Chapter 14 – Searching XAI collaborating with manager: bi-directional learning for Human-Tech applications

Michela Arnaboldi[a*] and Luca Brusasco[a]

[a] *Politecnico di Milano - Dipartimento di Ingegneria Gestionale, Milan, Italy*

**Corresponding author: Michela Arnaboldi,* michela.arnaboldi@polimi.it

Abstract

The attention of practitioners and policymakers on artificial intelligence (AI) has reached unprecedented levels, especially following the European Union's final approval of the Artificial Intelligence Act (AI Act). Although systems with unacceptable risks are outright banned, high- and limited-risk systems present critical challenges, particularly in terms of transparency and assignment of responsibility. This chapter focuses on an underexplored area of AI, referred to as "Human-Tech AI," which involves AI systems integrated into organizational processes where human accountability is necessary. The study explores a collaborative approach to build explanations and trust among decision-makers and users interacting with AI, emphasizing that explainability should be a bidirectional process of knowledge exchange rather than a one-way explanation from developers.

Keywords

Risks, Responsible regulations, Ecosystem, Stakeholders, Black-box, Trust

14.1. Introduction

The attention of practitioners and policy makers on AI have never been so high (Fosso Wamba et al., 2021). This attention has been recently augmented by the adoption of the European Union's Artificial Intelligence Act (AI Act), approved on 13th March 2024. The approval comes after a long and tortuous path, started in 2021, which entails a new type of regulation based on risks categories where the "human" is put at the forefront as in the quote from the regulation highlights:

> "The purpose of this Regulation is to improve the functioning of the internal market by laying down a uniform legal framework in particular for the development, the placing on the market, the putting into service and the use of artificial intelligence systems (AI systems) in the Union, in accordance with Union values, to promote the uptake of human centric and trustworthy artificial intelligence (AI) while ensuring a high level of protection of health, safety, fundamental rights as enshrined in the Charter of Fundamental Rights of the

> European Union (the 'Charter'), including democracy, the rule of law and environmental protection, to protect against the harmful effects of AI systems in the Union, and to support innovation. This Regulation ensures the free movement, cross-border, of AI-based goods and services, thus preventing Member States from imposing restrictions on the development, marketing and use of AI systems, unless explicitly authorised by this Regulation." (p. 1, AI Act).

The EU has set human rights and trustworthiness as key. The EU links risk and transparency closely. In fact, the AI Act classifies AI systems according to the level of riskiness and establishes prohibition or transparency obligations for each level. Such categories are: unacceptable-risk, high-risk, limited-risk, and minimal-risk.

Unacceptable risk includes applications that dangerously affect individuals' physical and psychological spheres or strongly limit the human rights acknowledged by the EU. These risks are totally banned. High-risk comprehends applications that adversely impact people's safety and fundamental rights. They require a conformity assessment with detailed procedures and documentation. Limited-risk encompasses applications that may have light negative outcomes on individuals. They require only transparency obligations. Finally, minimal-risk applications are systems that do not affect individuals at all. They do not have any requirements, but a code of conduct is anyhow strongly suggested.

Although unacceptable risk for AI applications are the most critical, the "solution" is easier given that they are forbidden, with no need to create trustworthiness in users. However, the other categories are critical for two reasons: first, the possible ambiguity in assigning AI to a category, favoring the potential action of developers to "move" AI into the least risky category possible; second, the mainstream attention is on the final users overlooking applications where the AI system is integrated within organizational processes. For instance, think of an AI application which augments the calculative possibility to detect fraud and then prosecute people. The final decision and action to prosecute will need human responsibility, i.e., an accountable person.

These applications, that will be called "Human-Tech AI" in this chapter, outline the need to create trustworthiness not only in the final user but also in managers, officers who will operate and act on the AI outputs. Indeed, previous studies in other fields (e.g., Quattrone, 2016) have evidenced that two possible behaviors might happen. When human responsibility in the Human-Tech AI is clearly defined, a lack of control over results leads to not use AI. On the other hand, there might be uncritical behaviors, especially when final responsibility is distributed among more than one accountable person.

In this chapter, we will focus on this overlooked area of Human-Tech AI. Specifically, we aim to analyze the experimentation of a collaborative approach in building explanations and then trust for

decision-makers and users operating with the AI system. The premise of this study, however, is that explainability is not a traditional one-way teaching process in which one actor, typically the developers, explains to the others. Instead, the premise is that explainability is a bidirectional process in which analysts and managers collaboratively exchange information and increase mutual and overall knowledge.

