# Empirical Research on Ensuring Ethical AI in Fraud Detection of Insurance Claims: A Field Study of Dutch Insurers

Martin van den Berg[1](✉), Julie Gerlings[2], and Jenia Kim[1]

[1] HU University of Applied Sciences Utrecht, Heidelberglaan 15, 3584 CS Utrecht, The Netherlands
martin.m.vandenberg@hu.nl

[2] Copenhagen Business School, Howitzvej 60, 2000 Frederiksberg, Denmark

**Abstract.** The insurance industry in the Netherlands applies artificial intelligence (AI) in different processes and acknowledges that AI should be implemented in an ethical and responsible manner. Therefore, the Dutch Association of Insurers supported the industry by publishing an ethical framework. However, the framework is a set of high-level requirements, and the question is how these requirements are translated into local practices. Our research question is how ethical requirements are applied by insurance companies when using AI systems to detect fraud in insurance claims. To answer this question, we conducted interviews with representatives of four different organizations. The study demonstrates the awareness amongst interviewees that AI needs to be applied in a responsible way. The ethical framework provides a good starting point for insurers to develop their own practical ethical guidelines. Empirical evidence confirms that accountability, safety, transparency, non-discrimination, and human agency are priorities in the process of AI implementation. The research shows that translation of the ethical framework into operational and actionable instructions is done in-house by each organization and requires a multidisciplinary approach and cooperation between teams.

**Keywords:** Ethical AI · Responsible AI · Insurance · Fraud detection

## 1 Introduction

The insurance industry applies artificial intelligence (AI) in different processes and acknowledges that AI must be applied in an ethical and responsible manner [1]. Therefore, the Dutch Association of Insurers published an ethical framework which is binding for its members [2]. However, the framework is a set of high-level requirements, and the question is how these requirements are applied in practice. An IBM report indicates a “disparity between intent and implementation of AI ethics” [3]. The World Economic Forum calls this the “intention-action” gap [4]. Our research question is how ethical requirements are applied by insurance companies when using AI systems to detect fraud in insurance claims. Our research indicates that insurance firms are aware of the risks, limitations, and challenges of applying AI and have ethical frameworks in place to

S. Nowaczyk et al. (Eds.): ECAI 2023 Workshops, CCIS 1948, pp. 106–114, 2024.
https://doi.org/10.1007/978-3-031-50485-3_9

mitigate these risks. They found ways to narrow the intention-action gap. The main contribution of this research is that it provides practitioners and researchers with insights on how to implement ethical AI. This paper is organized as follows: Section 2 provides a short overview of the process of fraud detection. In Sect. 3 the research method is discussed and in Sect. 4 the results. Finally, Sect. 5 contains the discussion, conclusion, limitations, and opportunities for future research.

## 2 Process of Fraud Detection

Fraud detection of insurance claims is the process of determining the risk an insurance claim is fraudulent and results in lower premiums for honest consumers [1]. The process of fraud detection has the following steps:

- A private policy holder submits a claim to the insurance firm where s/he has a policy.
- The insurance firm processes the claim in its systems. Part of this processing is to check the claim for suspicious or anomalous information that may indicate fraud. This involves checking if the claim has been submitted elsewhere and checking the claimant's history of insurance fraud.
- The claim is either automatically approved, or manually checked by a claim handler. Some insurance firms have a partly automated system to evaluate whether the claim should go to a claim handler or to direct pay-out. Some systems are based on business rules, others are a combination of business rules and AI.
- If a claim handler finds the claim of a certain level of risk or something that is out of context, s/he transfers the claim to a fraud investigator who investigates the claim in more detail. Fraud investigators operate and decide independently whether the claim is fraudulent or not.
- In the end, claims that have been found fraudulent are disapproved by the insurance company and can be reported to an external warning system. The insurers can report their fraud investigations and incidents to the Dutch Association of Insurers. This association also provides guidance in the form of frameworks and best practices, as well as fraud trend-analysis and alerts on modus operandi.

