QUINTEN RIPHAGEN

THE DEEPFAKE PROBLEM

THE DEEPFAKE PROBLEM

QUINTEN RIPHAGEN

Developing a novel maturity model for forensic image authentication research

Master Student
Business Information Technology
EEMCS
University of Twente

August 2022 – version 1.3

Dedicated to the loving memory of Annie de Goeij-Nijhof
1939 – 2022

Ohana means family.
Family means nobody gets left behind, or forgotten.

— Lilo & Stitch

An investment in knowledge always pays the best interest.

— Benjamin Franklin

ABSTRACT

The possibility and ease of manipulating evidence with deepfake technology poses a threat to the judicial system. Video evidence cannot be trusted on face value anymore and requires analysis for possible manipulation. Forensic research organizations around the world lack the tools to benchmark and improve their image authentication capabilities. In this paper, the image authentication process is explored and the first image authentication capability maturity model (IACMM) is proposed to provide these organizations with the tools to incrementally increase their image authentication maturity.

SAMENVATTING

Het gemak en de mogelijkheid om bewijs te manipuleren met deepfake technologie is een gevaar voor het Nederlandse rechtssysteem. Video bewijs kan niet meer genomen worden als waarheid en heeft uitgebreide analyse nodig voor mogelijke manipulatie. Forensische onderzoek organisaties over de hele wereld missen de middelen om hun beeldauthenticatie proces te keuren en hun capaciteit voor beeldauthenticatie te verbeteren. In dit artikel wordt het beeldauthenticatie proces onderzocht en het eerst beeldauthenticatie maturity model wordt voorgesteld om forensische onderzoek organisaties de middelen te geven om hun capaciteit incrementeel te verbeteren.

EXECUTIVE SUMMARY

The digital and bio-metric traces (DBS) department at the NFI may soon face a new challenge in the area of image authentication. Deepfakes are videos manipulated with AI and are becoming indistinguishable from real video. This presents a problem for the admissibility of video evidence in court. More video evidence will have to be authenticated by the NFI, which poses a problem if the NFI is not equipped to handle an influx in cases.

The aim of this research is to explore the rise of deepfake technology and it's effects on the judicial system. The perceived problem is evaluated and a model is designed to benchmark and improve forensic organization's ability to manage an influx in cases due to deepfake technology.

Deepfake technology poses a significant problem for digital forensic investigators in their image authentication process. Deepfake technology might soon become good enough to create video manipulations without any identifiable artifacts. Pixel level analysis may soon no longer be a viable method of analysis in the image authentication process. This raises the question for forensic organizations: “Is our image authentication sufficiently ready to deal with the advancements of deepfake technology?"

In this research I assess the risk for forensic organizations, identify the bottlenecks in the image authentication process, and provide a model to incrementally increase the image authentication capabilities of a forensic organization to deal with this problem. One of the more interesting solutions to investigate is deepfake detection software, offered by startups.

Through interviews with experts and literature review, factors that determine maturity in digital forensic organizations are evaluated and used to construct a maturity model. Deepfake detection was found to be too unreliable and unverifiable in detecting novel deepfakes. Therefore it could not be used as a standalone solution in the Image authentication process. However, it can be useful within the IA process as the explainable AI component of most deepfake detection software can be used in the local analysis of images. The results can also be factored into the calculation of the likelihood ratio.

The result of this research is a maturity model that can be used as a road map for improving image authentication process in case of an influx in case, and the resulting maturity level can be used as an indication for image authentication maturity. The NFI should look to implement some of these improvements to increase the image authentication maturity in the organization.

"If I have seen further it is by standing on the shoulders of Giants."

— Isaac Newton

ACKNOWLEDGMENTS

Firstly I would like to thank Zeno Geradts for guiding me through the thesis and the fun weekly meetings that always boosted my spirits and motivation for the project. I also would like to thank him for his understanding and advice in difficult times and making sure I had all the tools to my disposal to complete the thesis.

I would like to thank my Thesis supervisors Abhishta Abhishta and Jan-Willem Bullee for their continued support and confidence in my abilities as a student. When I hit a road block they were always available to listen to my problems and provide the necessary feedback for me to continue. And the bi-weekly meetings never failed to reinvigorate the passion for the project and or graduating the degree.

Special thanks to my study advisor Amaal Shamari for guiding me through the process of writing a thesis and providing me with the tools to structure and organize the road ahead.

I want to thank my study mate Ioana Miu for giving me a much needed confidence boost when I needed it. I also want to thank my friend Koen Oortgiessen for providing much needed motivation and distraction from the stress of writing a thesis in the form of daily coffee breaks, lunch walks, a listening ear, and interesting conversations about life.

Lastly I want to thank my parents for supporting my ambitions to go to university and pursue a Master's degree, without them it wouldn't have been possible. I feel privileged and humbled to have been given these opportunities and the faith that I would be able to figure out what was best for me.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part I

PROBLEM INVESTIGATION

This part describes the first phase of the thesis. In this phase, the problem that will be investigated in this thesis is explained, the scope of the research is explored and the methodology used to answer the research questions is determined.

1 INTRODUCTION

"In a time of deceit telling the truth is a revolutionary act."

— George Orwell

1.1 SOCIETAL CONTEXT

A video gets sent to the police showing security camera footage of someone committing a criminal offence. The suspect in the video is relatively visible. The provider of the video is the suspect's neighbour, and is insistent that his neighbour should be arrested. The police arrests the individual and the public ministry sets up a criminal case against the suspect. The video showing the suspect commit the crime is the main piece of evidence in the case. The evidence is analysed by the police and the suspect in the video matches the suspect in the case. The day arrives that the case goes to court, the public prosecutor lays out the case to the judge and brings in the evidence. When the case is laid out, the suspect's defense says that the evidence is manipulated and that the suspect they are representing is not the person in the video. In a time where deepfakes can be made using your mobile phone, with minimal knowledge of the technology, this claim is plausible enough for the judge to request a further investigation.

In most cases, where manipulation is plausible, this claim would need to be investigated by a forensic organization such as the Netherlands Forensic Institute (NFI). The NFI would get a request to investigate the claims of the video material, and will do so if they think the claims can be investigated.

1.2 PROBLEM STATEMENT

Currently the NFI handles all video evidence they receive in the same manner. The variety of questions asked by the court can vary. For example, a common question about video evidence is "How likely is it that the person in the video evidence is the same person as the suspect?" To which the NFI will give a statistical explanation based on the opinions of experts. This is a specialized process in which three digital forensic image/video researchers analyze a piece of evidence using some help of tools and give their own expert opinion on the question asked by the court. They statistically support their opinion with Bayesian statistical methods which take into account the prior and posterior likelihoods of an event happening. Such as the likelihood that the suspect in court is the same person as the suspect in the video evidence.

Sometimes the NFI also receives questions about whether or not video evidence has been manipulated. This question is likely to occur more in the future with the technological developments of deepfakes and an increase in public knowledge about deepfakes. As more people know about deepfakes, the more likely it is that it will be used as a defense against incriminating video evidence in court.

The current method of analyzing video evidence is unlikely to be sufficient if the amount of questions surrounding its credibility increases. The amount of claims that are made in court are likely to increase, since the knowledge surrounding deepfakes in general is also increasing. Deepfake technology is getting better and more people are aware of the existence of deepfakes, even if they don't know how to make one. The general knowledge about the existence of deepfakes combined with the lack of technical knowledge among judges and lawyers surrounding the difficulty of making one, makes this a difficult problem to solve.

The current process has not been formally documented and is done in a mostly ad-hoc manner. This means that the process is obscure and knowledge about the process could leave the organization when the employees decide to leave. In the first place, the current situation needs to be documented to contain the knowledge that is currently in the organization. Then, the process needs to be improved to be more capable of handling the increase in requests from court. The improvement of the process could be achieved through the implementation of the suggestions from a maturity model [1]. In the literature the current consensus seems to be that deepfake detection methods are unlikely to be sufficiently accurate to detect novel deepfakes for forensic analysis. There are standardized practices for the forensic analysis of images in general, but there are no guides specifically for image authentication research. Therefore, this research will provide the academic community with the tools to develop more solid image authentication processes. In addition, this research will investigate whether deepfakes are actually likely to become a problem in the future.

This research aims to provide advice on ways to solve this problem within forensic institutions. The researcher will develop a maturity model that outlines the factors for maturity in the image authentication process for forensic research institutes. This will prove to be a difficult task, since maturity models are usually built on an organizational level. Building a maturity model based on process level maturity has been done, but the overwhelming amount of maturity models is created at the organization level. Prior research showed that no maturity model has been made in the area of image authentication (IA), so this research will be an exploratory study into the creation of maturity models in IA.

1.3 RESEARCH SETTING

This research is done in collaboration with the Netherlands Forensic Institute (NFI) as part of my graduation from the Master Business Information Technology at the University of Twente.

The NFI is a research institute that specializes in conducting forensic analysis for criminal cases in the Netherlands. Besides forensically investigating cases, the institute is constantly developing novel methods for forensic analysis of cases. This is based on expectations of the future or cases requiring novel research methods. The research institute collaborates with foreign organizations on forensic analysis techniques as well. During the research project I was a part of the team *Image Research* from the division *Digital and Biometric Tracks (DBS)*. This team specialized in analyzing image and video evidence for criminal court cases. This involves anything that requires the analysis of video evidence. Such as investigating video evidence for manipulations and verifying the authenticity of video evidence. The research I conducted was limited to specifically verifying the authenticity of video evidence.

The analysis of digital evidence is often necessary in court cases as expert opinions are needed on video evidence. The questions that can arise during a court case about video evidence is broad and can vary wildly.

One of the processes is the analysis of evidence to determine whether the material provided is authentic and was not manipulated to mislead the court. If a suspect is identified in a piece of evidence and the defense claims that the evidence was manipulated, an independent organization needs to give an unbiased analysis on whether this claim could be true or not. As well as giving an insight to the court into the clues that might signal manipulation or why it is unlikely that the evidence was manipulated. This type of claim is currently not made very often, but the amount of claims may increase as awareness of deepfake rises.

The verification of video evidence was a business process that was undocumented internally. The only process definitions that were available were from the European Network of Forensic science institutes (ENFSI). To get a better view of the implementation of this process at the NFI, the process had to be documented.

1.3.1 *Societal implications*

The results of this research will have mayor societal implications. Determining the levels of maturity for forensic organizations in the IA process will allow for more effective improvement of the process. It will also allow customers of this process to benchmark different organization against each other, and choose the organization with the most mature IA process. This in turn will lead to a reduction of false conclusions in cases, a more efficient legal system, and a more robust digital forensic image research process. For the scientific community, this research has implications in Maturity model research in image authentication, as this study is the first of it's kind. Further research could lead to more elaborate breakthroughs in image authentication maturity.

1.3.2 *Deepfake definition*

There are several deepfake definitions of the word deepfake available on the internet and in the literature. All of the definitions are valid, it is just a matter of how broadly you would define the term. The original definition is: "Face manipulation in image through artificial intelligence" [2]. The other more broad definition often found in the literature is: "Any type of manipulation or synthesizing of image material through artificial intelligence" [2].

In the context of this research there are multiple types of deepfakes which can appear in court. It is necessary to properly differentiate between these types of deepfakes to avoid confusion. The types are as follows:

1. A deepfake was used to commit a crime, but whether a deepfake was used has no effect on the case
2. Video evidence was brought into a case, but the suspect claims the video evidence was manipulated to show them instead of someone else (faceswap)
3. Video evidence was brought into a case, but the suspect claims the audio was manipulated to make them say things they didn't say.

The first type of deepfake will not have to be investigated. Evidence brought into a case might have been manipulated, but the fact that it has been manipulated has no bearing on the burden of proof. For example, an audio deepfake may be used to scam a CEO out of money. The recording of this manipulated audio is evidence in the case, but the information surrounding the recording, such as the time of the phone call and it's origin, may be more important to the case than whether or not the audio was manipulated. If however the video evidence is of the defendant committing a crime, and the defendant says that it was not actually him and that the evidence has been manipulated. From the context of the video evidence it cannot be ruled out that the video has not been manipulated. This case would go to the NFI, since the burden of proof of the evidence is resting on whether the defendant is actually the same person as the suspect in the video. In any case which involves video evidence the defendant may claim that the evidence was manipulated to depict them doing things they were not doing. The most important part before this claim gets investigated by the NFI is to know all information surrounding the video evidence. Information such as:

1. When was the video taken?
2. Where was the video taken?
3. What device was used to record the video?
4. Is the video still in its original state or has it changed places? In other words, is it even possible for the video to have been manipulated?
5. Does the defendant have an alibi?

The NFI will most likely be dealing with the manipulation defence, in which the defendant will say that they are not the person in the video. This refers to the faceswapping scenario of deepfakes. Which according to the literature [2] and leading deepfake detection companies is also the type of deepfake that is used the most. Therefore, the scope of this project will reflect the face swapping definition.

1.4 scope

The scope of the research is centered around finding IT-based solutions to problems that are present in the process of determining the authenticity of video evidence. This includes locating where the problems lie in the process, determining the requirements of possible solutions, designing adequate solutions and providing an advice on how to implement the proposed solution. This project exists entirely in the sub-process of the analysis of evidence for authenticity purposes. Any other analysis of evidence done by the image-research team at the NFI is not included in this project.

1.4.1 *Scope extended*

Initially when I looked at the process that needed to be investigated for the NFI. The idea was that there could be a new process for the authentication of video evidence. When the initial interviews started, it became clear that the original scope was too simplistic, and the goals set out for the may be difficult to achieve. There currently are almost no requests for the authentication of video evidence at the NFI. Since each case is different and requires custom investigation methods, it is difficult to design a process that accounts for all different possibilities. Nevertheless, it is possible to model the capabilities that are needed to handle an increase in cases that accounts for the types of cases that are likely to be investigated. Therefore, I decided to redefine the scope of the problem and create better and more achievable goals for the research.

1.4.2 *Reasoning*

The scope of the problem extends beyond the NFI. From the initial interview with Manon Den Dunnen, the strategic innovation advisor of the Dutch police, it became clear that the problem of deepfakes extends beyond evidence manipulation and simple cybercrime. She expects that synthetic media will be the dominant type of media on the internet in the future. So she posed the problem, how do we properly value the information that is present in the media whether the media has been manipulated or not. The means that the scope of the problem is wider than just deepfakes, and that instead of designing a new process, a maturity model of the capabilities needed in the entire chain of evidence analysis would prove to be a more valuable artifact. Since my current knowledge of maturity models is limited, an integrated literature review is needed to further research this. This ties in with the interviews with Roel Klaar and Martijn Egberts (OvJ) where we explored the history of the "hack-defense" [3] , which can be seen in the interview notes in Appendix A. A defense which was often used in child pornography cases, in which the suspect would simply claim that they were hacked to avoid prosecution. The police/NFI would then be asked if there was evidence of a hack on the computer of the suspect, to which the answer would often be yes. Since it is relatively likely that some evidence of malware is found on computers of people who are not very tech-savvy. The question for the court remained whether the fact that evidence of malware was found on the computer had any bearing on the burden of proof. The malware could be on the computer and have no impact on whether the suspect was guilty of the crimes they were charged with. Judges in the past often had trouble dealing with the "hack-defense" since knowledge with judges around hacking was limited. Nowadays dealing with the hack-defense has become standard practice since the knowledge surrounding this defense has increased, and standardized models for dealing with it have been developed. A similar standardization could be realized for the evidence manipulation defense. This standardization could be integrated into the capabilities for the maturity model. Seeing as this new scope requires knowledge that was not acquired during the first systematic literature review, I propose an integrated literature review into the subject of maturity models in digital forensics.

1.5 RESEARCH GOAL

Develop a framework for Video analysis at the NFI based on the requirements of various stakeholders, available frameworks, available deepfake detection tools, and the forensic method for the analysis of digital evidence. The goal is to document the current process and make it more efficient to make handling multiple cases at once easier to manage. The current goal of the research is trending towards a design problem and not a knowledge problem. The end result of the research will be an artefact that can be applied to the problem context of the NFI.

1.5.1 *Research questions*

The research questions will be based on the steps in the design science methodology process. Each research question reflects a step in the design science process and will be answered using different methods.

The main research question that we will try to answer with this research is as follows:
MQ: *How to design a forensic image authenticity research maturity model that reflects current and future needs of forensic organisations?*

To answer the main research question, the following sub questions need to be answered:

1. What capabilities are needed in an optimized image authenticity research process?
2. What capabilities regarding forensic image authenticity research does the NFI currently have?
3. How is a capability maturity model constructed for digital forensics?
4. Does the design reflect the reality of the situation according to experts?

The implementation of the new capabilities of the designed maturity model is outside of the scope of this research. The goal of the research is to develop a novel maturity model that can be used to determine what level of maturity the NFI has in relation to forensic image authentication research and what capabilities they can add to get to a higher level.

1.6 THESIS STRUCTURE

The thesis is structured according to design science methodology. The chapters are structured in order of the research questions. Chapter 2 will cover the background of the thesis, which covers all information about deepfakes that is useful to know for the context of the research. The subsequent chapter, Chapter 3, will cover the research methodology and will describe exactly how the principles of design science methodology are applied in the research. Chapter 8 will cover the Conclusion. Chapter 9 will cover the discussion. In the subsection below, the research questions will be mapped to their respective chapter.

1.6.1 *Mapping of research questions*

The research questions SQ1 is a knowledge problem in the design phase, and will be solved using the interviews in Chapter 5.
SQ2 is also a knowledge problem in the design phase and will be solved using the interviews in Chapter 5 and the literature review in Chapter 4.
SQ3 is part of the design phase, the result of the treatment design is the artefact, and is achieved through analysis of the literature and interviews. This question will be answered in Chapter 6.
SQ4 is part of the evaluate design phase, in which the artefact is validated using expert opinion. This question will be answered in Chapter 7.

The main research question will be answered and reflected upon in Chapter 8.

1.7 CONTRIBUTIONS

In this research the following contributions are made to the literature of image authentication, deepfake detection, and digital forensics maturity.

- The threat of deepfakes for the judicial system and forensic organizations is assessed.
- A novel maturity model for the image authentication process is proposed. Including factors of maturity found through expert interviews and literature review.
- An overview of the image authentication process at the NFI is created and a new process view of a future image authentication process is made based on industry best practices.
- An assessment Matrix based on the maturity model for companies to determine their maturity level.

2 BACKGROUND

"He who does not learn from history is doomed to repeat it"

— Winston Churchill

In this chapter, the background of the Thesis is explored. The reasoning for the necessity for increasing the efficiency in the image authentication process will become clear in the following sections.

2.1 DEEPFAKE CREATION

Deepfakes are most commonly created using Generative Adverserial Networks. Goodfellow et al. [4] provides the first Mathematical model of the generative adversarial networks (GAN) that most Deepfake technology is currently based on. These networks are capable of making nearly undetectable generated media such as images, video and audio. Goodfellow et al. [4] explains the way these networks create these undetectable media is by pitting a generative model against an adversarial discriminatory model whose responsibility is to determine whether a given sample is from the training model or from the generative model. The generative model learns from the discriminatory model by which samples passed through detection and which were detected. In the following iterations, the features from the undetected samples are kept and expanded upon. This process is repeated for many generations until the discriminatory model can no longer accurately predict whether a given sample is real or generated. An example of the network structure can be seen in Figure 1.

Goodfellow et al. [4] explains this with an analogy of counterfeiters vs the police. The counterfeiters are trying to create the most realistic counterfeit money while the police keep working on new measures to detect counterfeit money. Competition in this field leads to nearly undetectable counterfeit currency. The same process applies in the creation of GAN generated deepfakes. Which makes detection by software measures a difficult problem.

