Manipulation Risks in Explainable AI: The Implications of the Disagreement Problem

Sofie Goethals (✉)[1][0000−0003−3784−826X], David Martens[1], and Theodoros Evgeniou[2]

[1] Department of Engineering Management, University of Antwerp, Antwerp, Belgium
sofie.goethals@uantwerpen.be
[2] Decision Sciences and Technology Management Department, INSEAD, Fontainebleau, France

Abstract. Artificial Intelligence (AI) systems are increasingly used in high-stakes domains of our life, highlighting the need to explain these decisions and to make sure that they are aligned with how we want the decision to be made. The field of Explainable AI (XAI) has emerged in response. However, it faces a significant challenge known as the disagreement problem, where multiple explanations are possible for the same AI decision or prediction. While the existence of the disagreement problem is acknowledged, the potential implications associated with this problem have not yet been widely studied. First, we provide an overview of the different strategies explanation providers could deploy to adapt the returned explanation to their benefit. We make a distinction between strategies that *attack* the machine learning model or underlying data to influence the explanations, and strategies that *leverage* the explanation phase directly. Next, we analyse several objectives and concrete scenarios the providers could have to engage in this behavior, and the potential dangerous consequences this manipulative behavior could have on society. We emphasize that it is crucial to investigate this issue now, before these methods are widely implemented, and propose some mitigation strategies.

Keywords: Explainable AI (XAI) · Manipulation · Responsible AI

1 Introduction

Artificial Intelligence (AI) is used in more and more high-stakes domains of our life such as justice, healthcare, and finance, increasing the need to explain these decisions and to make sure that they are aligned with how we want the decision to be made. However, the complexity of many AI systems makes them challenging to comprehend, posing a significant barrier to their implementation and oversight [4, 38]. Legislative initiatives, including the EU General Data Protection Regulation (GDPR), have recognized the 'right for explanation' for individuals affected by algorithmic-decision making, emphasizing the legal necessity of explainability [19]. In response, the field of Explainable Artificial Intelligence (XAI) has emerged, aimed at developing methods for explaining the decision-making processes of AI models [1, 46].

Nevertheless, the landscape of post-hoc explanations is diverse, and each method can yield a different explanation. Furthermore, even within a single explanation method, multiple explanations can be generated for the same instance or decision. This phenomenon, known as the *disagreement problem*, has been studied in literature [9, 22, 33, 37]. While the existence of the disagreement problem is acknowledged, the potential implications of this problem have not yet been extensively explored. Barocas et al. [6] already mention that the power to choose which explanation to return, leaves the providers with significant room to promote their own welfare. Aivodji et al. [2] discuss the possibility of fairwashing, where discriminatory practices can be hidden by selecting the right explanations, while Bordt et al. [7] argue that post-hoc explanations fail to achieve their purpose in adversarial contexts. Finally, Carli et al. [10] highlight how singular explanations can already be a source of manipulation as they can interfere with the users' natural decision-making process. However, an overview of potential misuses by the explanation provider is still missing from the literature, and we believe it is imperative to study the implications now, before explainability methods are implemented on a wide scale. The main contributions of this paper are:E

- Providing a comprehensive framework that outlines the different strategies that could be employed by malicious entities to manipulate the explanations.
- An overview of the different objectives these actors could have to engage in this behavior, and the potential implications.

This paper is structured as follows: We introduce the field of Explainable AI and the disagreement problem in Sections 2 and 3. In Section 4, we explore various strategies that providers could employ to manipulate the explanations according to their preferences. Additionally, in Section 5, we present specific objectives and scenarios that may drive providers to engage in such behavior. Finally, in Section 6, we offer discussion and potential solutions to address this.

2 Explainable AI

Within the field of Artificial Intelligence, providing insights into the decision-making process is crucial for various reasons. First, it establishes trust and compliance with stakeholders, as they can understand and validate the reasoning behind the model's output. Secondly, it enables improved domain insights, allowing practitioners and users to gain a deeper understanding of the problem space and uncover valuable knowledge. Lastly, insights derived from explanations can lead to model improvement, aiding in the optimization of AI systems [31, 46].

To reach these goals, various methods to achieve comprehensibility in AI models have been proposed. In general, there are two main approaches commonly used: inherently transparent models and post-hoc explanations. Inherently transparent models, such as small decision trees, are comprehensible by nature due to their simple structure, without the need for additional explanations [31]. However, in many real-world scenarios, data is becoming increasingly complex and black-box models are used due to their superior predictive performance [18]. These models lack inherent interpretability, and post-hoc explanations are used to provide insights into their decision-making process. This field of research is commonly known as Explainable Artificial Intelligence (XAI).

