Francesco Amato, Biagio Aragona, Mattia De Angelis

Factors and possible application scenarios of Explainable Ai

(doi: 10.53227/113110)

Rivista di Digital Politics (ISSN 2785-0072)
Fascicolo 3, settembre-dicembre 2023

Ente di afferenza:
()

Francesco Amato, Biagio Aragona, Mattia De Angelis

Factors and possible application scenarios of Explainable Ai

FACTORS AND POSSIBLE APPLICATION SCENARIOS OF EXPLAINABLE AI

The article focuses on the explainability of Artificial intelligence (Ai) algorithms used in public administrations. It presents agnostic and non-agnostic Explainable ai (Xai) frameworks with the main literature about their development and application and the advantages of the possible deployment of these frameworks to the sociotechnical system employed by public administrations. As a case study, we analyse the narratives of teachers' X users about the algorithms that assigned school-teacher positions from 2016 to 2023, an algorithmic system that has generated unexpected and potentially problematic effects on society. We argue that the Xai framework can be employed by stakeholders as a guideline for the design of transparent systems by design, to prevent or mitigate the negative effects of these technologies and provide methods and tools for inspecting the processes performed by the automated decision systems.

KEYWORDS *Explainable Artificial Intelligence, Transparency, Impact Evaluation, Automated Decision Systems, Public Administration.*

1. Introduction

Artificial intelligence (Ai) technologies have gone through significant growth in recent years, highlighting the need to examine the social impact of Automated decision-making systems (Adm) (Hess *et al.* 2017; Mackrill and Ebsen 2018; Aragona 2022). These technologies can influence crucial aspects

Francesco Amato, Department of Social Sciences – University of Naples Federico II – Vico Monte della Pietà, 1 – 80138 Napoli – email: francesco.amato2@unina.it, orcid: 0000-0003-4384-1646

Biagio Aragona, Department of Social Sciences – University of Naples Federico II – Vico Monte della Pietà, 1 – 80138 Napoli – email: aragona@unina.it, orcid: 0000-0001-8697-2932

Mattia De Angelis Department of Social Sciences – University of Naples Federico II – Vico Monte della Pietà, 1 – 80138 Napoli – email: mattia.deangelis@unina.it, orcid: 0009-0006-3961-6920

ISSN 2785-0072

of the lives of citizens and organisations, such as, for example, job applications and loan approvals (Aragona 2022). Digital transformation, together with the algorithmisation process, has permeated various sectors of Public administration (Pa), bringing with it a series of advantages and improvements (Landri 2018). Public infrastructures are equipping themselves with advanced algorithms and Ai technologies to simplify and automate a diverse range of processes (Madan 2023). This integration could support and could contribute to operational efficiency and the overall modernisation of the administrative system, considering that without control these tools reproduce bias and discrimination (Veale and Brass 2019). Although these technologies are often considered tools, they should instead be identified as actants enabled to act (Latour and Woolgar 1979; Callon 1986; Latour 1987; 2005; Hoch *et al.* 1987; Cozzens *et al.* 1989). These technologies exist within an intricate socio-technical assemblage, composed of multiple factors that act in their creation (Kitchin 2014; Kitchin and Lauriault 2014; Kitchin 2017).

The complexity of the socio-technical interweaving makes these technologies difficult to inspect. Interacting with the assemblage becomes challenging due to its inherent complexity. Consequently, digital technology, often likened to a black box, exhibits significant opacity in its operations (Pasquale 2015). These technological applications within Pa make it necessary to discuss the algorithmic risks associated with the implementation of these solutions (Aragona and Amato 2022a).

To mitigate the impacts of Ai and Adm it is necessary to take into account the ethical aspects of these technologies. In this regard, we discuss the Explainable ai (Xai) framework, which consists of several methods that expand and argue the outputs of machine learning-based systems (Doshi-Velez and Kim 2017; Gilpin *et al.* 2018). In the literature, some principles have been proposed transparency, interpretability, and explainability (Angelov 2021). These three concepts are the focus of the Xai framework (Došilović *et al.* 2018; Gilpin *et al.* 2018). Alongside the Xai framework there is the Artificial intelligence act (Ai act 2024), the regulation on Ai of the European union – Eu (2021/0106). This regulation has among its main objectives the construction of a regulatory framework common to Eu countries for the management of socio-economic risks involving the use of Ai, paying attention to the necessary coherence between the data acquired and the regulatory framework, and greater transparency, explainability and documentability of the actions performed by Adm. The Eu has also proposed Ai assessment guidelines through the Ethics guidelines for trustworthy ai (2019) which provide the ability for stakeholders to access and evaluate decision-making processes that implement automated processes or Ai applications.

In the second paragraph the main literature on the topic is presented and the Xai frameworks are shown. The third paragraph introduces the principles and objectives set by the Eu regarding Ai. The fourth paragraph develops by considering the relationship between Xai, administrative transparency and Pa. Added to this is the presentation of a case study regarding the algorithms used for the assignment of teachers in Italy, to highlight the potential benefits that the Xai framework could implement. Finally, in section 5 the decision-making process and the Xai framework will be related and a possible application will be proposed.

We argue that the Xai framework serves as a valuable guide for stakeholders in crafting systems inherently transparent, interpretable and explainable. This proactive approach aims to forestall or alleviate the adverse impacts associated with such technologies.

2. Explainable Ai, a framework for the evaluation of algorithmic systems

Rapid advancements in Ai have underscored the imperative to rendering machine decision-making processes comprehensible and interpretable (Nauta *et al.* 2023). Interest in Xai systems also began to grow in academia in 2018, attesting to its peak in 2023. Examining the documents within Scopus, it remains uncertain if the subject will sustain its growth in the future, aligning with the ongoing trend in publications. In 2023 alone, Scopus has 2639 documents for the extraction query[1] of, and 2221 in 2022 out of a total of 7329 since 2004. From 2004 to 2018, the term appeared in only 16 documents (Figure 1).

