Putting Accountability of AI Systems into Practice

Beatriz San Miguel, **Aisha Naseer** and **Hiroya Inakoshi**
Fujitsu Laboratories of Europe, United Kingdom
{beatriz.sanmiguelgonzalez, aisha.naseer, hiroya.inakoshi}@uk.fujitsu.com

Abstract

To improve and ensure trustworthiness and ethics on Artificial Intelligence (AI) systems, several initiatives around the globe are producing principles and recommendations, which are providing to be difficult to translate into technical solutions. A common trait among ethical AI requirements is accountability that aims at ensuring responsibility, auditability, and reduction of negative impact of AI systems. To put accountability into practice, this paper presents the Global-view Accountability Framework (GAF) that considers auditability and redress of conflicting information arising from a context with two or more AI systems which can produce a negative impact. A technical implementation of the framework for automotive and motor insurance is demonstrated, where the focus is on preventing and reporting harm rendered by autonomous vehicles.

1 Introduction

This paper focuses on one of the major requirements of Ethics on Artificial Intelligence (AI): accountability [HLEG, 2019; AI4People, 2019]. In AI domain, accountability is generally associated with:

- demonstrate outcomes, for example, thought explanations [Adadi and Berrada, 2018; Arrieta *et al.*, 2019]
- ensure responsibility of AI systems [Rieke *et al.*, 2018]
- produce fair decisions or redress unfair outcomes [Zemel *et al.*, 2013; Adebayo, 2016; Chouldechova and Roth, 2018]

Taking the definition given by the European Commission's High-Level Expert Group on Artificial Intelligence [HLEG, 2019], accountability includes auditability, minimization and reporting of negative impacts, trade-offs, and redress.

With the aim of putting the previous definition into practice, we propose the Global-view Accountability Framework (GAF) that analysis data received from different sources (auditability), identifies discrepancies of the data received, and tries to redress them in order to minimise potential negative impact. Moreover, the GAF provides several reports and information that could be used by third parties with different proposes (i.e., investigate an incident, look for liability, automate tasks, etc.).

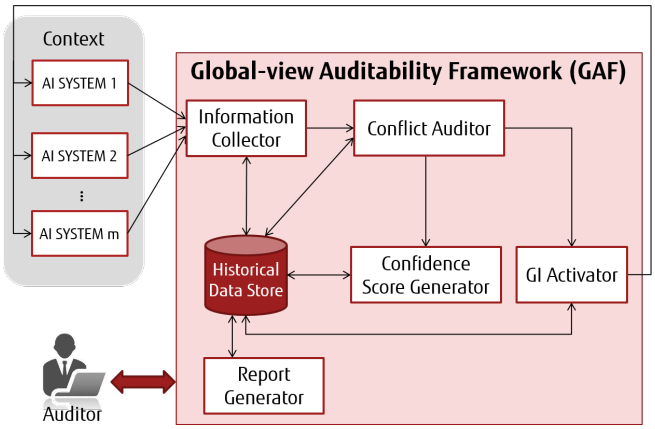

Figure 1: GAF Architecture

The GAF can be implemented for different fields and use cases where two or more AI systems participate. A few examples of its application could be: automotive, motor insurance, manufacturing, and home automation. In this demonstration, we propose the GAF's specialization for automotive (to avoid car accidents) and motor insurance (to make more efficient the claims assessment process).

2 GAF Architecture

Figure 1 depicts the general GAF architecture, highlighting its main components and its interactions with external elements: *m* AI systems in a context and an Auditor.

AI systems are smart elements equipped with different sensors and AI components that carry out autonomous decisions. These can send data about its perceptions (outputs of its AI components) to the GAF and receive orders to try to avoid negative impact(s) in the context.

The GAF consists of six components:

- **Information Collector:** receives information from a set of AI systems, processes the information (i.e., normalization and standardization), and resends it to the Conflict Auditor component.
- **Conflict Auditor:** identifies discrepancies among the

Figure 2: Scene for the demonstration

data received and, if discrepancies are detected, calculates a Global Information (GI). This GI tries to rectify and avoid incorrect perceptions of AI systems, taking into account the reputation and confidence of systems.