## 3 Research Method

The research has been conducted in a qualitative and explorative manner during the first half of 2023. To gain a practical understanding of status and challenges in applying ethical AI, we conducted five interviews with experts in the field from four different organizations in the Netherlands (see Table 1). The interviews were recorded and transcribed. Interviews were conducted by two researchers and lasted about one hour each. The transcripts were coded through axial coding [5] and analyzed with NVivo.

**Table 1.** List of experts.

| | Function | Organization | Year of experience in insurance |
|---|---|---|---|
| E1 | Manager Centre Against Financial Crime | A | 15 |
| E2 | Chief Analytics Officer | B | 19 |
| E3 | Head of Anti-Fraud | B | 14 |
| E4 | Ethicist | C | 7 |
| E5 | Actuary | D | 18 |

This study has limitations. First, the results are based on only five interviews with representatives of organizations in the Netherlands. Second, interviewees may be biased on their perception of the firm's practices. And lastly, the interviewees all belong to the managerial levels in their organizations and might not fully represent the challenges encountered by the employees who interact with the AI systems in practice (such as developers and end-users).

## 4 Results

An articulated and deep understanding of the ethical challenges in the process of fraud detection in general was seen across all interviewees. The leading guideline used by the companies is the ethical framework of the Dutch Association of Insurers which is binding for the association's members. This guideline is inspired by national and EU laws and regulations, with a more rigorous approach at times.

> *"In the ethical framework it says even if the National law or the European law allows something, and the ethical framework of the Insurance Association says no, we do not do that. We ask our members to follow the rules of the ethical framework, so we narrow our own boundaries, even if there's more possibilities within the (European) law." (E1)*
>
> *"We have a strong ethical framework, which is a nine-page legal document which explains to what type of things a model should adhere to. These consist of the seven principles of trustworthy AI from the high-level expert group of the EU that published this paper." (E2)*

Ethical guidelines have been incorporated in different ways at the firms interviewed. According to the interviewees, the incorporation of the framework is thorough and well thought about. The interviewed companies have typically started out with workshops to create awareness about the framework and guidelines in general.

> *"I've done some ethical workshops with our fraud department. And as we were implementing the ethical framework internally, we've looked at the fraud detection process within [Company] to see if there's any risks involved that touch upon points from the ethical framework." (E4)*

However, awareness is not sufficient when it comes to building responsible AI. The data scientists who work on developing and iteratively testing the model need practical instructions that translate the ethical principles into actionable tasks.

> *"…but if you're a data scientist, you want to have something much more practical. So, we created an AI assessment that covers all the seven principles in the ethical framework, but in a questionnaire type of way. It asks you what type of data you are going to use. Does it contain [personal identifiable information]? And if so, is your data protection officer involved? And did he or she check the baseline for data processing?" (E2)*

Moreover, it is not a one-time assessment; every iteration of the model demands a review of the data used and a possible update of the checklist.

> *"…the assessment starts and ends basically never because once it is in production, you also need to come back to the assessment every six months or every year, depending on the type of use case that you're doing, and you need to update this document." (E2)*

The ethical framework is based on the 'Ethics guidelines for trustworthy AI' [6] which contains seven principles that AI systems should meet to be deemed trustworthy. These principles were mentioned multiple times during the interviews, with special focus on accountability, safety, transparency, non-discrimination, and human agency.

## 4.1 Accountability and Safety

The interviewees indicated that they prefer developing their AI solutions in-house, to have full control and full accountability. They stress the importance of ensuring that the model is robust and safe and continually testing to see if it needs to be updated or retrained.

> *"…And the main reason for that [developing in-house] was to be in control yourself. To ensure we comply with our ethical framework, law, and legislation." (E3)*
>
> *"Everything is being tested repeatedly… sometimes we retrain the model based on the outcome of tests and, we did some shadow runs. We run the model for quite some time to do it in parallel but not in production and see what the performance would be." (E2)*

## 4.2 Transparency

Transparency towards internal stakeholders is regarded as very important by the interviewees, mainly as a means to gain employees' trust and acceptance and improve their understanding of the AI system. One such internal stakeholder is the managerial level, for whom it is important to understand how the model works, as they need to sign off on it and therefore are accountable for it. This need for transparency and explainability often drives the preference towards less complex, but more explainable, models.

