques should be brought under the category of prohibited practices. Conversely, its development is encouraged. Nonetheless, the request to assess and regulate its manipulative potential is considered imperative.

4 Conclusions and Future Works

This paper focused on claiming that XAI plays a central role in increasing transparency, ensuring a higher level of safety in the use of AI systems, and preserving users' autonomy during the interaction. Furthermore, it highlighted that in the context of persuasive and argumentative technologies, the explanations of the recommendations confer additional support to influence the behavioral change (pursuing a preset goal) users are the object of.

[5] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206
[6] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

However, current research on XAI does not take into account the possibility that the explanation is not only inadequate for the purpose for which it is given, but may even manipulate, rather than merely persuade, the recipient. Such a dynamic takes place if the users are induced to comply with the recommendation given in a way that interferes with their natural decision-making process and substitutes internal interests and purposes for those induced from the outside — without the users being aware of the ultimate reason that drives them to action.

In the case of persuasion, on the other hand, individuals would act aware of both the reasons on the basis of their choice and of the fact that they can recognize them as their own.

The aim is to draw attention to the possibility that explanation may be itself a manipulative practice and to provide desiderata on which to ground future research on the subject from both a technical and a legislative point of view. The central idea is that, given the technical limitations that still affect XAI models and the vulnerabilities inherent in the human cognitive and psychological structure, explanations should not be understood as a solution to the dangers possibly posed by AI systems. Conversely, they could represent a valid instrument to mitigate some of the concerns that persuasive devices still raise regarding people's physical and psychological integrity.

To this aim, some desiderata have been formulated. Namely: (i) the device should be able to adapt the explanation according to the needs arising from the interaction, and not according to specific profiling of the user's personality and inclination; (ii) the explanation should be granular and not unique for groups of recommendations or purposes; (iii) it should be possible to ensure that the users retain awareness of the actual paternity of their actions, rather than an only presumed autonomy of decision-making; (iv) the order of the explanations should follow the indications provided by neuroscience with regard to human perception and attention, on the one hand, and the system of balancing fundamental rights – which can possibly be violated – on the other hand; (v) explanations should be analyzed in the light of their potential manipulative effect and not only in the light of accountability and data security profiles. In doing so, the focus should be not so much on the effects of manipulation – which are difficult to quantify and certainly connect to their cause – but on the manipulative techniques, where, despite the consequences, action should be taken to correct the technical side and to compensate the victim.

These desiderata could represent a starting point to structure a further analysis of these aspects, which should be focused on individuating legal strategies to address manipulative techniques - possibly grounded on an efficient human rights impact assessment - and on the technical implementations - which may make the algorithm effectively safer and inherently non-manipulative for non-experts users, despite the ability it has to explain itself. Nevertheless, a clear definition of manipulation, especially with regards to new technologies, is missed. Starting from the brief juxtaposition between persuasion and manipulation here presented, it would be crucial to deepen the understanding of such dynamics – both from a technical/practical and conceptual point of view. This could be

essential to develop an aligned multidisciplinary approach to the topic, in the knowledge that a sectorial perspective cannot prove exhaustive and effective. At the same time, it would be useful to accompany such an analysis with a future investigation of liability profiles, which are certainly relevant in the regulation of new technologies.

Summarizing, a two-folded intervention is required: (i) at the system level – realizing constructs and mechanisms to analyze, filter, prune, or adapt the outcomes is required to comply with norms and regulations, and (ii) at the normative level – definition of clear conceptualization and boundaries enabling a loyal actualization of the point (i).

Acknowledgments

This work has received funding from the Joint Doctorate grant agreement No 814177 LAST-JD-Rights of Internet of Everything.

This work is partially supported by 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. AI, H.: High-level expert group on artificial intelligence (2019)
2. Albert, E.T.: Ai in talent acquisition: a review of ai-applications used in recruitment and selection. Strategic HR Review (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. Antonov, A., Kerikmäe, T.: Trustworthy ai as a future driver for competitiveness and social change in the eu. In: The EU in the 21st Century, pp. 135–154. Springer (2020)
5. Bertolini, A.: Insurance and risk management for robotic devices: Identifying the problems. Global Jurist **16**(3), 291–314 (2016)
6. Bjørlo, L., Moen, Ø., Pasquine, M.: The role of consumer autonomy in developing sustainable ai: A conceptual framework. Sustainability **13**(4), 2332 (2021)
7. Blumenthal-Barby, J.S.: Biases and heuristics in decision making and their impact on autonomy. The American Journal of Bioethics **16**(5), 5–15 (2016)
8. 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)
9. Calderai, V.: Consenso informato (2015)
10. 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)