Within XAI, a distinction can be made between global and local explanations. Global explanations aim to provide an understanding of the model's logic as a whole, allowing users to follow the reasoning that leads to every possible outcome. Techniques such as rule extraction [27] and Partial Dependence plots [13] fall under this category. On the other hand, local post-hoc explanations focus on explaining the logic behind a specific prediction or decision made by the model. Methods like SHAP (SHapley Additive exPlanations) [26] and LIME (Local Interpretable Model-agnostic Explanations) [36] are examples of post-hoc explanation that measure the impact of each feature for a given prediction score (feature importance methods). Another local technique, known as counterfactual explanations, describes a combination of feature changes required to alter the predicted class [45]. While this paper predominantly uses counterfactual explanations as an example, the findings and discussion presented are applicable to other post-hoc explanation techniques as well. At the moment, we do not see manipulation issues for inherently transparent models but this would be an interesting avenue for future research [7].

3 The Disagreement Problem

	Sex	Age	Residence time	Home status	Occupation	Job status	Employment time	Other investments	Bank account	Time at bank	Liability	Account reference	Housing expense	Savings account
Instance	2	16	22	1	2	6	7	0	0	0	0	1	1	125
CBR	2	16	0.25	1	2	6	7	0	1	0	0	1	1	125
DiCE	2	16	22	1	2	6	7	24	0	0	0	1	1	125
GeCo														
NICE(none)	2	34	0	3	3	10	1	0	0	0	0	1	2	136
NICE(plaus)	2	34	0	3	3	6	1	0	0	0	0	1	2	136
NICE(prox)	2	34	0	1	2	10	1	0	0	0	0	1	1	136
NICE(sparse)	2	16	0	1	2	10	1	0	0	0	0	1	1	136
SEDC	2	16	22	1	2	6	7	0	1	0	0	1	1	125
WIT	1	278	8	2	1	5	1	6.5	1	1	6	0	0	102

Table 1: Illustration of the disagreement problem for an instance of the Australian Credit dataset.

A known issue within Explainable AI is that the results of different explanation techniques do not always agree with each other. Even one explanation technique can generate many different explanations for one instance, which is known as the disagreement problem [22, 33, 37]. One of the reasons behind the disagreement problem is that a 'true internal reason' why the machine learning model comes to a certain decision, generally does not exist [7]. For example, for feature importance methods such as SHAP and LIME, there is no mathematically unique way to determine the importance of each feature to the decision of a black-box function [7, 43]. As a consequence, all feature importance methods rely on their own assumptions to approximate this [7, 43]. For counterfactual explanations, this issue also exists as the optimization problem to create the explanations can be set up in a different way in every implementation. Even a single counterfactual explanation method could lead to a large number of explanations, as the choice of parameters (such as the distance metric) has an influence on the explanations that are returned first [17]. The diversity of multiple counterfactual explanations, generated by the same counterfactual algorithm is also known as the Rashomon effect [31].[3]

Other authors already showed the level of disagreement between different post-hoc explanation techniques: Roy et al. [37] show disagreement between LIME and SHAP explanations, Brughmans et al. [9] illustrate this for different counterfactual explanation algorithms, and Bordt et al. [7] demonstrate the disagreement between SHAP, LIME, and counterfactual explanations. We illustrate the disagreement problem between different counterfactual explanation algorithms for one specific instance with an example in Table 1, in line with Brughmans et al. [9]. This table demonstrates the disagreement problem for one instance from the Australian credit dataset, where the target variable indicates whether a person should be granted a loan or not. The depicted instance was not awarded credit and asks for a counterfactual explanation to know which features to change to receive a positive credit decision. Table 1 shows the explanations returned by 10 different counterfactual algorithms, which vary widely. [4] This example illustrates that every feature can be included in the explanation by switching between explanation algorithms. Brughmans et al. [9] verify this for multiple datasets and classifiers, and establish the feasibility of both including and excluding specific features across different scenarios. Note that the potential for manipulation of explanations extends beyond switching between different counterfactual explanation algorithms. In Section 4, alternative strategies that can be employed for manipulation are explored. Currently, a consensus on how to resolve this ambiguity has not yet been reached. Research indicates that most developers rely on arbitrary heuristics, such as personal preferences, to choose the final explanation [22].

This plurality is not necessarily a bad thing. Bordt et al. [7] distinguish between a cooperative and an adversarial context. In cooperative contexts, where stakeholders have the same goal, this plurality can be beneficial as it is expected

[3] The Rashomon effect means that an event can be explained by multiple causes, and is named after a Japanese movie that tells multiple (contradictory) stories about the death of a samurai [31].