14.2. XAI for Human-tech applications

The concept of trustworthy AI is a common node for regulators, particularly in the EU, and even academic studies, through the approach of explainable AI (XAI).

At the practitioner level, the High-Level Expert Group on Artificial Intelligence (HLEG) offers its vision of trustworthy AI by originally dividing this notion into three main components: lawful AI, ethical AI, and robust AI (HLEG, 2019a). Whereas the first component tends to have a focus on the technical aspects of laws and regulations, the other two components deal respectively with ethical principles/values and with the prevention of technical and social harm. By focusing on ethical and robust AI, the HLEG introduces a set of seven nonexhaustive key principles (the HLEG calls them "requirements") whose collective satisfaction promotes the practical implementation of trustworthy AI. The seven categories of requirements are human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, nondiscrimination and fairness; societal and environmental wellbeing; and accountability. These requirements are based on human rights, besides the principles of respect for human autonomy, prevention of harm, fairness and explicability (HLEG, 2019a). Their assessment should be frequent and happen in every portion of the AI lifecycle. The definitions are provided through their dedicated webpage (HLEG, 2019b).

The AI lifecycle is central also when we move to academic studies where the need of trustworthy AI is referred to and then implemented through Explainable AI (XAI). This is seen as a proper strategy to engender both trust and accountability, particularly in front of the increasing demand for trustworthy AI (Barredo Arrieta et al., 2020). Literature has many contributions stressing the relationship between explainability and trust. For instance, Haque et al. (2023) identify trust as the first of five dimensions within the effect of explainable AI, whereas Grimmelikhuijsen (2023) experimentally confirms with street-level bureaucrats the importance of explainability for trust in AI compared to the simple transparency or accessibility of an algorithm (i.e., public availability of source code, model and/or data). Analogous observations elevate explainability as a paramount mediating factor for accountability (Kim et al., 2020; Busuioc, 2021; Colucci et al., 2023).

Explainability gains popularity with the spreading of nontransparent black-box models which fail to engender trust in humans, particularly when explanations have to "make sense to the users" (Saeed

and Omlin, 2023). The main reason lies in the trade-off between accuracy and explainability (Lee and Shin, 2020): AI models constitute a frontier where the maximum accuracy sacrifices the explainability of the model and, consequently, trust and acceptance. This is true to the point that some techniques are discarded ex-ante by organizations because of the lack of explainability (Barredo Arrieta et al., 2020).

Early XAI approaches date back to before the AI and machine learning boom and focused on explaining the results of rule-based systems, such as expert systems (Swartout et al., 1991; Xu et al., 2021). The field of XAI has moved from explaining rule-based systems to new self-learning systems such as machine learning. Whereas rule-based systems are based on human-built rules, self-learning AI systems (e.g., machine learning and deep learning) are based on learning from data and sophisticated statistical models (Belle and Papantonis, 2021; Domingos, 2012). The majority of these self-learning AI models are often referred to as "black-box models" due to the lack of information about their inner workings (Hassija et al., 2024; Meske et al., 2022; Pasquale, 2015; Xu et al., 2021). Thus, rule-based systems have predetermined rules from the start, whereas in self-learning models, data is the starting point.

Explainability for AI and ML refers to helping the future user understand why models behave the way they do (Bhatt et al., 2020). The results of the models and the reasoning process behind these results (i.e., the relationship between input and output) should be explainable and thus understandable to humans. Therefore, opening the black box and sharing internal information is at the core of XAI approaches for these types of models (Barredo Arrieta et al., 2020). One reason for the growing interest in XAI seems to be the increasing application of self-learning AI algorithms in critical contexts such as health (Sun and Medaglia, 2019) or the use of AI to detect tax fraud (Hartmann and Wenzelburger, 2021). The decision to investigate citizens indicated by an AI as potentially fraudulent can have far-reaching consequences for these citizens. In such cases, an explainable AI model can create confidence in these decisions (Belle and Papantonis, 2021).

14.3. Method

This section presents the methodology followed to study how an XAI approach is adopted towards the internal actors who will operate the system upon release of an application. XAI versus managers and employees is key for creating awareness of the "black box," and highlights their residual human responsibility in the augmented process. In doing so, an action research project was undertaken with an interdisciplinary team. The two authors who were involved in the project had diverse roles: one was actively involved in the implementation; the other was a participant observer not involved in the implementation. This role configuration allowed us to have a dual perspective on the results.