2.2 DEEPFAKE DETECTION

There are many methods currently available to forensic analysts to prevent the negative effects of deepfakes. Current deepfake detection methods seem to be an effective tool for detecting fake videos. The machine learning models can reach high detection accuracies on large datasets of high-quality deepfakes. However, most state-of-the-art de-

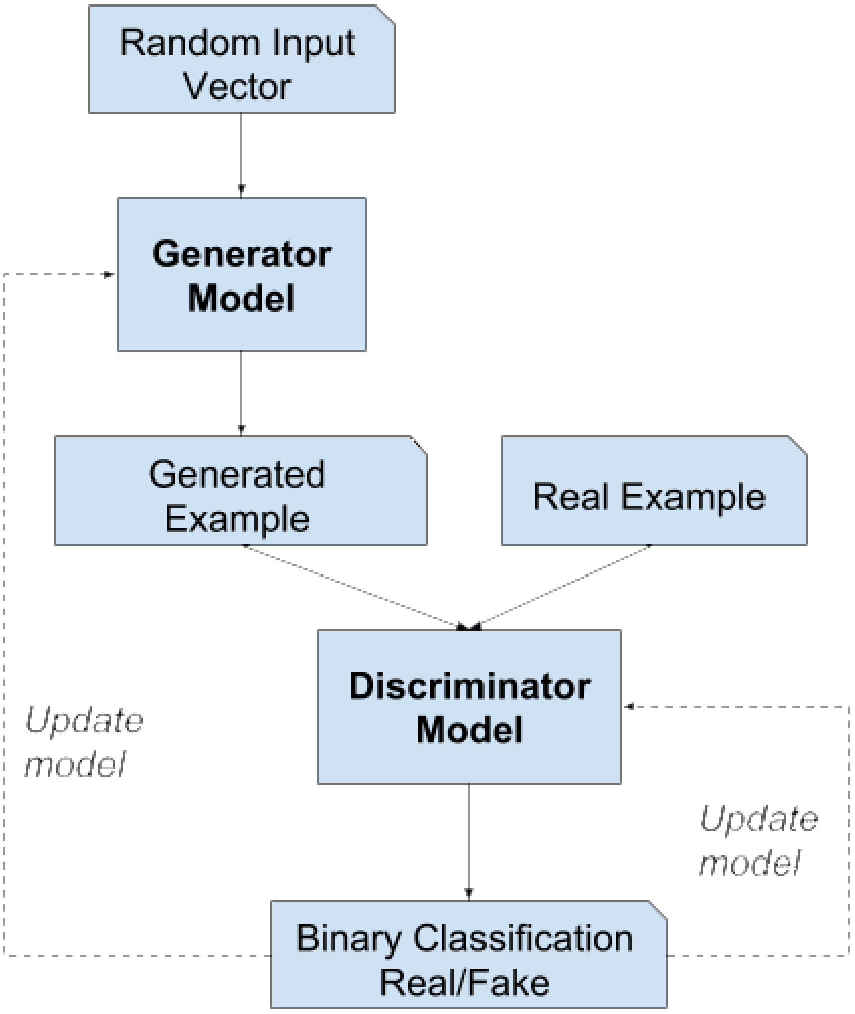

Figure 1: Example of a GAN. Retrieved from: Brownlee [5]

tection methods are not equipped to handle out-of-domain data and deepfakes in the wild. Which is supposed to be the main use of deepfake detection algorithms. Currently, the detection algorithms, save a few specifically designed for deepfakes in the wild (DFWs), are hampered by the way they are trained and the public datasets that are used to validate these detection methods. It is a difficult problem to handle out-of-domain data, although some detection methods such as Tariq, Lee, and Woo [6] and Khalid and Woo [7] are better equipped for this type of detection. GANs allows for the easy creation of relatively high quality deepfakes. The research into detection techniques also fuels the development of better technology which is able to avoid detection. This arms race in detection and creation has enabled high quality, hard-to-detect deepfakes to emerge. These are not easily created and have to be made by someone with experience and expertise on the subject, but the development of the GANs will allow for high-quality deepfakes to be made by anyone. The existence of these high-quality deepfakes which might be hard to detect means that not only will videos have to be proven fake, it will also be necessary to be able to prove that videos are real. Which current most of the current detection methods are unable to do due to the lack of transparency in how the model got its results. Current artificial intelligence (AI) methods often don't offer an explanation as to what clues it used to determine why a frame is considered fake. Deepfake detection will always be behind the deepfake creation, as the detection techniques are reactionary by definition. Anyone with malicious intent could create a very realistic deepfake that easily evade detection in modern state of the art methods. Carlini and Farid [8] have also shown that current detection methods are still very vulnerable to specific attacks such as adverserial fakes. Solving this problem will prove difficult if we keep training and benchmarking the detection methods in the same way. Generalizability should therefore be the most important research topic. Blockchain and smart contract methods may be a solution, although current implementations of this are not scalable and severely lacking in the tools needed to execute this type of proof-of-authenticity platform. In addition, the required amount of adoption of such a system to make it work will not be realistically reached anytime soon. In conclusion, the current available methods to prevent the negative effects of deepfakes are not sufficient. It is still hard to prove why a video is real, which is relevant in a forensic scenario. Authenticating content when it is created might be a solution in the future. However, current implementations are not scalable, will require widespread adoption, and offer no solutions to deepfakes in the short term.

2.3 THREATS FOR ORGANIZATIONS

Deepfakes will most likely be used as a part of social engineering attacks such as phishing attacks. Since most types of social engineering attacks rely on some form of impersonation, deepfakes are the perfect addition for these types of attacks. According to the security company Tessian "Deepfake generation adds new ways to impersonate specific people and leverage employees' trust" [9]. Social engineering is an effective method to gain access to a company's network or private data.

In March of 2021 the FBI issued a report to private industries notifying them that deepfakes and synthetic media will be used by cyber crime actors within the upcoming 12-18 months [10]. It was issues to help cybersecurity experts and companies identify the threats of deepfakes and to provide them with some basic best practices to protect their organization from deepfakes. The report states that the FBI expects that malicious cyber actors will use AI techniques broadly across all of their operations. Mostly as an extension of already existing social engineering and spearfishing campaigns. They also define a new attack vector called Business Identity Compromise (BIC). Which is defined as "[sic]The use of content generation and manipulation tools to develop synthetic corporate personas or to create a sophisticated emulation of an existing employee" [10].

2.3.1 *Financial threats*

The most important threats for organizations to be aware of are the financial threats. Since this is what affects their bottom line the most. Outlining these threats also make it more likely that organizations do something to mitigate the threats. In a report by Bateman [11] on the threats of deepfakes in the financial system. The author states that deepfakes should not be a threat to mature, healthy economies or the stability of the global financial system. But can cause varying degrees of harm to individuals, governments, and businesses that are targeted by malicious actors. In the report the author describes 10 different scenarios in which synthetic media (deepfakes) can play a role in financial fraud for individuals, companies, and markets. The author makes a distinction between narrow cast synthetic media and broadcast synthetic media. The difference being who the synthetic media is targeting. Narrow cast synthetic media is made for small individual targets and usually delivered through private channels such as email or a phone call. While broadcast synthetic media has a larger target audience and is delivered through public channels such as social media. Both types can cause financial damage to organizations.

Through all scenarios examined by the author 3 key malicious techniques keep reappearing, these are vishing (voice phishing), synthetic social botnets, and fabricated private remarks. All of these techniques rely on some form of social engineering. Deepfake vishing is a type of narrow cast synthetic media and can be used in a variety of ways, including identity theft, fraudulent payment schemes, and imposter scams. This technique often does not require a perfect recreation of the person as the victim of this type of scam will often attribute any inconsistencies to a faulty connection or phone line due to the high-pressure situation the scammers create for the victim [11].

2.3.1.1 *Audio impersonation*

The first scenario described in Bateman [11] for organizations is payment fraud through audio deepfake, a narrowcast synthetic media scenario. The umbrella term for tricking a business into transferring money illegitimately is Business Email Compromise (BEC), because criminals often hack or spoof the account of a CEO and contact a financial employee to trick them into wiring money or purchasing a gift card. This scenario would be expanded by spoofing the voice or likeness of a CEO in a video-call or voice call. The video scenario would require live deepfakes which are still being developed, but will be easy to achieve in the future [11]. This scenario has already happened as can be seen earlier in this literature review in the article by Stupp [12]. Where a CEO of a U.K. based energy company was tricked into transferring money by a vishing call from his boss at its German parent company. The CEO said afterwards that he recognized his boss' voice with a slight German accent and believed the call to be genuine [12].

2.3.1.2 *Extortion*

The second scenario is cyber extortion, the scammer extorts an individual by threatening to release sensitive information to the internet or friends and family, often sexual in nature. Another scenario that uses narrow cast synthetic media is mostly aimed at individuals instead of organisations, but scammers could target specific individuals in organizations with certain authentication and make them give access to company funds or networks. Deepfakes would make this type of extortion easier as hackers do not need to obtain sensitive materials since they can just forge it [11].

The following scenario's described by the authors are all scenario's.

2.3.1.3 *Stock manipulation*

The first broadcast synthetic media scenario is the manipulation of stock by fabricating events using deepfakes. By fabricating false events, malicious actors could lower or raise a company's stock price, profiting of the swing in price. This type of manipulation would likely only affect the stock price of a company for a short while, as when the information is proven false, the price would likely return to a normal level. Nevertheless, the reputation of an organization can be permanently damaged by such a fake. This type of threat is especially dangerous for larger organisations which have a public facing CEO which generates lots of video material for scammers to use in the creation of deepfakes [11]. For example, a deepfake featuring the CEO of a company committing a crime or saying racist slurs has the potential to swing the stock price of that company [13]. This type of media might also be difficult to disprove, and a CEO might have to call upon his good reputation to discredit the video. And even if the video is disproven, it would likely still have long term consequences for a company's reputation. As many people only see the first video without the discrediting of the video. Another factor is that a significant amount of individuals will still believe a video to be true even if it is labeled as false [14].

2.3.1.4 *Negative sentiment creation*

In the second broadcast synthetic media scenario, synthetic media would be used in the creation of more believable social media bots. These bots can be used to express a negative sentiment about a company on social media, impacting the stock price as modern trading algorithms take consumer sentiment into account. The addition of deepfaked media to these bots could increase their credibility and make it harder for social media companies to detect. Although no cases of this has been documented as-of-yet, social media accounts using images generated by GANs such as https://thispersondoesnotexist.com/ have been seen spreading misinformation online [11].

2.3.1.5 *Fake rumours*

In the third broadcast synthetic media scenario, fake rumours about instability of the financial system or banking sector may cause bank runs. This type of manipulation is often only possible in places which already have low trust financial stability of a country. Like in the previous example, synthetic botnets could express a negative sentiment around the financial stability and instigate a bank run. Alternatively, a deepfake depicting a bank manager expressing concerns of liquidity problems could be used to instigate a bank run. This type of threat is mostly relevant for banking and governmental organizations [11].

The remaining scenarios described in the report mainly affect financial regulators and markets as a whole, which organizations have little control over. Therefore these scenarios will not be addressed here.

2.3.2 *Which organizations are most vulnerable?*

The types of threats discussed above are not all equally effective on all types of organizations. Some organizations might be more vulnerable to different types of threats. In this section, the differences between these threats and to which types of organizations they are applicable will be addressed.

Large organizations with public facing CEOs generate a lot of video and audio content around that CEO and are therefore more likely to be affected by deepfakes involving the CEO doing things that never happened or saying things they have never said. There is more content available for malicious actors to create these deepfakes and since these organizations are more well-known by the public, a scandalous deepfake involving the company will result in more reputation and financial damage to such an organization [11].

Smaller organizations such as small-to-medium enterprises (SMEs) are more likely to be infiltrated by a vishing scam since these organizations often invest less in good cyber security practices. A survey done by the National Center for the Middle Market in 2016 found that only 45% actually had an up-to-date cybersecurity strategy [15]. While a survey by Bisson [16] 95% of professionals working at SMEs believed their firm had an above average cybersecurity strategy. This overconfidence and lack of investment in properly assessing the threats posed to these organizations may leave them vulnerable to vishing scams.

2.3.3 *How to protect against the threats?*

2.3.3.1 *General cybersecurity measures*

It is always advised to have an up-to-date cybersecurity strategy, since the world of security is constantly adapting to new threats such as deepfakes. One of the most important methods to protect against cybersecurity threats is proper employee training [17]. Security awareness training and media literacy training can help employees verify whether certain information is authentic. It is recommended that organizations update their security strategy bi-annually and assess whether their company is resistant to new threats through security audits. Viña [18] says that protecting data integrity is the most important measure to prevent deepfakes from affecting your company. While previously, data confidentiality has been the main target for hackers through data breaches. Data integrity is now also a big target for hackers and companies should adapt their security measures to include this. Current fraud detection teams can be adapted to include video and audio content verification.

Most broadcast scenarios described by Bateman [11] will be hard to prevent or for an organization to prepare for. So the focus of protecting against threats should be mostly centered around preventing the narrow cast synthetic media scenarios from happening or causing damages.

2.3.3.2 *AI systems*

In larger organisations, AI systems could be used tot prevent phishing attacks and attacks involving deepfakes. An example of this is discussed in Lee [17], AI systems could be used to abstract transaction data when hackers are trying to elicit transfers from financial employees. This system would enter the transaction data into a legal compliance framework, which needs to be approved by another human. AI can also be used to flag irregularities in video or audio messages by quickly comparing the footage to previously recorded audio and video to make sure the clips are not manipulated [17]. Deepfake detection systems that are integrated are currently not available, but with current developments there might be such an AI system in the future that could protect organizations.

2.3.3.3 *Authentication*

Another measure that organizations can use is authentication of content on the company networks. If all content is present in a decentralized authenticity system such as block chain, fake content can be quickly proven fake and will thus not cause any damage [17]. Identity verification can also be a part of this, in which an employee must verify their identity before they can execute an action. [9] recommends to have some sort of system in place in which employees can verify information. Verification is a viable method for preventing conventional vishing (phone fishing) attacks, and can also be useful when handling deepfakes.

2.3.3.4 *Insurance policies*

For the last measure, Viña [18] outlines several other steps that companies could take to prevent the harmful effects of a deepfake attack. Such as cyber insurance policies, which are expanding to include damage from deepfakes as well. This would allow organizations that fall victim to deepfakes to recover some of the financial loss. Such an attack would need to be investigated by an outside company, data may need to replaced, and damaged reputation may be managed by public relations experts. A good insurance policy can cover all of the costs associated with these actions and thus prevent the negative effects of these attacks [18]. In addition, certain crime policies also help companies that fall victim to transferring funds to criminals via phishing or in this case vishing attacks. These insurance policies would cover the lost funds [18].

2.3.4 *Summary*

Deepfakes pose genuine cybersecurity threats to both large organizations and small-to-medium enterprises. As deepfakes become easier to make, the threats of them being used in attacks becomes larger. Companies which lack the proper means to prevent or deal with the fallout of these attacks might soon find themselves being targeted by vishing attacks, synthetic botnets, and other deepfake related attacks.

Larger organizations have more resources available to properly address the threats, but should still train their employees and regularly update their security strategy. SMEs should invest more in their cybersecurity strategy and realize that their lack of investment in proper measures is leaving them exposed.

AI systems on corporate networks can be useful in authenticating content. Although fully automated detection methods are not available yet, the expectation is that deepfake detection companies will offer such services in the future. Which will aid organizations properly address the threats. Until these types of systems are realized, the best measures organizations can take are building verification and authentication into the business processes and properly training their employees with the latest cybersecurity strategies.

Insurance against the fallout of attacks using deepfakes can be useful if the expected cost of preventing the attacks is greater than the cost of the insurance [18]. Larger organizations should aim to have both mechanisms in place for prevention and insurance for when these attacks do happen to minimize the cost.

Since organizations are currently mostly unprepared for attacks with deepfakes, it is reasonable to assume that more legal cases will appear in the long-term.

2.4 ACTORS

Up until this point a mostly negative view of deepfakes was given, but deepfakes can also be used in legitimate situations. The differences between legitimate uses and illegitimate uses of deepfakes will be explored in this section, to illustrate why a straight up ban is not the solution to the problem.

2.4.1 *Education*

Deepfakes also open up new possibilities in the realm of education. They can be used for a variety of educating purposes which can have a positive effect on the effectiveness of the message they are trying to bring to students. For example, deepfakes using historical figures can be made to make otherwise boring lectures more interesting for students [19]. Also, with today's era of online lectures, deepfakes can also be used to automate parts of lectures, as the content of the lecture can be presented by an AI clone while the lecturer answers the questions.

2.4.2 *Deepfake Detection Companies*

Another actor which operates in the space of deepfake technology legitimately are Deepfake Detection companies. As the technology grows the demand for reliable deepfake detection methods will only grow larger. Companies which offer deepfake detection services therefore have an incentive in the growth of the technology. While it is highly unlikely that there are companies advancing deepfake technology in order to remain relevant, the incentive for deepfake technology to grow in order to sustain these companies should be considered. Currently, several different companies have been offering deepfake detection services as part of their repertoire.

An example of this is DuckDuckGoose [20], the company offers a deepfake detection software to companies and government agencies. They are working together with several Dutch institutions to further the development of deepfake detection software. As of now they offer a deepfake detector which shows a percentage representing likelihood that a given image is a deepfake. Which is explained through a visual representation of what the algorithm detected. This explainable AI principle adds to the trustworthiness of the algorithm. In addition to the standalone detector, they offer a browser plug-in which automatically notifies the user of possible manipulated content. As well as deepfake creation software which can create deepfakes to use for penetration testing [20].

Another company which offers services to combat Deepfakes is Truepic. The company raised funds from Microsoft, Adobe, and Sony's venture funds. They offer content authentication at the moment the content is captured, they use metadata from the photo or video at the moment it is captured and use cryptography to protect the content from tampering. The company does not aim to detect tampered photo's and videos as they do not believe that it is not a viable or scalable option [21]. The technology offered by the company is already being used by more than 100 companies such as financial service providers, online marketplaces, real estate companies, insurance agencies and many more. According to the CEO, any company which relies on visual media for daily business processes can benefit from Truepics technology [21]. This shows that content authentication is a viable alternative to detection and it seems as if the use cases and technology will only continue to grow. It also helps that the service offered by Truepic is a lot broader than just deepfakes as Truepic can detect any manipulations to images and videos. Since media manipulation is a growing risk for corporations, truepics software will play an important part in many corporations fraud management strategy. It might also proof to be a valuable tool in the fight against disinformation in the future.

In 2020, Microsoft released a video authenticator tool aimed at detecting deepfakes. The software was not released publicly but was offered through a third-party organization to news agencies and political campaigns free of charge. This was done to prevent the detection software to fall into the hands of deepfake creators which could use it to avoid detection when creating a new deepfake creation technique. The tool was just meant to be used ahead of the 2020 US election to prevent the spread of disinformation and the sight offering the software seems to be online. However, Microsoft have said that they will continue to develop the software and techniques to combat deepfakes in the future [22].

Zemana, a Turkish anti-virus software company has released an open-source deepfake detection tool called Deepware. The tool allows you to upload or input a link to a video and the tool will tell you whether it is a deepfake or not. The company has made the tool free-to-use and open source as they feel communities should be working together to solve the deepfake detection problem [23].

Defudger is a Deepfake detection company that offers content authentication, deepfake detection, and keeps track of previously detected deepfakes. They offer a SaaS-package to news agencies and digital platforms at a rate of $2500[24].

2.4.3 *Movie studio's*

Movie studios are obvious actors for legitimate deepfake use. In 2016, disney studios released Rogue One: A Star Wars Story in which deceased actors were brought back to life using CGI models on which special effects artists worked many months to create only a few seconds of footage. Even after these many hours of work viewers of the movie were still critical of the 'uncanny valley' effect that the CGI models evoked. Which is the effect that trying to recreate human faces and behaviour leads to Another example of this could be seen in the Mandalorian by Disney studios were the actor playing Luke Skywalker was de-aged to play his younger self using CGI [25]. While fans were excited of the CGI, deepfake maker Shamook uploaded his own version of the scene using deepfake technology to enhance the realism after only a few days. This problem could be solved by using high quality deepfakes, saving months of time in the editing process and allowing for more realistic faces. Recently, Disney's own research studio has published a paper examining the possible uses of deepfake in movies and proposing an algorithm for fully automated neural face swapping in images and videos [26]. This shows that movie studio's are at least interested in the technology.

Another company in the movie industry is Flawless, a start-up that uses deepfake technology that aims to retain the original performance of the actors when dubbing the movie in another language. Using AI in this way can save a lot of costs and create a more realistic dub, allowing for the movies to be available to a wider audience while remaining as immersive as the original [27].

However, deepfake technology in film still has controversial applications. In 2020, a director planned to use a deepfake of the deceased actor James Dean in one of his new movies [28]. This would entail a fully new movie which has no continuity issues due to an actors age, but would use the image of a famous actor in a new role. This has obvious moral and legal implications, such as the fact that the actor itself cannot consent to his image being used in this new film. As well as the fact that the character in the movie could be played by another living actor. This would allow movies to be created without expensive actors, by replacing living actors with doubles that will be deepfaked over later, diminishing the chances for upcoming actors to be discovered and creating a career [28].

2.4.4 *Scammers*

Scamming people is a big business, the introduction of the internet only broadened the possibilities for malicious actors to execute their scams. Scammers have a massive financial incentive in the advancement of deepfake technology and to take the time to learn the ins and outs of the tools that are used to create them. AI technologies such as Deepfakes and especially audio deepfakes opened up new ventures for scammers to invest in. In the first chapter we already established that deepfakes were used to steal large sums of money from big companies by using audio deepfakes to imitate the company director. The existence of these technologies allows for new and improved scams. Currently, the lack of good quality audio deepfakes means that these scams are relatively rare and don't end up successful most of the time[29]. A report from network company RSA estimated that in 2018, 1% [30] of all phishing scams involved some form of audio deepfake. However, it is likely that since the technology has continued to developed, it is also being used more.

According to Panda Security [29] center, there are several types of scams that can use deepfakes. Examples this are ghost fraud, new account fraud and new identity fraud. In ghost fraud, the scammer steals the data of a deceased person to impersonate them. They can then use this to access financial services of the deceased for financial gain. New account fraud defines the fraud where criminals use stolen or synthetic identities to open new bank accounts. They can then use this for financial gain by maxing out credit cards or applying for loans that won't be paid back. In new identity fraud, the criminal does not steal a certain persons identity but creates an entirely new identity of a person who does not actually exist. They can use this identity in any sort of way they choose, such as opening new bank accounts. An example of this can be seen in Vincent [31], where a spy used fake LinkedIn accounts to create the look of a legitimate network to try and infiltrate professional networks in order to steal information.

Scammers use deepfakes not only to impersonate an individual for financial gain, but also use it to create material to blackmail them with. In India scammers are using deepfakes in video calls to pose as women wanting to have sex with the victim and entice the victim to masturbate during the video call. Afterwards the victim will receive an extortion call demanding money or the scammers will release the video of the victim to his friends and family. It is unclear how widespread this scam is, since most people who get scammed are ashamed to speak up about it. However a 46-year old businessman from India did not do anything obscene during the video call and still got the extortion call, he decided to make the story public [32]. If the victim does not cooperate, the scammers sometimes use the face of the person in the video chat and deepfake them into pornographic videos. Which is then sent to the victim's friends and family if the victim does not pay. According to the Indian police, these types of scams happen hundreds of times a day as they get a lot of these cases every day [33].