11. Ciatto, G., Schumacher, M.I., Omicini, A., Calvaresi, D.: Agent-based explanations in ai: towards an abstract framework. In: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems. pp. 3–20. Springer (2020)
12. Confalonieri, R., Coba, L., Wagner, B., Besold, T.R.: A historical perspective of explainable artificial intelligence. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **11**(1), e1391 (2021)
13. Contissa, G., Docter, K., Lagioia, F., Lippi, M., Micklitz, H.W., Pałka, P., Sartor, G., Torroni, P.: Claudette meets gdpr: Automating the evaluation of privacy policies using artificial intelligence. Available at SSRN 3208596 (2018)
14. Coons, C., Weber, M.: Manipulation: theory and practice. Oxford University Press (2014)
15. Craven, M., Shavlik, J.: Extracting tree-structured representations of trained networks. Advances in neural information processing systems **8** (1995)
16. Crawford, K., Schultz, J.: Big data and due process: Toward a framework to redress predictive privacy harms. BCL Rev. **55**, 93 (2014)
17. De Jong, R.: The retribution-gap and responsibility-loci related to robots and automated technologies: A reply to nyholm. Science and Engineering Ethics **26**(2), 727–735 (2020)
18. Directive, C.: 88/627/eec of 12 december 1988 on the information to be published when a major holding in a listed company is acquired or disposed of. OJ L348 pp. 62–65 (1988)
19. 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)
20. 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)
21. Emilien, G., Weitkunat, R., Lüdicke, F.: Consumer perception of product risks and benefits. Springer (2017)
22. Fischer, P., Schulz-Hardt, S., Frey, D.: Selective exposure and information quantity: how different information quantities moderate decision makers' preference for consistent and inconsistent information. Journal of personality and social psychology **94**(2), 231 (2008)
23. Fox, M., Long, D., Magazzeni, D.: Explainable planning. arXiv preprint arXiv:1709.10256 (2017)
24. Gandy, O.H.: Coming to terms with chance: Engaging rational discrimination and cumulative disadvantage. Routledge (2016)
25. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
26. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)
27. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.Z.: Xai—explainable artificial intelligence. Science Robotics **4**(37), eaay7120 (2019)
28. Hasling, D.W., Clancey, W.J., Rennels, G.: Strategic explanations for a diagnostic consultation system. International Journal of Man-Machine Studies **20**(1), 3–19 (1984)
29. Hellström, T., Bensch, S.: Understandable robots-what, why, and how. Paladyn, Journal of Behavioral Robotics **9**(1), 110–123 (2018)

30. Hoffman, R.R., Klein, G., Mueller, S.T.: Explaining explanation for "explainable ai". In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. vol. 62, pp. 197–201. SAGE Publications Sage CA: Los Angeles, CA (2018)
31. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable ai: Challenges and prospects. arXiv preprint arXiv:1812.04608 (2018)
32. Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable ai systems for the medical domain? arXiv preprint arXiv:1712.09923 (2017)
33. Jones, M.L.: The right to a human in the loop: Political constructions of computer automation and personhood. Social Studies of Science **47**(2), 216–239 (2017)
34. Kool, W., Botvinick, M.: Mental labour. Nature human behaviour **2**(12), 899–908 (2018)
35. Kroll, J.A.: The fallacy of inscrutability. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **376**(2133), 20180084 (2018)
36. Kroll, J.A.: Accountable algorithms. Ph.D. thesis, Princeton University (2015)
37. Lam, S.K., Frankowski, D., Riedl, J., et al.: Do you trust your recommendations? an exploration of security and privacy issues in recommender systems. In: International conference on emerging trends in information and communication security. pp. 14–29. Springer (2006)
38. Lanzing, M.: The transparent self. Ethics and Information Technology **18**(1), 9–16 (2016)
39. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
40. Leonard, T.C.: Richard h. thaler, cass r. sunstein, nudge: Improving decisions about health, wealth, and happiness (2008)
41. Li, Y.: Deep reinforcement learning: An overview. arXiv preprint arXiv:1701.07274 (2017)
42. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
43. Mackenzie, C., Stoljar, N.: Relational autonomy: Feminist perspectives on autonomy, agency, and the social self. Oxford University Press (2000)
44. Margalit, A.: Autonomy: Errors and manipulation. Jerusalem Review of Legal Studies **14**(1), 102–112 (2016)
45. Margetts, H.: The internet and transparency. The Political Quarterly **82**(4), 518–521 (2011)
46. Margetts, H., Dorobantu, C.: Rethink government with ai (2019)
47. Matulionyte, R., Hanif, A.: A call for more explainable ai in law enforcement. In: 2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW). pp. 75–80. IEEE (2021)
48. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267**, 1–38 (2019)
49. Mualla, Y., Tchappi, I., Kampik, T., Najjar, A., Calvaresi, D., Abbas-Turki, A., Galland, S., Nicolle, C.: The quest of parsimonious xai: A human-agent architecture for explanation formulation. Artificial Intelligence **302**, 103573 (2022)
50. Obar, J.A., Oeldorf-Hirsch, A.: The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. Information, Communication & Society **23**(1), 128–147 (2020)