- **Confidence Score Generator:** calculates a confidence score for each perception of each AI system using a multi-method approach that considers historical data, majority criterion, closest neighbours, etc.
- **GI Activator:** uses the GI calculated to send technical orders to specific AI systems that have wrong perceptions and thus, try to avoid negative impact(s).
- **Historical Data Store:** stores all data received, processed and sent by the GAF.
- **Report Generator:** allows third parties to retrieve relevant data that can be used to monitor and repair AI systems or to look for liability when negative impact occurs.

3 Demonstration

We demonstrate the GAF in the fields of automotive (to increase visibility of vehicles and avoid car accidents) and motor insurance (to make more efficient the claims assessment process in case of an incident). In this context, referring to Figure 1, AI systems are vehicles that detect objects, its positions, its nature (or classification in a set of classes such as "car", "pedestrian", etc.), among others. These can send and receive data to and from the GAF. Moreover, the Auditor is a human considering to make a decision on the insurance claims and needing information related to the vehicles in a specific moment.

In particular, we show a scene where a set of vehicles are traveling, and an accident happens in a road junction around a pedestrian who wants to cross a main road. Due to the extended capability of the GAF, we observe how the vehicles increase their visibility and perceptions using the knowledge of others.

For example, in the scene (Figure 2), there are 5 vehicles (4 cars and 1 van) and a pedestrian (`person_1`). The lines represent the field of view for vehicles, and each square is the perception of the pedestrian (`person_1`) for a specific vehicle. Here, in practice, the car called `car_2` (green car) can not visualize the pedestrian (`person_1`) because he/she is behind `car_1` (yellow car) and `car_4` (purple car). However, the `car_2` receives data from the GAF with the perception of the `person_1` (green square) and, therefore, `car_2` is aware of there is a pedestrian behind the cars `car_1` and `car_4`.

As readers will have realized, there are some perceptions for the `person_1` (squares in Figure 2) that are not well-located. This could be because some vehicles have wrong perceptions (`car_1` and `van_1`) or the perception received from the GAF is incorrect (`car_2`). These wrong perceptions could cause a negative impact such as a traffic accident (as the one shown in the demonstration).

Specifically, if the position of the pedestrian is referred as (x,y), there is a moment before the traffic accident where:

- `van_1` detects the `person_1` in position (650, 400), which is a wrong perception for the position of the pedestrian.
- `car_1` detects the `person_1` in position (400, 400), which is a wrong perception for the position of the pedestrian.
- `car_3` detects the `person_1` in position (600, 400), which is the actual position of the pedestrian.
- The GAF, and in particular the Conflict Auditor component, detects discrepancies in the feature that represent the coordinate x of the pedestrian and therefore, this calculates a GI for this feature.

  For the GI calculation, a method inspired by the barycentric coordinates is designed and applied, following the general equation:

$$\frac{\sum_{k=1}^{m} cs_{\mathrm{k,1}} * value_{\mathrm{k,1}}}{m - \sum_{k=1}^{m}(1 - cs_{\mathrm{k,1}} * value_{\mathrm{k,1}}, where cs_{\mathrm{k,1}} != 1)}$$

  Where:

  - *m is the number of AI systems that have information about the feature 1*
  - *$cs_{k,1}$ is the confidence score of the AI system identified by k for the feature 1*
  - *$value_{k,1}$ is the value of the AI system k for the feature 1*

  To apply this equation to the feature x of the pedestrian's position in the moment described, we have to consider 3 vehicles (*m* in the equation), and the values of their perceptions (650, 400 and 600). Moreover, the GAF relies on confidence scores of AI systems to calculate the GI ($cs_{k,1}$). All systems start from having the maximum confidence score, and this is updated whenever auditability and redress processes are carried out. For simplicity in the description, we assume that there is no historical data about the vehicles, and thus, they have the maximum confidence score associated (value 1).