[4] The counterfactual algorithm GeCo was not able to find a counterfactual explanation for the given instance.

that the explanation provider will choose the explanation that is in both parties' best interest. For example, when data scientists are debugging a model for their own company, this plurality of explanations can be useful. Other researchers suggest combining multiple explanation techniques to provide a more accurate *meta* explanation [30]. However, in adversarial contexts, the interests of the explanation provider and the data subject are not necessarily aligned, and the explanation providers will be incentivized to choose the explanation that best fits their own interests. An example of such an adversarial context is a loan application where the customer was denied the loan and wants to flag the decision as being discriminatory [7]. In this case, the bank might want to conceal this discriminatory practice by returning a different explanation. This phenomenon is known as *fairwashing*, and has received significant attention [2]. While fairwashing is the most extensively studied objective, we will explore additional scenarios for misuse in adversarial contexts in Section 5. However, even in adversarial contexts, this plurality can be used in a positive way. For example, Bove et al. [8] do mention that in settings such as loan applications, the plurality of explanations can benefit the user if they are provided with multiple explanations.

4 Manipulation Strategies: How can explanation providers exploit the disagreement problem?

The manipulation of explanations by explanation providers is not limited to the mentioned example of switching between explanation algorithms, but can occur at various stages throughout the pipeline, as depicted in Figure 1. We specifically focus on the manipulation that takes place in the post-processing stage, where the explanations are generated, as we imagine that the explanation provider may not always possess the authority to modify the machine learning model or underlying data (the explanation provider is not necessarily the same entity as the model owner). Nevertheless, it is important to note that manipulations directly to the data or model are still feasible, and we discuss some relevant literature exploring this below.

Manipulating the training data to result in different explanations, is related to the area of *data poisoning attacks*. Data poisoning attacks usually involve injecting manipulated data into the training set to compromise the performance of the machine learning model, and while the main focus in literature is on model behavior, its goal might also be manipulating the explanations. Baniecki et al. [5] illustrate that it is possible to attack Partial Dependence plots by poisoning the training data. Bordt et al. [7] highlight the important role of the reference dataset, and show how changing this set influences the resulting SHAP explanations. With regard to changing the model, Slack et al. [42] demonstrate the possibility of modifying biased classifiers in such a way that they continue to yield biased predictions, while the explanations generated by LIME and SHAP will appear harmless. Other authors show the possibility of fine-tuning a neural network to conceal discrimination in the model explanations [11, 21]. Finally, in the domain of images, Dombrowski et al. [12] present evidence showcasing the

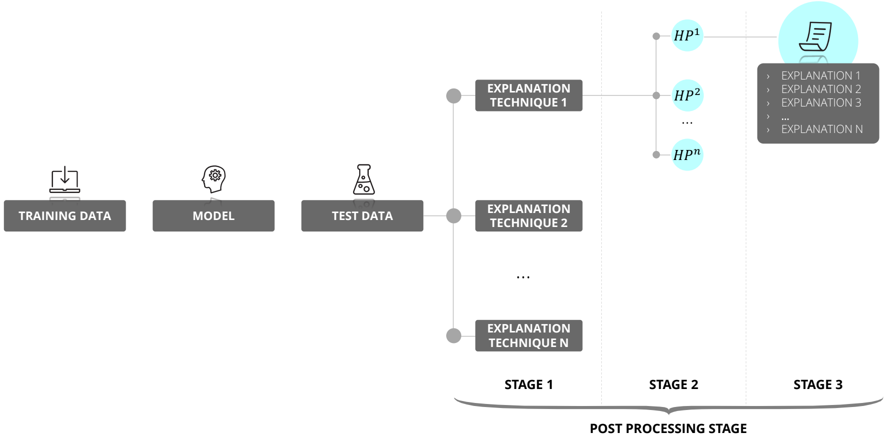

Fig. 1: Strategies the explanation providers could deploy to manipulate the explanations

manipulation of explanations through the application of nearly imperceptible perturbations to visual inputs. In this case, the test data, for which the prediction needs to be explained, is altered. These perturbations would not change the output of the machine learning model, but could result in drastic changes in the explanation map. [5] Additionally, Slack et al. [41] focus on modifying both the model and the test data, such that slight perturbations to the input data can lead to more cost-effective recourse for specific subgroups, while giving the impression of fairness to auditors.

As mentioned, we focus on strategies to alter the explanation in the post-processing stage, without making any alterations to the used data or the underlying machine learning model. We foresee three main strategies the providers could deploy in this stage:

1. **Change the explanation technique**
   Many different post-hoc explanation techniques exist, both local and global, as outlined in Section 2. Consequently, a first evident strategy entails switching to a different explanation technique. For example, when the surrogate model reveals patterns the explanation provider wants to conceal, he might switch to using Partial Dependence plots as an alternative if these patterns do not manifest clearly in those plots. However, on a local level, using different explanation techniques between instances may attract greater attention than the strategies described below, as the output could have a significantly different format (e.g., feature importance plot versus a counterfactual explanation).

[5] One could argue that altering the test data in an imperceptible way will be mostly applicable to image data, as in tabular data these changes may be more noticeable.

