Opening Statement

Interviewer:
Good morning. First of all, I would like to thank you for being willing to participate in the interview. My name is Imane, I am currently doing my master's degree at the University of Amsterdam in Information Studies. I came to you through [Name]. This interview is part of my thesis research at [Company Name] The research is about how artificial intelligence can be effectively integrated into cybersecurity risk management, especially at financial institutions. The interview takes about 60 minutes, but I think it's more like 30-45 minutes. Everything is processed anonymously and only a transcript is created, which is then deleted.

The interview has two parts. First is a short introduction of about 5 minutes and then I go to the key questions.

Respondent:
Great that sounds good, also nice that you had sent the briefing.

Section 1: Background of the interviewee

Interviewer:
Q1.1: Can you briefly tell us about yourself and your role within the organization?

Respondent:

My name is [Name]. I studied in Tilburg and at Erasmus University, with a focus on financial business, economics and ICT. Then I moved to Utrecht where I now live with my girlfriend and two children. And in Utrecht I completed my RE training. Then I worked at [Name of company] for about six years. After that period I made the step to entrepreneurship. I now lead a team of ten people that focuses on red teaming, pentesting, social engineering and related security testing. We put ourselves in the shoes of the attacker, so to speak, to better protect organizations. Our development team also creates -solutions, where we use AI as an input source for intelligence and findings, and the model itself does not run on AI, but we use it to support our analysis.

Interviewer:
Q1.2: How much experience do you have with cybersecurity, risk management or AI?

Respondent:

Yes, I have quite a bit of cybersecurity experience by now. I've been in the business for about 18 years, starting when I worked at [Name of Company] and later at [Name of Company]. Since then, it has always been a part of my job. As for AI, my experience is a bit more recent, which also makes sense given the development of the technology.
But by now we do a lot with it , including tools like ChatGPT and Claude, which we compare among ourselves. We even have our own LLM running here as a test environment. We are also involved in a WBSO project, subsidized through the RVO, through which we further explore and develop AI within our organization.

Interviewer:
Q1.3: To what extent are you familiar with LLMs or other AI technologies in your work context? Respondent:

My experience with AI is relatively recent, which makes sense given the still young nature of the technology. In our organization, we now actively work with tools such as Claude and ChatGPT . We have also set up our own Large Language Model as a test environment, which we are experimenting with. This project is partly supported through a WBSO subsidy from the RVO. Although I would not call myself an AI expert, I am certainly familiar with LLMs and see how they can be applied within our work, for example in generating insights and supporting red teaming activities.

Interviewer:
Okay great so you do know what the Large Language Models are. That's nice because that's actually the AI technology that I use.

Those were the introduction questions. The next questions are more substantive and about the framework.

Section 2: Thematic depth

Interviewer:
Q2.1: In what areas within cybersecurity risk management do you see the most potential for LLMs to add value?

Respondent:

I see that LLMs in particular can add value in three areas. First, they are good at establishing relationships between concepts and tokens; they can recognize connections within complex clusters of information. That makes them powerful in analyzing risks and linking those risks to other control frameworks or existing data sources. Second, they can help generate predictions, for example, by suggesting additional risks based on previously observed patterns. And finally, LLMs are strong in text creation. That makes them suitable for automatically creating reports, explanations or recommendations within risk management processes.For me, the real added value is mainly in content support, pattern recognition and text generation.

Interviewer:
You already mentioned two types of tools in the introduction, namely Claude and ChatGPT .

Q2.1a: Are there any other types of LLMs that you find most suitable for application in your sector, and why?

Respondent:

We currently use Claude, and I personally like it better than ChatGPT . This is mainly because Claude often gives more relevant and substantive answers. Of course that also depends in part on how you formulate the prompt, but in our experience Claude is more consistent. In addition, we are in the process of developing our own LLM. That model is based on an open-source LLM, to which we have then added our own functionalities. Still, I must honestly say that our own model does not (yet) come close to what Claude or ChatGPT can do. For now, the performance of the large, commercial models remains superior.

Interviewer:
It is interesting that you both work with Claude and develop your own model.
You also mentioned that you have added your own functionalities based on an open-source model. That raises the question for me how you guys organize and control the deployment of those AI applications.

Q2.2:: How are AI systems overseen in your organization? Are there specific governance procedures?