2.4.5 *Disinformation Agents*

Actors which are definitely operating in illegitimate realm are the individuals and organizations using deepfakes to spread online disinformation. The motivations these actors have to spread this disinformation can vary wildly, but most seem to have some sort of financial incentive to their agenda. An example of this is vaccine disinformation, the center for countering digital hate (CCDH) found in a report on vaccine disinformation that just 12 people were responsible for 65% of all vaccine disinformation [34]. The disinformation dozen, as the CCDH has dubbed them, had a combined following of over 59 million across multiple platforms such as twitter and Facebook. The incentives of the disinformation dozen are varied, but most of the actors have a financial incentive by selling alternative medication to COVID-19 [34]. Some of the actors are not selling anything, but still gain a larger following by spreading this disinformation, which they can benefit from financially. Other actors just seem to spread the disinformation because of their political ideology or personal beliefs, although there is no checking whether these actors actually believe the disinformation they generate.

Deepfakes may be used to spread disinformation, but as a report by Hwang [35] found, most disinformation actors are pragmatists and will spread disinformation with minimal effort. This currently seems to be just as effective as more elaborate campaigns, so deepfakes are not that common in the spread of disinformation. Realism does not equate to a more believable message. In fact, disinformation campaigns that use low quality editing such as the slowed down version of Nancy Pelosi appearing inebriated were very successful despite the low-quality nature of the disinformation Hwang [35]. However, as machine learning technology continues to develop and deepfakes can be created with a lot less effort than now, they might take a presence in the future of online disinformation.

3 METHODOLOGY

"A goal without a plan is just a wish."

— Antoine de Saint-Exupéry

This chapter will cover the research methodology. This research operates in the realm of design science, the research goal reflects that. The goal is to improve a problem using a designed model. We are not looking for an objective statistically supported truth like in knowledge based research. An artefact is developed and applied in a problem context based on subjective truths and opinions of experts. So according to the principles of design science outlined in Wieringa [36], we are dealing with a design problem.

3.1 DESIGN SCIENCE RESEARCH

Since the research area is more of a design problem than a knowledge problem, the methodology for gathering information surrounding the topic has to reflect that. We can reformulate the research problem into a design science problem by using the methodology suggested in Wieringa [36]. The original research question: "How do we design a forensic image authenticity research maturity model that reflects current and future needs of forensic organisations?" will be translated using a template. The template may not be fully filled in at the start of the project since some of the information may be missing. The original research question might also be translated into multiple design problems. The template is as follows:

- Improve <a problem context>
- by <(re)designing an artefact>
- that satisfies <some requirements>
- so that <stakeholder goals>.

If we translate our research problem into this template, we would get the following:

- Improve <*the efficiency of the analysis of video manipulation*>
- by <*designing a Maturity Model for forensic image authentication*>
- that satisfies <*validation from experts in the field*>
- so that <*the NFI can measure their maturity in the image authentication process*>

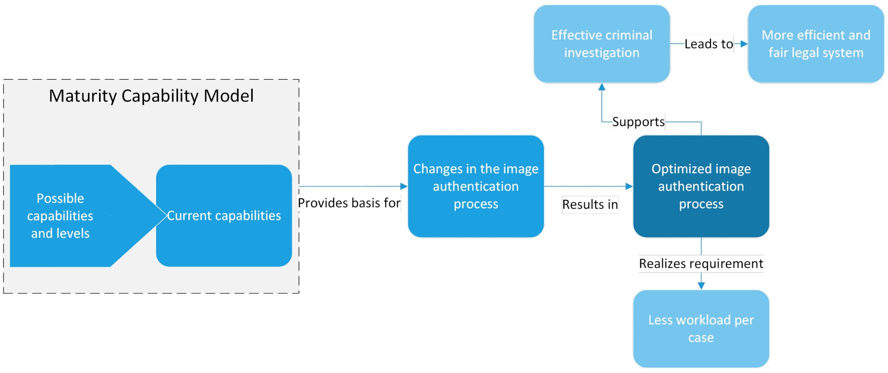
Maturity Capability Model
Possible capabilities and levels
Current capabilities
Provides basis for
Changes in the image authentication process
Results in
Optimized image authentication process
Supports
Effective criminal investigation
Leads to
More efficient and fair legal system
Realizes requirement
Less workload per case

Figure 2: Conceptual research framework

3.1.1 *Artefact*

The artefact as outlined in the previous section is a maturity model that models the capabilities that are needed in the entire chain of evidence analysis to warrant a certain level of maturity. The level of maturity reflects the readiness to deal with the new influx of manipulation claims, and would account for the different types of evidence that would need to be analyzed. Reaching a higher level of maturity in the model would also increase the efficiency of the entire process. This will provide the NFI, it's partners and similar forensic research institutions a roadmap with goals that can be achieved when the necessity arises for a more efficient IA process. And new capabilities to reach a new level of maturity could be implemented incrementally over the entire evidence investigation chain. The maturity model also includes a benchmark to show the level that the IA process is currently at by filling out the respective assessment matrix.

3.1.2 *Conceptual research framework*

Wieringa [36] also describes creating a conceptual framework of the research to contextualize the problem, the proposed artefact, and the interaction of the artefact on the problem. Realizing a conceptual framework will make the purpose of the artefact in the specific context it is applied in more clear. The conceptual research framework can be seen in Figure 2. The maturity model will map the current capabilities and the level of maturity of the NFI in addition to capabilities needed to reach a higher level of maturity.

3.2 RESEARCH DESIGN

Since the artefact for this design problem is a maturity capability model, the specific steps in the design science method have to reflect that. Maturity capability models can fall victim to being based

Define Scope	Design model	Evaluate model	Reflect evolution
Focus/breadth	Maturity definition	Subject of evaluation	Subject of change
Level of Analysis	Goal function	Time-frame	Frequency
Novelty	Design process	Evaluation method	Structure of change
Audience	Design product		
Dissemination	Respondents		
	Application method		

Table 1: Decision parameters maturity model development [1]

on a poor theoretical basis and without sufficient evidence to justify their use [1]. In order to solve this issue several different methodologies have been proposed to help with these issues when developing maturity models. Therefore, this research will follow the design science method for maturity models by Mettler [1]. This methodology is commonly used in maturity model development and has proven to be an effective method at producing high quality maturity models. Mettler [1] describes four phases of the design science process.

1. Design scope
2. Design Model
3. Evaluate model
4. Reflect evolution

Mettler [1] also describes the decision parameters that are relevant within each phase. They help the researcher in creating a stable framework to base the maturity model on. Each decision parameter requires the researcher to decide between 2-4 options to narrow down the research framework. For example, in the first phase, the decision parameter focus/breadth has the options: "General issue" vs "specific issue" [1]. If we apply these decisions to our problem, the general focus would be "Digital forensics in forensic organizations", while the specific issue focus would be "Forensic image authentication research in forensic organizations". Since the specific issue reflects the problem better, and there are already a few maturity-like models for digital forensics available [37], [38], the focus/breadth decision falls on the specific issue parameter. This process is repeated for all decision parameters.

The decision parameters described in Mettler [1] can be seen in Table 1. The options of the decision parameters are not visible in the table but will be elaborated on further.

The first phase, define scope, allows the researcher to narrow down the scope of the maturity model. The scope was partly determined in the problem investigation in the introduction, but will be further constrained here. In this case, the focus of this maturity model is providing forensic organizations an overview of capabilities needed for a mature forensic image authentication process. The focus is therefore a specific issue forensic image authentication research in forensic organization. The level of analysis is group-decision making, the novelty is emerging, the audience is both management oriented as technology oriented, and the dissemination is open.

The second phase covers the design of the model. This phase will be split up into two sections. One of the sections will cover the theoretical basis for the model from the literature, while the other section will be from interviews with stakeholders and experts that are working in digital forensics and deal with image authentication. Each section will cover a chapter. The decision parameters for this phase are as follows.

The definition of Maturity will be a combination of process focused maturity and people focused maturity. Increasing the efficiency of the process is important, since a reduction in time spent in the process leads to better throughput times, and this will lead to an increase in process focused maturity. Increasing the knowledge of the people that work in the process is also important, since the effectiveness of the investigation is directly linked to the knowledge of the people working in the process. Increasing knowledge in this area will lead to more people focused maturity. The goal function is multidimensional, since the goal is to both increase the efficiency of the image authentication process while also increasing it's credibility. The design process will be a combination of theory-driven design and practitioner based design, since practitioners can give a lot of valuable information about the process, but might lack a vision of the possible capabilities needed, these can be retrieved from the literature.

The literature can also support the ideas and statements of the practitioners. The design product will be a textual description of form and functioning. The application method of the data collection is based on self assessment. And the people involved in the design process are staff and business partners of the NFI. A full overview of the decision parameters chosen in this phase can be seen in Table 2

1. The first part is an integrated literature review looking at existing maturity models in digital forensics. In addition, the literature will also be reviewed for capabilities related to image authentication research and technologies that might aid in this process. This section of the research and the specific methodology for determining which papers are included can be found in Chapter 4.

2. The second part of this phase will cover semi structured interviews with stakeholders who are involved in the image authentication process at the NFI and will include additional ex-

Design model	**Decision parameter**
Maturity definition	Combination of process-focussed and people focssed
Goal function	Multi-dimensional
Design process	Combination of practioner-based and theory-based
Design product	Textual description
Respondents	Staff & Business Partners
Application method	Data collection based on self assessment

Table 2: Decision parameters Design Model Phase [1]

perts who work with image research and deepfakes. The interviews will mostly involve gathering knowledge about the current image authentication process and the capabilities within them. This can be used to determine the current level of knowledge within the process and will look into the possibilities for technological capabilities. The type of interview will be semi-structured, this allows for asking standardized questions to achieve consensus between different experts, while also allowing for asking spontaneous questions that arise during the interview to dig deeper into a subject the interviewee is knowledgeable about. The main goal of the interview is to gather information about the problems within image authentication research and deepfakes and their avenues to solve them. The full interview protocol, along with an analysis of the interview materials can be found in Chapter 5.

The third phase is the model evaluation. According to Mettler [1], this phase is concerned with the verification and validation of the maturity model. The process of verification is determining that the maturity model represents the conceptual description of the model [1], and validation is determining whether the model represents the real-world scenario in which its relevant [36]. The validation of the model is done through an interview with an expert on the image authentication process. This phase can be found in Chapter 7.

In the fourth phase, the evolution of the model and the development of the model will be reflected upon. This phase is outside of the scope of this research, although some reflection of the development of the model can be found in Chapter 9.

A full overview of the phases and which chapters covers them can be seen in Figure 3.

3.3 INTEGRATIVE LITERATURE REVIEW

The integrated literature review in Chapter 4 is a literature review based on the systematic literature review (SLR) method by Kitchenham and Charters [39]. The paper describes the following three stages

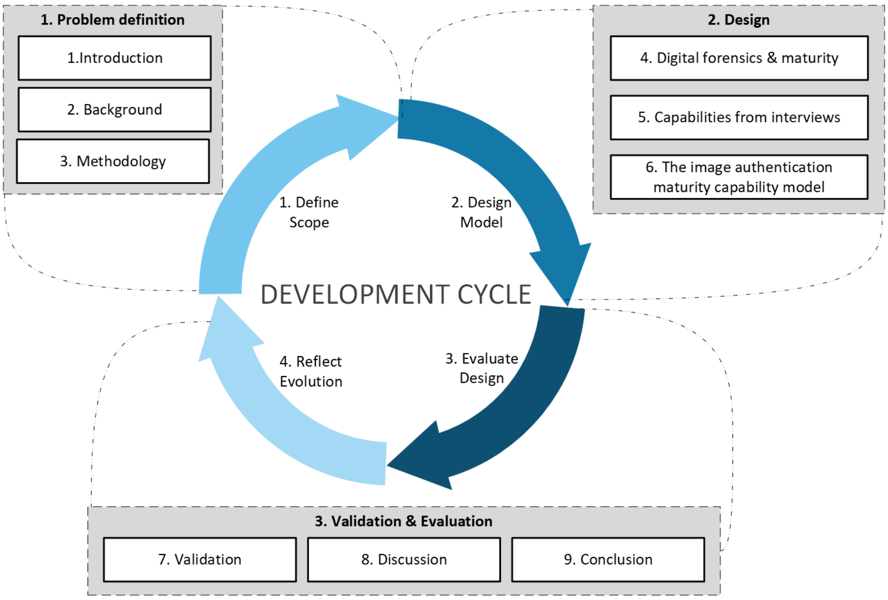

Figure 3: Development cycle adapted from Mettler [1]

of an SLR; planning, conducting and documentation. First we define the research topic in the planning stage along with the review protocol. In the conducting stage we define the inclusion and exclusion criteria, select the article databases and execute the search according to the protocol. The topic that is explored in this literature review is maturity models in digital forensics. The guiding question is: *How is a maturity model constructed for business processes in digital forensic science?*

3.3.1 *Search strategy*

For the search process a the following research databases will be used:

- Scopus
- Web of Science

Each search consists of keywords and will be queried on each database. The search of keywords is focused around the Title, Abstract and keywords. The search phrases are in order of the research questions. **Query**: *WoS:(TS=(Maturity) OR TS=("capability model")) AND (TS=("digital forensic*") OR TS=("image authentication") NOT TS=(readiness))*

3.3.2 *Inclusion & Exclusion Criteria*

Kitchenham and Charters [39] suggested using the following inclusion criteria, which were chosen because of their relevance. The inclusion and exclusion criteria were based on reading the title, keywords, and abstract.

	RQ1
Scopus	13
WoS	13
Total	26

Table 3: Study selection

- The article is in English
- The article is peer-reviewed and published
- The article is from a journal or conference proceeding
- The article provides an answer to a research question

The exclusion criteria are as follows

- The article is not available on the University of Twente library
- There is a newer version of the same paper
- The article is not relevant to the research question

To apply the inclusion and exclusion criteria, the title and abstract of each paper are screened. In addition, if the title and abstract are not conclusive to either inclusion or exclusion, the introduction is also screened.

3.3.2.1 *Study selection*

As previously outlined, the selection of articles is performed following the defined steps. The process is described below and the results can be seen in Table 3. The searches were performed between the 4th of June 2022 and the 7th of June 2022.

1. The searches are performed on the selected databases.
2. 26 articles were found, after the removal of duplicates we are left with 23 distinct articles
3. After the application of the exclusion and inclusion criteria, 15 articles were removed. 1 article was not available through the university library, 2 articles were not in English, and 12 articles did not provide an answer to a research question.

4. After applying the inclusion and exclusion criteria. A total of 8 articles were left for the quality assessment.
5. 2 articles were later added from outside of the search string as they were deemed relevant to the topic.

3.3.2.2 *Quality assessment*

Research articles will be assessed for quality through a set of quality metrics. We decided to use Kitchenham and Charters [39] as a guide and use their set of quality metrics for assessing the quality of a paper.

An article with no clear purpose will not lead to an informative conclusion and might not add anything to existing literature. If the author defines the goal or purpose of the article in the introduction, such as for example: To explore the effects of deepfake on online disinformation. Then the article can clearly work towards that goal. In contrast to research that just states that it's doing research into deepfakes and online disinformation. Which cannot be properly scoped. Therefore, quality assessment question 1 is as follows:

QA1: Is the purpose of the research clear?

Some articles are written by foreign researchers whose native language is not English. Most authors can write English in a grammatically correct and understandable way. However, sometimes the information in the article is unable to be properly understood due too poor English. This can be missed in the inclusion exclusion part of the research since only the title, abstract, and keywords are read. Therefore the following quality assessment metric was also included:

QA2: Is the text written grammatically correct and clear?

Finally, one of the important aspects for determining the quality and usefulness of an article is whether the author addresses the questions or achieves the goal they set originally for the article. This does not mean that the question has to be answered in the article. But the author at least has to call back to the original goal and conclude whether the research was successful in achieving this goal.

QA3: How well does the evaluation address the original aims and purpose? Does the author achieve the goal he set out to do?

3.4 INTERVIEW PROTOCOL

This section of the methodology describes the interview protocol used for the interviews for the analysis of Chapter 5. The results of the analysis of the interviews can be found in this chapter, while the interview notes of the interviews themselves can be found in the appendix. .

3.4.1 *Research Frame*

Due to the complexity of the problem and the little amount of literature available on the topic of maturity models in image authentication. It was deemed necessary to collect data from outside of the literature. Since it is a complex process in which many different actors are involved. These actors all have their own perspective on the business process and will be able to give an insight into how the process works and how the process of image authentication is supposed to work.

The knowledge about the process is also currently too low to craft a questionnaire [40], in addition the actors that are involved in the process are in relatively high functions in their organization and are therefore unlikely to respond to a questionnaire [40]. Therefore a series of interviews will most likely yield the best results as the circumstances of the research call for qualitative data gathering methods and the potential respondents are more likely to be responsive to interviews than to other methods of qualitative data gathering.

3.4.2 *Guiding Interview Questions*

1. What are the current capabilities of the NFI regarding the image authentication process?
2. What capabilities should a forensic institution have to have an efficient image authentication process?
3. What are the options for improving the capabilities of forensic institutions?

3.4.3 *Interview design*

The interviews are meant to create a good basis of knowledge surrounding the image authentication process to create a model of the process and find the requirements of the capabilities for the maturity model. It is important to get insights and opinions of as many stakeholders as possible in order to create a good basis of knowledge to base the model on.

3.4.3.1 *Documentation*

The interviews will be documented using interview notes instead of a full transcription. A full transcription was deemed to be unnecessary for the purpose of this research. Since the interviews are mostly to extract knowledge from the experts, and sentiment analysis or a method where subtle

Role & Organization	Goals of the interview
Digital Forensic Image Expert(NFI)	- Get current capabilities of the NFI - Find out about the current image authentication process - Explore options for the needed capabilities
Digital Forensic Image expert(NFI)	- Expand the view of the current process and capabilities - Get his view on the capabilities needed in the final process
Digital Forensic Image Expert(NFI)	- Confirm capabilities of the previous interviews
Strategic specialist Digital Transformation (Dutch police)	- Discuss the digital crime strategy of the police - Figure out the conditions for when evidence is sent to the NFI - Discuss how to deal with deepfakes on a wider level - Figure out the extend of the evidence analysis capabilities of the police
Law Professional (Court of The Hague and Cybercrime knowledge centre)	- Requirements for the processing time of a piece of evidence - Explore how often the manipulation defense is used - Explore what the level of knowledge is surrounding the manipulation defense
Public Prosecutor (Public Ministry)	- Explore the input of the image authentication process - Explore which cases need to be analyzed
Deepfake detection professional (Duckduckgoose)	- Figure out the future of deepfake detection - Find out whether deepfake detection might be usable for forensic image authentication - Explore the level of explainability that the software will have - Explore the future accuracy of deepfake technology
Deepfake detection professional	- Get his perspective on the future of deepfake detection - Discuss the application in forensic image authentication
Forensic image analyser (Dutch Police)	- Figure out the image authentication process at the police - Discuss bottlenecks - Discuss the strategy that is used when analysing evidence

Table 4: Approached Respondents

3.4.3.2 *Interviewee selection*

The interviewee's are selected based on their level of expertise and their experience and weather they were involved in the image authentication process. The full list of interviewee's can be seen in Table 4. Stakeholders who are dependent on the results of the image authentication or are involved in the process will be approached. The interviewee's will be approached through email, an example of the email that is sent can be seen at the end of Appendix A. In total 14 people were approached with the question to do an interview, of which 10 people responded. In total 9 people were willing to participate in the interview.

3.4.4 *Interview Length*

Rowley [40] states that researchers should aim for around 12 interviews of around 30 minutes or 6-8 interviews of around an hour to reach data saturation. Since the topic is very complex and some actors will know varying amounts of information about the process, 30 minutes to both answer the pre-determined questions and the spontaneous questions that arrive from context information seems to be too short. Therefore, interviews are expected to take around 45 minutes with a 10 minute margin to cover about 12-15 questions from each respondent. Since this is longer than 30 minutes and each respondent is expected to have more relevant expert knowledge in the process, 9 interviews with respondents from different domains is considered sufficient. The interviews will mostly be held online as this is becoming the norm.

3.4.5 *Interviews*

The interviews were multi-purpose, and the results of the interviews will be used in answering multiple research questions. The interviews were used in the problem investigation, process investigation, and proposing and extracting capabilities needed in the maturity model. The capabilities extracted during the interviews and those found in the literature will both be used in the maturity model.

The interviews were conducted with three groups of practitioners, the practitioners are all considered stakeholders of an image authentication maturity model:

- The first group consisted of forensic image specialists and researchers in the digital forensic domain. These practitioners are referred to as "image researchers".
- The second group included people with knowledge of the viability of deepfake detection. With stakeholders that held views from both sides of the argument. Referred to as "digital innovation specialists".
- The last group included people with knowledge of the law surrounding deepfakes and the problem with evidence analysis at this level. These are referred to as "Law practitioners"

The type of information that is extracted from each of these groups is different. The focus of the interviews in the first group is threefold:

1. Figuring out how big the problem of deepfakes is
2. What does the current process look like, what are the bottlenecks? What kind of capabilities are there? How often is the process revised?
3. What possibilities for improvements are there?

The focus of the interviews in the second group was figuring out the viability of using deepfake detection in a forensic image authentication setting. The digital innovation specialists were knowledgeable in this field and were able to provide an insight into the practicality of implementing these techniques in different types of situations.