> *"When I look at the senior managers and directors, they also want to understand. They tend very much towards the less complex models for the time being. Maybe in time it will change. Yes, but for the time being when I look at it and I see how the people at the top think, I think they are quite careful…" (E5)*

Other important stakeholders are the internal end-users of the model, i.e., the claim handlers and the fraud investigators. Since they need to work with the outputs of the model, they need to understand what these outputs are. In addition, they need to be prepared to answer questions about these outputs from the customer, if such questions arise.

> *"Before we started this, we assessed all risks. And there's one risk we described. We must be clear about what the outcome of this model means. We must be clear that the claim handler must understand, but also the fraud investigator must understand how this model works and what they are seeing." (E3)*
>
> *"…you can explain the model well, but it is sometimes too in-depth for the claim handlers. That's why I came up with competence, to be able to understand such a model properly. For the current colleagues who work in claims, they have learned things in a different way… But because they do not yet have that competence, they need to get to know those AI models well, but they also need to know how the score is arrived at… if the claim handler does not understand why he is asking for certain information and the customer asks, why are you asking this? Yes, then it will be difficult. So, you need some kind of further training… The customer wants a good explanation." (E5)*

There is, however, a sensitive aspect to the transparency principle, which has to do with how much information can and should be disclosed to the customer. On the one hand, the companies have a moral (and sometimes legal) obligation to disclose the use of an algorithm in their fraud detection process and to explain what the algorithm does. On the other hand, full transparency about proprietary in-house algorithms is problematic in terms of competition between firms, and it also creates a risk of gaming the system.

> *"They must inform clients when they are processing their personal data. But that will not mean you have to tell them all the details of what you're doing in your process. But you must explain why something is taking up a little bit more time before they get a decision on their claim, for instance. But it's always difficult." (E1)*

An additional tension is found between the pros and cons of providing detailed explanations about the outputs of the model to internal users. Some companies provide the claim handler with the risk score outputted by the model, as well as a detailed explanation in natural language about the features that contributed to this score. The advantage of this approach is that it gives the claim handler an indication on what is suspicious in the claim and where s/he should look first.

> *"Very important thing we built in. So, the model, of course, gives a score. But to the person who receives the claim, there's an explanation. You received this*

> *claim to be handled manually because XYZ and then it gives the explanation in human language…For instance, a highly unusual price for a claim like this or a combination of certain factors. This same claim amount has been issued before, or an email address or this bank account was used in a similar claim before, but with another policyholder… So, there are different rules in the claim process." (E2)*

However, this level of explainability also has some potential disadvantages, as it might create a bias or a tunnel vision of the handler. Therefore, some of the interviewed companies chose not to provide detailed explanations; instead, they order the cases by levels of risk, so that the most suspicious cases are handled first, but they expect the handlers and investigators to do the investigation "from scratch" to avoid potential bias by the model.

> *"So, it might give a score to a certain case and that case might be prioritized. And then the human comes in and starts to do their own research……we talked about in our explainable AI workgroup, how important it is for the human not to just see all the factors that the AI has determined as fraudulent because that might already bias them in a certain direction. It might already color their judgement." (E4)*

The level of explainability in models such as random forest or boosting (XGBoost) may seem simple on a general level, however reasoning through the decision from a single claim evaluation can be very difficult. Therefore, firms have introduced SHAP and LIME as explainable components in their model framework. These explainable frameworks can assimilate an instance (case) and show which features are most likely to have the highest impact on the evaluation.