2. **Change the parameters or used implementation of an explanation technique**
   Even within a single explanation algorithm, significant leeway exists for manipulating the explanations, contingent upon the selected parameter configurations. For example, LIME explanations depend on the number of perturbed instances and the bandwidth [7, 15], while for Shapley values, there is a multitude of ways to implement them and each operationalization yields significantly different results [43]. Global methods, such as surrogate modeling, are heavily influenced by the choice of architectural design (e.g., linear models, decision trees, etc.) and the complexity of the surrogate model. In the case of counterfactual explanations, as shown in Table 1, the used implementation exerts a substantial influence on the returned explanations, with the number of potential implementations proliferating at a rapid pace. Additionally, even within one counterfactual algorithm, there often exist many modifiable parameters that influence the results.
3. **Exploit the non-deterministic component of some explanation algorithms**
   Some explanation algorithms such as DICE [32] inherently provide multiple possible explanations for one instance. In such cases, the explanation provider can simply select an explanation from the available options without requiring any modifications. Furthermore, certain explanation algorithms are not designed in a deterministic way and may return different explanations across runs. For example, when using LIME, the randomness introduced during the sampling and perturbation process can lead to variations in the generated explanations for each execution [25, 47]. Additionally, Mazzine et al. [35] show that multiple counterfactual algorithms do not generate consistent results over multiple runs, when the same model, input data and parameters are used. In this scenario, the explanation providers can repeatedly execute the explanation algorithm until an explanation that aligns with their preferences is returned.

In the scenario we describe, we assume explanation providers deliberately choose the explanation out of all the possible explanations that best aligns with their interests. The returned explanation will still be technically correct, it will just not necessarily be the explanation that will be in the best interest of the user. It is important to note that we are not referring to situations where explanations chosen by the explanation provider are not in the best interest of the user ‘by accident’ due to differences in knowledge background or a lack of awareness of the user’s preferences [8, 16]. Instead, we are concerned with cases where the explanation provider knowingly opts for an explanation that serves their own agenda, despite knowing that it may not be the optimal explanation for the end user. Note that in described strategies, the providers maintain a partial ethical stance by delivering explanations that retain technical correctness. However, providers have the potential to further exploit the situation by offering spam explanations, containing superfluous features [20], or by deliberately presenting entirely false explanations that are fabricated. The complexity of the pipeline depicted in Figure 1 demonstrates the extensive potential for manipulation and, consequently, the fragility of explanations.

5 Manipulation Objectives: Why would explanation providers want to exploit the disagreement problem?

Which objectives could the providers have to engage in this behavior? We outline them in Figure 2, and discuss various scenarios for each objective in the subsections below. At the moment, we see mitigating liability, implementing their beliefs and maximizing their profits as the main objectives. This list may not be exhaustive yet as the way that technology is used in society is constantly evolving and new objectives may emerge.

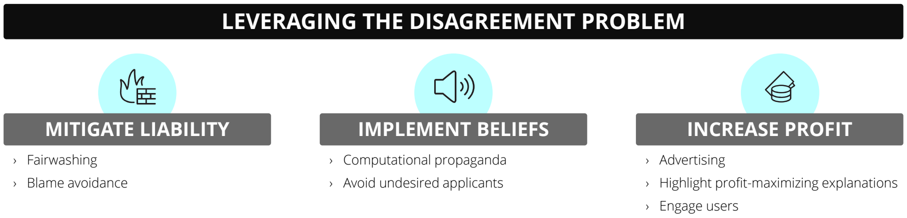
LEVERAGING THE DISAGREEMENT PROBLEM
MITIGATE LIABILITY
› Fairwashing
› Blame avoidance
IMPLEMENT BELIEFS
› Computational propaganda
› Avoid undesired applicants
INCREASE PROFIT
› Advertising
› Highlight profit-maximizing explanations
› Engage users

Fig. 2: Main objectives to leverage the disagreement problem

5.1 Mitigate liability

The model could be unethical or suboptimal in several ways and model explanations could reveal this. Explanation providers could manipulate the explanations to avoid these issues coming to light.