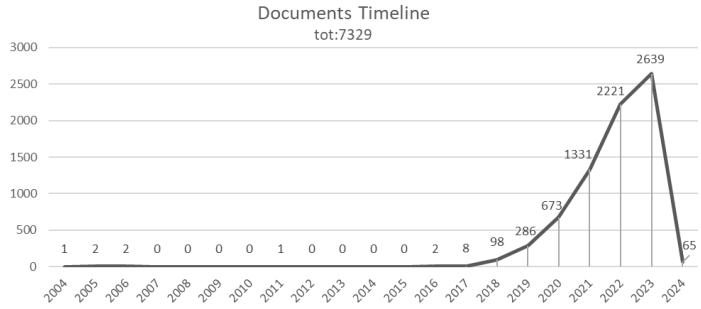

Fig. 1. Xai documents timeline.
Source: Elaboration by the authors on Scopus[2].

[1] Scopus extraction query «Title-abs-key({explainable artificial intelligence} or{explainable ai}or {explainable ai (xai)}or{explainable artificial intelligence (xai)})».

[2] Search query on Scopus «https://www.scopus.com/term/analyzer.uri?sort=plf-f&src=s&sid=3796eb0c813f0817cf2b1f967322631a&sot=a&sdt=a&sl=145&s=Title-abs-

The number of funder sponsors increased from 6 before 2018 to 159 after. These actors are both public and private, and considering the number of papers per funding sponsor, it can be seen that the first two places are occupied by funds provided by the Eu (Figure 2).

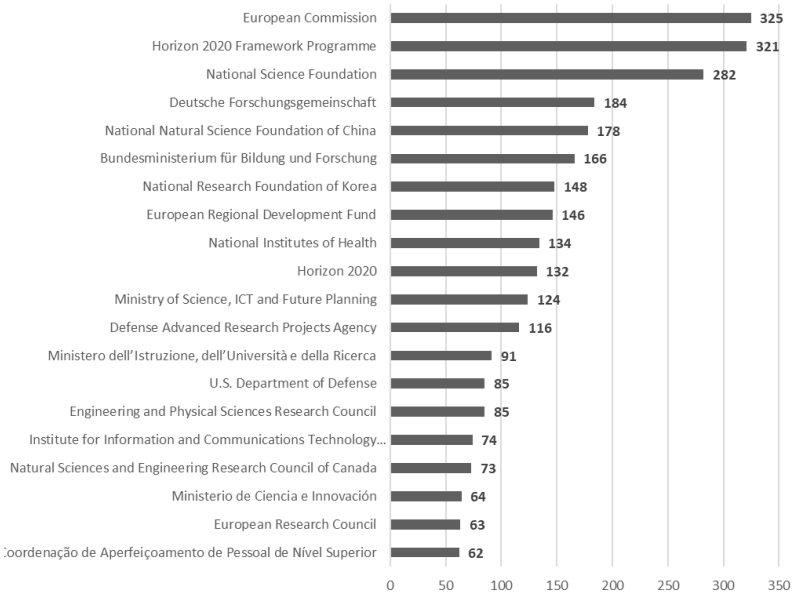

Fig. 2. Top 20 funding sponsors on Scopus.
Source: Elaboration by the authors on Scopus[3].

In the same time-frame, the Xai framework assumes paramount importance in the context of Ai, striving to elucidate the decision-making procedures employed by algorithms.

In the scientific literature, Xai systems have different types of applications, depending on the interaction an Xai framework has with the application domain, one can have the specific Xai frameworks and the agnostic Xai frameworks (Adabi and Berrada 2018; Rawal *et al.* 2021). Specific frameworks refer to applications closely related to the system to be explained, while agno-

key%28%7bexplainable+artificial+intelligence%7d+or+%7bexplainable+ai%7d+or+%7bexplainable+ai+%28xai%29%7d+or+%7bexplainable+artificial+intelligence+%28xai%29%7d%29&origin=resultslist&count=10&analyzeResults=Analyze+results».

[3] Search query on Scopus «https://www.scopus.com/term/analyzer.uri?sort=plf-f&src=s&sid=3796eb0c813f0817cf2b1f967322631a&sot=a&sdt=a&sl=145&s=Title-abs-key%28%7bexplainable+artificial+intelligence%7d+or+%7bexplainable+ai%7d+or+%7bexplainable+ai+%28xai%29%7d+or+%7bexplainable+artificial+intelligence+%28xai%29%7d%29&origin=resultslist&count=10&analyzeResults=Analyze+results».

stic frameworks are at a higher level and apply to more technologies (Arrieta *et al.* 2020; Brasse *et al.* 2023).

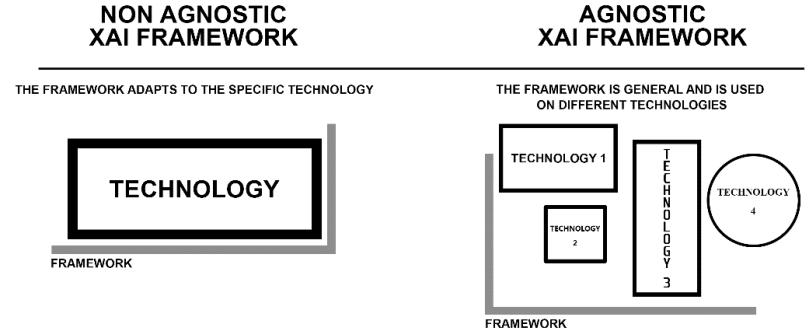

Fig. 3. Xai frameworks.
Source: Elaboration by the authors.

As shown in Figure 3, the agnostic Xai framework is used for different technological applications, therefore, it cannot be exhaustive for all technologies. The non-agnostic framework, on the other hand, considers a specific technology and adapts to it, making it an effective and usable tool for making that technology explainable.

These two types refer to the approach with technologies, but such systems arise in a human-machine relationship, and it is necessary to also take into account the audience to whom the explanation must be made understandable. Three types of audiences have been identified, which we can define as: the audience of experts and those who act directly in the management and implementation of Ai technologies, the audience of those familiar with the applications who do not deal directly, and lastly the users impacted passively by the Adms (Ribera and Lapedriza 2019; Wang *et al.* 2019).