> *"We use a relatively easy simple machine learning algorithm where you can get quite good results with SHAP or LIME with it." (E2)*

In combination with simpler models that do not involve deep learning, firms overcome the challenge of extracting information about the reasoning of the ML models choices. Now, the challenge is to ensure understanding from the stakeholders who need the information.

> *"I talked with a colleague who also worked with these models, and he said yes, you can explain the model well, but it is sometimes too in-depth for the claim handler…" (E5)*

Though SHAP and LIME plots have been extensively promoted as explainable and interpretable, they still cause confusion to many stakeholders outside the data science domain since they are not contextual to the people who receive them. Moreover, claim handlers and fraud investigators tend to be analytical people who seek information until they understand in detail what is going on. Therefore, the plots can be too detailed, or may show the wrong context, to be useful for these stakeholders. One firm has generated indicators based on the plots, which are formulated in natural language to overcome this challenge.

> *"So, the claim handler sees on his screen the claim. Based on our model the claim gets a risk score of High, Medium, or Low. Our model will also add a simple explanation in three to five lines. So not just red, orange, green or a difficult explanation or code, but explanations like: 'watch this invoice or look at this address, it's known in another case. See claim number x.' So, the data scientists must make a translation from the code to send it to the claim handler to make it clear for them how to interpret this risk." (E3)*

### 4.3 Non-discrimination

The interviewees indicated that they prioritize the clients and their experience rather than solely focusing on detection of more fraud. Using ML to identify suspicious claims and ending up wrongly accusing someone of fraud can have tremendous consequences for the individual. Moreover, it can tear the image of a company and the entire industry down. Therefore, firms have high standards for what data is being fed into the models to minimize the risk of discrimination or bias towards specific groups. For example:

> *"For the detection of fraud, area codes are a no go. You cannot create any fallout of your straight through process just on an area code. I do know that you can use it for risk management, for risk evaluations. And for instance, my car insurance premium is a bit lower than two zip codes to my left. But that's a risk assessment issue and not a fraud assessment issue." (E1)*

One example of the complexity of practically implementing ethical guidelines is how to eliminate discriminatory features from the data going into the model. The basics of supervised ML start with learning from historic data and build upon that to establish a probability of a new claim falling into one of the categories. According to the interviewees, the features going into the AI-model are carefully chosen to minimize the risks of discrimination and biases. Therefore, some features, such as 'country of origin' or 'nationality' might be excluded or altered before they go into the model. However, some features are less easily identified as problematic, as they do not seem discriminatory by themselves, but they do serve as a proxy for a discriminatory feature.

> *"We're putting a lot of effort into bias detection. We created some tools ourselves to detect whether there is a statistical bias for the model to affect certain people who are vulnerable. So, either based on a religion, sexual orientation, a social class… there are 25 attributes that are prohibited to use because they are discriminatory. These are clear for everyone. The true harm is in the proxies of those 25 attributes. So, we are now in a late phase of deploying also this bias detector based on features that might be a proxy to discriminatory features." (E2)*

In the example presented by the interviewee, it turned out that even though 'country of origin' was excluded from the data, there was a proxy feature for this information hidden in the 'marital status' feature, since one of its values was 'Married outside of the Netherlands' (a proxy for a foreign country of origin). This was discovered by a dedicated bias detection tool built by the company.

> *"If you are married, you both take a mortgage. It has some impact on the product. So, you are allowed to ask that: married? Yes or No. But in this case the bias detector discovered that there was a strong proxy to this marital status attribute. And it was just because we had different categories in this attribute. It could be Yes, it could be No, but it could also be Yes, married outside of the Netherlands, which was a different category, which was not being used by us deliberately in the model. But marital status was part of the model. And now, potentially this could be a proxy for ethnical background... So, we did a recode of this attribute to simple Yes or No." (E2)*

### 4.4 Human Agency

Human oversight is another crucial aspect of implementing ML in the fraud detection process. The model is used only to assess the risk and output a score; the rest of the process, which includes the investigation, and the final decision is always performed by a human expert. This is also expressed in how the model is being named and talked about, and it is part of ensuring the intended use of the model. As the quote below shows, there is a deliberate distinction between 'fraud detection' and 'fraud risk', which emphasizes that it is the investigator, and not the model, who detects fraud.