51. Phillips, P.J., Przybocki, M.: Four principles of explainable ai as applied to biometrics and facial forensic algorithms. arXiv preprint arXiv:2002.01014 (2020)
52. Rai, A.: Explainable ai: From black box to glass box. Journal of the Academy of Marketing Science **48**(1), 137–141 (2020)
53. Raz, J.: The morality of freedom. Clarendon Press (1986)
54. Regulation, P.: Regulation (eu) 2016/679 of the european parliament and of the council. Regulation (eu) **679**, 2016 (2016)
55. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
56. Rudinow, J.: Manipulation. Ethics **88**(4), 338–347 (1978)
57. Sadek, I., Rehman, S.U., Codjo, J., Abdulrazak, B.: Privacy and security of iot based healthcare systems: Concerns, solutions, and recommendations. In: International conference on smart homes and health telematics. pp. 3–17. Springer, Cham (2019)
58. Skouby, K.E., Lynggaard, P.: Smart home and smart city solutions enabled by 5g, iot, aai and cot services. In: 2014 International Conference on Contemporary Computing and Informatics (IC3I). pp. 874–878. IEEE (2014)
59. Smuha, N.A.: The eu approach to ethics guidelines for trustworthy artificial intelligence. Computer Law Review International **20**(4), 97–106 (2019)
60. Strünck, C., Arens-Azevêdo, U., Brönneke, T., Hagen, K., Jaquemoth, M., Kenning, P., Liedtke, C., Oehler, A., Schrader, U., Tamm, M.: The maturity of consumers: A myth? towards realistic consumer policy (2012)
61. Susser, D., Roessler, B., Nissenbaum, H.: Technology, autonomy, and manipulation. Internet Policy Review **8**(2) (2019)
62. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
63. Timan, T., Mann, Z.: Data protection in the era of artificial intelligence: Trends, existing solutions and recommendations for privacy-preserving technologies. In: The Elements of Big Data Value, pp. 153–175. Springer, Cham (2021)
64. Towell, G.G., Shavlik, J.W.: Extracting refined rules from knowledge-based neural networks. Machine learning **13**(1), 71–101 (1993)
65. Union, E.: Directive 2003/6/ec of the european parliament and of the council of 28 january 2003 on insider dealing and market manipulation (market abuse). Official Journal of the European Union **50**, 16–25 (2003)
66. Veale, M., Borgesius, F.Z.: Demystifying the draft eu artificial intelligence act—analysing the good, the bad, and the unclear elements of the proposed approach. Computer Law Review International **22**(4), 97–112 (2021)
67. Wick, M.R., Thompson, W.B.: Reconstructive expert system explanation. Artificial Intelligence **54**(1-2), 33–70 (1992)
68. Zarsky, T.: Transparency in data mining: From theory to practice. In: Discrimination and privacy in the information society, pp. 301–324. Springer (2013)
69. Zhang, Y., Chen, X., et al.: Explainable recommendation: A survey and new perspectives. Foundations and Trends® in Information Retrieval **14**(1), 1–101 (2020)
70. Zhang, Y., Liao, Q.V., Bellamy, R.K.: Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 295–305 (2020)