Pa processes involve the presence of heterogeneous actors - who have different roles and different levels of skills - and must consider this heterogeneity, precisely in the design, production, adoption, and use of explainable systems.

Due to their skills, the three identified groups who relate differently to technology should be able to find a common level in the interaction with the Ai services implemented by the Pa.

Following these considerations, it appears that the Xai framework applies not only to technologies but also to the context and nature of the communicative and explanatory process that exists between technology and human actors. Therefore, it requires a multidisciplinary approach that guarantees its specific or agnostic applicability (Miller 2019).

The Xai framework is not only applicable to Ai systems but is effective for all Automated decision systems (Ads) which, just like Ai, are black-boxed. Xai aims to overcome the opacity often associated with advanced machine learning models, Ads and Ai by enabling users to understand how and why a system reaches a certain conclusion (Rai 2020). This transparency is especially crucial in high-accountability industries, such as healthcare and finance, where automated decisions can have significant impacts on human lives and the economy (Wen Loh 2022; Weber 2023). Xai emerges as an important component in the debate on future developments of responsible Ai, seeking forms of balance between advanced performance and social accountability (Arrieta 2020).

The Xai framework is designed to provide an approach to make the decision-making processes of Ai algorithms transparent and understandable (Ali *et al.* 2023).

In scientific literature, the Xai theme encompasses various aspects, including explanation, interpretation, and transparency, juxtaposed against black box systems (Angelov 2021). The topic of transparency through Xai applications is mainly discussed here and will be related to administrative transparency.

The explanation and transparency could improve the capacity to empower users who want to apply and use Ads enabling them to independently interpret the outcomes suggested by the system without other intermediary assistance. This main theme is linked to the sensitivity of the model underpinning the operation of automated systems. In instances where these frameworks and technologies lack self-evidence, prevent inexperienced individuals from grasping the system's internal workings, their action is limited. The Xai application and human action capacity are essential for building a type of application and an application context in which the technology is more efficient (Bento *et al.* 2021). Explainable systems that are more intuitive and comprehensible can also mitigate the complexity for non-experts, enabling them to actively contribute to the production process (Silva 2022). This collaboration between different actors involved in the production process of automatic systems gives access to further interventions that define the Xai framework such as, for example, the application of research methods to analyse the decisions made by the framework and evaluate their effectiveness and coherence concerning the expected objectives, subject of our discussion. This could enhance the clarity of the system's operation, benefiting not only end-users but also all stakeholders engaged in the production of digital technologies. By incorporating these elements during the design and development phases of automated and Ai systems, ethical principles can be seamlessly integrated into the models employed by

algorithmic systems. This integration aims to guarantee that the system's explanations align with prevailing social norms and regulations.

Unified within an Xai framework, these elements strive to strike a balance between the intricacy of Ai algorithms and the necessity to articulate their decisions in a lucid and comprehensible manner for heterogeneous users.

3. The European regulation on Ai between transparency, explainability and documentability

The graph of funding sponsors (Figure 2) denotes how in terms of investment there is an interest by European institutions concerning explainable artificial intelligence.

The Eu has been vigilant about the societal ramifications of digital technologies. Initially, through the enactment of Eu regulation 2016/679, known as the General data protection regulation (Gdpr) (Eu Agency for fundamental rights 2019), a reform was initiated concerning data protection and its digital dissemination. Subsequently, there is an ongoing effort through the regulatory process, specifically targeting the regulation of Ai applications and Ads.

In response to the growing spread of Ai, the European parliament produced a regulation called the Ai act. This is crafted to establish a unified regulatory framework among Eu nations, aiming to adeptly tackle the socio-economic risks that may arise from the extensive integration of Ai-based systems (Ai act, article 1). The initiative underscores an escalating recognition of the imperative for a cohesive regulatory strategy, ensuring the ethical and safe development and utilisation of Ai (Ai act, article 1, 5). A key focal point of the Ai act revolves around prioritising transparency, explainability, and documentation of actions executed by automated systems (Ai act, recital 38). This emphasis underscores the Eu's commitment to ensuring that both citizens and organisations possess a comprehensive understanding of the processes guiding decisions made by Ai-based systems. Recognising transparency is pivotal not only for instilling trust in emerging technologies but also for addressing potential biases or discrimination arising from non-transparent or opaque algorithms (Kizilcec 2016).

This regulation aims to establish a unified approach across the Eu, enabling member states to collaboratively address emerging challenges associated with Ai. The goal is to prevent regulatory fragmentation that could impede the advancement and adoption of intelligent technologies.

Moreover, the Ai act holds the potential to profoundly impact the future trajectory of Ai development. By urging the scientific and industrial communi-

ties to integrate ethical principles and transparency standards into the design and implementation processes of Ai-based systems, this regulatory initiative fosters a more responsible and thoughtful governance of Ai in the Eu (Li 2023). It aligns with common Eu values, promoting innovation in an ethical manner.

The Eu has introduced the concept of «trustworthy Ai» as a cornerstone of its strategy to regulate and steer the development and deployment of Ai within the Eu. In 2019, the European commission unveiled guidelines for trustworthy Ai. The trustworthy Ai framework rests on seven fundamental pillars, referred to as the «ethical Ai requirements».

These pillars include:

- respect for fundamental rights;
- good governance;
- transparency;
- diversity, non-discrimination and equity;
- safety;
- interoperability;
- responsibility.

These ethical requirements are designed to guarantee that Ai development and utilisation in the Eu adhere to principles of trustworthiness, safety, and respect for human rights, aligning with the Eu's core values.