Fairwashing The first, and most studied, reason for explanation providers to engage in this behavior, is *fairwashing* [2, 3, 39]. Fairwashing is defined as '*promoting the false perception that a machine learning model used by the company is fair while this might not be so*' [2]. In a fairwashing attack, the explanation provider will manipulate the explanations to underreport the unfairness of the machine learning model. This has a significant impact on the individuals that received a negative decision based on unfair grounds, as this will deprive them of the possibility to contest it [3]. The relative easiness with which fairwashing can be executed has already been shown in the literature. [3, 39]. Imagine a bank that decides it prefers people from a certain demographic group, and predominantly gives out loans to this group (without a justified reason to do so). It could easily mask this behavior by choosing a different explanation. For example, instead of returning the explanation '*If you would have belonged to a different demographic group, you would have received the loan*', it could return as explanation '*If your income would be double as high, you would have received the loan*', even if the latter explanation is less plausible. Some counterfactual algorithms such as DICE [32] even have as an input parameter the features that can be part of the explanation, so if sensitive features such as demographic attributes are removed from this list, these counterfactual explanations will never flag discrimination. We use counterfactual explanations as illustration here, but this objective extends to other explanation techniques as well. All the mentioned techniques in Section 2 have the potential to reveal bias within a model (for example a feature importance ranking where the sensitive attribute has a very high score). Manipulating the explanations in this way can mislead human users into trusting a problematic black box [24], and it undermines the core principles of alogrithmic fairness.

Blame avoidance Explanation providers can also take advantage of the plurality of explanations to shift blame or evade responsibility for controversial or erroneous decisions made by Artificial Intelligence (AI) systems. Nissembaum et al. [34] already mention that placing accountability in a computerized system can be a very obscure process due to the '*problem of many hands*' (many actors and factors contribute to the process, and is not clear which factor ultimately led to the decision). This issue is reflected in the explanations, where different explanations can point to different actors or circumstances. For example, in the case of autonomous vehicles, AI systems make critical decisions that impact passenger safety. Malicious model owners, such as manufacturers or operators, may downplay system failures or accidents caused by their vehicles. They could selectively present an explanation that attributes the fault to external factors or human error, and as such divert attention from potential design flaws or inadequate safety measures. Similarly, in the field of healthcare, this exploitative behavior can manifest when mistakes by surgeons or flaws in operating machines are concealed to avoid accountability. These actions not only endanger lives but also run contrary to our ethical values. Placing the entire blame on parties that are only partially responsible for an incident contradicts the principles of fairness and accountability. The appropriate distribution of responsibility is crucial for ensuring that the errors are properly addressed and the necessary improvements are made.

5.2 Implement beliefs

Explanation providers may use the explanations to promote their belief system, either by influencing people through propaganda or by excluding applicants that they deem unworthy, despite the machine learning model not sharing this perspective.

Computational propaganda The power to choose an explanation that best fits its interest, can be used to exert an influence on the public opinion. Propaganda itself is defined as '*the expression of opinion or action by individuals or groups deliberately designed to influence opinion or actions of other individuals or groups with reference to predetermined ends*', while computational propaganda is defined as '*propaganda created or disseminated using computational (technical) means*' [28]. Note that propaganda does not necessarily have to lie; it could simply cherry-pick the facts, which is exactly the option explanation providers have to their disposal [28]. By selectively presenting explanations that align with their preferred ideology or desired narrative, explanation providers can amplify certain perspectives while downplaying or ignoring others. For example, in the realm of political campaigns, AI systems are used to analyze public sentiment, create targeted messaging, and influence voter behavior. Imagine an entity with access to an AI model that predicts the likelihood of successful integration for immigrants based on various factors like employment, language proficiency, and government support. The entity firmly believes in the principle of stricter requirements for immigrants, and they could selectively highlight specific factors such as language proficiency or employment history, while downplaying or omitting other important factors such as government support and community involvement. By presenting the AI model's predictions as mainly being driven by these selected factors, they could frame the narrative that successful integration is mainly due to language proficiency, and engaging in employment. The goal is to shape public opinion regarding immigration policy and generate support for stricter language and employment requirements for immigrants. Evidently, machine learning models cannot perfectly mimic the actual world, so even if a machine learning model could be perfectly explained, such an explanation would not constitute a perfect explanation of the real world. However, the concern here lies in the fact that people may still perceive machine-generated explanations as accurate depictions of the actual world, and consequently, the cherry-picked explanations have the potential to influence and shape their understanding of the world at large. Additionally, if the power to generate the explanations would be in the hands of a few actors, they would have the potential to wield significant influence over a large number of people. In this context, the manipulation of explanations can have far-reaching consequences for public opinion and democratic decision-making, and could promote the spread of misinformation.

Avoid undesired applicants In this scenario, the explanation provider, who is using a machine learning model, has the ability to engage in discriminatory practices without directly manipulating the model itself. Instead, they alter the quality of the explanations given to certain population groups, thereby introducing discrimination. In algorithmic decision-making, explanations are often provided to users (the explanation recipients) to help them understand the factors that influenced the decision and potentially take corrective actions (*algorithmic recourse*). Counterfactual explanations are most often used here, as they guide users in modifying their input data to achieve a desired outcome.