The case for this chapter concerns a large public government agency, here called WORK to ensure anonymity. WORK protects workers against physical and economic damage. Each employer is required to submit a self-declaration of the tasks performed by its workers and the relative amount of remuneration. Based on these data and an estimate of each task's risk, diverse tax rates are set by WORK. Inspections controlling the truth of an employers' declarations are carried out on a regular basis to verify that there are no irregularities in the self-declared data. This inspection process is complex and expensive due to the need for in-presence controls. Therefore, WORK is prevented from controlling every single employer.

The goal of WORK in the action research project was to "augment" the process of the control of real data which precedes the physical inspection (preinspection process). Augmentation had to be achieved by implementing tools that exploit AI and Machine Learning. It was agreed since the beginning that it is paramount not to replace existing processes and related crucial players, but rather to develop systems and tools that assist human beings in the decision-making process, improving their efficiency and effectiveness. A collateral objective (but still relevant) was to experiment with the use of data held by WORK in order to feed into AI and Machine Learning processes and systems, with a view to increasing the use of these technologies within, and in support of, WORK's decision-making processes.

These objectives are reflected in the following requirements: (1) make the process more objective and data-driven; (2) leverage AI and automation to scale the preinspection process, focusing human inspectors' attention on the most important cases; (3) study the impact on the organizational structure of the use of AI-based tools; (4) sharing each phase of the project and decisions in order to experiment with an XAI in collaboration with the managers and officers involved in the preinspection process.

In the implementation of the project, we adopted a well-defined methodology to maximize the efficiency and quality of the results. This methodology is based on four fundamental principles:

- Process Augmentation. Our approach does not aim to completely replace the existing process, but rather to make the most of it. We intervened in a targeted manner on the most critical parts, using Artificial Intelligence algorithms to improve their performance.
- Human in the Loop. We designed the flow of using AI algorithms with an explicitly predicted human presence. This means that human decisions are integrated into the process, ensuring constant control and supervision of automated operations.
- Modularization. Our methodology is based on modularization, where AI models and algorithms are focused on solving specific problems vertically. Subsequently, these specialized solutions and tools can be integrated into the overall process, creating a cohesive and high-performance system.

- Data Centricity. We recognized that data quality and understanding are crucial factors for the success of the project (as well as the majority of projects that involve the use of AI and ML within a complex business process). Therefore, we have paid special attention to data management, ensuring that it is accurate and suitable for use with AI algorithms. The performance of the final solution depended largely on the quality and completeness of the data used.

14.4. Results

The results are broken down according to the following diagram (Fig. 14.1), in which four circles are highlighted. These circles highlighted the appropriateness of fine-tuning the XAI approach according to the level of internal stakeholder involvement and relationship with the application of AI. The following section illustrates results following the circles.

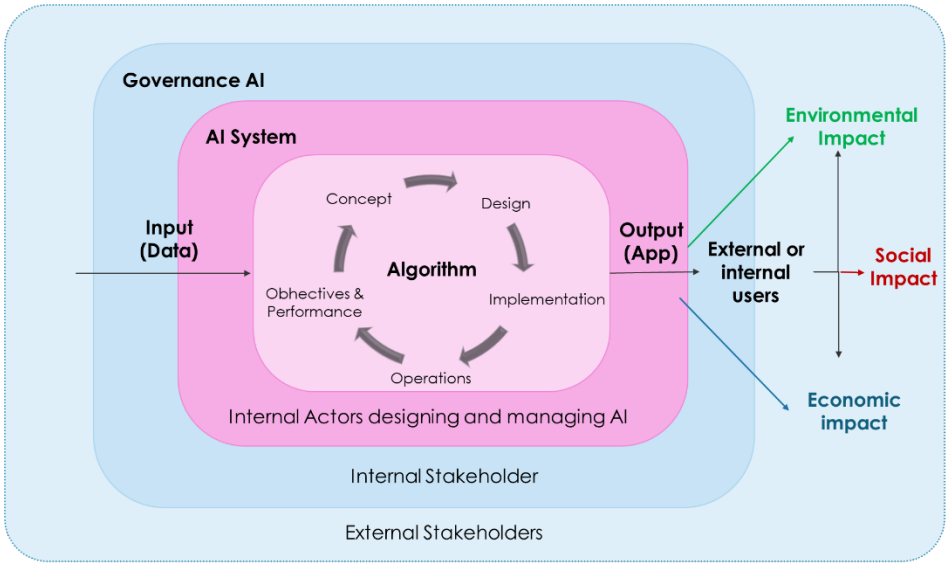

Figure 14.1 – The reference framework: producing the output of an application (i.e., App).