The first pillar underscores the importance of respecting fundamental rights, emphasising the preservation of human rights, privacy, and social justice throughout all stages of Ai development. The second pillar, focusing on good governance, strives to establish a clear framework of responsibility and control to ensure the responsible and informed management of Ai advancements. Transparency, the third pillar, emerges as a pivotal element, necessitating understandable explanations for decisions made by intelligent algorithms, thereby fostering public trust. The fourth pillar centres on diversity, non-discrimination, and equity, aiming to eradicate bias and disparities, ensuring that Ai is developed and employed in an ethical and inclusive manner. The fifth pillar, security, concentrates on safeguarding Ai-based systems from threats and attacks, fostering a secure digital environment. Interoperability, the sixth pillar, encourages collaboration between intelligent systems and the harmonious coexistence of diverse technologies. Lastly, the seventh pillar places a significant emphasis on responsibility, delineating the roles and responsibilities of the entities involved in Ai development and usage, ensuring clear and transparent governance. Together, these pillars constitute a complete ethical and regulatory framework which, implemented in the Xai framework, becomes a fundamental guide for the evaluation of digital technologies, in particular Ai ones. This contributes to the social impact evaluation of these technological systems.

4. Evaluation of algorithmic systems through the Xai framework

The preceding discussion underscores the potential of integrating the Xai framework into the development of digital technologies, specifically Ai, to enhance transparency. Application of this framework can elucidate automated processes, providing clarity not only to public decision-makers but also to those affected by the Ads.

Recent historical trends reveal a swift and diverse digitalisation wave across public and private sectors, resulting in technologies and products with adverse effects on individuals and society (Androniceanu *et al.* 2022). In recent years we have witnessed various automated systems that have generated unexpected and potentially problematic effects on society such as, for example, the algorithm used for assigning teachers in Italy in the 2016/2017 school year (Aragona 2022) and the Automatic image recognition system (Sari) used in Italy in 2018 whose real-time image recognition application was blocked by the Italian data protection authority (negative opinion, March 25, 2021, n. 9575877).

Some digitalisation processes have prompted critical discourse on observed black-box phenomena (Pasquale 2015) (Figure 4).

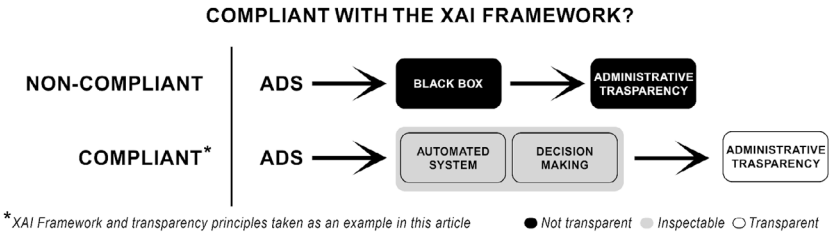

Fig. 4. From Xai to administrative transparency.
Source: Elaboration by the authors.

According to Xai framework, fully black-boxed systems will not be transparent while grey and white box systems that allow inspection will be compliant with the framework and administratively transparent (Figure 4).

In the following section we delve into a pertinent case study illustrating tangible adverse effects on individuals and society, attributable to a lack of administrative transparency stemming from the use of automated technologies that deviate from the principles of explainability.

The adoption of a case study aims to elucidate how the integration of the Xai framework into these processes can contribute to mitigating the adverse impacts of technologies employed in Pa. We employed two research methods

to reconstruct the context of deployment of the technology. Firstly, we used Search as research (Salganik 2018; Aragona and Amato 2022b), conducting research on Google trends and the Google search engine to identify keywords associated with the topic. Subsequently, these keywords were used on X to extract relevant hashtags (Marres and Weltevrede 2013). This sequential process enhanced our comprehension of the closely linked issues surrounding the investigated case. Furthermore, it broadened our investigative scope by retrieving additional information from newspaper articles, dedicated web pages, television broadcasts, and videos. The combination of these diverse sources enabled us to reconstruct the chronological sequence of events that led to the utilisation of the algorithm in teacher selection, and the subsequent protests.

Case study: the algorithm behind the regulation of school teacher rankings

In Italy, during the 2016/17 school year, the Ministry of education (Miur) employed an algorithmic system for teacher mobility operations. However, this implementation sparked extensive discussions and had adverse effects on numerous Italian teachers (Aragona 2022). Challenges arose in connection with the algorithmic system, particularly concerning the assignment of destinations. Destinations were not allocated based on the candidates' scores but rather prioritised the destination listed as the first choice by the candidate. This resulted in instances where teachers with higher scores, who had chosen a specific location as their second preference, were surpassed by teachers with lower scores. The determining factor was the priority given to those who had listed that city as their first choice. This issue gained national prominence as teachers expressed their grievances through protests, and the issue was extensively covered in both newspapers and national public service television.

The algorithm in question was created in response to a Ministry of education tender for overseeing It processes, allocated a budget of 117 million euros. Despite being implemented at a total expenditure of 444 thousand euros and developed by globally acclaimed It firms, experts called upon to evaluate the algorithm highlighted that «the most fundamental programming criteria known to apply were not observed» (Salvucci *et al.* 2017). They further remarked that the reasons behind the programmer's choice to create a system described as «pompous, redundant, and not oriented towards maintainability» were not clear (Salvucci *et al.* 2017).

Although this algorithm was introduced with the aim of improving the process of assigning teachers to schools, making it more transparent and based

on objective criteria, it is a clear example of how the development and implementation of the system did not achieve the initial objectives of the project.

The implementation of this algorithm has been the subject of controversy and criticism. Not only schools and teachers but also computer scientists summoned as experts and asked to express their opinions on the system have raised concerns regarding the lack of clarity in the functioning of the algorithm and the possible distortions generated in the assignments (Salvucci *et al.* 2017).

To gather date, we pinpointed the time frame during which the algorithm was active and the period when its effects became apparent. We employed Search as research on Google trends, utilising the terms «school»[4] and «algorithm». The period we selected for the research is 2016, the year the software was used. In this time frame, we focused mainly on the summer months when the teacher-graded list would be published. On Google trends, we compared the two selected terms to detect common peaks and from there began to identify themes and associated words. The Google trend search highlights how the term «algorithm» at that time was not yet in common use in searches, and the software clearly highlights this by showing how it is associated with a wide heterogeneity of terms. The most relevant peak that the two terms share is that of September 28, 2016, when, continuing the search on Google search, results in the publication of a television report aired on the La7 broadcaster entitled «Good school-in words. Chaos and desperation in reality»[5] in which specific reference is made to the teacher selection algorithm and the problems it has caused. On Google trends, the insight into the term «algorithm» on that day sees «meaning of algorithm» as the most frequently associated query. Going to look at the most associated queries about the term «school» there is the word «eligible». By changing the search parameters and employing «eligible» and «algorithm» jointly the peaks increase and the term «ghost» emerges. With this information, we moved back to Google search to expand the documents available to us and prepare for data extraction on social media X.