In this case, the explanation provider treats different population groups unequally by manipulating the quality of the explanations provided to them. The preferred population group is given explanations that are concise, actionable, and easily implementable. For example, they might receive suggestions such as adjusting the loan amount slightly or making small changes to their reported income. These explanations empower the preferred group to take specific actions that could potentially improve their chances of receiving a positive outcome. On the other hand, the disadvantaged demographic group is given explanations of lower quality. These explanations are designed to be difficult or even impossible to act on. They might involve suggesting large changes to their income or modifying their age, which are factors that applicants typically have limited or no control over. By providing such explanations, the explanation provider creates a significant imbalance in the recourse options available to different society groups. Note that the discriminatory practices described in this scenario are not related to the machine learning model itself, but to the post-processing stage where explanations are generated and shared with applicants. This issue is related to fairness in algorithmic recourse, where fairness is assessed by measuring the distance between the factual and the counterfactual instance [23, 40], and highlights the need for fairness assessments not only during the modeling stage but throughout the entire decision-making pipeline, including the provision of explanations.

5.3 Increase profit

Explanation providers might feel incentivized to capitalize on the explanations. They could return the explanation that would be the most profitable for them, and for this we envisage several scenarios.

Advertising One possibility discussed in previous work, is the integration of algorithmic explanations with advertising opportunities, creating an '*explanation platform*' where advertisements are served alongside the explanation [20]. An example of this could be, that during a job application you receive the following explanation: '*If your CV would have included Python, you would have been invited for the next round*'. This explanation would then be accompanied by an advertisement for an online Python course, which would be a convenient solution for users to reach their goal [20]. This approach allows the explanation provider to select the explanations that have the potential to generate the highest revenue in the advertising market.

Highlight profit-maximizing explanations However, monetization avenues can go beyond advertising. Explanation providers can also exploit the plurality of explanations to direct users towards actions that would maximize their own profits directly. This is related to the advertising scenario, but in this case the actions of the decision subject would directly lead to an increase in profit for the provider. For example, in the domain of healthcare diagnostics, AI systems are increasingly used for the identification of diseases and treatment recommendations. Malicious explanation providers, such as healthcare providers or insurance companies, may strategically choose explanations that prioritize certain measures or specific treatments. In this context, the goals of healthcare providers and insurance companies may diverge. Healthcare providers may have incentives to promote more expensive treatments, while insurance companies may prefer cost-saving measures and cheaper treatment options. However, by favoring explanations that are not necessarily the best or most appropriate, these providers can exert influence over medical decisions and potentially compromise patient care. This scenario could also happen in other domains than healthcare: for example, in the realm of credit scoring, AI systems are employed to evaluate an individual's creditworthiness. Barocas et al. [6] already mention that decisions (and therefore explanations) in this scenario are not simply binary. The provider gives the decision subject a counterfactual that results in a *specific* interest rate, and as such it can choose the interest rate that is likely to maximize its profit [6].

Engage users In line with *Computational Propaganda*, discussed in Section 5.2, providers could also choose to return the explanations that reinforce the ideologies of the data subject itself. In this case, the explanation provider would be a platform, and the goal would be to maximize the revenues of the platform by keeping users as engaged and satisfied as possible (for many platforms daily/monthly active users is an important objective in their financial reports). An example of an explanation in this case, could be the same as in the scenario of propaganda, but in this case different society groups would receive very different explanations, depending on their beliefs. It is known that presenting them with content and information that is likely to resonate with their interests is a way to achieve this (in line with filter bubbles in content recommendation systems). However, this could lead to different groups in society receiving vastly different explanations for the same phenomenon, and consequently to *epistemic fragmentation* [29].[6] . By reinforcing filter bubbles and echo chambers, these platforms exacerbate polarization and hinder constructive dialogue between different groups in society.

Introducing a profit motive into the generation of explanations at all seems contradictory to the initial goals of Explainable AI. An explanation recipient should not have to wonder whether the selected explanation was chosen for its profit-making potential rather than for its ability to accurately explain the situation [20].

6 Discussion

The examples discussed in Section 5 shed light on potential ethical concerns, even though they may not necessarily involve illegal activities. In these scenarios, the

[6] Epistemic fragmentation refers to the tendency for different people to have different sources of knowledge and different, often conflicting, understandings

generated explanations remain factually correct but are selectively hand-picked by the explanation provider to serve their own interests. At the moment, this process is completely unregulated, but could have very serious consequences, as outlined in the scenarios above. In scenarios listed in Section 5, we assumed the explanation providers had malicious incentives, but obviously, this will not always be the case. In fact, some providers may be motivated to manipulate the explanations for the social good. For example, they might explicitly avoid providing explanations that reinforce biased stereotypes, in an attempt to promote fairness and equity. Nevertheless, even though their motives might be aligned with societal goals, it remains questionable whether unregulated entities without the required authority should have the power to make this call.