14.4.1. The Algorithm

To address the problem posed, the team developed two innovative solutions. After presenting the first one and receiving feedback on the subject, a second version was proposed, which, after being approved, was presented to the general management.

The first solution starts from the assumption that evasive behavior follows similar patterns in different sectors, and depends on risk rates. The solution is based on two main components. The first is generating physical inspection candidates. The process starts with a potential candidate for a physical inspection. Using advanced data analysis and AI techniques, the application identifies companies that have significant similarities with the candidate under consideration. This allows the creation of a list

of similar companies that have been previously inspected. The included dimensions are the behaviors regarding self-declarations, any infringements committed and identified by the inspections, as well as the real distribution of the wage frequency given the risk rate.

With the list of similar companies, the application proceeds with the processing of their wage statements. This step is crucial because it allows reviewers to understand the financial and salary behavior of similar companies. Based on these detailed analyses, the application provides the user with an informed assessment of whether or not a physical inspection is necessary for the original candidate. This assessment is based on evidence and is a key factor in decision-making.

The implementation of this solution lasted six months. The duration was higher due to the desire to test the XAI approach, intended as a higher interaction between the developers' team and the managers at diverse levels.

After the presentation of this solution, there was a moment of realization that employers' industries count. Therefore, another road was undertaken, in the form of a second solution. The second solution assumes that different companies within the same sector (identified by a code) have a limited number of possible patterns. It is based on two main components. The first component is the construction of an identikit for a sector code: starting from the data accumulated so far, a ``description" of the declarative behavior of each sector is built, which is easy to read for a user. This was a challenging part where a clustering of data was needed using K-means. Once 1 to K clusters have been generated, graphs that compare the difference between a greater or lesser number of clusters are produced. To do this, two metrics were used: the first is the Silhouette Score, a measure to assess the cohesion and separation of clusters. In general, a higher Silhouette Score indicates better clustering quality, with well-defined and separate clusters. In addition to this measure of intercluster variability, an intracluster variability (highlighted mean) is also calculated. This variability is defined as the volume of the intersection between the hypercube defined by the confidence intervals that are created for each entry, and the simplex within which the data resides, compared to the total volume of the simplex.

The second solution has a second step which is the extraction of companies of interest: given the declarations of many new companies, a list of companies is extracted that may be interesting to analyze in detail. The operation in this case is simple and intuitive, using a simple ranking algorithm. At this point, to get a sublist of companies of interest, simply take the first ones based on the ranking. In both solutions, humans have a key role and maintain the final responsibility to decide whether a company, and which company, will be inspected.

In the following section, we illustrate how the XAI has been carried out along with the conceptualization and design of solutions.

14.4.2. AI system: algorithms and human actors

The inner circle is at the heart of XAI, where explainability has been set up as highly interactive, thus highly challenging both temporally and cognitively for the WORK staff ("AI system," Fig. 14.2). The involved staff include: top managers setting the intended goals and expectations; selected middle managers in charge of areas involved in the future operative management of the AI application; and officers with specific competencies of the risk process or information systems and datasets.