On this topic, we extracted the social media platform X users' posts from July 1, 2016, to December 31, 2016[6] (Table 1). The keywords used for the extraction were «algorithm» and «school». From the extraction, we derived 341 posts that in total were re-shared 838 times. The most frequently

[4] The keywords used for data extraction were in Italian, for the article they were translated into English.

[5] Title translated from Italian by the authors: Buona scuola – a parole. Caos e disperazione nei fatti. Retrieved November 13, 2023: https://www.la7.it/la-gabbia/video/buona-scuola-a-parole-caos-e-disperazione-nei-fatti-29-09-2016-194166.

[6] Search query on X «scuola algoritmo until:2016-12-31 since:2016-06-01».

used hashtags were «#algorithm» 86 times, «#school» 74 times, «#mobility2016» 73 times, and «#eligibleghost» 42 times (Figure 5).

The users' posts highlighted the frustration of those teachers who, having failed to score a suitable score to be selected for their first choice because they were overtaken by competitors with the same location who had scored higher, were passed over by teachers with lower scores who had selected as their first choice the one they had listed as their second.

Tab. 1. *Table of X scraping records*

Number of records	341
Number of authors	173
Number of accounts	224
Re-tweets	838

Source: Elaboration by the authors.

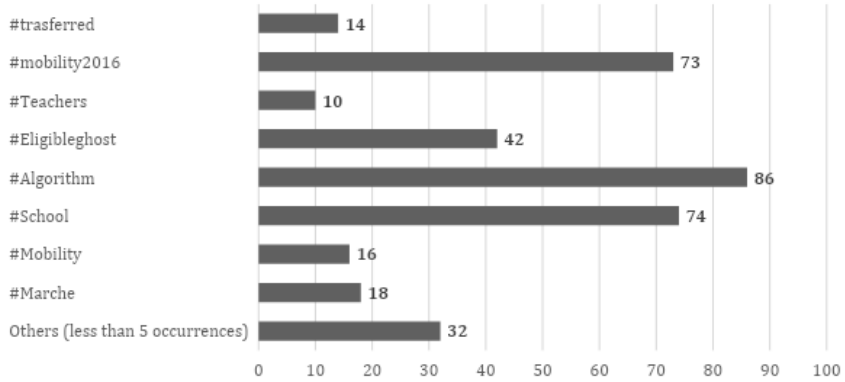

Fig. 5. Frequency of hashtags.
Source: Elaboration by the authors.

Despite protests, complaints, and media prominence, to date, it is reported that the teachers' trade union still denounces the negative harm that the teacher selection algorithm, continues to cause (Redazione scuola 2023).

Teachers suffered concrete and tangible negative effects from the algorithmic system, as, in the face of low pay, they were moved to distant regions due to system malfunctions. The hashtag «#eligibleghosts» is an indicator of the problem.

In the following years, the Miur continued to use algorithms which, at various levels, generated other problems for teachers and aspiring teachers. This further highlights how those principles of transparency and accountabi-

lity described within regulations may be abstract, but then do not find full integration and application within Pa processes, entailing the possibility of taking legal action (in this regard see the Council of State rulings 8472/2019, 2270/2019, 881/2020, 7891/2021). The focus on the issues arising from the use of algorithms for assigning positions is still active in 2023, becoming a current topic within the public sphere. Some more critical issues emerged by analysing the evolution of the case from 2016 to 2023.

In 2017 Miur declared that it used a different algorithm, but the system had a malfunction that prevented the loading of the teacher's applications. We observed new hashtags related to the case: «non-resident teachers united» 129 times, «increase mobility quota» 101 times.

In 2018, the main hashtag observed was «exiled teachers 107» 48 times.

In 2020 the «istanze online» platform for the provincial rankings for substitute teachers had problems in the user access phase. This same event also happened in 2021, when the algorithm used for the teacher rankings for the Gps excluded vulnerable groups. Furthermore, in September the rankings were cancelled and remade following other malfunctions.

Since 2021, following a tender worth 5.4 million euros, the Miur has relied on a new algorithm called National resident matching program match (Nrmc) for the assignment of teachers who, in addition to presenting the critical issues that have already emerged with the previous system discussed here, also has new problems like the exclusion of vulnerable groups and the inadequate match between teacher and positions.

In the three years 2021-23, there were numerous posts on X from users complaining about the negative effects of this new algorithm. Table 2 shows the number of posts (776) obtained from scraping that contain the terms «school» and «algorithm» from 2017 to 2023.

Tab. 2. *Table of scraping records 2017-2023*

Number of records		776	
Number of authors		583	
Number of accounts		679	
Yearly breakdown of tweets	Year	Tweets	Re-Tweets
	2017	59	1302
	2018	80	308
	2019	86	216
	2020	76	328
	2021	131	697
	2022	240	700
	2023	104	357

Source: Elaboration by the authors.

The case highlighted the challenges of using algorithms in decision-making, especially when it comes to sensitive issues such as human resource management in educational institutions. It emphasised the importance of careful implementation of algorithms, considering the Ai act, the trustworthy Ai framework and its seven fundamental pillars, and the involvement of stakeholders to avoid possible bias or misunderstanding.

It is our opinion that the that the application of the Xai frameworks could have helped in showing both the designer and the developer potential unexpected effects. and mitigate their impacts.