The focus of the interviews in the last group was figuring out how big the deepfake problem is from a law perspective. These interviews also tried to find out how the process works outside of what the NFI does in the process of investigating evidence.

3.4.6 *Interview implementation*

This section gives a brief description of the data collection and analysis methodology to ensure that the data is collected and analyzed in an unbiased manner.

3.4.6.1 *Data collection*

The data retrieved from the interviews will be analyzed based on the interview notes. In order to manage time constrains of the thesis, interview notes are used in data collection instead of a full transcript. The use of interview notes instead of a full transcript also allows for the data to be analyzed quicker, since the data is already filtered for the most important information [41]. A full transcript of the interviews is also unnecessary, since the interviews are mostly to gather information about the current process of digital image authentication, and not to gather opinions and feelings about this process. In which case a sentiment analysis of the interview data would make more sense.

To facilitate the taking of interview notes, the interviews were also recorded, so that the researcher could listen back after the interview and ensure that the notes are correct. Since it is difficult to write notes during the interview, as it might interrupt the flow of communication with the respondent [41], and the interviewer might miss details of the conversation. This problem is alleviated by creating a summary of the interview right after it is done, and rewriting the notes along with the audio recording.

3.4.6.2 *Data Analysis*

After the interview notes have been documented, the notes needed to be analysed to find the themes in the data. The specific themes that are going to be investigated will be discussed in Chapter 5. There are 2 types of interviews that have been done; Interviews with participants who are directly involved in the IA process (NFI employees), and interviews with participants who are indirectly involved (business partners). The purpose of the interviews in each situation is different, so the interview questions and topics that arise are also different. The interviews notes were analyzed retrospectively since the researcher did not determine the data analysis method beforehand. The methodology can be seen as a form of informal thematic analysis [42]. The researcher determined the themes that were important in the interviews, and linked each question to a certain theme. Afterwards, the answers from the respondents were informally read to extract information regarding the categorized topic. A few interesting quotes were extracted as well during this process to highlight the insights and opinions of the respondents regarding that topic. The results of this process can be seen in Chapter 5, where each topic is explored in depth.

3.5 SUMMARY

This chapter covered the methodology that is followed in this paper. The next chapter will feature the integrative literature review which methodology is described in this chapter. The chapter after that will cover the interviews an the subsequent chapter will cover the design of the model. All of the methodology that is used in these chapters can be found in this chapter.

Part II

DESIGN PHASE

This part describes the second phase of the Thesis in which the main work was done. It includes the literature review in the second chapter, the methodology of the project in the third chapter, the results from the interviews in the fourth chapter, and the design and validation of the model in the sixth chapter.

4 INTEGRATED LITERATURE REVIEW

"Study the past if you would define the future."
— Confucius

4.1 MATURITY MODELS IN DIGITAL FORENSICS

This section of the integrated literature review looks at the literature on Maturity models in the context of digital investigation. The goal is to get an insight into other maturity models in this area and find out how they are constructed, what capabilities are often present in the literature, and find a theoretical basis for the construction of a capability maturity model for image authentication. An overview of the papers and models discussed in this section can be seen in Table 5 at the end of the chapter.

4.1.1 *Capability Maturity Models*

In Paulk et al. [43] the first comprehensive capability maturity model (CMM) is presented. It is a framework for software organizations and was used to measure the maturity of the software development process and to provide these organizations the tools to continuously improve their capabilities in this area. It allows organizations to move from ad-hoc processes to highly structured and effective processes.

Paulk et al. [43] describes that in immature organizations, (software) processes are generally improvised, even if the correct processes are defined. Theses organizations have no methods to measure the product quality and throughput times of processes are hard to predict. The author says that mature organizations have the ability to monitor and manage their processes on an organizational level. The processes are standardized and any changes are documented, the costs are known since historical data on the process is available and the organization can effectively plan for future improvements. The CMM allows organizations to measure their maturity of in this case the software development process, and gives them the tools to reach higher levels of maturity. Each level in the maturity model establishes different components of a process, which results in an increase in the capability of an organization.

The CMM is comprised of five different aspects [43]:

- `Maturity levels`: There are five levels of maturity in the model, the level of maturity of an organization is an indication of the process capability of an organization.
- `Key Process areas:` Maturity levels consist of key process areas show the areas in which an organization should focus it's improvement efforts. The areas identify the issues that have to be tackled to reach a certain maturity level. The key process areas outline goals that are considered to be important to reach maturity in this area. A capability is defined as the level of maturity of an organization in that specific key process area. The level of maturity of an organization is the average of the maturity of all capabilities in the key process area's[43].
- `Common features:` According to the authors, common features are the attributes that are indicative of whether the implementation of a key process area is effective, repeatable, and lasting. Each key process area has the five common features:
    1. Commitment to perform
    2. Ability to perform
    3. Activities performed
    4. Measurement and analysis
    5. Verifying implementation
- `Key practices`: Each key process area consists of key practices that contribute to satisfying the goals of the key process area. They describe infrastructure and activities that help achieve the key process area. The key practices only describe what to do, not how to do it.
- `Goals:`Goals are the summary of key practices, and determine whether an organization has successfully implemented a key process area.

The CMM outlines the 5 levels of maturity in order: Initial, repeatable, defined, managed, and optimizing. These same levels can be applied in other maturity models although the names and descriptions for them may vary slightly.

INITIAL At the initial level, an organization does not provide a stable environment for supporting it's business processes. The processes are undefined and business processes are executed ad-hoc. The success of these processes is inherently dependent on the competence and experience of the people doing them. If they would to leave the company, these processes would no longer be able to be finished. The business processes in these organizations cannot be repeated unless the same competent person is doing them. This often leads to delays in projects and unpredictable costs. Level one has no key process area's to be aware of.

REPEATABLE At this level the policies for managing the processes are established and procedures on how to implement them are also established. Processes at this level have basic management controls and commitments are made based on the results of previous processes. Management processes are able to be monitored and can be executed by anyone who has access to the documentation. There is some variance in the throughput times of cases but significantly less than in the ad-hoc situation. Key process areas at this level are focused on basic project-management controls. The ones described in the article are specific to software development so it's not relevant to mention all of them here [43].

DEFINED At this level, in the case of this paper, the software processes for developing and maintaining software are also documented next to the management processes [43]. These processes are also integrated into a coherent whole business process. This standardization of the full spectrum of different processes within an organization allows the organization to work more efficiently and predictably. The variability of project times goes down significantly, but large outliers are still present. The key process areas at level 3 are aimed at solving project and organizational issues [43].

MANAGED At level 4, the organization sets quantitative quality goals for the processes and the products that come from them. It also equips processes with well-defined measurement instruments to continuously monitor the efficiency and quality of there processes. All products that are produced are of a predictable high-quality. Data about the processes is automatically generated and collected and is used t. The throughput times of processes goes down and any variation in the throughput times is minimal, any meaningful outliers can be investigated so that they can be distinguished from random outliers. The capability at this level can be described as being quantifiable and predictable, since the processes are measured and are operated within measurable limits. The key process areas at this level focus on creating a quantitative understanding of the processes in the organization [43].

OPTIMIZING At level 5, the entirety of an organization is focused on continuously improving the business processes. The organization can identify weaknesses in their business processes. Data generated in the business processes is used for cost-benefit analysis of new technologies to add to the processes. Teams in optimizing organizations are self-improving and can determine the causes of defects. At this level of maturity, the waste of resources in unacceptable and the goal is to remove any waste from the process. The capability of mature organizations can be described as continuously improving. The key process areas at this level are focused on organizational and project issues that are in the way of continuous improvement [43].

Applying these levels in a digital forensics process, specifically the image authentication process, does not translate one-to-one. A deep dive into the specifics of the process and how to achieve maturity in this process at different points in the process remains a challenge. We can take a look at other maturity models that have been created in the domain of digital forensics for inspiration on how to apply the concepts of the CMM in the context of a digital forensics process.

Proença and Borbinha [44] found that modern state of the art maturity models are mostly focused on highly complex and specialized tasks, that are being performed by competent assessors in an organizational context. Image authentication done by image experts at the NFI fit the description of this kind of process perfectly.

4.1.2 *Maturity in Digital investigations*

Kerrigan [45] summarizes multiple standards and best practices for digital investigations and combines the most important parts into a single maturity model. The author noted that while a lot of standards were available to guide digital investigations. No maturity models to assess the readiness of an organization to execute digital investigations were available. The author looks into all of the standardized models for digital forensics that were available at the time. Most notably the ENFSI best practices manual. The author also describes a foregone patent by Krutz [46] for a computer forensics Capability maturity model(CMM). This model also described a series of processes and best practices for performing digital forensic investigations on the computer, but is limited in its approach to forensic investigation. So it should only be regarded as part of a bigger digital investigations CMM.

The ENFSI best practices manual that is examined in this article is different from the one explored later in this literature review.

Kerrigan [45] then proposes the Digital Investigation capability maturity model. The Maturity model is aimed at forensic organizations and other organizations that want to measure their current capabilities for digital investigations. The CMM, like most maturity models, consists of 5 maturity levels. From the bottom to the top: Ad-Hoc, Performed, Defined, Managed, and Optimized. At the first level, most of the digital investigations happen in an ad-hoc manner, meaning that there are no formal processes for digital investigations. At the second level, the digital forensic team is seen as a provider of a technical service. At the third level, the digital forensics team has developed a track record of providing good services and are viewed as experts

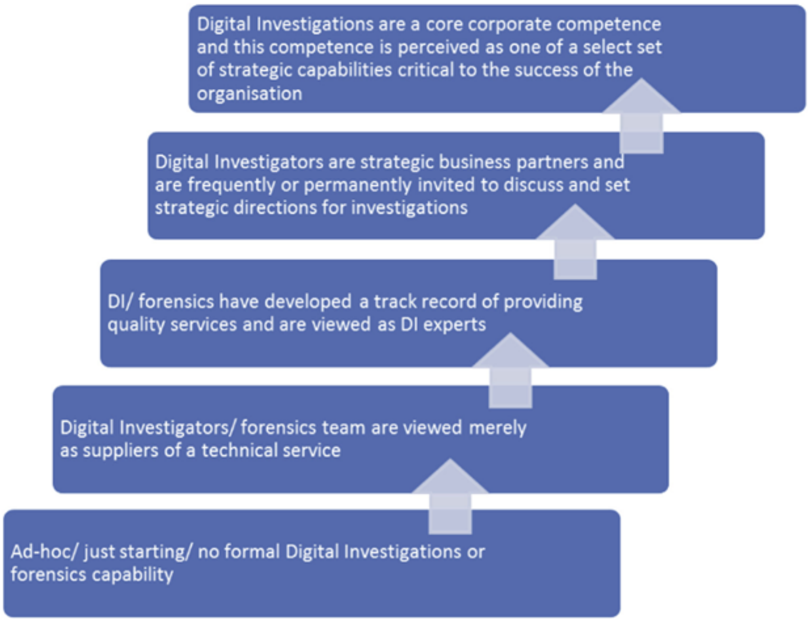
Digital Investigations are a core corporate competence and this competence is perceived as one of a select set of strategic capabilities critical to the success of the organisation
Digital Investigators are strategic business partners and are frequently or permanently invited to discuss and set strategic directions for investigations
DI/ forensics have developed a track record of providing quality services and are viewed as DI experts
Digital Investigators/ forensics team are viewed merely as suppliers of a technical service
Ad-hoc/ just starting/ no formal Digital Investigations or forensics capability

Figure 4: 5 levels of CMM by Kerrigan [45]

within the organization. At the fourth level, digital investigators are seen as integral business partners and take part in the planning of strategic directions of investigations. At the final level, digital investigations are viewed as a capability critical to the competitiveness and success of a forensic organization. Every digital investigation department should strive to attain this level within their organization. The full levels of the CMM by Kerrigan [45] can be seen in Figure 4.

`Example:` *readiness can be improved by making sure traces are left when systems are accessed.*

Kerrigan [45] provides a good maturity model to improve the digital forensic readiness in organizations other than forensic organizations themselves. Readiness in this case refers to an organizations capacity to do digital forensic investigations internally. Either through an outside company or internal forensics team. So many of the capabilities outlined in this paper do not reflect the capabilities needed in digital forensics organizations themselves.

4.1.3 *DFaaS*

In Van Baar, Beek, and Van Eijk [47] a new way of offering digital forensics services is presented. The method has been implemented at the Netherlands Forensic Institute 3 years prior to the paper and the authors are reporting on it's effects on the organization in this paper. They outline the method as Digital Forensics As a service (DFaaS), which is supposed to be more efficient than traditional digital forensics methods. The full traditional forensic investigation steps can be seen in Figure 5.

The authors analyze the traditional process in which they found a number of factors which have a high impact on efficiency. Resource management is a big factor since digital investigators used to be responsible for keeping up with data storage, back-ups, security, system administration, imaging disks, network captures, unknown file research etc. While it is useful that the digital investigator is capable of doing all these things, it is not useful to spend a lot of time doing menial tasks that could also be spent doing forensic research.

The second thing is questions, according to Van Baar, Beek, and Van Eijk [47] there are three types of questions that could be asked about evidence. The first type is not really a question, it's just a request for general information. Such as for example: "Give me all information related to drugs" [47]. These types of questions lead to a lot of work for the investigator, since the implication of the question is not clear, and a lot of the question is left undefined. The next type of question is still kind of vague, but is a bit more specific than the first type of question. For example, questions like: "What was the suspect looking for?" or "What are the origins of this document". While these questions narrow down the search a bit, the investigator still has a lot left over for interpretation and work. With the last type of question, the customer has a specific hypothesis that they want tested. An example would be "Was the video taken with this type of digital camera?". This type of question has a clear research goal and can be answered using a statistical test using a null hypothesis.

While this last type of question allows the researcher to work efficiently on a problem, it also narrows the window of discovery. The digital investigator usually only has the evidence and is missing much of the case context. They might come across court admissible evidence, but since the right question was not asked, will not be able to identify it as such.

Another factor that determines the efficiency of the process according to the authors is the time frame. Within the first days of the investigation usually the hypotheses are formed and tested. Usually this is not enough time for the digital investigator to do a full investigation. The amount of data is growing much harder than the amount of digital investigators, so the digital investigators are always busy with the first steps of the investigation, namely collection and authentication. When the digital investigator is done with these steps, the next case is already a high priority, leaving little time to actually do any forensic investigation.

Collaboration is also identified by the authors as being difficult in the traditional digital forensics process. While collaboration between digital investigators can be very valuable. The last factors which impacts efficiency is the R&D of digital investigators. Some digital investigators might have to do research for a specific problem, the knowledge gained through this research can be useful in other cases, but

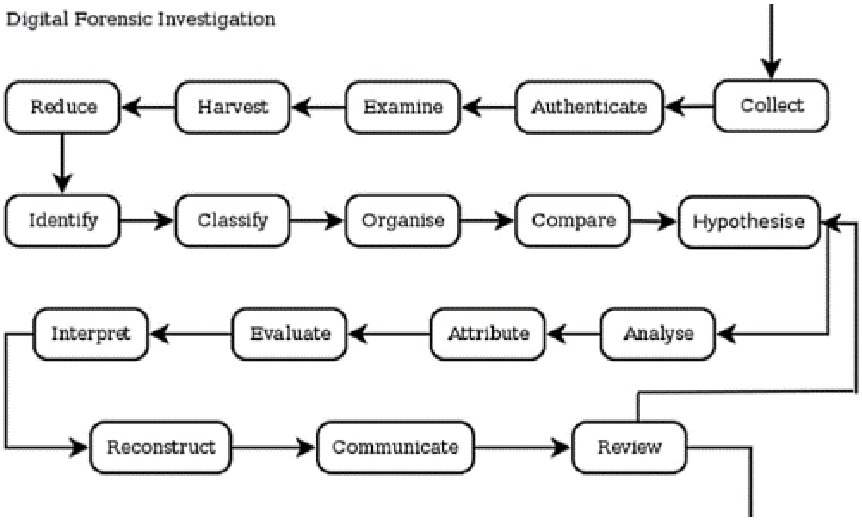

Figure 5: Digital forensic investigation process from Van Baar, Beek, and Van Eijk [47]

it is hard for digital investigators to share this knowledge and keep up with developments from other research. So sometimes a lot of redundant work is done for similar cases.

In the proposed DFaaS solution, some of these issues are solved through more efficient processes and well-defined roles within the system. The solution also includes the use of a new closed-source system called Xiraf in which different parts of the process of digital investigation can be executed by multiple different people, and some parts of the process are automated. This reduces the time spent in some parts of the process to mere seconds. This system is centralized, so instead of each department having their own system, all the data is saved in a centralized system. The system also allows the computing power, storage space, investigative tools, back-ups, and other resources to be shared. This results in departments no longer having to buy these things individually and the centralization allows for more effective security and storage measures. The centralized system also allows for an increase in processing power to automate certain parts of the process. In the previous model it did not make sense for a department to have a powerful system to quickly extract, index, and analyze the data. Since the system would most likely be idling most of the time. Centralized computing allows for the sharing of this computer power, which means that the system will spend more time actually processing cases [47].

For example, in the case of resource management, the digital investigators are no longer responsible for keeping up with system administration, making back-ups, etc. This is done by system operators, who can upload images to central storage and index them. During this process all of the metadata including timestamps is extracted and keyword indexes are created [47].

The fact that detectives can now directly query the digital material also increases the efficiency. They no longer have to ask the digital investigators specific questions, can now directly use their expertise in multiple cases and create hypotheses based on the data they receive. The fact that they can access the data themselves saves the digital investigators a lot of time.

The authors found that the centralized system has also increased the sharing of information between digital investigators. The time freed up due to the system has also led to an increase in in-depth research into specific problems. And this information is more easily shared as multiple digital investigators can access multiple cases based on their categories. In the new system, digital investigators are also expected to specialize on one or more forensic tasks. Such as creating forensic images, interpreting results found by detectives, doing investigations at the byte-level, and doing investigation into specific traces found by detectives.

As a result of the implementation of DFaaS, case work is done more efficiently, research into new forensic methods has increased, collaboration is encouraged, and backlogs of cases have been reduced. The tools have increased productivity across all parts of the process, and the automation allows for digital investigators to spend more time doing what they are actually hired to do. Data from cases can actually be used to form hypotheses now that the time frame has been reduced instead of just being used to test hypotheses [47].

4.1.3.1 *Hansken*

In Beek et al. [48] the article by Van Baar, Beek, and Van Eijk [47] is followed up by providing an insight into the implementation of the DFaaS platform Hansken in the forensic organization of the NFI. The authors define the use of digital traces in Hansken as follows: "A trace represents a digital artifact and consists of (a link to its) data and corresponding meta data." [48] Traces can be investigated using a variety of tools, that are integrated into the system and usable for a variety of different use cases. The authors note that Hansken is not used to collect evidence, it was merely used to search through and provide insight in already available evidence. The original evidence is stored separately from the traces that are worked with in the system. Allowing the original evidence to stay intact while still allowing for a thorough investigation of the traces. At the time of writing the article, Hansken had been used in over a 1000 cases.

Within Hansken, a variety of improvements have been made to the traditional digital forensics process. Several processes that used to be done by a human have been automated, and artificial intelligence is employed to classify the evidence more efficiently. However, currently no automated image authentication mechanism is present in the Hansken platform.

The platform is available to a variety of different actors, it is not solely used by digital forensic investigators at the NFI. It is also used by the digital investigators of the Dutch police.

4.1.4 *Service levels*

Horsman [49] defines different "service levels" that digital forensic science (DFS) organizations can define for clients who want to know what capabilities they can expect in which cases. This allows outside clients to know what level of service they will receive for a certain type of case, and what the possibilities for investigation are in these cases. According the author, defining these service levels is necessary as the demand for digital forensic investigation continues to rise, while the field has finite resources. This rise is mostly due to every case having some kind of digital evidence nowadays. Evidence such as phone messages, photos, video material, etc. But an increased digital footprints of suspects and an increase in the amount of data collected has made it more difficult the determine the value of the information of each piece of data. Defining which kind of investigation is appropriate in which situation will allow the DFS organization to manage their resources effectively. The current level of resources and available infrastructure in most DFS organizations is not high enough to keep up with the increasing demand. The problem will not be solved by simply throwing more money at it, more carefully planned techniques are needed [49].

As in Van Baar, Beek, and Van Eijk [47], Horsman [49] considers DFS to be a service offered to clients. From this perspective, the DFS organization must fulfill the requirements of its clients efficiently. Since the service-demand is so high, and resources are limited, defining appropriate services in specific cases ensures that resources are used effectively and efficiently. Therefore, the author suggests that DFS units may benefit from a resource assessment to define their current capabilities. These capabilities can then be aligned with a client's case requirements. When this is applied correctly, it should allow for a higher case throughput and the necessary resources to be applied in each and every case. The author list a few reasons for specifically define service levels to solve these issues. These are as follows: Consistency of service, Regulation of client expectations, transparency of capability, effective allocation of resources, flexibility, workforce deployment & internal productivity, and assessment of a DFS unit's ability.

Horsman [49] then defines 7 service levels, some with sub-service levels (SSL). The first service level is service level 0, which involves a consultation session with the client. In this service level, the investigative needs of the client are determined, an appropriate DFS strategy is developed, and the hypotheses are set up. This service level consists of three sub-service levels, SSL 0.1 for standard devices, SSL 0.2 for all non-standard device consultations, and SSL 0.3 for exploratory research, testing and capability testing of devices unknown to the DFS unit. At this service level it is required for all investigative strategies used to be documented, any additional work carried out in SSL 0.3 should also be documented to be used in future comparable cases. The result of this service level will likely be the selection of an additional service level, specific to the needs of the case determined in the consultation.