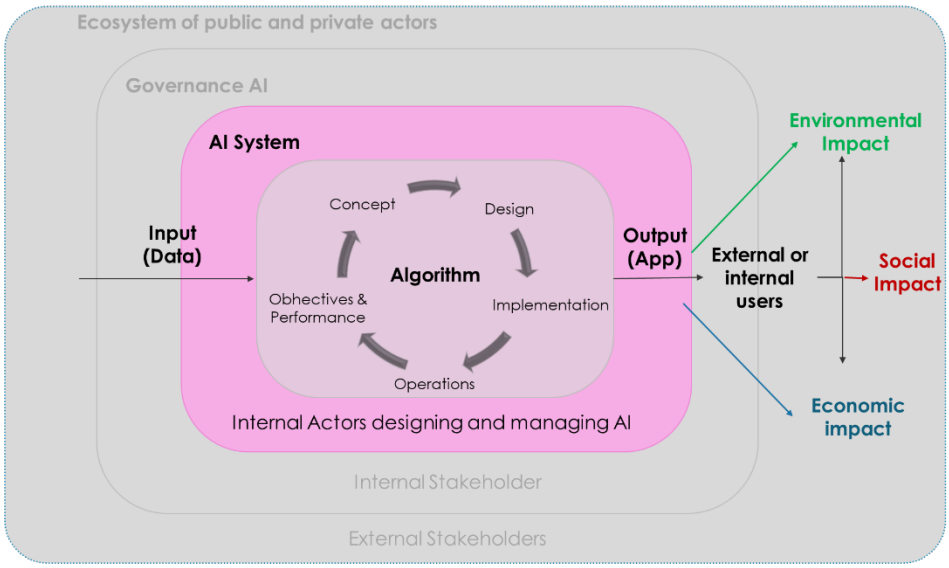

Figure 14.2 – The AI System.

The high commitment required was agreed with the CEO and the top management. The agreement was formalized in a document and then presented in an official kick-off meeting. These are the principles shared:

1. Leverage AI and automation to scale and make the preinspection process more efficient and effective;
2. Provide an intersection in the process between managers and the AI application, creating knowledge and trust in the overall AI system;
3. To study the impact on the organizational structure and process of WORK due to the use of the AI system;
4. Enhance WORK's data sources, also by cross-referencing them with external data.

Those principles were presented at the kick-off meeting, with the presence of the CEO, top managers and the selected "core" staff.

Knowing the context of the application (risk monitoring) was the first step of the project, which revealed it to be a training ground for explainability. The process to be augmented was in fact

complicated and spread over several organizational units; learning for the implementer was incredibly beneficial in understanding both the risk processes and the preparation of people on AI.

When the conceptualization of the application began, the developers' team scheduled almost one meeting per week to share progress and choices. At this stage, the explanation started to go into technical details; and the interactions with managers slowed down and became longer than expected. The preparedness of managers in the field of AI was not very high, with a few exceptions. Yet the curiosity to know was on average high in all and ready to read and study material. This was beneficial for the XAI, but it affected the efficiency of the project.

The true value-added of the XAI, however, emerged in the second half of the development, when the first solution was presented. The staff could test the application and simulate the new context of their operations. The core staff of WORK here realized that some choices suggested initially to the developers resulted in being inadequate for the scope (i.e., preselection of a candidate company for inspections). Mutual learning was at its peak here, shifting to specific responsibilities within the preinspection process, and also between them and the AI developers. Where the responsibility of one and the other ends and begins was unknown. But taking a concrete example, the sectoral aggregation taken in Solution 1 was too broad, leading to the inclusion of companies that were not risky at all. To be even more clear in their explanation to the developers' team, a statistician expert of industries' data was involved, who prepared a short presentation to illustrate the pitfalls of Solution 1.

The developers then started to design Solution 2, enriched by learning about specific responsibilities. During the implementation of the second solution, both parties were more focused on formulating the right *questions*.

To summarize, the core system is where XAI is best implemented. Evidence from experimentation points to a fundamental evolution of the approach (*how*): in the first part the focus is on "explaining," while in the second part, it is on "asking" for the right things at the right time. Moving on to "what," i.e., the object of the explanation, in the initial stages it is necessary to focus on the application *context and technical aspects*, while in the final stages, having noted the first two, it is possible to descend to the *specific responsibilities* of the people involved. This result is a key point in view of the evolving nature of AI applications. Within this responsibility was included also the investigation of how the *workload* of people now involved will change.

14.4.3. Governance system actors

The XAI approach emerged not to stop with the sole AI system. On the contrary, it was extended to include what we call AI Governance (Fig. 14.3). AI Governance stands for the set of human actors in

the organization who are involved in the management of AI with a portfolio scope, namely, without looking at a specific application. Their role is within the general governance of AI.