Potential applications of the Xai framework

The adoption of agnostic frameworks for Xai emerges as a promising strategy for designing automated decision systems that integrate diverse digital technologies. These frameworks, crafted to be adaptable and versatile across various contexts, seek to establish a robust conceptual foundation for addressing the challenge of Ai explainability (Miller 2019). Despite the advantage of a unified starting point offered by agnostic frameworks, it is crucial to acknowledge that, during development, distinct technical and contextual nuances of each technology may surface. Digital technologies exhibit considerable diversity in their characteristics and functionalities, defined in the literature as affordances (Norman 1999; Ehsan 2023). Put differently, the intricacy of different digital platforms may demand the incorporation of specific features within the Xai framework tailored to the features of the system or the specific context to which a given technology pertains. Consequently, while the utilisation of agnostic frameworks serves as a pragmatic outset (Arrieta *et al.* 2020; Brasse

et al. 2023), it is frequently concluded that, for enhanced explainability, the development of Xai specifications tailored to each context becomes a necessity (Adabi and Berrada 2018; Miller 2019; Rawal *et al.* 2021).

Employing frameworks for enhancing transparency in digital technologies involves a trade-off. This compromise entails weighing the features and the affordances of digital technology against the specifications of the intended implementation framework (Miller 2019; Angelov *et al.* 2021). The Xai framework necessitates customisation to align with the reference technology and the economic, social, and cultural context of its application.

5. An implementation of Xai in the policy cycle

Following the setting of the policy cycle (Howlett *et al.* 2020) and how this model is applied to the digital context, we considered the possibility of relating it to the Xai framework. To show this application of Xai to the policy cycle, we follow the approach presented by Höchtl, Parycek and Schöllhammer (2016) in which the relationship between big data and the e-policy making cycle is shown (Johnston 2015).

The Ai act and the developments in discussions around Ai highlight the importance of assessing the impact of Ads. However, this assessment is not directly linked to achieving greater explainability of these Ai systems. The media and scholars' discussions of issues related to Ai and the risks associated with digital technologies, can lead to a more active awareness regarding the functioning of these systems. The Xai supports the technological assessment by guaranteeing more transparent information on the functioning of the Adms, allowing the media, experts and citizens to understand more clearly the functioning of a specific Adm. Figure 6 shows how the Xai framework could fit into the policy cycle.

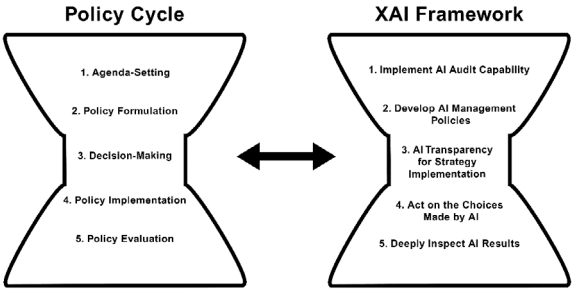

Fig. 6. Implementation of the Xai framework in the policy cycle.
Source: Elaboration by the authors.

As Agcom has already done about the Gdpr (2016) and the adoption of the rules regarding cookies and the cookie banner, a designated authority could also act as a guarantor for Xai instances. This could be achieved through the development of a framework and metrics to initiate a case study evaluation of Ai technologies in use in Pa. Furthermore, it could act as an evaluating authority for future applications for which the presence of Adm is expected. To analyse and evaluate not only the levels of risk and impact but also the explainability and transparency of these technologies.

Considering the presence of an authority capable of evaluating and implementing Xai instances, as happened with the Gdpr and its effective application, can significantly lengthen the timeline for the implementation of explainable solutions within the Adms. Otherwise, different solutions could, on the one hand, take into consideration the establishment of professional figures who, placed within the Pa offices, can manage Xai instances from within, or, on the other, equip these same offices in the implementation of these tools, the skills necessary to accommodate Xai instances.

Like the e-policy cycle that integrates continuous assessments between process phases, the same may be necessary for Xai integration. The characteristics that define Xai are closely connected to cultural and social values that differ based on context and can change over time. For these reasons, it is important to take into consideration how the needs for such requests today can exclude elements that could become essential tomorrow.

6. Conclusions

Pa increasingly employs digital technologies, so frameworks such as the Xai plays a crucial role in the design of transparent systems from the earliest stages of development, to prevent or mitigate the adverse effects of technologies, especially Ai based. The ability to explain Ai results is critical to ensuring regulatory compliance and addressing ethical concerns related to the use of decision-making algorithms. In this context, the development of Xai emerges as an important pillar for the future deployment of Ai in Pa, creating a balance between advanced performance and social accountability (Panigutti *et al.* 2023).

The case study considered in this article highlights the critical issues faced by a national context in the development of algorithmic systems that have significant degrees of discretion such that the actions performed can have normative effects on individuals.

The Xai framework applied to Adms requires being inside Pa processes. This contribution, not being able to enter the black box of technology, instead

took into consideration a well-known case to highlight how the lack of those principles that define the Xai framework could produce negative effects on society.

The posts that we had the opportunity to analyse, as well as the documentary sources that we took into consideration, brought various questions to the attention of the media, public decision-makers, and judicial bodies. Firstly, was identified a growing demand for transparency and control on the part of the institutions involved in the teacher selection process. This request, not being applied in the years following the use of the algorithm discussed here, led to less trust in the decision-making process, in the institutions, and in the technological tool itself. The digitalisation process has highlighted a discrepancy within the Pa which, although it adheres to european norms by proposing to defend the values of transparency and accountability of administrative processes, ends up not implementing those principles within the processes in use. The discrepancy between the objectives and the results highlights how these tools are not used appropriately. One of the effects generated by these practices was to increase appeals to judicial bodies which, through legal actions, took charge of the unfair effects of algorithmic decisions.

We are aware that the sole introduction of the Xai framework is not decisive, because greater participation and attention would be needed from the stakeholders involved in the production and use of Adms by Pa. But it could guarantee that basic knowledge, currently precluded by the presence of unexplainable systems (black boxes), so that multiple subjects, having the opportunity to exercise the principle of knowability, can assess digital technologies more central topic in public opinion.