The second service level is Level 1: Data extraction & Packaged data. This level involves extracting all of the submitted device's data and supplying this data to the client without any internal data examination or interpretation. The third level, level 2, involves data extraction, screening and packaging the results. It encompasses all tasks in level 1, in addition to applying data screening criteria. Which means that the data extracted is refined, and a smaller sample size is extracted. The fourth level, level 3, involves the use of device triage and preview examination. This can provide the client with an insight into which data is present on a particular device. This service level is meant to avoid full investigation into device's which may not need such an extensive process. At service level, the standard device is subjected to standard examination, and an investigative report is created. Level 5 consists of the same, but then for non-standard devices. Level 6, the last service level, is characterized by a full expert evaluation. This service level is the most resource intensive and will often take the longest amount of time. This level is most likely offered most of the time in the case of image authentication at the NFI [49].

To aid clients with choosing an appropriate service level, the author has also created a Service Level Allocator (SLA) decision model, in which clients answer questions in a flowchart to determine the required level of service. Combining the defined service levels and the SLA will provide the client with transparency of service and the DFS organization with effective resource allocation.

4.1.5 *ENFSI Guidelines*

ENFSI is the European Network for Forensic Science institutes, of which the NFI takes part. The ENFSI has several working groups in which forensic institutes all across Europe collaborate on guidelines and best practice manuals for forensic research. They have also released a best practice manual for image authentication. This best practice manual contains a lot of information regarding the entire image authentication process, summarizing all of it would result in practically the same document. So the parts that are most relevant to the maturity models and scope of the research have been extracted, but we recommend readers who are interested in applying the methods described in the Maturity model and are working on increasing the efficacy of the IA process to read the BPM themselves. Below are some important definitions from the BPM that are needed for the next section.

- `Image authentication:` Assessing the extent to which supplied questions and claims concerning the genesis and life-cycle (provenance) of digital image data can be supported or answered [50].
- `Auxiliary data:` The file system information of the file, any other external information about the image file and any data contained in the image file beside the pixel data [50].
- `Context Analysis:` The process of verifying that the context in which the image is placed is consistent and coherent with the image itself [50].
- `Integrity analysis:` The process of examining for the presence (or absence) of traces that can be due to possible file modifications (either intentional or unintentional) after the acquisition [50].
- `Local Manipulation detection:` The task of locating manipulated areas within a questioned image. By "manipulated area", it is meant any region of the image that underwent some processing operation that was not applied to the rest of the image [50].
- `Pixel-level analysis:` Includes technical visual inspection (e.g., shadows, perspective, geometry, discontinuities) and techniques based on global features(e.g., compression level analysis, PRNU analysis) and local features (e.g., correlation map, clone detection) [50].
- `Processing Analysis:` The process of examining for the presence (or absence) of traces that can be due to possible global or local modifications of the visual content of the image [50].
- `Source analysis:` The process of classifying, identifying, or verifying the source device [50].

The ENFSI Best practice manual (BPM) can be used as a framework for the procedures, quality principles, training processes and approaches to the forensic examination of images. The document establishes best practices procedures for forensic laboratories in the field of forensic image authentication (IA). The methods are based on the scientifically accepted practices at the time of creation [50]

For IA, several different types of analysis can be used to support or deny the authenticity of an image. These types are as follows: context analysis, source analysis, integrity analysis, processing analysis

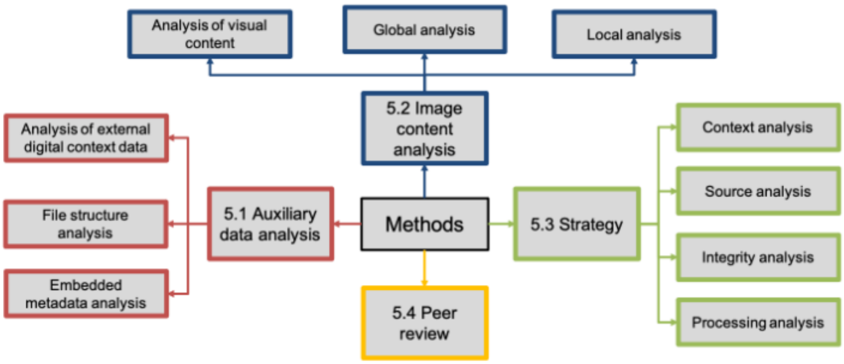

Figure 6: ENFSI Analysis methodologies

and manipulation detection. These methods include the image content and auxiliary data, through both algorithmic methods and visual inspection. The BPM describes the resources needed to properly conduct IA processes, and the requirements that these resources should adhere to. Such as personnel qualifications, software tools, hardware requirements, lighting requirements, confidentiality considerations, reference images, etc. For example, the examiner (digital forensic expert) must be able to demonstrate a competence in the following things:

- How images are created
- How images can be manipulated
- Image processing theory
- An advanced understanding of the authentication techniques which are used during examination.

The next section describes the different methods that can be used for IA, and provides a good overview of them, which can be seen in Figure 6.

Auxiliary data analysis covers the analysis of all of the data that is related to the image file, not the image itself. The embedded metadata, file structure and file system metadata can all reveal hints about the authenticity of an image. For example, if the file size is outside of the expected range then this can indicate manipulation. If the image location does not match up with similar images, this can indicate manipulation. MAC (Modified, Accessed, Deleted) timestamps can show when the image was originally made or modified, which can be used to verify the timeline of the case.

Other non-image related metadata should be searched as well. Such as browsing history, which might contain searches for image manipulation techniques, image processing software on the device which might contain log files, looking for related imagery, identical file fragments in free space, etc.

After all of the auxiliary data is analyzed, methods for analyzing the content of the image can be used. According to ENFSI, there are 3 areas of analysis, Analysis of visual content, global analysis, and local analysis. The first type of analysis looks into the visual features of a scene, and whether the depiction of this scene depicts reality. Images of a real scene must be consistent with the physical constraints on the scene. The size of rigid objects, the rules of optics and geometry, shadows, reflective objects, and other indicators must be consistent for the image to be authentic. Notable distortions in the image, called artefacts, can also be a sign of manipulation. This is traditionally most often the case with deepfakes, since the deepfake model might not be fully capable of synthesizing the image in a consistent way, leaving artefacts. However, this might not be the case in future deepfake technologies.

The second type analysis is the global analysis. These methods often provide compact descriptions with aggregated statistics, that the forensic expert has to interpret. The BPM goes over several different types of global analyses and what they analyze. Including but not limited to: chromatic aberration analysis, Photo-response non-uniformity (PRNU) analysis, fixed pattern noise analysis, JPEG ghosts analysis, histogram analysis, and Fourier analysis.

Then local analysis methods are explained, in which specific locations in an image are examined after traces of manipulations are found there. Local analysis methods usually follow global analyses, if traces of manipulation are found. Local analyses are also especially relevant in the case of deepfakes, since deepfakes usually only affect a certain part of an image.

Aside from explaining the different analysis methods, the BPM also provides a guide on when to use them and provides several different workflows which detail which analyses can work together. Specifically the strategy element of the model provides analyses types to use when specific types of questions are asked. For example, when the location the image has been taken in is being questioned, the BPM details several methods to verify the consistency of the location in the image. Such as examining the available geolocation data in the embedded metadata, or examining the consistency of images taken in roughly the same place. Another example would be when the source of the image is being questioned, usually the device make can is registered in the image metadata, if this does not add up with the known PRNU values of the image, then there is an indication that the image or the metadata has been manipulated. These methods are very specific to questions asked about the evidence, so choosing the right methodology to answer questions from the court is paramount [50].

The guide specifies that no certain method will give conclusive evidence of manipulation. Methods may support the hypothesis that the image is pristine or the inverse hypothesis. Combining several different analysis methods and creating support for their results is the most important part of building a case. The results of the research should reflect the cumulative results of different types of analysis in the form of a likelihood ratio on which the court can decide whether it is deemed likely enough that the hypothesis is correct.

The BPM goes into several different things to look out for, with guides on how to verify them.

4.1.5.1 *Quality assurance*

The [50] BPM also describes how to ensure that the results of the image authentication are of good quality. It describes a set of quality controls that should be implemented into the process in order to mitigate against bias within the examination. The following measures are described:

- Initial assessment and communication should be delegated to a different person than those who do the examination
- Having multiple examiners for conducting the examination independently or for critical findings checks
- Using an arbiter to settle the differences in opinion between examiners
- Establishing a peer review system for the reporting

Aside from these quality controls [50] also advocates for frequent proficiency testing of the IA process and conducting collaborative exercises with the ENFSI's expert working groups. These will ensure that the organization is proficient in AI and that it's employees are up-to-date on the latest IA methods. Lastly the BPM outlines the importance of collecting data for the monitoring of the processes within the process. Since the development in this area of research happens so fast, it is necessary to maintain and review statistics about the success rate, applicability and efficiency of new methods.

4.1.5.2 *Examination sequence*

Aside from the different methods described earlier in the BPM, [50] also describes a sequence of examination which take into account the elements of the process that happen outside the analysis itself. This includes creating an initial overview of the evidence to review and available resources. It also includes estimating the evidential value of each item in order to properly prioritise the evidence to review. This is important in cases where a lot of items are submitted, since it is very time consuming to analyse them all. I will shortly cover all of the steps described in the BPM.

PREPARATION: Prepare cases by selecting items to analyse with the customer. This can be done through random sampling or meetings with the customer. Establish whether there is an evident connection between the items. Connections like:

- Are images taken at different points in time or in similar conditions?
- Do the images seem to originate from the same device?

PRIORITISATION: Optimizing the cost/benefit ratio through prioritising the right items is essential to reach comprehensive results in a limited time frame. Proper prioritisation of case items will depend on a few elements like the items, request, available resources, and constraints in customer time frame and cost. In order to prioritise the items, the following concepts should be considered [50]:

- Expectation that examination of the questioned item may yield very strong support towards one of the propositions
- Value of the evidence in relation to estimated complexity
- Value of the evidence in relation to cost
- Value of the evidence in relation to time

WORKFLOW The BPM emphasizes that these are general guidelines, since strict rules for the sequence of examination cannot be given due to the amount of different possible cases. But the following sequence should provide a good basis for most cases. The sequence described in [50] is as follows:

1. Initial assessment
2. Reconstruction
3. Methods:
   - Analysis of external context data
   - File structure analysis
   - Embedded metadata analysis
   - Analysis of visual content
   - Global analysis
   - local analysis
4. Evaluation and interpretation
5. Presentation of results

In addition to this list, the BPM describes an example of a case to illustrate this workflow.

4.1.5.3 *Reconstruction*

The BPM also describes the process of reconstruction within the image authentication process. According to the authors, the creation, detection and use of reference material plays a central role in image authentication. This involves extracting features like statistics or single values from the items to investigate as well as from reference items. These results can be compared for similarity, which can support propositions that the images are from the same source or have the same processing history.

Reconstruction is the process of creating new images in similar conditions to which the supposed images in question were made. This would involve using the same source device in the same location under similar circumstances, in order to reconstruct the image to use in the comparison.

4.1.5.4 *Evaluation and interpretation*

Aside from the entire process, [50] also describes the method for interpreting the individual findings in the IA process. They describe setting up a pair of propositions as hypotheses that illustrate the degree of support of a finding based on the discriminating power of the method used to examine these findings. For example, setting up to propositions in a case in which the customer wants to know whether a suspect in an image is real or pasted into the picture. The hypothesis would be formulated as follows:

- H1: The person has been pasted into the image after it was captured by the camera
- H2: The person was in the scene the moment the image was captured by the camera

A variety of methods can be used to investigate these claims. A method that could be used in this case is local noise analysis, in which the area of the person is investigated to see if the level of noise in this part of the picture matches the expected noise based on the rest of the picture. A finding could be that the level of noise at that part of the picture does not match the rest of the picture, which can support H1. However, other variables such as the clothing the person is wearing in the picture could affect the level of noise in this part of the image. This should be taken into account when interpreting the result of this analysis. This is why it's important to create reference images in similar conditions as the original picture, so that the results can be compared and the discriminating power of the method can be determined. Another method is to investigate the performance of the known method on a dataset of images of which the results are already known. This way you can evaluate the discriminating power of the method.

This is also the reason a single method is not enough to prove or disprove a proposition posed by the researcher. The researcher should aim to combine a variety of methods, which might give evidence towards a certain hypothesis and add the results of these methods and their discriminatory power together to formulate the likelihood of that proposition to be true. Finding this likelihood is very difficult and often impossible for a variety of reasons, which is the way the likelihood ratio's are illustrated is left up to the organizations. And is also why illustrating the reasoning behind how the likelihood ratio was constructed by the researcher is important. In addition, the way the results are formulated don't draw any conclusions regarding whether the suspect is guilty or not. The results simply say that either the forensic findings provide **strong** support *for* the proposition (H1) rather than to the alternative proposition (h2) [50]. Or that the forensic findings provide **strong** support *against* the proposition rather than to the alternative proposition [50].

`Strong` representing a variety of different words with varying weight that can be given based on the likelihood ratio that was determined. This is often backed up with a numerical likelihood such as 1:2000.

In general, the ENFSI IA best practice manual is a good place to start when designing a new IA process. A lot of the combined knowledge of several different research institutes is present in this manual. We recommend any reader who is designing a image authentication process to read the entire manual, as it could not be fully included here.

4.1.6 *DFOCC*

DFOCC is an acronym for Digital Forensic Organization Core Capability Framework, which was first described in Almarzooqi and Jones [38]. The framework is meant to be used as a tool for standardizing the creation and improving the capabilities of digital forensic organizations. Essentially the purpose of the framework is similar to a maturity model: To define the success factors for an organization, function as a universal benchmark for organizations in that sector, and provide a road map to improve the organizations capabilities. Although the author says the framework differentiates itself from traditional Maturity models by also looking into the key factor of (iv) policy, next to the key factors (i) people, (ii) processes, (iii) tools. In the case of this framework, they are labeled as (i) policy, (ii) people, (iii) infrastructure(tools), and (iv) investigations(process). The authors also specify that the framework does not provide answers to developing and management of the capabilities, they merely provide a tool to measure them.

The authors represent the framework as an equation between the key success factors discussed earlier. The capabilities of an organization can be measured using a multiplication of the policy with the other key factors. In this framework the key success factors are not viewed separately, but have to be viewed with the policy multiplier.

The author says that incorporating policy is important, since it is not possible for these capabilities to succeed without policies to support them. For example, infrastructure capability is impossible to achieve without the right policies to govern the use of software, it's maintenance, access controls. The authors emphasize that the role of policy is essential to the DFOCC framework in all aspects of the capabilities within a digital forensic organization. Nevertheless, the authors do not provide any policies in the DFOCC framework, stating that this has to be determined with factors such as an organization's size, budget, and scope.

Almarzooqi and Jones [38] found in a survey that the foremost reasons for the inadmissibility of evidence were:

1. qualifications of the expert
2. authenticity of evidence in unbroken chain of custody
3. preservation of digital evidence during investigations

The reasoning provided by the author for the inadmissibility of evidence was that digital forensics organizations lacked the policies to prevent the above events from happening and to ensure that the evidence is admissible. This is an interesting finding as ensuring the authenticity of evidence appears to be a problem in more forensic organizations, even before deepfakes became a thing.

The author states that the framework can be used as a cheap replacement benchmark for smaller organizations instead of being accredited to ISO 17025 or 27001.

4.1.7 *DF-C^2M^2*

The only capability maturity model that is directly aimed at digital forensics organizations in the literature is the DF-C^2M^2, which is short for Digital forensics Comprehensive Capability Maturity Model. The maturity model is introduced in Al Hanaei and Rashid [37] and was aimed was meant expand the ISO 17025 standard in the area of quality management and basic competency management in digital forensics labs. It's meant to be more future proof than the ISO 17025 standard as this standard did not account for the growth of storage capacity in digital devices. The maturity models allows for the measurement of maturity in the three key dimensions: people, processes, and tools. It can be applied in any type of forensics organization. Al Hanaei and Rashid [37] has also introduced a tool that can be used by an organizations to evaluate their level of maturity in the area of digital forensics and even provides a roadmap for possible improvements. Unfortunately, the author does not outline the specific capabilities that a digital forensics organization should have to reach a certain level of maturity. Also, from the context of the paper it seems that this maturity model does not include the process of image authentication, which is the area of digital forensics most relevant to this research.

4.1.8 *Conclusion*

In summary, while some factors of maturity for digital forensics is available in the literature. Most of these factors do not apply specifically to the image authentication scenario. One reason for this could be that most of these papers were written at a time (<2017) when deepfakes was not a relevant topic, and therefore the question of image authenticity was not as big of a subject. In this time ENFSI has released a best practice guide to applying image authentication in forensic organisations, but while the manual provides good guidelines to image authentication, it provides no tools for forensic organizations to measure their level of maturity in this regard. In addition, the manual misses a roadmap for implementing new capabilities in the image authentication process.

Framework	Description	Features	Article
Capability Maturity Model	Described the original maturity model developed for software development companies	-Standard 5 level, 3 dimension model. Strictly used for software development - This model introduced the concept of maturity in organizations - Defines maturity as the organizational efficiency at a core competency.	Paulk et al. [43]
ENFSI Best practices	Guide to implementing image authentication by all of the european Forensic Science institutes	- Gives various approaches to image authentication - Describes methodology from start to finish - Process based improvement - Best practices determined by a variety of Forensic science institutes	[50]
DICMM	The Maturity model is aimed at organizations that want to measure their readiness to do digital investigations	- Mostly deals with readiness for digital investigations in other organizations - Not really applicable to Digital forensic organizations	Kerrigan [45]
DFOCC	Digital Forensic Organization Core Capability Framework	- Added the policy dimension to the traditional 3 dimension model - Defines the success factors for an organization - functions as a universal benchmark for organizations in that sector - provides a road map to improve the organizations capabilities	Almarzooqi and Jones [38]
DF-C^2M^2	Digital forensics Comprehensive Capability Maturity Model	- Meant to expand on ISO 17025 - Provides a tool and roadmap for improving DF capabilities	Al Hanaei and Rashid [37]
DFaaS	Digital Forensics as a service is a method of seperating responsibilities within the DF process	- Method to apply Digital forensics in a more efficient way - Developed at the NFI - Features a DFaaS platform called Hansken that offers DF services	Van Baar, Beek, and Van Eijk [47], Beek et al. [48]
Service Levels	Introduces the concept of service levels to digital forensics	- Features a service level allocator - Helps manage customer expectations - Describes using service levels for a more effective allocation of resources	Horsman [49]

Table 5: Overview of different models discussed in the ILR

5 INTERVIEWS

In this chapter, the results of the interviews are discussed. Firstly the general takeaways from the interviews are discussed in the interview results section. The respondents will be numbered so it is easier to refer to them, the numbers corresponding to the respondent role and organization can be seen in Table 6. The respondents of the interviews are in chronological order based on when they were interviewed. In the subsequent sections, the relevant findings from the interviews will be discussed. The findings will be supported by references to which interviewee talked about it, and might be backed up with a direct quote from the interview.

5.1 PROBLEM INVESTIGATION

During the problem investigation part of the interviews, the following questions were explored:

1. Are deepfakes currently a problem for forensic organizations? Why, why not?
2. How big of a problem are deepfakes going to be in the future for the NFI?
   - How difficult will it be to create deepfakes in the future?
   - Is it likely that deepfakes will be submitted for evidence and that forensic experts can't spot the manipulation?
   - What types of crimes are likely to be committed using deepfakes?
3. Should the solution to a potential problem be solved at the NFI or somewhere else in the evidence chain?
4. What is the level of awareness surrounding deepfakes in the legal community?

5.1.1 *Definition*

It is important to establish a common definition at the beginning of the interviews to ensure that when the conversation is about deepfakes, the interviewer and respondent are talking about the same thing. The definition of deepfakes that all participants agreed upon when asked was along the lines of "audiovisual material that has been manipulated using artificial intelligence". This definition aligns with

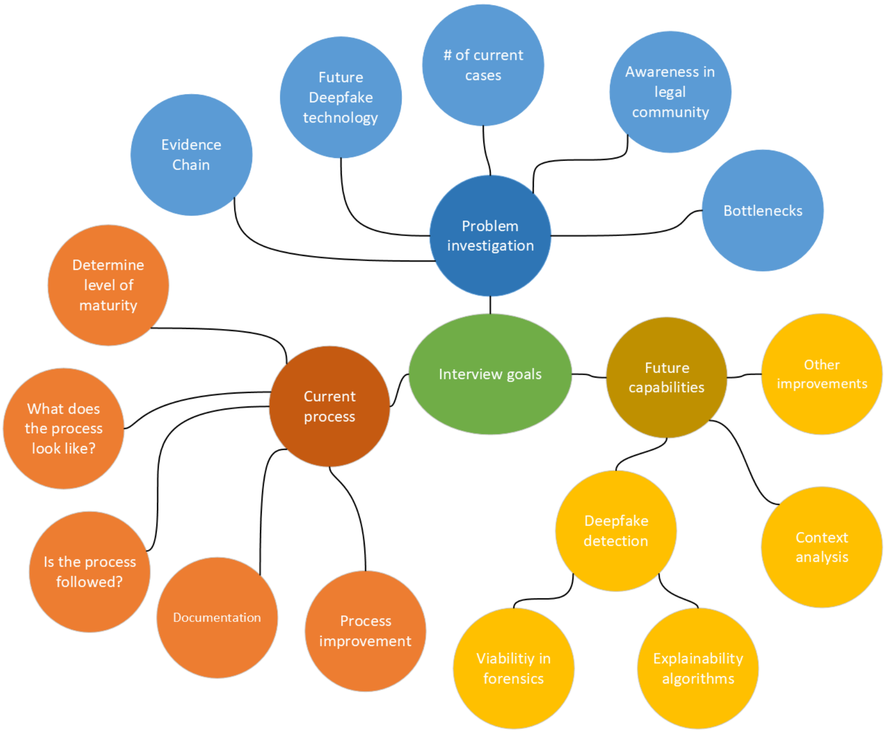

Figure 7: Mindmap interview topics

#	Respondent role	Organization
1	Forensic image researcher	NFI
2	Forensic image researcher	NFI
3	Legal professional cybercrime	Court of The Hague
4	Strategic digital innovation specialist	Dutch Police
5	CEO & CTO startup	Deepfake detection startup
6	Public prosecutor	Public ministry
7	CEO/developer	Deepfake detection startup
8	Forensic image researcher	NFI
9	Digital forensics expert	Dutch Police

Table 6: Interview respondents

the definition used earlier in this article and the one found in the literature [51].