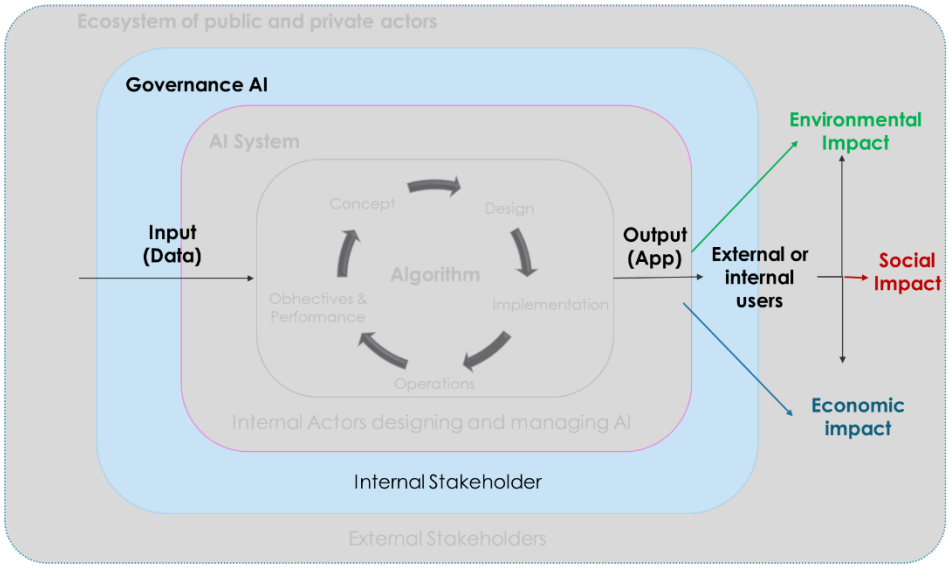

Figure 14.3 – The Governance of AI.

Despite the AI application having its own specificities both due to technical and managerial reasons, it has to be compliant with rules set internally in order to be managed in a homogeneous way compared to previous technologies introduced within the organization. With reference to the case of the present chapter, the AI application had to first of all be compliant with the technical requirements of the IT portfolio, entailing the subsequent easy and fast integration into the existing IT architecture. Likewise, the AI application had to comply with the stringent requirements in terms of privacy (due to the European General Data Protection Regulation, GDPR). These two examples were both managed with the initial cycle of interviews, letting the developers know how to build their application by respecting the internal IT (and AI, by extension) Governance. In the same way, the final document sent to WORK contained all of the details to ensure, communicate, and ultimately explain the full compliance of the latest model. Thus, it emerged that a level of explainability for AI Governance was nonnegligible for the project's success.

The XAI approach at this level was developed differently. Whereas, at the core level, the actors were interested in understanding everything about the application, here the request for information from the AI Governance actors, e.g., WORK's privacy expert, was linked mainly to the team's approach and their need to make sure her/his role was clear, to be a kind of gate before release. When the team tried to apply the same pattern of interaction adopted in the core, there was no openness; on the contrary, they wanted to maintain distance from the design and more specific choices. They wanted to act as in a *super partes* role (i.e., impartial).

Another emblematic example is that of the Innovation Manager whose task is to monitor and give a homogeneous imprint to AI projects. The meeting with him was also fruitful in understanding another perspective in the AI design methodology; he objected to the methodology presented and suggested some changes in the direction of his approach (which he presented to us with slides). Although the two stakeholder engagement methodologies were practically identical, he emphasized the differences in terminology and the desirability of conforming to his scheme.

To summarize XAI at this level, it was less intense and more interactive; in term of "what," the object of the explanation was "approaches, methodological choices and rules compliance."

14.4.4. Wider Ecosystem of actors

The last circle is the wider ecosystem of actors (Fig. 14.4). This is becoming increasingly relevant and needs more attention because it overpasses the boundaries of what companies can control.