The importance of addressing future challenges and implementing solutions that promote the balance between technological innovation and rights protection from an inclusive perspective is emphasised, ensuring a future in which Ai can contribute positively to society. Following the call for a multidisciplinary approach to the Xai framework in Miller's (2019) article, we believe that research in the social sciences can contribute constructively by providing methods and approaches necessary for the design and evaluation of explainable artificial intelligence systems. Integrating these processes into technology production facilitates the incorporation of essential Xai framework characteristics into product design from the early stages (Mulvenna *et al.* 2017). Consequently, new technological applications are developed with transparency and explainability principles integrated, promoting a fairer and less opaque functioning of these digital technologies.

References

Adadi, A. and Berrada, M. (2018), *Peeking Inside the Black-Box: A Survey on Explainable Artifcial Intelligence (XAI)*, in «IEEE Access». doi: 10.1109/ACCESS.2018.2870052.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Del Ser, J., Diaz-Rodriguez, N. and Herrera, F. (2023), *Explainable Artificial Intelligence (XAI): What We Know and What is Left to Attain Trustworthy Artificial Intelligence*, in «Information Fusion». doi: 10.1016/j.inffus.2023.101805.

Androniceanu, A., Georgescu, I. and Sabie, O. M. (2022), *The Impact of Digitalization on Public Administration, Economic Development, and Well-Being in the EU Countries*, in «Central European Public Administration Review». doi: 10.17573/cepar.2022.2.01.

Angelov, P.P., Soares, E.A., Jiang, R., Arnold, N. I. and Atkinson, P. M. (2021), *Explainable Artificial Intelligence: an Analytical Review*, in «Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery». doi: 10.1002/widm.1424.

Aragona, B. (2022), *Algorithm Audit: Why, What and How?*, London, Routledge.

Aragona, B. and Amato, F. (2022a), *Rischi algoritmici e strumenti di mitigazione*, in «Riskelaboration», 3(1), pp. 41-46.

Aragona, B. and Amato, F. (2022b), *Retracing Algorithms: How Digital Social Research Methods Can Track Algorithmic Functioning*, in F. Comunello, F. Martire and L. Sabetta (eds.), *Frontiers in Sociology and Social Research*, Springer International Publishing.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A. and Herrera, F. (2020), *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible Ai*, in «Information Fusion», 58, pp. 82-115.

Bento, V., Kohler, M., Diaz, P., Mendoza, L. and Pacheco, M. A. (2021), *Improving Deep Learning Performance by Using Explainable Artificial Intelligence (XAI) Approaches*, in «Discover Artificial Intelligence». doi: 10.1007/s44163-021-00008-y.

Berente, N., Gu, B., Recker, J. and Santhanam, R. (2021), *Managing Artifcial Intelligence*, in «MIS Quarterly». doi: 10.25300/MISQ/2021/16274.

Brasse, J., Broder, H. R., Förster, M., Klier, M. and Sigler, I. (2023), *Explainable Artificial Intelligence in Information Systems: a Review of the Status Quo and Future Research Directions*, in «Electronic Markets». doi: 10.1007/s12525-023-00644-5.

Callon, M. (1986), *The Sociology of an Actor-Network*, in M. Callon, J. Law and A. Rip, *Mapping the Dynamics of Science and Technology*, London, Palgrave Macmillan.

Cozzens, S. E., Bijker, W. E., Hughes, T. P. and Pinch, T. (1989), *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*, in «Technology and Culture». doi: 10.2307/3105993.

Doshi-Velez, F. and Kim, B. (2017), *Towards a Rigorous Science of Interpretable Machine Learning*, in «ArXiv». doi: 10.48550/arXiv.1702.08608.

Došilović, F. K., Brčić, M. and Hlupić, N. (2018), *Explainable Artificial Intelligence: a Survey*, paper presented at the 41st International convention on information and communication technology, electronics and microelectronics (MIPRO), IEEE.

Ehsan, U., Saha, K., De Choudhury, M. and Riedl, M.O. (2023), *Charting the Sociotechnical Gap in Explainable Ai: A Framework to Address the Gap in XAI*, in «Proceedings of the ACM on Human-Computer Interaction». doi: 10.1145/3579467.

European Commission (2023), *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts.*

European Union Agency for Fundamental Rights (2019), *The General Data Protection Regulation. One Year on: Civil Society, Awareness, Opportunities and Challenges*, in «Publications Office». doi: 10.2811/538633

Gillespie, T. (2014), *The Relevance of Algorithm. Media Technologies: Essays on Communication, Materiality, and Society, Cambridge*, MIT Press.

Gilpin, L. H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M. and Kagal, L. (2018), *Explaining Explanations: an Overview of Interpretability of Machine Learning*, paper presented at the IEEE 5th International Conference on data science and advanced analytics (DSAA).

Hess, M., Garside, D., Nelson, T., Robson, S. and Weyrich, T. (2017), *Object-Based Teaching and Learning for a Critical Assessment of Digital Technologies in Arts and Cultural Heritage*, in «The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences», 42, pp. 349-354.

Hoch, P.K., Mackenzie, D. and Wajcman, J. (1987), *The Social Shaping of Technology,* in «Technology and Culture». doi: 10.2307/3105489.

Höchtl, J., Parycek, P. and Schöllhammer, R. (2016), *Big Data in the Policy Cycle: Policy Decision Making in the Digital Era*, in «Journal of Organizational Computing and Electronic Commerce», 26(1-2), pp. 147-169.

Howlett, M., Ramesh, M. and Perl, A. (2020), *Studying Public Policy: Principles and Processes*, Toronto, Oxford University Press Canada.

Italian Data Protection Authority (2021), *Opinion on the Sari Real Time system - 25 March 2021* [9575877], https://www.asaps.it/downloads/files/GarantePrivacy-9575877-1_2.pdf (last accessed on 11th January 2019).

Johnston, E. W. (2015), *Governance in the Information Era: Theory and Practice of Policy Informatics*, London, Routledge.

Kitchin, R. (2014), *Big Data, New Epistemologies and Paradigm Shifts*, in «Big Data & Society». doi: 10.1177/2053951714528481.