  Considering the above, the GI for the feature x is the arithmetic mean of all the perceptions, yielding the next equation:

$$GI_{\mathrm{x}} = \frac{650 + 400 + 600}{3} = 550$$

  Thus, the GI for the feature x (GI_x) is 550, and all the vehicles should perceive the position of the pedestrian as (550, 400). The GI Activator component sends this

GI to the vehicles with different perceptions or without perceptions (`car_2`) of the pedestrian to try to avoid a negative impact(s). However, some situations could be unavoidable like the scenario presented in the demonstration.

- After the processes of audit and redress of discrepancies, a new confidence score for the feature x of each vehicle is calculated. For the example, the confidence scores of `van_1` and `car_1` will be decreased, and on the contrary, the confidence score of `car_3`, which has a perception close to the GI, will be slight increased or maintained.

  The confidence score is calculated using a multi-method approach that is based on historical data (to reward or penalize previous perceptions), majority vote (to take into account what is perceived by other systems), and closest neighbour (where closer values have more relevance than more distant values).

It is important to highlight that the aforementioned steps describe the GAF operation in an initial stage. In this situation, there are no historical data and the confidence score of all vehicles is the same. This can provoke failures in the calculation of the GI (as shown, the actual position is 600 and not 550). However, the confidence scores are updated frequently and after several executions they will reflect the reputation and trustworthiness of vehicles to obtain an optimal GI.

Lastly, if a car accident occurs, motor insurance companies would be in charge of investigating who should be held responsible. This process is labour-intensive, time-consuming, and usually involves the need of many data sources such as customer claims, car repair bills, police reports, etc. The GAF allows motor insurance companies to obtain different reports to analyze what was doing and perceiving each car at any time. It is important to highlight that the GAF acts as a Decision Support System for domain experts that facilitates the assessment process, assisting them, but does not make final decisions related to who is/are responsible.

In conclusion, the GAF provides the following functionalities for the demonstration considered:

- Traceability of all AI systems presented in a context. This includes registering and logging all data detected and perceived by each vehicle (other systems could be considered such as cameras, traffic lights, etc.).
- Auditability of the data perceived by all AI systems to detect discrepancies and if any is found, try to redress it, calculating a GI that will try to avoid a negative impact such as a potential car accident. Moreover, this GI could be sent to systems that do not perceived specific features and thus, increase the visibility of these systems.
- Calculation of confidence scores based on historical data, majority vote, closest neighbours, etc.
- Generation of a report with the previous information that could be used by insurance companies to analyze a specific scene and take a decision about who was responsible of a car accident.
- Generation of user accident statements that are automatically created and sent to the insurance companies (with the previous user authorization).

4 Conclusion

The GAF presented is a first attempt to put into practice the definition of accountability given by the European Commission [HLEG, 2019]. According to it, accountability includes auditability, minimization and reporting of negative impacts, trade-offs, and redress.

Currently, the framework is being tested and validated in a simulated environment that allows the design of multiple scenarios with vehicles and pedestrians that can travel and move in different directions and with chosen speeds.

We focus the demonstration to address discrepancies on the positions perceived of specific objects (i.e. pedestrians), but other features could be equally implemented. In the future, we plan to trial the GAF in real-world scenarios for the purposes of evaluation.

References

[Adadi and Berrada, 2018] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[Adebayo, 2016] Julius Adebayo. *FairML: ToolBox for diagnosing bias in predictive modeling*. PhD thesis, Massachusetts Institute of Technology, 2016.

[AI4People, 2019] AI4People. AI4People's ethical framework for a good AI society: Opportunities, risks, principles, and recommendations, 2019.

[Arrieta *et al.*, 2019] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 2019.

[Chouldechova and Roth, 2018] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint cs.LG/1810.08810*, 2018.

[HLEG, 2019] HLEG. European Commission's High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI, 2019.

[Rieke *et al.*, 2018] Aaron Rieke, Miranda Bogen, and David G Robinson. Public scrutiny of automated decisions: Early lessons and emerging methods, 2018.

[Zemel *et al.*, 2013] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.