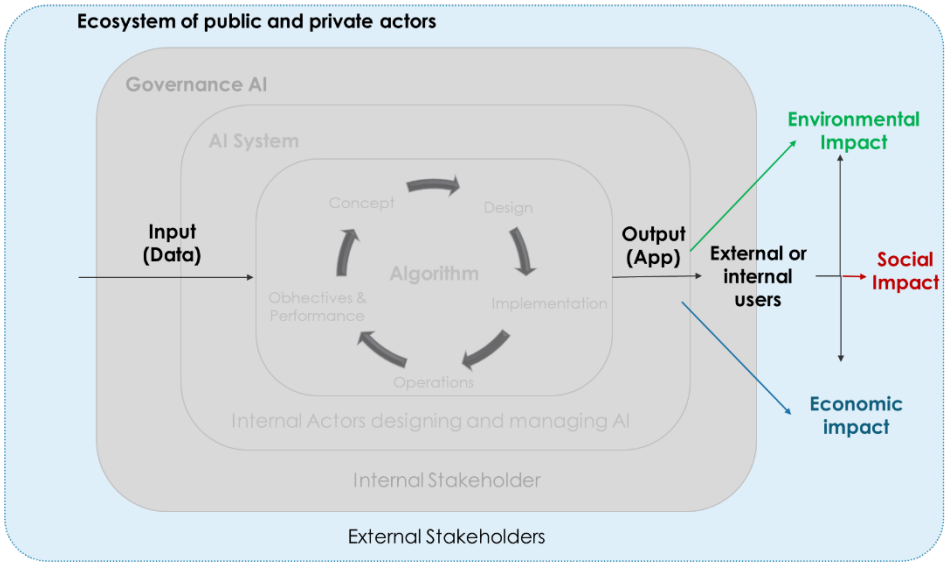

Figure 14.4 – The wider ecosystem.

In this study, with the focus on internal managers and stakeholders, an explanation was not implemented to the external stakeholders, but this circle entered in the XAI approach in any case. The XAI approach has in fact been characterized by outlining the entire lifecycle in the design phase, explaining how solutions could generate opportunities or problems for external stakeholders.

The developers' team, in addressing this part, started by further clarifying that the output of the system is evolutionary because it is based on data that will be continuously updated and integrated. The stakeholders of WORK, very sensitive to reputational and political issues, gave a lot of importance to this part and precisely to the need to understand the mechanisms of an operation, first of all, of AI in general. The designers have therefore provided synthetic and accessible material to an average

trained user. In addition, during the presentation of the solutions, many questions were asked about possible system biases.

In terms of "how," an interaction at this level has been emerging for two reasons. The first is linked to seminal knowledge in one's own areas of expertise, but also to experience, deriving not only from knowledge of processes and technicalities, but also from experience in the field.

In terms of "what," three major areas were touched upon. The first is linked to the end user, who in the case of the application can be considered a direct recipient (i.e., the inspector) or indirect (i.e., the company receiving the inspection). The XAI was used here as a training ground for the parties as a risk analysis. The second area is related to compliance and interpretation of regulations and trustworthiness. The AI regulation in particular offers some margins of freedom in interpretation for medium-low risk applications; but that leaves the question: "Which positioning to choose for WORK"? An XAI for managers also involves touching upon the organization's "risk appetite." Finally, the last area is related to the interface with other public organizations involved in the provision of data, or those from downstream of the preinspection process.

14.5. Conclusion and future development

The results presented in the previous section show that XAI stands together with a multistakeholder perspective. These multiple stakeholders belong to different levels depending on their stake proximity to the single and specific AI application. Thus, they require different levels of explainability which are vary in terms of the frequency of explanations and level of detail.

In this final section we illustrate the key points.

The first key point is the importance for the developer and the developers' teams that plan to use the XAI to know the "who" to address in an organization's real and wider environment realm. Here the reference scheme organized by level of stakeholders (Fig. 14.1) offers an answer, which was enriched by the case we presented. The experimentation highlighted that for this search, the developers need to ask the right question for understanding the responsibility they have now and will have at the release of the AI application.

The explanation has to be more intensive, interactive and open with actors at the core, who have a direct responsibility and involvement in the application. The more we go from the inner to the periphery of the XAI's framework, the "how" is and has to be less interactive and intense. Also, the "what" of the explanation changes; at the core, every detail is shared and needs to be understandable; at the periphery selected issues become relevant in the XAI application.

A second issue of difference was the performance relevant to the different levels of stakeholders. At the central level, top management actors were interested in a broader and more comprehensive

contribution of the AI to WORK's performance, considering both social and economic performance. Managers and officials were also interested in their sustainability when the application was released. The governance actors, on the other hand, were mainly concerned with compliance with external and internal regulations and standards.

Finally, some limitations and future developments are outlined. This study focused on conceptualization and design; the proposed framework could be used to analyze what changes when moving to other phases of an AI lifecycle. A second limitation concerns the type of application, where end-users (inspected companies) are mediated by managers and internal inspectors. How will the approach change?

To conclude, our chapter contributes to an outcome to bring the attention to the importance of creating trust, trustworthiness not only to the final users, but also to internal stakeholders, especially when AI is nested with human responsibility.