5.1.2 *Problem*

Almost all of the interview participants that were asked the question of whether they saw the coming of deepfake technology as a problem, indicated that they thought it would be a problem that needed to be solved. It was therefore surprising to see that this was not really a problem yet. Respondent #1 indicated in the interview that the question of authenticity in image material is still relatively rare, stating that on average they may get one case per year. Respondent #2 confirmed that the question of authenticity was not something they have often had to deal with. The interviewee stated that if they get questions like this, then it's usually from international courts like the International Criminal Court (ICC). Respondents #4, 5, 6, 7 also weren't aware of any cases in which deepfakes were relevant.

Respondent #3, the legal professional, answered that it's a very real possibility that the amount of cases in which deepfakes are claimed increases, and also indicated that he was surprised that the current amount of cases was low. The respondent personally was not aware of any cases, with the exception of one, in which deepfakes were a topic, while working in an advisory role regarding cybercrime for all of the Dutch courts.

Respondent #3 did say that he knew of a case that featured the topic of deepfakes. In this court case it was related to banking, someone had allegedly opened a bank account on a popular online bank in the Netherlands, and using a deepfake as his identity. At least, this was being claimed. This is a case in which whether a piece of media was a deepfake is directly relevant. If the suspect was able to pass the automated identity detection using a deepfake, then this is a technical problem for the bank. At the time of the interview, this case was not public knowledge yet, so the ending to the case can unfortunately not be discussed.

Respondent #2 did call back to a scenario in the 90s in which virtual child pornography was legal and they had to do a lot of investigations into this. The amount of cases that had to be analyzed back then proved to be too much for the NFI. So the law surrounding virtual child pornography was changed, and this was banned. This is relevant to this research as a similar situation could arise if the problem does end up being big.

5.1.3 *Explanations*

As for reasons for why deepfakes don't seem to be a problem, no person could really give a definitive answer. All of the respondents said that they expected it to be a bigger problem, and expect it to become a bigger problem in the future.

This could be due to low awareness in the general population, the respondents that were interviewed during this study where all aware of deepfakes, and this awareness might have influenced the perception of knowledge in the general population. This phenomenon is generally known as the curse of knowledge, in which a person who knows about something is incapable of imagining what it is like for someone who does now know that thing. A paper by Cochran and Napshin [52] in 2021 found that in a population consisting of mostly young tech-savvy individuals in the United States, the percentage of people not aware of what deepfakes were was 51%. It is reasonable to assume that in a relatively technologically developed nation such as the Netherlands, the awareness is around the same. Which is likely to be even lower than that for the general population, given that the study featured mostly young people, who are generally more technologically aware.

Another explanation for the relatively low amount of cases was uncovered in the interview with respondent #1. Was that a lot of the video evidence that is used in most criminal cases is not possible to be manipulated due to the context of the case. Camera footage that comes directly from a security camera in a supermarket is so difficult to manipulate without significantly altering the file structure of the video. This makes claiming that a video is manipulated in these cases useless, since it can be immediately ruled out during context research.

The current limitations of most consumer grade deepfake software could also offer an explanation. The options for the quick creation of deepfakes don't offer the desired level of realism to be used as evidence in a case. Creating a really good deepfake currently still takes a lot of time, effort, and planning. In most criminal cases it is simply not reasonable to assume that someone had the foresight, skills, time, and resources to be able to create a deepfake that will be hard to detect in a case. It is likely that this will change in the future as the technology will become more widely available, large neural networks are developed, and more training data is put online. This could make it easier to manipulate evidence for people with a relatively low level of knowledge on the overall technology, but it is hard to say when this will actually be the case.

5.1.4 *Legal Community*

In an interview with respondent #4, who was also responsible for providing information about cybercrime to judges and courts upon request, said that the general awareness in the legal community is low. In addition, they added that aside from a few of the bigger courts in the cities in the Randstad the knowledge surrounding almost all forms of cybercrime is low. The organization they worked for was busy increasing the knowledge surrounding among judges. The interview with respondent #6 confirmed this, they said that the general knowledge in the legal community surrounding evidence manipulation is low.

Respondent #3 advised that from a legal perspective, the manipulation defense should be handled like the hack-defense. Which is a common form of defense that is often found in criminal cases of digital nature, in which the suspect claims that they were hacked and thus not the one doing the illegal activity [3]. This type of defense is very common and used to cause big problems due to the discrepancy between the communication between the investigators and the judges. The judge would ask the investigator if the claim by the defense that the suspect was hacked could be true. The investigators would then have to try to find evidence of a "hack" on the computer. The problem was that the definition of a "hack" is relatively broad, lots of traces could be evidence of a "hack". But this could be unrelated to the criminal offence that was committed by the suspect. The lack of digital knowledge in most judges and legal professionals made it difficult to explain this definition, and therefore a lot of investigation time was committed to investigating things that were unrelated to the case. To combat this, a standard way of working was created that made it easier to specify exactly what kind of hacks needed to be included in the investigation, and the investigator could more deliberately do the investigation and give a more direct answer to the question.

Respondent #3 argued that we needed a similar approach for the manipulation defense. The manipulation defense can legally be categorized either as an the defense providing an alternative scenario or attacking the reliability of the evidence. In the case of the alternative scenario, the defense says that the evidence is valid but creates an alternative scenario that is supposedly also supported by the evidence in question. And in the reliability defense, the reliability of the evidence against the suspect is called into question.

5.1.5 *NGRD register*

Regarding the qualifications for a forensic image researcher. Respondent #6 mentioned that the forensic analysis can be done by any person which is registered at the Nederlands Register Gerechtelijk Deskundigen (NRGD) in the Netherlands. Which roughly translates to Dutch Register Judicial Experts. This means that forensic cases are not only picked up and analysed by the NFI, but can also be done by another organization or independent forensic researchers. This has implications for this research, as the image authentication can also be done by independent digital forensic experts. The maturity model can also be used by this type of organization.

5.2 PROCESS

For the image authentication process part of the interviews, the following questions and subjects were explored:

1. What does the current evidence investigation process look like
2. What parts of the process are documented?
   - Is the documentation used and the process followed?
3. How often does the process get improved?
4. What feedback mechanisms are there for customers of the process?

5.2.1 *The current process*

After the initial interviews with the forensic researchers at the NFI. The following business process model was made based on the descriptions of the respondents. The business process model was later validated during an informal conversation with the head of R&D at the NFI. After which some minor changes were made to the model, such as the addition of the possibility for in court explanations. The business process model shown in Figure 8 describes the generic process of image investigation.

5.2.2 *Before analysis*

Part of the business process happens outside of the NFI. This is the part where the evidence is collected and analyzed by the police and public prosecutor. This part of the process was discussed with respondents #3, 4, 6, 9. The evidence collection happens the moment the suspect is in custody. Respondent #9 said that digital evidence is stored in the system and images of the digital devices are created the moment that they are brought in. Digital devices are also stored in a mobile Faraday cage the moment they are taken. This prevents any manipulation of deletion of evidence on the device.

Respondent #3 indicated that if evidence goes to the NFI, then a lot of analysis has already been done on the evidence. Respondent #9 confirmed this by stating that the forensic analysis done at the police is actually relatively high level, the NFI will only get the cases that the forensic investigators at the police are unfamiliar with.

According to respondent #6, the public prosecutor decides which evidence will be brought into the case. The evidence that has been decided upon usually will have been analyzed for their value to the case. So only the evidence which has a strong evidential value will be taken into the case. Which makes it more unlikely that manipulated evidence, since material that can be manipulated easier will have less value.

5.2.3 *Analysis*

There is a sub-process called execute investigation in which the investigation is done based on the needs of the case. The respondents #1, 2, 8 declared that the level of standardization in this process was low and that a high variability of throughput times is to be expected due to the nature of the cases being handled by the NFI. Aside from the NFI, most of the generic forensic investigations that have to be done are done at the level of the police who has their own in house forensic experts, including tools to execute most standard cases. This was confirmed in a later interview with a digital forensic expert from the Dutch Police. Given that only the cases that cannot be handled in a standardized manner go to the NFI, it is to be expected that investigating these cases is mostly done ad-hoc based on the needs of the case and therefore have a high variance in throughput times. The custom requests can take between 10 and 400 hours. This large variance in throughput times is usually an indication of an immature organization. However, the cases that can be standardized and improved upon to be handled by less skilled researchers have all been removed from this process. A lot of cases that used to be handled by the NFI can now be handled by forensic experts at the police, this was confirmed in an interview with respondent #9. These cases being handled directly by the police lead to a bias in the case data. The actual variance of cases would be much lower if we could account for the standardized cases. The removal of these cases from the daily tasks of the NFI also shows the capability of self-optimizing, which is the direct opposite of ad-hoc.

All the functionalities developed for digital investigations by the NFI are available in the digital forensics platform Hansken. The concept of providing digital forensics as a service through Hansken is also elaborated on in Beek et al. [48]. Which described the development and deployment of the digital forensics as a service and originated at the NFI. Currently, no form of automated image authentication is available on the Hansken platform. Some form of support in the image authentication could be a valuable addition to the platform as this would allow key partners such as the police to use image authentication tools to aid the process. Which will allow them to take on more cases, which results in less cases going to the NFI.

5.2.4 *Documentation*

In a mature organization the processes are documented and can in theory be executed by anyone who has access to this documentation

The methods that are used to analyse the evidence in new cases are documented after they have been applied so that it can be repeated in future cases. During an interview with respondent #7, it became clear that cases are continuously being documented in the quality management system of the NFI. The respondent indicated that every interaction with the customer is documented. All the methods used to investigate the material in the case is documented in the system. However, during a demo with the respondent of the system, in which they showed the system and where to find the image authentication manual, it became clear the respondent was not that familiar with where everything was located, suggesting that it didn't get used very often. Of course the system should serve a purpose and not be a bureaucratic slog, but the unfamiliarity with the system suggested that it was not something that was habitually used.

5.2.4.1 *Peer review*

According to respondents #2 and 8, the process of image research at the NFI is always done by multiple researchers. No single researcher will give the final conclusion about a single case. Three separate image researchers will do research independently of each other, and will share the results at the end. The results will then be discussed, and if there are disagreements between the researchers, a peer review will be done by another forensic image researcher. In the end, the report is setup based on the agreement between the researchers about a piece of evidence.

5.2.5 *Improvement*

Regarding the continuous improvement of the processes that are used in the investigations, the respondents were positive in the efforts that they expended to continuously improve investigation techniques. According to respondent #8, they decide what kind of research to do based on the results of student thesis'. If the thesis' provide promising results, then they will expand upon that research. In addition, according to respondent #8 if they get a case for which they don't have the right investigation method yet, research will be done to develop the methods needed to do a case.

5.3 IMPROVEMENTS

For the improvement part of the interviews, the following questions and subjects were explored:

1. Deepfake detection
2. Improvements outside of the internal process

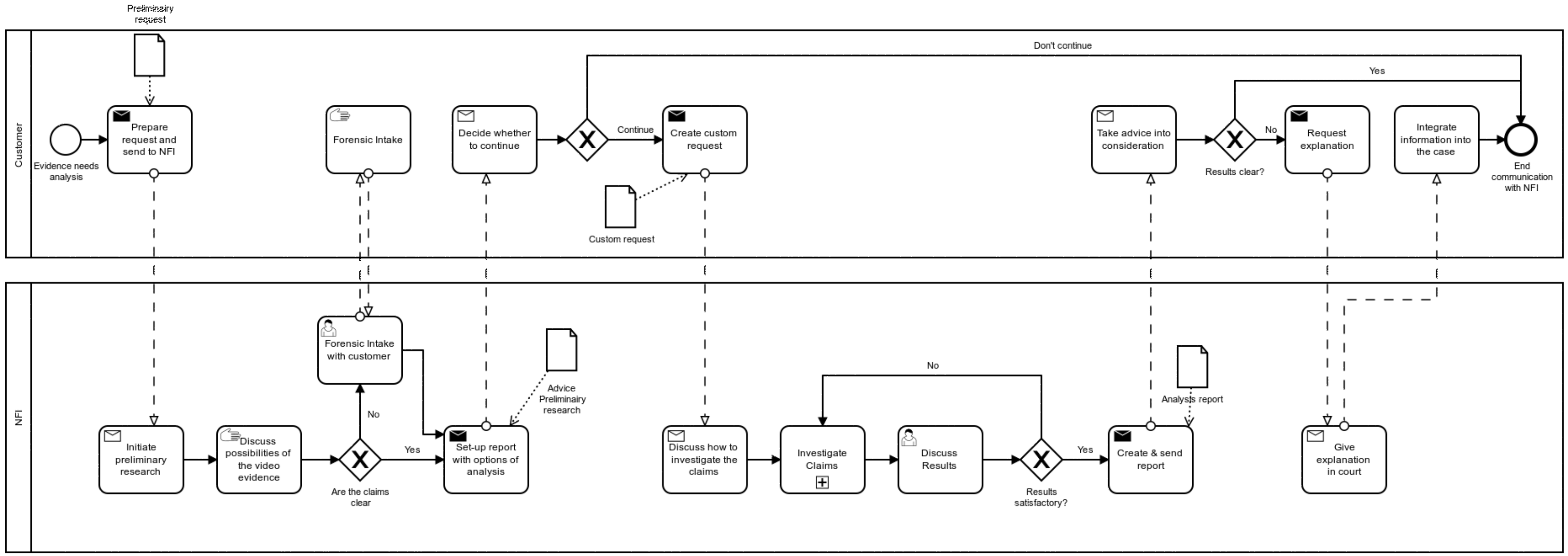
Customer
NFI
Preliminairy request
Evidence needs analysis
Prepare request and send to NFI
Forensic Intake
Decide whether to continue
Continue
Create custom request
Custom request
Don't continue
Yes
Take advice into consideration
Results clear?
No
Request explanation
Integrate information into the case
End communication with NFI
Initiate preliminary research
Discuss possibilities of the video evidence
Are the claims clear
No
Forensic Intake with customer
Yes
Set-up report with options of analysis
Advice Preliminairy research
Discuss how to investigate the claims
Investigate Claims
Discuss Results
Results satisfactory?
No
Yes
Create & send report
Analysis report
Give explanation in court

Figure 8: Evidence investigation process

5.3.1 *Deepfake detection*

The possibility of using deepfake detection in the image authentication process was one of the first avenues for research in this project. An extensive review of the past and current developments in the scientific literature on deepfake detection techniques was examined first to get an idea into the viability of using such methods. The conclusion of the review was that deepfake detection technology could be use in certain scenario's, but that the general accuracy of these methods was too low on the deepfakes-in-the-wild. Which is the type of deepfake that would most likely be relevant in a criminal case. A summary of the review of this literature can be found in the background chapter, the full review is available upon request.

The interviews would also include a component of deepfake detection to get an idea from the experts on what the requirements for using deepfake detection during the image authentication process would be, and in what way it could be useful. The experts that were talked to included deepfake detection companies and forensic image researchers.

The forensic image researchers did see the merit in using deepfake detection software to aid in the IA process. However, they did lay out certain requirements when inquired further. The use of just deepfake detection to get to a direct result for the case, was out of the question. A human expert always has to make the final decision regarding whether the video was manipulated or not. Another thing to take into consideration was that the scenario's in which deepfake detection can actually provide a useful input it very small, since the uses of deepfake software are currently constrained to just portrait videos. Most of the video's that are used as evidence in criminal cases are not in the form of the regular deepfake format, if in the future the technology improves to include more different types of angles, then this could pose more of a problem.

During the first interview with a deepfake detection company, the creators were very assured of the applications of their technology. Interestingly to note was that both of the startups used deep learning methods to detect deepfakes instead of directing their algorithm to look for certain features in the data. While respondent #7 noted on novel detection methods in the literature: *"most of the time with AI the simpler approaches are more effective because there are less assumptions that it's based on. So I think in most cases the latest ideas are just a bit too complex to make it work in practice."* Due to this, they try to combine the simpler earlier ideas from the literature with deep learning methods. The deep learning methods automate a lot of the work that used to be done by pre-selecting certain features.

In the literature the most common form of neural network was a convolutional neural network, which can be a type of deep learning network if sufficiently deep enough. The convolutional neural network is commonly used in image processing since it is especially efficient at these types of tasks. Both of the startups were unwilling to give more specifics into what type of neural network structure they used in the detection process. Aside from the fact that they both used temporal features in addition to frame-based features.

Respondent #7 saw the usefulness of deepfake detection in the forensic scenario more negatively. The respondent explained that deepfake detection software definitely has merits in specific scenarios, but that the forensic scenario was not one of those. Saying specifically: *"in the forensic situation I don't think you want an automated decision, you need a human in the loop"*. They explained that deepfake detection is very effective in cases of automated identity verification, which is a widespread method of digitally verifying someones identity. This can be required while applying for loans online, or opening bank accounts. This type of automated verification is very sensitive to deepfakes, since the deepfake only needs to fool a machine and not a human. Simple deepfakes that would not be able to fool a human can be used to fool the verification if no deepfake detection is in place. Implementing deepfake detection including the use of input controls, such as forcing the attacker to use your software to capture an identification video instead of allowing anyone to upload videos onto your platform, can significantly reduce or even eliminate the amount of identity fraud attacks. Comparing this to the forensic scenario, the respondent said the following: *"Forensic is more difficult. When you do forensic modeling and you want to know how difficult it is for an attacker to do something, which are the things that they can do in a specific situation. If you force them to use your system, the things they can do is very limited and can be automated against. In the forensic analysis this is much more difficult since you don't know how the video was made, someone could have worked for 2 months on a certain video and you wouldn't know."* They added later on in the interview that despite this deepfake detection can still be useful: *"It[deepfake detection] is a signal that you can use in combination with other information"*

5.3.1.1 *Explainable AI*

One of the main requirements of the use of deepfake detection software voiced by respondents #1,2 and 8 was that the results are explainable. In the interviews the deepfake detection companies were both aware of this requirement and had implemented some form of explainable artificial intelligence into their system. In the case of the first company the implementation of explainable AI meant that the algorithm would feed you the area of the image that it used to base it's result on. While this information could be useful in the investigation, it is not the type of explain ability that could describe what exactly tipped the algorithm off. Which is what you would need if your were to use this information during a court case. However, if this software would be used in a forensic case, this area could be used by the researcher for further investigation to find evidence of manipulation. This evidence can then be used to construct a solid likelihood ratio and evidence rapport for the court. They both argued that while it is hard to justify the answers of the deepfake detection algorithm in a forensic context, it can aid the researcher by providing an initial overview of potential evidence, and detecting older deepfakes immediately, similarly to anti-virus software instantly recognizing older viruses.

Respondent #7 was aware of explainable AI methods in the deepfake detection software, but was also more aware of it's limitations. The respondent said that explain ability is a very broad concept, and that it's use in deepfake detection is very hard to do properly. The respondent highlighted that it is very difficult to explain a deepfake if there are no artifacts. Saying: *"You can explain why the algorithm took a decision, but if you plot it on a video humans might still not be able to see it."* The algorithm can show you the area it used to base a decision on, but if there are no identifiable artifacts in the image, humans will still not be able to see it. This is why explain ability is a difficult thing to properly implement.

Implementing this comes with the possibility that the algorithm is over fitted to a certain type of deepfake. Since in deep learning algorithms, we don't know what features the algorithm selects to base it's decisions upon, we can't know for sure whether the algorithm will notice all evidence of manipulation as evidence of manipulation. The model might be blind to certain traces while noticing other traces more easily. This is why when implementing deepfake detection, it is important to not just focus on the area that the algorithm selected for manipulation, but also to keep in mind that some area's which feature manipulation traces may be invisible to the algorithm, while visible to a human.

5.3.1.2 *Research*

If an algorithm is used during the image authentication process, it is important that it can detect most types of deepfakes. Both of the deepfake detection companies said that they ensure that their algorithm is state-of-the-art by keeping up with recent literature on deepfakes and deepfake detection. This is the baseline requirement in a world were new technological developments happen so quickly. Both of the respondents from the detection companies drew the comparison to anti-virus software, they can't guarantee that they will catch all viruses, since a skilled hacker can create an exploit if he put enough effort into it, but as soon as the virus is localized, the software will be changed to be able to detect it. This means that the companies can't guarantee that no deepfakes will be able to go through, but it can guarantee that it will significantly lower the possibility of a deepfake going through.

Another issue that was brought up by the deepfake detection startups and respondent #4 was that the speed at which deepfake technology is developing requires that this should be continuously researched to keep up with the state-of-the-art deepfake software. If the capabilities of deepfake software is known then better estimates can be made for the likelihood of manipulation in the evaluation of the results.