As we are at the forefront of the XAI revolution, it is crucial to address this issue now, before these methodologies are implemented on an even wider scale. Currently, a substantial portion of AI power is concentrated among a few tech giants. If we also grant them the authority to control the explanations generated by AI models, they would possess yet another means to exert significant influence over society. To mitigate this concentration and potential misuse of power, it becomes imperative for government institutions to collaborate and establish agreed-upon standards and tools for XAI. In particular, in adversarial contexts where interests may clash, it should not be left solely to the explanation providers to create and choose the explanations. Instead, we argue that governments and policy makers should take the matter into their own hands, and agree on a framework that should be used as soon as possible. The key question here is "*What should be the process to make this decision, and what tools are needed to support this process?*". Similar to the no free lunch theorem, that indicates that there is no algorithm that always outperforms all others, there likely will also not be an universally superior explanation method. An agreement on which method to use in which scenario should be established, and this should be done democratically by allowing those affected by XAI to voice their opinion [44], in line with the 'democratic principles of affected interests' [14].

It will take some time to reach a global consensus on the procedures that should be used, and therefore as a short-term solution, regulation should demand full **transparency** in the used explainability method, and settings. This would remove some flexibility for the explanation provider to change the explanation technique continuously, but not remove all potential for manipulation as the providers could still exploit the non-deterministic component of some explanation algorithms or simply lie about the used parameters. Therefore, to ensure adherence to ethical values, we also foresee that it would be mandatory to have **external auditors** conducting audits of AI systems, explanations, and decision-making processes. These auditors should be independent entities without a vested interest in the outcomes, similar to how audits are conducted in other industries. Furthermore, in high-stakes contexts, where transparency is of paramount importance, we argue that the the use of white-box models needs more attention [18], given the manipulation risks surrounding explanations. To conclude, we believe that implementing these measures can ensure that AI systems are developed and deployed in a manner that aligns with societal values.

Acknowledgements

This research was funded by Research Foundation—Flanders grant number 11N7723N. The authors would like to acknowledge Dieter Brughmans for providing the example depicted in Table 1, and would like to thank Dieter Brughmans, Ine Weyts, Camille Dams and Travis Greene for their feedback on the paper.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE access **6**, 52138–52160 (2018)
2. Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., Tapp, A.: Fairwashing: the risk of rationalization. In: International Conference on Machine Learning. pp. 161–170. PMLR (2019)
3. Aïvodji, U., Arai, H., Gambs, S., Hara, S.: Characterizing the risk of fairwashing. Advances in Neural Information Processing Systems **34**, 14822–14834 (2021)
4. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion **58**, 82–115 (2020)
5. Baniecki, H., Kretowicz, W., Biecek, P.: Fooling partial dependence via data poisoning. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part III. pp. 121–136. Springer (2023)
6. Barocas, S., Selbst, A.D., Raghavan, M.: The hidden assumptions behind counterfactual explanations and principal reasons. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. pp. 80–89 (2020)
7. Bordt, S., Finck, M., Raidl, E., von Luxburg, U.: Post-hoc explanations fail to achieve their purpose in adversarial contexts. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 891–905 (2022)
8. Bove, C., Lesot, M.J., Tijus, C.A., Detyniecki, M.: Investigating the intelligibility of plural counterfactual examples for non-expert users: an explanation user interface proposition and user study. In: Proceedings of the 28th International Conference on Intelligent User Interfaces. pp. 188–203 (2023)
9. Brughmans, D., Melis, L., Martens, D.: Disagreement amongst counterfactual explanations: How transparency can be deceptive. arXiv preprint arXiv:2304.12667 (2023)
10. Carli, R., Najjar, A., Calvaresi, D.: Risk and exposure of xai in persuasion and argumentation: The case of manipulation. In: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems. pp. 204–220. Springer (2022)
11. Dimanov, B., Bhatt, U., Jamnik, M., Weller, A.: You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. (2020)
12. Dombrowski, A.K., Alber, M., Anders, C., Ackermann, M., Müller, K.R., Kessel, P.: Explanations can be manipulated and geometry is to blame. Advances in neural information processing systems **32** (2019)

13. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics pp. 1189–1232 (2001)
14. Fung, A., Wright, E.O.: Deepening democracy: Innovations in empowered participatory governance. Politics & society **29**(1), 5–41 (2001)
15. Garreau, D., Luxburg, U.: Explaining the explainer: A first theoretical analysis of lime. In: International Conference on Artificial Intelligence and Statistics. pp. 1287–1296. PMLR (2020)
16. Gilpin, L.H., Paley, A.R., Alam, M.A., Spurlock, S., Hammond, K.J.: " explanation" is not a technical term: The problem of ambiguity in xai. arXiv preprint arXiv:2207.00007 (2022)
17. Goethals, S., Martens, D., Calders, T.: Precof: counterfactual explanations for fairness. Machine Learning pp. 1–32 (2023)
18. Goethals, S., Martens, D., Evgeniou, T.: The non-linear nature of the cost of comprehensibility. Journal of Big Data **9**(1), 30 (2022)
19. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a "right to explanation". AI magazine **38**(3), 50–57 (2017)
20. Greene, T., Goethals, S., Martens, D., Shmueli, G.: Monetizing explainable ai: A double-edged sword. arXiv preprint arXiv:2304.06483 (2023)
21. Heo, J., Joo, S., Moon, T.: Fooling neural network interpretations via adversarial model manipulation. Advances in Neural Information Processing Systems **32** (2019)
22. Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., Lakkaraju, H.: The disagreement problem in explainable machine learning: A practitioner's perspective. arXiv preprint arXiv:2202.01602 (2022)
23. von Kügelgen, J., Karimi, A.H., Bhatt, U., Valera, I., Weller, A., Schölkopf, B.: On the fairness of causal algorithmic recourse. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 9584–9594 (2022)
24. Lakkaraju, H., Bastani, O.: " how do i fool you?" manipulating user trust via misleading black box explanations. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 79–85 (2020)
25. Lee, E., Braines, D., Stiffler, M., Hudler, A., Harborne, D.: Developing the sensitivity of lime for better machine learning explanation. In: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications. vol. 11006, pp. 349–356. SPIE (2019)
26. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems. pp. 4768–4777 (2017)
27. Martens, D., Baesens, B., Van Gestel, T., Vanthienen, J.: Comprehensible credit scoring models using rule extraction from support vector machines. European journal of operational research **183**(3), 1466–1476 (2007)
28. Martino, G.D.S., Cresci, S., Barrón-Cedeño, A., Yu, S., Di Pietro, R., Nakov, P.: A survey on computational propaganda detection. arXiv preprint arXiv:2007.08024 (2020)
29. Milano, S., Mittelstadt, B., Wachter, S., Russell, C.: Epistemic fragmentation poses a threat to the governance of online targeting. Nature Machine Intelligence **3**(6), 466–472 (2021)
30. Mollas, I., Bassiliades, N., Tsoumakas, G.: Truthful meta-explanations for local interpretability of machine learning models. arXiv preprint arXiv:2212.03513 (2022)
31. Molnar, C.: Interpretable machine learning. Lulu. com (2020)

32. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. pp. 607–617 (2020)
33. Neely, M., Schouten, S.F., Bleeker, M.J., Lucic, A.: Order in the court: Explainable ai methods prone to disagreement. arXiv preprint arXiv:2105.03287 (2021)
34. Nissenbaum, H.: Accountability in a computerized society. Science and engineering ethics **2**, 25–42 (1996)
35. de Oliveira, R.M.B., Martens, D.: A framework and benchmarking study for counterfactual generating methods on tabular data. Applied Sciences **11**(16), 7274 (2021)
36. 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)
37. Roy, S., Laberge, G., Roy, B., Khomh, F., Nikanjam, A., Mondal, S.: Why don't xai techniques agree? characterizing the disagreements between post-hoc explanations of defect predictions. In: 2022 IEEE International Conference on Software Maintenance and Evolution (ICSME). pp. 444–448. IEEE (2022)
38. Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.R.: Explainable AI: interpreting, explaining and visualizing deep learning, vol. 11700. Springer Nature (2019)
39. Shahin Shamsabadi, A., Yaghini, M., Dullerud, N., Wyllie, S., Aïvodji, U., Alaagib, A., Gambs, S., Papernot, N.: Washing the unwashable: On the (im) possibility of fairwashing detection. Advances in Neural Information Processing Systems **35**, 14170–14182 (2022)
40. Sharma, S., Henderson, J., Ghosh, J.: Certifai: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 166–172 (2020)
41. Slack, D., Hilgard, A., Lakkaraju, H., Singh, S.: Counterfactual explanations can be manipulated. Advances in neural information processing systems **34**, 62–75 (2021)
42. Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 180–186 (2020)
43. Sundararajan, M., Najmi, A.: The many shapley values for model explanation. In: International conference on machine learning. pp. 9269–9278. PMLR (2020)
44. Vermeire, T., Laugel, T., Renard, X., Martens, D., Detyniecki, M.: How to choose an explainability method? towards a methodical implementation of xai in practice. In: Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2021, Virtual Event, September 13-17, 2021, Proceedings, Part I. pp. 521–533. Springer (2022)
45. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. **31**, 841 (2017)
46. Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., Zhu, J.: Explainable ai: A brief survey on history, research areas, approaches and challenges. In: Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8. pp. 563–574. Springer (2019)
47. Zhang, Y., Song, K., Sun, Y., Tan, S., Udell, M.: " why should you trust my explanation?" understanding uncertainty in lime explanations. arXiv preprint arXiv:1904.12991 (2019)