References

Barredo Arrieta, A., Díaz-Rodríguez, N., del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2020. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012.

Belle, V., Papantonis, I., 2021. Principles and practice of explainable machine learning. Frontiers in Big Data, 4. https://doi.org/10.3389/fdata.2021.688969.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M.F., Eckersley, P., 2020. Explainable machine learning in deployment. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 648-657. https://doi.org/10.1145/3351095.3375624.

Busuioc, M., 2021. Accountable artificial intelligence: holding algorithms to account. Public Administration Review 81 (5), 825-836. https://doi.org/10.1111/puar.13293.

Colucci, S., Donini, F.M., di Sciascio, E., 2023. On the relevance of explanation for RDF resources similarity. In: Model-Driven Organizational and Business Agility. MOBA 2023, pp. 96-107. https://doi.org/10.1007/978-3-031-45010-5_8.

Domingos, P., 2012. A few useful things to know about machine learning. Communications of the ACM 55 (10), 78-87. https://doi.org/10.1145/2347736.2347755.

Fosso Wamba, S., Bawack, R.E., Guthrie, C., Queiroz, M.M., Carillo, K.D.A., 2021. Are we preparing for a good AI society? A bibliometric review and research agenda. Technological Forecasting & Social Change 164, 120482. https://doi.org/10.1016/j.techfore.2020.120482.

Grimmelikhuijsen, S., 2023. Explaining why the computer says no: algorithmic transparency affects the perceived trustworthiness of automated decision-making. Public Administration Review 83 (2), 241-262. https://doi.org/10.1111/puar.13483.

Haque, A.B., Islam, A.K.M.N., Mikalef, P., 2023. Explainable artificial intelligence (XAI) from a user perspective: a synthesis of prior literature and problematizing avenues for future research. Technological Forecasting & Social Change 186, 122120. https://doi.org/10.1016/j.techfore.2022.122120.

Hartmann, K., Wenzelburger, G., 2021. Uncertainty, risk and the use of algorithms in policy decisions: a case study on criminal justice in the USA. Policy Sciences 54 (2), 269-287. https://doi.org/10.1007/s11077-020-09414-y.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., Hussain, A., 2024. Interpreting black-box models: a review on explainable artificial intelligence. Cognitive Computation 16 (1), 45-74. https://doi.org/10.1007/s12559-023-10179-8.

HLEG, 2019a. Ethics guidelines for trustworthy AI. https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf.

HLEG, 2019b. Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Kim, B., Park, J., Suh, J., 2020. Transparency and accountability in AI decision support: explaining and visualizing convolutional neural networks for text information. Decision Support Systems 134, 113302. https://doi.org/10.1016/j.dss.2020.113302.

Lee, I., Shin, Y.J., 2020. Machine learning for enterprises: applications, algorithm selection, and challenges. Business Horizons 63 (2), 157-170. https://doi.org/10.1016/j.bushor.2019.10.005.

Meske, C., Bunde, E., Schneider, J., Gersch, M., 2022. Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. Information Systems Management 39 (1), 53-63. https://doi.org/10.1080/10580530.2020.1849465.

Pasquale, F., 2015. The Black Box Society. Harvard University Press. https://doi.org/10.4159/harvard.9780674736061.

Quattrone, P., 2016. Management accounting goes digital: will the move make it wiser? Management Accounting Research 31, 118-122.

Saeed, W., Omlin, C., 2023. Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems 263, 110273. https://doi.org/10.1016/j.knosys.2023.110273.

Sun, T.Q., Medaglia, R., 2019. Mapping the challenges of artificial intelligence in the public sector: evidence from public healthcare. Government Information Quarterly 36 (2), 368-383. https://doi.org/10.1016/j.giq.2018.09.008.

Swartout, W., Paris, C., Moore, J., 1991. Explanations in knowledge systems: design for explainable expert systems. IEEE Expert 6 (3), 58-64. https://doi.org/10.1109/64.87686.

Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., Liu, X., Wu, Y., Dong, F., Qiu, C.-W., Qiu, J., Hua, K., Su, W., Wu, J., Xu, H., Han, Y., Fu, C., Yin, Z., Liu, M., et al., 2021. Artificial intelligence: a powerful paradigm for scientific research. The Innovation 2 (4), 100179. https://doi.org/10.1016/j.xinn.2021.100179.