5.3.2 *Context analysis*

During the interviews with most of the respondents, the importance of the information surrounding the evidence came forward. Respondent #4 put it like this: *"for example meta-data of the video that includes location data or the IP-address that it was uploaded from. They have several other things surrounding it, so this information also has to be taken into account when investigating such a claim."*

Other respondents such as 1, 2, 8 and 9 mentioned that context analysis should be a very big part of the image authentication process. This can help decide what evidence is even worth analyzing. This type of analysis will also get more valuable as artifacts made by deepfake techniques become more rare. As the value of the image authentication itself decreases due to technological constraints. What this means for the police is that in the collection phase of evidence, a lot of information surrounding the evidence should be recorded. This is also seen in [50] where all different types of content analysis are laid out.

These are some questions that are worth considering before analysing the evidence.

1. What is the value of the evidence in the video? Is it very incriminating, or is it just an addition to a list of other types of evidence?

2. What is the chain of evidence? Is it straight from a security camera system or did it come from the internet? How difficult would it be to manipulate a video and restore it to it's original file image?

3. How much time is there between when the video was made and the evidence was collected?

4. Does the system were the evidence was localized allow for manipulation?

5. Are the angles/quality in the video suitable for manipulation?

These types of questions help the researcher determine what kind of investigation is necessary. A standard process for the context analysis could be beneficial in this context. This will be made based on the results from the literature review. And can be found in the maturity model chapter.

5.3.3 *Value of evidence*

One subject that was brought up by respondent #4 was the necessity for processes that can be used to determine the value of a certain piece of evidence. Determining the value of the evidence beforehand can safe a lot of time in cases that feature a lot of different. This is also reflected in the ENFSI IA best practice manual. Evaluating the evidential value before analysis and prioritizing bases on evidential value is

5.3.4 *Other prevention methods*

Aside from the above subjects, one subject that was briefly explored was the use of blockchain content authenticity systems. This was a subject that arose during the literature review phase of the research. Respondents were asked if they saw any merit in using this kind of system and if they thought it would be possible to implement.

Respondent #4 raised a valid point regarding blockchain methods, which is that it only works if it is implemented on the entire internet. Adding hashes to evidence sent to the police themselves won't accomplish anything. They said: "But it[authentication through blockchain] is very difficult to do with deepfakes, someone might send us a video, but if we create a hash at that moment there is no point to the hash, it doesn't tell us if it was manipulated before the video came to us." The respondent did mention a content authenticity initiative which seeks to automatically authenticate content coming from journalists in oppressive regimes. The respondent said that: "they want that their information can be used as evidence for war crimes. So how do you add a sort of authenticity stamp when they create this evidence on their phone which includes crucial information about the evidence, without revealing too much about the creator of the evidence and potentially putting them in danger."

A prevention method that was explored by respondents #5 was mostly aimed at raising awareness of deepfakes. The same respondent said that "We have also thought about using blockchain to capture authenticity at the moment the video is made. But that was a bit too complex." The consensus among experts working in the deepfake space seems to be that block chain authentication is currently too complex to be implemented.

5.4 CONCLUSIONS

The general consensus among the tech-savvy respondents is that tooling will not provide a good solution alone to the Deepfake problem. Tooling can be useful in some cases, but general policy and process based improvements must also be made in addition to the implementation of an image authenticity tool. The people participating in the process must also be aware of recent developments in deepfake technology and hold the image authentication software developers accountable for developing new image authentication techniques based on these new technologies. Awareness of the people should be ensured using policy to make it the organizations responsibility for keeping up with recent developments. A cooperation between the forensic organization and the providers of the image authentication software is also paramount to it's success. As the providers of the image authentication will be more knowledgeable about recent developments, due to a more focused research perspective.

6 THE MATURITY MODEL

> *"Knowing is not enough; we must apply. Willing is not enough; we must do."*
>
> *— Johann wolfgang von Goethe*

In this chapter the design of the maturity model will be elaborated upon. The factors that came from the literature and the factors that were uncovered in the interviews will be included.

6.1 THE IMAGE AUTHENTICATION PROCESS

From the interviews and the literature, several parts of the image authentication process have been identified. According to Mettler [1], three different perspectives should be taken into account when designing maturity models. Technology based maturity, people based, and process based. Almarzooqi and Jones [38] added the dimension of policy to this framework. In order to explore all different dimensions of maturity for image authentication, we are going to take all of these perspectives into account. The policy will be multiplicative to the other 3 dimensions, so that perspective will be taken into account for each dimension.

6.1.1 *Model Structure*

Similarly to Paulk et al. [43], we will create 5 different levels for the maturity model. Consisting of several key process area's for each maturity level in which improvements can be made to increase overall maturity in image authentication. The key process area's will be categorized into one or multiple of the 4 dimensions. The key process area's describe the area's to focus the efforts of improvement at that level of maturity. The levels of maturity will be named in order of low to high maturity; Ad-hoc, Defined, Managed, Controlled, and Optimizing.

6.1.2 *People*

One of the key dimensions of maturity are the people participating in the process. More qualified and experienced people will lead to higher organizational efficiency and maturity. In the case of Image Authentication, several different stakeholder are involved in the process:

1. Police (collection of evidence)
2. Digital forensic investigators
3. Public prosecutor or customer
4. Image researchers (NFI)
5. Judges

However, a forensic organization will not have control over the qualifications of all of these people in the process, the efforts should therefore be focused on the people that participate in the process at the level of the organization. What the organization can control is how they work together with the key partners (Police & public prosecutor) to decrease the workload by researching techniques that can be used by the police to do IA themselves. This tactic has already been employed by the NFI with other digital investigations. This reduces the need for specialized IA experts and decreases the workload of the forensic organization.

Working together with the collectors of evidence at the police to ensure that the right information is being recorded is another area of improvement. Information such as the source of the evidence, what the original system of the evidence was, how soon after the crime the evidence was collected, etc. This is all valuable information for the context analysis of the evidence, which will play a huge part in the full IA process. For Dutch forensic organizations, that means that experts are certified in the Dutch Register Judicial Experts (NRGD). This is a requirement for the evidence being admissible in court. The NFI has their own register of experts which is setup through formal examination of their own people.

6.1.3 *Process*

Everything from defining the process to continuously optimizing the process in different ways is included in this dimension of maturity. In the case of IA, this can mean several things, such as implementing a documentation system, creating standardized processes, ensuring that the defined processes are followed, improving upon existing process definitions, implementing feedback mechanisms into the process, etc. All things that can be improved on the basis of process improvement falls into this dimension.

6.1.4 *Technology*

This dimension of maturity encapsulates everything that increases Maturity through the implementation of new technology. In the case of IA, this means implementing DFaaS infrastructure into an organization to optimize a digital forensics researchers time spent doing actual research. As papers like Beek et al. [48] and Van Baar, Beek, and Van Eijk [47] have shown, the implementation of DFaaS can lead to a more effective digital forensic research process. Integrating deepfake detection technology into existing DFaaS platform is also a goal to strive for in achieving a higher level of maturity. As this will make working with the technology easier, and allow partners such as forensic investigators at the police to use the technology. This is the baseline for a forensic organization to have effective IA processes. Another thing that can be implemented within the IA process is deepfake detection with explainable AI elements. While deepfake detection is not a standalone solution, as we found out through the interviews and literature review, it can aid in the IA process by providing area's of investigation for the image expert to investigate. As deepfake technology gets more advanced and deepfakes can get undetectable, this is still useful to be used for the local analysis of the images.

6.1.5 *Policy*

The policy dimension should also be considered while implementing improvements from the previous dimensions. As found in Almarzooqi and Jones [38], policy can make or break a new improvement. All of the previous dimensions are supported by the right organizational policy to ensure that the tools, people, and process improvements are implemented and used as they are intended. The policy dimension is assumed to be an implicit part of the maturity model, so when a factor in another dimension is implemented it is assumed that there is policy supporting this improvement. The policy dimension will feature in the assessment matrix.

6.2 THE IMAGE AUTHENTICATION CAPABILITY MATURITY MODEL

The model is aptly called the Image Authentication Capability Maturity Model (from now on IACMM). In this section, I will present the model design, including general descriptions on the functioning of an organization at that maturity level. The general representation of the IACMM can be seen in Figure 9. This general representation in an overview of the levels including the key process area's (KPA's) that should be worked on at the level to reach a higher maturity level.

6.2.1 *Descriptions of the maturity levels*

AD-HOC: At this level, image authentication is rarely needed in the forensic organization. The moment a question regarding the authenticity of an image comes up, methods are figured out on the spot by a forensic digital researcher. This researcher is not specialized in image research, but more in digital forensic research all together. No time indication can be given to the customer, as the processes aren't documented and no historical data on the duration of previous cases are present. The research is carried out by a single researcher, no peer review mechanism is present. At this level, the forensic researcher might not be registered with the NRGD, which means that the results of the image authentication process cannot be used in court. Process definitions are not present and cases are executed spontaneously as the need arises.

DEFINED: At this level, the general guidelines of processes in the area image authentication are defined, but not necessarily followed. While the processes are defined, the execution of the process is often reactive instead of proactive. For image authentication, this means that no specific order of investigation is followed. Auxiliary data research might be done on request after pixel data research, instead of in the right order. A documentation system for older cases is present, and cases are documented, but no mechanism is available to re-use the techniques used to solve older cases. Personnel is trained in general image forensics, but not necessarily specialized to image authentication. The personnel is qualified and registered with the NRGD. The usage of tools at this level is not standard practice and may be implemented when the researcher deems it necessary.

MANAGED: At this level, the processes are defined and process monitors are implemented. The organization has some ideas as to how long a case will take due to historic case data. This historical case data includes the process as to how the result of the case was determined. So that it can be used in future cases. The forensic image researchers are aware of how to process should go, and the process is executed proactively. Context analysis is considered the main part of IA process. Feedback mechanisms with the customer are in place. The processes are executed by multiple researchers, as to have adequate peer review measures in place. Deepfake detection technology may be used, but it isn't standardized into the process and does not contribute to the result of the case.

CONTROLLED: At this level, processes are defined at multiple levels, the key partners are aware of the requirements of the evidence that can be sent to the forensic institute. The key partners work together with the forensic institute to optimize the input and output of the IA process. The image researchers are experts in IA and are comfortable working within the predefined processes. The documentation systems are use to record cases, use techniques from previous cases, and test new techniques on old cases.

Deepfake detection technology with explainable results is a standard part of the process, this is used to get indications for area's of interest. The results of the deepfake detection software can be used in the final evaluation, but has a very low weight compared to the analysis of the experts.

OPTIMIZING: At this level, image authentication is one of the digital forensics organizations main processes. And several different departments are vigilant to optimizing the process. Collaboration with key partners is standard practice, most cases of IA can be handled by the forensic experts at the police. Only specialized cases get through to the forensic organisation. This means that the majority of the time of the forensic image researchers can be spent on doing research on novel IA techniques to improve the efficacy and efficiency of the IA. This collaboration also entails the sharing of these new techniques with the key partners.

Deepfake detection software is implemented into the process and into the DFaaS platform, and the software is continuously updated with the latest papers on deepfake technology and deepfake detection. The deepfake detection company providing the license is considered a key partner and is expected to collaborate fully with the forensic image researchers on improving the algorithm and implementing new detection techniques.

6.2.2 *How to use the model*

The model can both be used as a benchmark for current forensic organizations to measure their level of maturity and as a road map of things to improve to reach a higher level of maturity in the key dimensions of the model. The model describes a few key process area's at each maturity level, these are the area's to focus on to progress to the next level of maturity. The key process area's differ for each maturity level as different things have to be implemented to progress at that level. The model in Figure 9 is the overview of the model, this can be shown to management to give them an idea of the things that should be achieved at each level to increase the maturity level within the organization. This is accompanied by the Table 7 in which the specific improvements at each specific dimension can be seen in more detail.

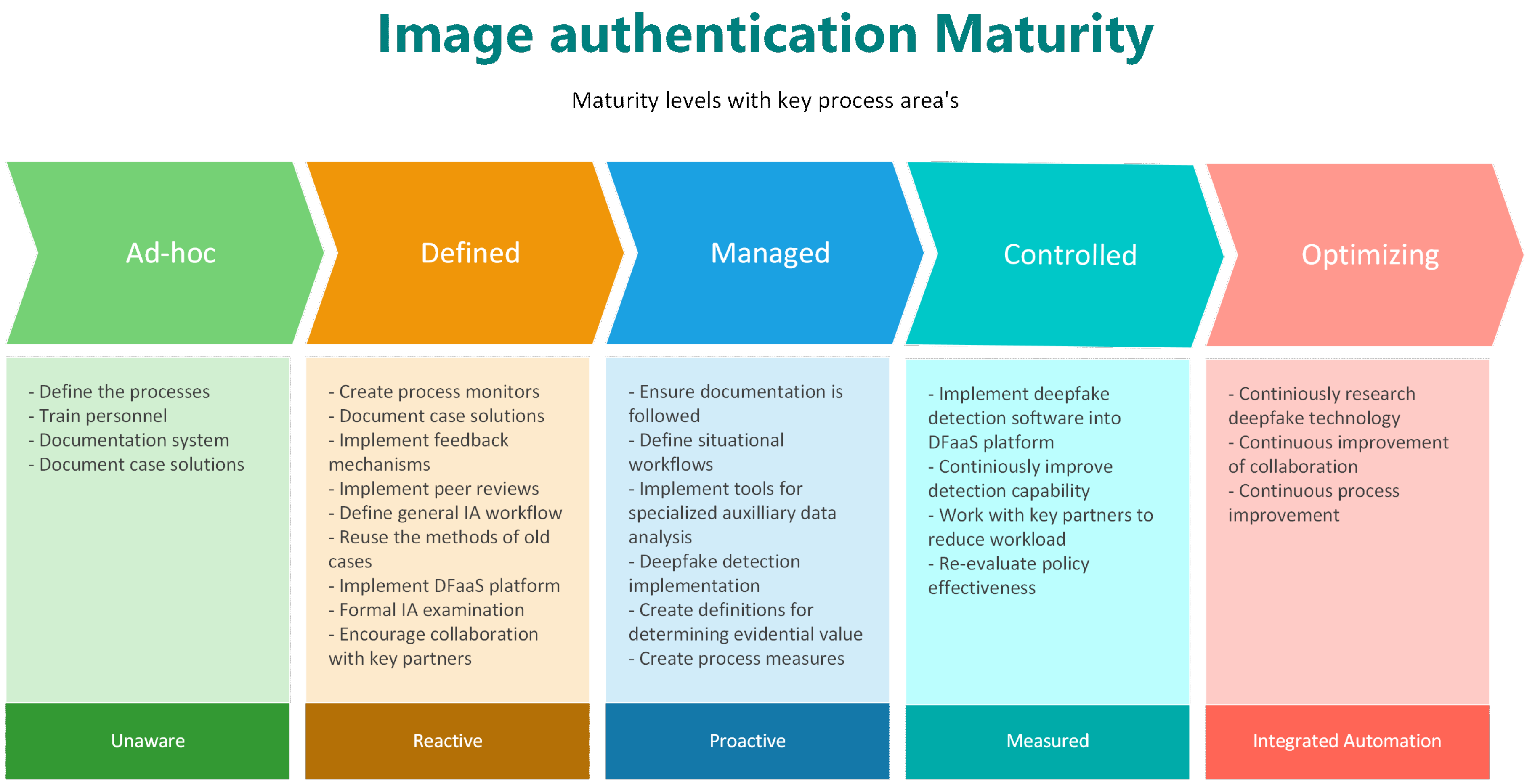

Image authentication Maturity

Maturity levels with key process area's

Ad-hoc	Defined	Managed	Controlled	Optimizing
- Define the processes - Train personnel - Documentation system - Document case solutions	- Create process monitors - Document case solutions - Implement feedback mechanisms - Implement peer reviews - Define general IA workflow - Reuse the methods of old cases - Implement DFaaS platform - Formal IA examination - Encourage collaboration with key partners	- Ensure documentation is followed - Define situational workflows - Implement tools for specialized auxilliary data analysis - Deepfake detection implementation - Create definitions for determining evidential value - Create process measures	- Implement deepfake detection software into DFaaS platform - Continiously improve detection capability - Work with key partners to reduce workload - Re-evaluate policy effectiveness	- Continiously research deepfake technology - Continuous improvement of collaboration - Continuous process improvement
Unaware	Reactive	Proactive	Measured	Integrated Automation

Figure 9: Image Authentication Capability Maturity model

6.3 ASSESSMENT MATRIX

In this section, the assessment matrix that can be use by organizations to benchmark their current level of maturity is explored. An assessment matrix is essentially a tool containing the questions that can be answered about capabilities an organization possesses in order to calculate a value which represents the level of maturity within that organization. The matrix shows the key process area's at each level for each of the four dimensions. The assessment matrix can be found in Appendix B. This matrix also features the origin of the factor for maturity, either the interviews or the literature.

The calculation of the assessment matrix is as follows. At each dimension a few questions have to be answered regarding the level of maturity of an organization in that specific dimension. The scale of the assessment matrix is from 1-5, with 5 being the highest level achievable. The questions in the assessment matrix are a direct application of the factors of maturity in the maturity model. The questions are binary, so either yes or no. The score that is output by the assessment matrix is calculated by calculating the level of maturity at each dimension, adding the results together and dividing by 4. The maturity level for a dimension is determined by the minimal level of maturity achieved and adding the amount of KPA's achieved at the next maturity level. In order to achieve a maturity level, all of the key process area's described at that level should be present. For example, if I have all of the key process area's in the people dimension at maturity level 2 and 3. So PEO2a, PEO2b, PEO3a, PEO3b and PEO3c, then my maturity level for the dimension of people is 3. If I also have 1/2 of the key process area's for the maturity level for, say PEO4b, then I would have a maturity level of $3\frac{1}{2}$, or 3.5.

When applying the assessment matrix to the NFI, the maturity score that the matrix outputs is: 3.1666. Which means that the organization is at the controlled level of maturity. The assessment matrix is only applicable to forensic organizations that have a digital forensics department.

6.4 ENFSI PROCESS

Based on the ENSFI BPM, a business process containing an envisioned scenario for image authentication was also made. This vision can be seen in Figure 10. The stakeholders were intentionally left out in this view, as well as the entire architecture layer. The architecture layer is left out due to missing information on the part of the author. It is simply outside of the scope of this research to examine the architecture layer, it is assumed that the hansken system is functional and the relevant architecture is available. The stakeholders are left out since the image authentication process that happens inside of the

NFI is irrelevant of the stakeholders requesting this authentication. The only relevant actors in the model are the image researchers.

	Ad-hoc	**Defined**	**Managed**	**Controlled**	**Optimizing**
People	No qualifications	Formal examination	Encourage collaboration	Work with key partners	Standard collaboration
		Research done by digital forensic experts	Research done by image researcher experts	Specific IA experts	Continuous development of experts through research
Process	No process definitions	Defined processes	Processes are followed	Processes are monitored	Processes are continuously improved
		Case solutions are documented	Average throughput time is known Feedback mechanism integrated into process	Processes are improved bi-anually	Processes are predictable and outliers are rare
Technology	Spontaneous use of tools	Documentation system	Use of DFaaS platform	Implement DF detection	Integrate deepfake detection into DFaaS
			Occasional use of deepfake detection		Continuously develop deepfake detection algorithm
Policy	No IA policy	Require personnel qualifications	Require process documentation	Require process monitors	Policy to facilitate contiuous development of IA capabilities
		Require documentation of cases			Policy to facilitate collaboration
Result	Org. is unaware of the IA process	Reactive processes Processes are unpredictable	IA is executed proactively Processess are predictable, but outliers are common	IA is a core business process Highly effective IA process Predictable outomes of cases	IA Maturity through: Continuously improving IA Highly specialized processes Implementation of IA technology Standard collaboration with key partners

Table 7: Maturity model Table representation

6.5 CONCLUSION

In this chapter, the different parts of the design of the maturity model have been explained. The maturity model, the assessment matrix and the architecture view of the future process of image authentication are all part of the artefact that applies to the problem context of image authentication at the NFI.