Respondent:

Within our organization, we have a clear understanding that sensitive data should never be entered into public AI systems. To ensure that, we work with placeholders as standard. So instead of customer names, for example, we use terms like "company" or "friend name," and other sensitive information is anonymized in the same way. If it is really necessary to work with sensitive data, we use our own LLM. This is implemented in a secure way and works through the principle of MCPs , maybe you know it? Those are like interfaces that allow you to have an LLM communicate with other LLMs or services. We are now experimenting with that so that we can work in a controlled way with sensitive or valuable data within a protected environment. So in some cases we use our own LLM in combination with an external system, but always through a secure route. Our own model also runs in complete isolation at the network level. In fact, it has no Internet access and cannot be accessed from the outside. We try to limit the use of sensitive data as much as possible anyway, unless it is really necessary to achieve impact. But even then, we take technical and organizational measures to do that securely.

Interviewer:
Q2.3: How do regulations such as GDPR, DORA, NIS2 or the AI Act affect the application of LLMs in cybersecurity?

Respondent:

If you look at it from a geopolitical perspective, you see clear differences. In Asia and the United States, privacy is given considerably less importance than here in Europe. As a result, developments in the field of AI can happen much faster. There is more room for experimentation and large-scale application, while here we are bound by stricter laws and regulations such as the GDPR and the AI Act. As a result, innovation in Europe is slower. Compare it to the discovery of carbon: once it's there, you can't "discover" it. It's the same with AI, you can't suddenly decide to stop training it or using it for certain applications. That development will continue anyway, regardless of regulations.

That said, I understand very well the reason behind those regulations. It is right that we in Europe try to regulate AI responsibly and transparently. But it does have the consequence that our innovativeness may lag behind regions where regulations are less strict.

*Interviewer***:**
You make a good point about how regulation can inhibit innovation, but is also necessary to manage AI. That actually brings me right to another important dimension of responsible AI use, which is the ethical side of it.

Q2.4: What ethical risks do you see when deploying LLMs, for example around bias, discrimination or misclassification?

Respondent:

Yes, there is definitely something in that. Look, an LLM ultimately just does what you teach it to do. You put human logic or behavior into it so it looks like it can reason, but it's still a reflection of the data it's trained on. If that data comes primarily from, say, white western authors, then that perspective is automatically baked into the model. What you're also seeing more and more now is AI generating new content based on previous AI content. You then get a kind of self-fulfilling loop: old data feeds new output, and that new output is used again as input. So you reinforce certain assumptions or blind spots without perhaps realizing it. And this is not just about ethnicity or skin color, but also cultural background, sexual orientation or geographic bias. The Western world is simply overrepresented on the internet, so countries like Botswana then hardly show up in the datasets, even though those are also valuable perspectives. And frankly, I also worry about how people handle it. Sometimes I see developers blindly copying code from ChatGPT and pasting it into their system without really understanding what it does. You have to keep thinking. AI is a useful tool, but it shouldn't replace your thought process. From time to time, just write that management summary yourself, and think critically about what a model provides you. Use it smartly, but keep steering yourself.

*Interviewer***:**
Okay great you mentioned that people need to keep thinking. And that is what my next question is about, namely:

Q4.a: Do you think human control is always needed, or can certain decisions be fully automated?

Respondent:
Well, I think if you really want to use AI effectively, you should build in as few moments of control as possible. You should only apply control where it's really necessary. That depends a lot on your operations, your policies and what you consider "actionable" or not. In the beginning, there will probably be a lot of controls, simply because it's new and people still need to build trust. But as time goes on and trust grows, those checks will naturally diminish. So yes, definitely controlling where necessary, but not overcompensating.

Interviewer:
So what I understand from this is that human control is necessary anyway but it shouldn't be overused. Okay, that's clear. Further, you had also included some about data input:

Q2.5: How important do you consider the quality, provenance and control of the data used in AI applications such as LLMs? And what approach does your organization take in this?

Respondent:
Data quality and provenance are absolutely important, precisely because in practice we often have little direct control over it. Within our organization, we try to overcome that by running models in a controlled, shielded environment and structuring or filtering input data where possible. But really in-depth control at the source level is usually not feasible. This is mainly due to scale and resources. Certainly smaller organizations or start-ups often do not have the capacity to train a full in-house LLM from scratch. It simply takes too much time, money and computing power. Therefore, most parties build on existing open-source models, such as DeepSeek or other foundation models, and develop their own applications on top of them.

The downside is that we have all developed a kind of blind trust in them. We assume that those models are well trained, and then when something crazy comes out of it, we often don't realize until later that we ourselves have very little visibility into the original training data. You can see tokens and relationships, but not the exact source content. We trust the reputation of the creator, but true transparency is lacking. Ethically and scientifically, you should really be training your own models. But that is simply not realistic for most organizations. For example, look at how long it took to develop GPT, which is unfeasible for most companies.