Kitchin, R. and Lauriault, T. (2014), *Towards Critical Data Studies: Charting and Unpacking Data Assemblages and Their Work. The Programmable City Working Paper 2*, pre-print version of chapter to be published in J. Eckert, A. Shears and J. Thatcher (eds), *Geoweb and Big Data*, University of Nebraska Press. Forthcoming, https://ssrn.com/abstract=2474112 (last accessed on 10th January 2024).

Kitchin, R. (2016), *Thinking Critically about and Researching Algorithms*, in Information, «Communication & Society». doi: 10.1080/1369118x.2016.1154087.

Kizilcec, R. F. (2016), *How Much Information? Effects of Transparency on Trust in an Algorithmic Interface*, in Proceedings of the 2016 CHI conference on human factors in computing systems, pp. 2390-2395.

Landri, P. (2018), *Digital Governance of Education: Technology, Standards and Europeanization of Education*, London, Bloomsbury.

Latour, B. and Woolgar, S. (1979), *Laboratory Life: The Construction of Scientific Facts*, Princeton, Princeton University Press.

Latour, B. (1987), *Science in Action: How to Follow Scientists and Engineers through Society*, Cambridge, Harvard University Press.

Latour, B. (2005), *Reassembling the Social: An Introduction to Actor-network-theory*, Oxford, Oxford University Press.

Li, Z. (2023), *Why the European Ai Act Transparency Obligation is Insufficient*, in «Nature Machine Intelligence». doi: 10.1038/s42256-023-00672-y.

Mackrill, T. and Ebsen, F. (2018), *Key Misconceptions When Assessing Digital Technology for Municipal Youth Social Work*, in «European Journal of Social Work», 21(6), pp. 942-953.

Madan, R. and Ashok, M. (2023), *Ai Adoption and Diffusion in Public Administration: A Systematic Literature Review and Future Research Agenda*, in «Government Information Quarterly». doi:10.1016/j.giq.2022.101774.

Marres, N. and Weltevrede, E. (2013), *Scraping the Social? Issues in Live Social Research*, in «Journal of Cultural Economy», 6(3), pp. 313-335.

Miller, T. (2019), *Explanation in Artificial Intelligence: Insights from the Social Sciences*, in «Artificial Intelligence». doi: 10.1016/j.artint.2018.07.007

Mulvenna, M., Boger, J. and Bond, R. (2017), *Ethical by Design: A Manifesto*, in «Proceedings of the European Conference on Cognitive Ergonomics», pp. 51-54.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., Van Keulen, M. and Seifert, C. (2023), *From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable ai*, in «ACM Computing Surveys», 55(13s), pp. 1-42.

Norman, D. A. (1999), *Affordance, Conventions, and Design*, in «Interactions». doi: 10.1145/301153.301168.

Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., Scalzo, S., Mazzini, G., Sanchez, I., Soler Garrido, J. and Gomez, E. (2023), *The Role of Explainable Ai in the Context of the Ai Act,*

paper presented at the ACM Conference on Fairness, Accountability, and Transparency, doi: 10.1145/3593013.3594069.

Pasquale, F. (2015), *The Black Box Society: The Secret Algorithms That Control Money and Information*, Cambridge, Harvard University Press.

Rai, A. (2020), *Explainable Ai: from Black Box to Glass Box*, in «Journal of the Academy of Marketing Science». doi: 10.1007/s11747-019-00710-5.

Rawal, A., Mccoy, J., Rawat, D., Sadler, B. and Amant, R. (2021), *Recent Advances in Trustworthy Explainable Artifcial Intelligence: Status, Challenges and Perspectives*, in «IEEE Transactions on Artifcial Intelligence». doi: 10.1109/TAI.2021.3133846.

Redazione Scuola (2023), *Gilda: L'algoritmo per le nomine dei supplenti sembra non aver funzionato*, «Il Sole24Ore», https://www.ilsole24ore.com/art/gilda-l-algoritmo-le-nomine-supplenti-sembra-non-aver-funzionato-AFBpozw?refresh_ce=1(last accessed on 10th January 2024).

Ribera, T. M. and Lapedriza, A. (2019), *Can We Do Better Explanations? A proposal of User-Centered Explainable Ai,* https://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-12.pdf.

Rogers, R. (2013), *Digital Methods*, Cambridge, MIT Press.

Salganik, M. J. (2018), *Bit by Bit: Social Research in the Digital Age*, Princeton: Princeton University Press.

Salvucci, A., Giorgi, M., Barchiesi, E. and Scafidi, M. (2017), *Perizia tecnica preliminare sull'analisi dell'algoritmo che gestisce il software della mobilità docenti per l'a.s.2016/2017*, https://www.gildavenezia.it/wp-content/uploads/2017/06/Perizia-tecnica-preliminare2017.pdf (last accessed on 10th January 2024).

Silva, A., Schrum, M., Hedlund-Botti, E., Gopalan, N. and Gombolay, M. (2022), *Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of xAI on Human-Agent Interaction*, in «International Journal of Human–Computer Interaction». doi: 10.1080/10447318.2022.2101698.

Van Dijk, J. A. G. M. (2020), *The Digital Divide*, Cambridge, Polity.

Veale, M. and Brass, I. (2019), *Administration by Algorithm? Public Management Meets Public Sector Machine Learning*, in K. Yeung and M. Lodge (eds.) *Algorithmic Regulation*, Oxford, Oxford University Press.

Wang, D., Yang, Q., Abdul, A. and Lim, B. Y. (2019), *Designing Theory-Driven User-Centric Explainable Ai*, in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI '19: CHI Conference on Human Factors in Computing Systems. ACM. doi:10.1145/3290605.3300831.

Weber, P., Carl, K.V. and Hinz, O. (2023), *Applications of Explainable Artificial Intelligence in Finance—a Systematic Review of Finance*, paper presented at Information Systems, and Computer Science literature». doi:10.1007/s11301-023-00320-0.

Wen Loh, H., Ping Ooi, C., Seoni, S., Barua, P. D., Molinari, F. and Acharya, R. (2022), *Application of Explainable Artificial Intelligence for Healthcare: a Systematic Review of the Last Decade (2011–2022)*, in «Computer Methods and Programs in Biomedicine», 226, doi: 10.1016/j.cmpb.2022.107161.