> *"We call it a fraud risk model because the model itself doesn't detect fraud. It's always the human who must assess this risk and must decide if it is a possible fraud or not. An important thing in the development of our tool was a human in the loop. So first, the system presents to the claim handler, these are the fraud risks identified. Then the claim handler must look at it and must assess these risks. He might ask some questions to the client or ask for additional information ...and then he says, well, I don't trust this claim to be valid. Maybe it's fraud. Then it goes to the fraud investigator. And then he also looks at it. Are there enough indicators for fraud? If so, okay, we take over this claim and start a fraud investigation. The investigation has to point out if it is possible fraud or not. So, it's a human who always makes the decision." (E3)*

## 5 Discussion and Conclusion

Our research provides an up-to-date overview of the practical use of AI in fraud detection of insurance claims in the Netherlands. Based on the five interviews we conducted, we conclude that:

- Interviewees acknowledge the limitations of the AI and determine its place in the whole process accordingly, so that the cooperation between the model and the human experts is optimal.
- The implementation of AI is taken seriously: it is a long process, and a lot of effort is put not only in the technical aspects but also in the human and organizational aspects.
- There is a lot of awareness among interviewees of the ethical principles that need to be met to implement AI responsibly. The Dutch Association of Insurers provides an ethical framework. Translation of the ethical framework into operational and actionable instructions is done in-house by each company.

- Compared to extant literature where the intention-action gap was described [e.g., 7, 8], this study indicates that the insurance industry in the Netherlands is actively and seriously working on ways to narrow the gap and to implement ethical AI in practice.

The main takeaway from this research is that the implementation of AI in fraud detection is a business transformation that requires many ethical and organizational considerations. Education and inclusion are crucial to ensure a successful integration of AI into the fraud detection process, and an optimal human-machine cooperation. All interviewees are aware of the risks, limitations, and challenges of applying AI and insurance firms have ethical frameworks in place to mitigate these risks. This research sheds light on the way insurance firms are implementing ethical AI and how they use ethical frameworks.

Further, and more detailed research is necessary to identify which factors, such as education, contribute most to a successful implementation of ethical AI and in what manner. Moreover, research is needed to learn how certain tools, such as bias detection tools, can help narrow the intention-action gap.

## References

1. EIOPA (European Insurance and Occupational Pensions Authority). AI Governance Principles towards ethical and trustworthy AI in the European insurance sector. https://www.eiopa.europa.eu/eiopa-publishes-report-artificial-intelligence-governance-principles-2021-06-17_en. Accessed 26 Aug 2023
2. Verbond van Verzekeraars. Ethisch kader. https://www.verzekeraars.nl/branche/zelfreguleringsoverzicht-digiwijzer/ethisch-kader-datatoepassingen. Accessed 26 Aug 2023
3. Goehring, B., Rossi, F., Rudden, B.: AI ethics in action. An enterprise guide to progressing trustworthy AI. IBM Institute for Business Value (2022). https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/ai-ethics-in-action. Accessed 26 Aug 2023
4. Guszcza, J., Skeet, A.: How businesses can create an ethical culture in the age of tech. World Economic Forum (2020). https://www.weforum.org/agenda/2020/01/how-businesses-can-create-an-ethical-culture-in-the-age-of-tech/. Accessed 26 Aug 2023
5. Williams, M., Moser, T.: The art of coding and thematic exploration in qualitative research. Int. Manage. Rev. **15**(1), 45–55 (2019)
6. European commission: Ethics guidelines for trustworthy AI (2016). https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Accessed 26 Aug 2023
7. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci. Eng. Ethics **26**(4), 2141–2168 (2020)
8. Georgieva, I., Lazo, C., Timan, T., Van Veenstra, A.F.: From AI ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience. AI Ethics **2**(4), 697–711 (2022)