As for specific models like DeepSeek, you can see that opinions are divided. Some people are reluctant because of the origin (China), others use it without objection. My view is: as long as you isolate the model in a private, secure environment without Internet access, you can deploy it responsibly, provided you are aware of the broader geopolitical context. That context is becoming increasingly important anyway. During the Cybersec fair in Brussels, I already noticed that geopolitical preferences are also increasingly influencing choices for AI models, software and cloud vendors. That's only going to increase in the coming years.

Interviewer:

Yes, clear. It also seems logical that many organizations rely on existing models since almost everyone is already using it. But then, of course, you run into transparency about the origin of the data. That immediately brings me to the next topic: the way those models are managed and controlled within the organization.

Q2.6: In what ways does your organization control how LLMs are deployed, updated or modified? And are there agreements about which models can be used when?

Respondent:

Yes, look, we just monitor that tightly. Everything that happens with LLMs, we do it in a closed environment. So if something weird happens, it stays in there. It just runs separately from everything else, so it can't do crazy things in the primary process. We actually always test first: what does the model do, where are the risks, how does it respond to certain prompts. So we always just give input, so to speak. We use it to write inventions, for example. That just saves an awful lot of time. So if he starts putting really weird things in there all at once, we see that before that goes to a bigger customer. And if it doesn't feel right, then it doesn't go live. But there's always a human check about it anyway. Never automatically to the outside or towards the customer.

In terms of models, we do have some agreements. If it's sensitive data, we either use placeholders or our own LLM. And that is also fully protected. We have linked that via MCPs. So yes, we don't have a formal policy document with ten layers of governance, but in practice there is just a working system behind it. We don't do anything at random, everything is shielded, tested, and there is always a human being on top of it.

Interviewer:

Those were basically all the key questions. You had already answered a few questions in another question so obviously I'm not going to ask those again. The next questions are about the framework:

Q2.7: What would you need to have confidence in an AI-supported framework for cybersecurity risk management? And how would such a framework ideally be tested or improved in practice?

For me, a framework should be practically applicable and aligned with how organizations are already working with AI. So not just stuck at a high level of abstraction, but really show what technical measures you can take.

Think about isolation of models, the use of audit trails, and for example measures such as MCPs that allow you to securely link between systems. In addition, the framework must also be able to move with technological and geopolitical developments. You see that changes are coming faster and faster in how we use AI and how regulations are evolving. So a framework has to be flexible enough to keep up with that. What also helps for trust is if the framework has been tested in multiple contexts: at both larger banks and smaller organizations. And that it uses realistic scenarios, so you can see what really happens in cases of prompt injection, misclassification or data leakage via input, for example.

Interviewer:
Okay great so what I take from this is that a framework should be tested anyway with small and large companies. And that it should be able to keep up with the times, so all the changes that will take place. The next question is actually about the preliminary framework itself.

Q2.8: After going through the preliminary framework, which parts do you find most relevant or problematic in your practice? Are you missing certain elements or do you see opportunities for improvement?

Respondent:

Let's see here. Yes, of course it depends on how you deploy it. In your framework I do see the technical aspect coming back a little bit, in the bottom right corner for example, but that still remains quite general. From my experience, and I do look at it through technical glasses, I notice that the technical side of AI often remains underexposed in these kinds of discussions. While many practical measures are possible there. Think, for example, of isolating models, or of using MCPs. Such measures are enormously valuable: they protect the confidentiality of data and make AI more usable in sensitive environments.

I am not saying that you should immediately make a whole new framework part of it, but it would be good to explicitly name it. After all, a sound technical implementation directly supports other pillars of your framework, such as compliance, ethics and governance. If you get it right technically, you avoid having to explain away all kinds of risks later. You could incorporate this under 'AI lifecycle and LLM-specific controls' so that it is included, without making it unnecessarily complex. Another thing I miss is explicitly naming risk detection as a separate component. In practice, this is often one of the first applications of AI in cybersecurity: the detection of deviations, anomalies, unusual behavior. That comes back somewhat implicitly now, but really deserves its own place in the framework as far as I'm concerned. Not only because it is relevant, but also because it presents unique challenges, think of false positives, explainability and impact on incident response All in all, I think the framework strong in terms of content, but with these additions it becomes even more practical and applicable in daily reality.

Closing

*Interviewer***:**
This also brings us immediately to the end of the interview. Thank you again for your time and sharing your insights, it is incredibly valuable for my research. As stated earlier, all responses are completely anonymized and kept confidential. I will only use the recording to develop the transcript and analysis in my thesis. After that, everything will be deleted. In the second week of June, you will receive another email from me with a Google Survey. This is necessary for the validation of the final framework and will be sent along with the deadline at the same time.

Do you have any further questions or comments?

Respondent:
Yes, when do you have to turn it in?

Interviewer:
Good question, my deadline is June 27 and a week before that I will send my final draft. Because then I still have enough time to adjust something