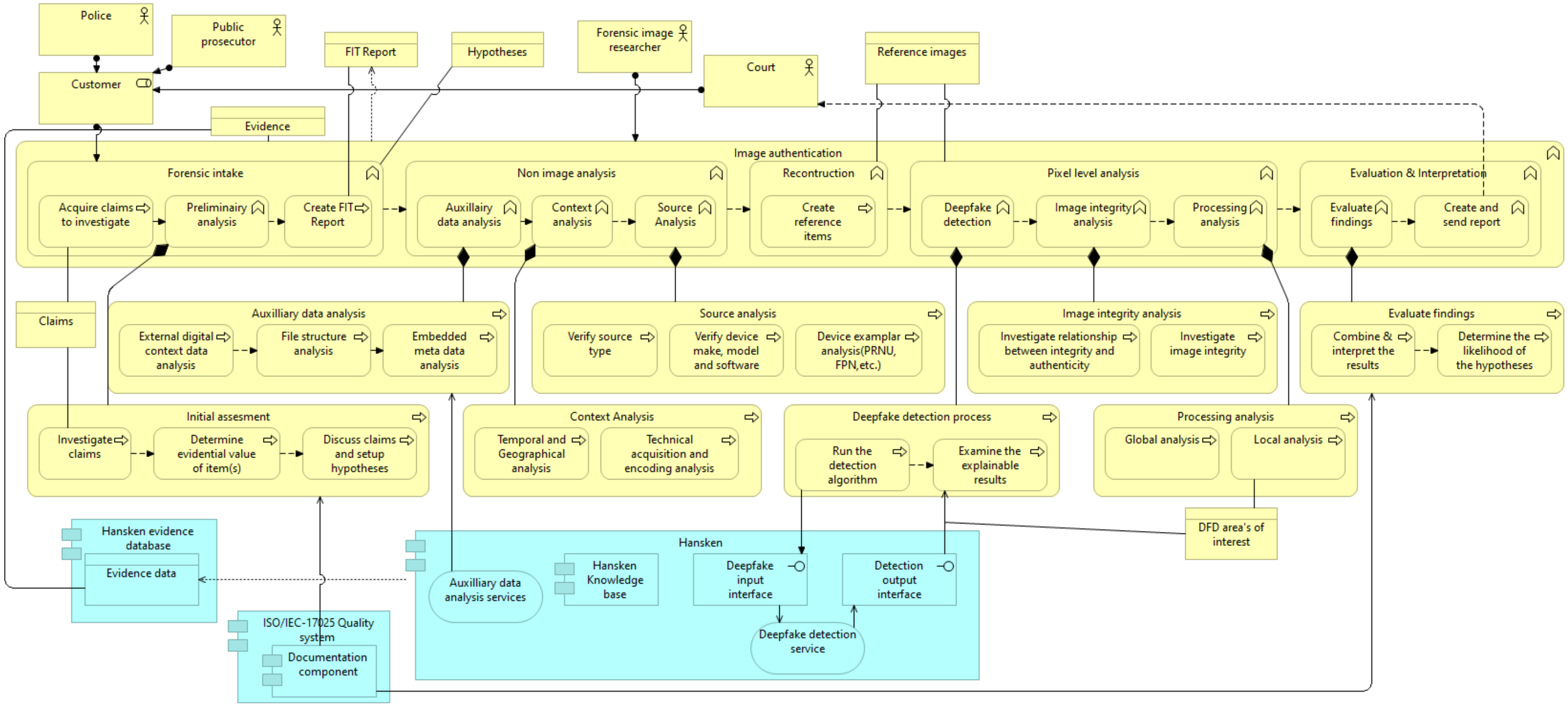
Police
Public prosecutor
FIT Report
Hypotheses
Forensic image researcher
Court
Reference images
Customer
Evidence
Image authentication
Forensic intake
Acquire claims to investigate
Preliminairy analysis
Create FIT Report
Non image analysis
Auxillairy data analysis
Context analysis
Source Analysis
Recontruction
Create reference items
Pixel level analysis
Deepfake detection
Image integrity analysis
Processing analysis
Evaluation & Interpretation
Evaluate findings
Create and send report
Claims
Auxilliary data analysis
External digital context data analysis
File structure analysis
Embedded meta data analysis
Source analysis
Verify source type
Verify device make, model and software
Device examplar analysis(PRNU, FPN,etc.)
Image integrity analysis
Investigate relationship between integrity and authenticity
Investigate image integrity
Evaluate findings
Combine & interpret the results
Determine the likelihood of the hypotheses
Initial assesment
Investigate claims
Determine evidential value of item(s)
Discuss claims and setup hypotheses
Context Analysis
Temporal and Geographical analysis
Technical acquisition and encoding analysis
Deepfake detection process
Run the detection algorithm
Examine the explainable results
Processing analysis
Global analysis
Local analysis
DFD area's of interest
Hansken evidence database
Evidence data
Hansken
Auxilliary data analysis services
Hansken Knowledge base
Deepfake input interface
Detection output interface
Deepfake detection service
ISO/IEC-17025 Quality system
Documentation component

Figure 10: Image Authentication architecture view

Part III

VALIDATION & CONCLUSIONS

7 MODEL VALIDATION

This chapter will cover the validation of the model. Validation in design science is essentially seeing if the designed artifact is applicable in the real world. This can be done through interviews with a panel of experts, or in some cases, directly implementing a model of the artefact in a model of the problem context [36]. In this case, the implementation of the model is outside of the scope of this research. So an expert opinion is needed, unfortunately a full panel of experts was not able to be arranged due to time constraints.

7.1 VALIDATION INTERVIEW

Due to time constraints, the model will be validated through a single validation interview with the Digital forensics expert and image researcher Zeno Geradts from the NFI. While this is not enough information to fully validate the interview. It will give some insight into the efficacy of the model for an organization like the NFI. In future research into image authentication capability maturity models the validation of the model should be a larger part of the research.

The validation interview will feature a short explanation of the model. Followed up by the following interview questions. The expert has the option to answer these questions at a later moment.

7.1.1 *Interview questions*

1. Is the purpose of the maturity model clear?

   R: The model describes the different levels of maturity from unaware to expert level for digital forensic research institutes. Maybe change the description to something different than Digital forensics organization, since the police are also digital forensics experts, just in a different way. They handle a lot more cases then us, only the really complex ones go get through to us, so in their own way they are also mature. They might take it the wrong way if you bunch them together with us and say they are less mature than the NFI.

2. Do you think the Model accurately describes the different levels of maturity in Image authentication?

   R: There are always different levels within an expert organization?

3. What do you think about the overall applicability of the model in an organization like the NFI?

   R: It can provide some insight into the general process of image authentication in an organization like the NFI. Although a lot of know how and tacit knowledge is not described.

4. How do you feel about the applicability of the model on forensic organizations outside of the NFI?

   R: Perhaps, although you might find some resistance of these models at the Police, for the aforementioned reasons.

5. Do you feel like there are capabilities missing from the model?

   R: Something that is missing is the know how and education needed to be an actual image expert, it takes years before you can actually call yourself an expert on this area. This is underrepresented in the model. The know how is learned intrinsically while doing the job, a lot of the processes are not really documented since it's sometimes unnecessary work. Even if you have the process models, it is still not possible for someone who hasn't been trained to do the job.

6. Are the key process area's properly described?

   R: I think so, although they need to refer to our quality assurance model ISO 17025/17020

7. Does the model help with the current challenges in forensic image authentication?

   R: Partly so, since image authentication is a moving target.

8. Is the total view of the future IA business process clear?

   R: Yes, though there are some issues.

9. Is the order of the business process in the architecture correct?

   R: I'm not necessarily familiar with this type of model, so I assume so.

10. Are there important IA processes missing from the architecture view?

    R: Our quality assurance model with the ISO 17025 and the ENFSI best practices.

11. Is the assessment matrix applicable on digital forensics organizations?

    R: Yes, partly. The assessment matrix is geared more towards digital forensics research institutes, so saying digital forensics organizations might be a bit too broad.

12. Are things missing from the assessment matrix that should be taken into account?

    R: The NFI register of experts, the NRGD register and the formal education of image researchers.

7.1.2 *Takeaways*

These questions were answered prior to the interview. After which a meeting took place to discuss the answers. Overall he liked the insight the model gave into the process, but he had some reservations about the applicability of the new IA process architecture view. Mostly because the view did not consider a lot of the "know how" that experts acquire while working through the process for multiple years. All the tacit knowledge that is intrinsic to expert work that is usually only learned as a new employee when working through the process with another expert. This is of course relevant in every complex process, but part of the purpose of the architecture view is to get rid of the need of this knowledge. To ensure that the process is clear to people who are outside of it. This will also be different in most organizations, so integrating that into the architecture view will not improve the generalizability of the model.

However, since not all of the tacit knowledge in the image authentication process is organization specific, more research should be done into integrating this information and "know how" into the process. This is outside of the scope of this research, but could be researched at a later date.

Zeno also raised concerns about the use of the word Digital forensics organizations, since according to him the digital forensics department at the police can also be considered a digital forensics organization. While they might not score high on the Maturity Model that was made here they are mature in different ways. They don't necessarily do their own advanced research in the way that an organization like the NFI would do. But overall they are very efficient in handling complex cases, and the level of automation in the analyses are high. This is a fair point from Zeno, this more consideration should be taken for the different types of Digital Forensic organizations. This model is mostly aimed at Digital forensic research organizations, and doesn't take into account organizations that do routine image authentication.

However, the maturity model is created as a measure of maturity in the image authentication process. If a customer had to decide where to analyse their digital evidence and they could choose between two organizations, in this case the police and the NFI. The NFI would score higher on the maturity model than the police, since it can handle more complex cases than the police. Although we don't know the score of the police, since we lack the necessary information to input into the model. If this is the measure that is most important to the customer, then they would be right to choose the NFI, even if this is potentially dismissive of the maturity in the IA process at the police.

Zeno also fairly pointed out that the evaluation part was missing in the architecture view, so this was added after the validation. This is a big part of the IA process and leaving it out of the model would generate confusion to those who know what the process is like. And it would lead to newer organizations implementing the process without is.

7.2 FURTHER VALIDATION

Since the current validation is not a complete validation of the model, the following steps can be taken to further validate the model. These steps should provide the researcher with enough feedback to further develop the model to the needs of digital forensics organizations.

1. Validate the model by facilitating a group discussion with experts in image authentication from multiple organizations.

2. Send the model to different forensic research organizations along with a questionnaire regarding validity.

3. Review feedback on the model and integrate this into a better version.

4. Do another round of validation on the new model with a group of 5 experts in digital forensics

5. Finalize the model based on the last round of discussion with the experts.

7.3 CONCLUSION

In general, Zeno liked the maturity model, assessment matrix and the associated process model said that it provided a good basis for creating an image authentication process. In addition it would provide a good basis for further research into this topic. The informal nature of the conversation that was used as the validation meeting might have interfered with the amount of valuable feedback that was given by the expert. As more area's of the model could have been highlighted for improvement. However, the expert was aware of the time left to finish the research and might have opted to be less harsh with the feedback. This diminishes the value of the validation and in turn diminishes the value of the model. Nevertheless, given the exploratory nature of the research, the model is considered to be validated for the purpose of answering the research question.

8 CONCLUSION

"I may not have gone where I intended to go, but I think I have ended up where I needed to be."

— Douglas Adams, The Long Dark Tea-Time of the Soul

In this chapter, the thesis will be concluded by providing answers to the research questions based on the earlier chapters.

This research was originally aimed at solving a problem regarding deepfakes for court cases. The initial idea of the research was to see if deepfake detection algorithms could be used in the process of image authentication. After initial interviews and research into deepfake detection, it became clear that the accuracy of deepfake detection algorithms is not high enough to warrant it's full use in the process. Additionally, the algorithms were too constrained to specific deepfake scenario's to make the software generally applicable and the full IA process was too complex to replace into a single deepfake detection algorithm. Due to these reasons the decision was made to shift the scope of the research to a closely related yet wider topic: Maturity in the image authentication process.

8.1 GENERAL CONCLUSIONS

Outside of the research questions, a few findings regarding the image authentication model and the problem of deepfakes for forensic organizations like the NFI were made. The problem of deepfakes causing a flood of new cases for an organization will not cause big problems if the process is automated well. The initial idea of using deepfake detection in the image authentication process is possible, but with several caveats. Deepfake detection cannot be a standalone solution to the problem, the accuracy of deepfake detection algorithms is too low to be used in that way. The accuracy of these algorithms on deepfakes in the wild is also unverifiable. Which makes it difficult to create likelihood ratio's for the results of the algorithm. So instead of using it as a standalone solution or as an additional researcher in the process, the deepfake detection software can be used to find area's of interest for local pixel analysis of the picture, given that the software has sufficient explainable components to highlight these area's of interest. This could save time for the researchers as signs of manipulation can be investigated earlier without.

8.2 RESEARCH QUESTIONS

In the beginning of the thesis, the following research questions were defined. Each of the research questions was partly answered in their respective chapters. In the section below a short summary of the conclusions for each research question will be given.

1. What capabilities are needed in an optimized image authenticity research process?

   This research question was answered through an integrated literature review and partly using the interview answers. The capabilities that were most relevant indicators of maturity were varied. Yet some interesting similarities in the capabilities mentioned in the interviews was observed. The most frequent factors of maturity were integrated into the maturity model. The most important factors to come out of the study were increasing the efforts in the context analysis part of the process, facilitating continuous process improvement, and integrating deepfake detection into the authentication process. The full results of this research question can be found in Chapter 4 and Chapter 5 .

2. What capabilities regarding forensic image authenticity research does the NFI currently have?

   This research question was answered during the interviews. The current process regarding image authentication was examined and modeled based on the way people working in the process explained it. The current process did involve a lot of the factors that were outlined in the ENFSI Best practice model. The full current process at the NFI can be found in Chapter 5 in Figure 8

3. How is a capability maturity model constructed for digital forensics?

   This research question was answered with the creation of the maturity model. The maturity model was designed based on the principles uncovered earlier in the literature review. The model consists of the 5 maturity levels that is standard in maturity models. The key process areas that were most relevant in the image authentication scenario were determined through interviews and literature review. The key process area's, people, technology, process, and policy, were determined for each of the 4 dimensions and ranked based on the level of importance for maturity. The current model is a representation of the most important factors at each maturity level in the IA process. The full explanation of the Maturity model and key process area's can be found in Chapter 6. The model can be seen in Figure 9 and the associated assessment matrix can be seen in Appendix B.

4. Does the design reflect the reality of the situation according to experts?

   This research question was answered with the validation of the model. Unfortunately, the question remains mostly unanswered, as not enough experts were able to give their opinion on the model or the architecture view of an optimized process. The model was validated using the best available method within the time constraints. There was some feedback on the model by the expert, and this feedback has been taken into consideration, some changes were made based on this feedback. The full validation can be found in Chapter 7.

The main research question has also been answered using the thesis:

MQ: *How to design a forensic image authenticity research maturity model that reflects current and future needs of forensic organisations?*

The entire thesis is the answer to this research question. As it goes in depth into how to construct a maturity model for image authentication in the problem context of the NFI. The model can be applied in the context of the NFI, it has been validated by an expert working at the NFI and they now have a road map of changes that they can implement should they want to improve their image authentication process. The model can also be used by other digital forensic organizations to improve their image authentication process.

In conclusion, the thesis provides answers to the sub-questions stated in the introduction, and the combination of these sub-questions answers the main question of the thesis. The main research goals have been achieved through the design of the model.

9 DISCUSSION & LIMITATIONS

"By three methods we may learn wisdom: First, by reflection, which is the noblest; Second, by imitation, which is easiest; and third by experience, which is the bitterest."

— Confucius

This chapter will feature a discussion of the results and a reflection on the limitations of this research. The researcher acknowledges that a lot of mistakes have been made during the research, reflection on these mistakes is the only way to truly learn. The end of the chapter will feature some recommendations for future research.

9.1 DISCUSSION

This was a first of it's kind study into the application of maturity models on the image authentication process. Usually, maturity models are applied onto a full organization and takes into account all of the organizations processes. Maturity models of specific processes are rare, since maturity is mostly judged organization wide.

The maturity model provides a good basis for the continuation of research into IA process improvement and maturity. Since no earlier Maturity models had been made on the subject. Still, the need for a maturity model in this process seems to be low, as currently the amount of cases is very low.

Deepfake technology is definitely getting more advanced very quickly, and the answers to the problems this causes aren't exactly clear or straightforward. More research into the effects of deepfakes on the judicial system are needed. Currently the amount of cases in which deepfakes are a factors is low, but given the fast development of neural networks and deepfake technology, it is highly likely that this will become a bigger problem in the future.

Deepfake detection technology is unlikely to keep up with the rapid development of deepfake technology to be a viable option in forensic investigations. Like anti-virus software, it has it's places in some systems for cybercrime prevention, but this is more a case for the individual companies having to implement this into their system than forensic organizations doing something about this. The deepfake detection companies can make the software work well for known technologies in a controlled environment, such as ones that require live video.

Improving the context analysis part of the image authentication process seems to be the most effective method for preventing manipulated images. Researching more effective forensic techniques to do the context analysis is recommended. Next to implementing all of the different context analysis methods that are described in the ENFSI best practice manual. Which is the most comprehensive piece of literature available for image authentication.

One of the things that was not a big topic in this thesis was the lack of knowledge in the legal community regarding all things cybercrime. This poses a big problem for properly dealing with cybercrime related issues legally. This includes deepfakes. One of the respondents was directly working on increasing the knowledge of legal professionals regarding cybercrime.

9.2 LIMITATIONS

This maturity model was the first of its kind, it took a while for the research to take the shape of it's current form, therefore the setup of the interviews was slightly flawed. More attention could have been spent on gaining the specific factors that could be relevant for the maturity in image authentication in digital forensics organizations.

9.2.1 *Integrative literature review*

The literature on maturity models in digital forensics was rare, the field is relatively new. The origin of maturity models was in software engineering companies. Which might have something to do with the lack of maturity models in this area. Another reason could be that digital forensics is a highly specialized area, with not a lot of different organizations offering services in this field. In the Netherlands, there are only a few independent companies outside of the NFI offering digital forensic services.

9.2.2 *Interviews*

In hindsight, the interview methodology was not fleshed out well enough to get the results needed for the maturity model. Initially this was due to poor preparation on the part of the researcher and a lack of knowledge into qualitative research methods. After the initial interviews, in which it became clear that the original idea for the research was not going to work in the way I expected it, the scope was redefined to it's current form. The interviews would have some of the information I needed for this to work, but also missed a lot of information due to the difference in scopes. This was kind of remediated in later interviews, but not entirely, since the basis for the research methodology was not prepared well enough. This led to problems later on in the research, as the analysis of the research was difficult without a predefined analysis method. The lack of prior research into digital forensics maturity models also contributed to this, since there were limited examples available on how to do similar qualitative research.

In future research, the research method of the qualitative data should be prepared better. To get more accurate empirical results.

9.2.3 *Maturity model Design*

The current design of the maturity model is lacking in theoretical basis, due to the limited amount of literature available on the subject and the ad-hoc approach to the interviews. Nevertheless, for an initial analysis of the factors influencing maturity in image authentication processes in digital forensics organizations, the model is a good start. Some of the factors should be supported more and be defined a more measurably. It was difficult to put numbers to this since the research was fairly general and did not go in depth on the aspects that maybe needed a bit more.

9.2.4 *Validation*

The validation part was limited due to time constraints, in an ideal scenario, a full expert panel would be used to validate the model. Unfortunately, it was not possible to arrange this within the available time frame. Nevertheless the validation should be sufficient for the application of the model in the problem context.

9.3 future research & recommendations

In future research, the model can further be evaluated for it's effectiveness and more factors of maturity in image authentication can be added. While the factors that are currently in the model are most likely valid, the validation of the model and the fact that this was the first time a model like this had been made for image authentication all suggest that more factors have to be researched for the model. In this research, different digital forensics organizations can be taken into account to more accurately reflect the maturity factors in the industry. The current model might be too limited to the information given by the NFI, this limits the amount of perspectives that the model can provide.

Since this is qualitative research, some of the factors that were determined to have a positive effect on image authentication maturity have been determined based on the opinions of experts. What is currently missing in the research is the empirical results of the benefits these factors provide to a digital forensics organization. This makes it difficult to determine the business value of implementing these capabilities in the organization. A cost-benefit analysis should be done of the cost of the factor versus the expected benefit to the image authentication process.

I recommend take the thesis result as a guide in implementing new parts of their image authentication process as they are needed. The factors should be evaluated based on their effectiveness before they are implemented.

Part IV

APPENDIX

A

APPENDIX A

A.1 INTERVIEWS

As part of the research, interviews where conducted with various stakeholders of the business process to determine the requirements of a new situation and to get a clear picture of the current situation. The purpose of the interviews can be found in the methodology. The emphasized are indicates the response of the respondent while the emphasized Q represents the interviewer asking additional questions.

A.1.1 *Image Researcher #1 - NFI*

This was an interview with an image researcher at the NFI. The most important information that can be gained from this interview is an insight into the current image authentication process. And to get an idea what the current capabilities are for the capability model.

Interview Notes

1. Can you tell me a bit about yourself and your background?

   R: I am a forensic analyst at the NFI. I graduated from a Master in Aerospace Engineering at the TU Delft in 2003 and have been working at the NFI as a forensic analyst since.

2. What do you do at the NFI?

   R: I work in the digital forensics department as an forensic analyst. One of my specialties is image research. I am one of the 4 image researchers in our department. Another one of my specialties is retrieving video material at bit-level.

3. How does the process of analyzing evidence work?

   R: It depends on the claim made by the prosecutor or defense of a case. Every case is different and requires different strategies and resources to solve.

4. What information do you receive from a court case?

   R: We usually receive a request for investigation, which includes the video material and the claim made about the material. We discuss this request within the team to see if we can do anything with it. If we think we can analyze this piece of evidence then we return a rapport about how they submit a customized request and give suggestions as to which claims they can investigate.

5. What information are you expected to return to the court?

   R: If the customer decides to submit this request then we will do the research and return a report containing an answer to the claims being made. Since for most requests the forensic scientist can't answer a definitive yes or no to a question, the NFI uses likelihood terms to describe the probability of a certain claim being true. For some types of research we give a likelihood ratio based on two hypothesis that were set at the beginning of the research. For example, hypothesis 1 could be that the person in the video is the same person as the suspect and hypothesis 2 that it is not the same person as in the video. The answer in the report would then be: It is extremely more likely that hypothesis 1 is true than when hypothesis 2 is true(h1 is more than 1,000,000 times more likely than h2). Or it is about as likely that hypothesis 1 is true as hypothesis 2 is true(h1 is about 1 -2 times as likely as h2). (Vakblad waarschijnlijkheidstermen)

   Some types of research only have 1 hypothesis, the calculation is the same but the way we phrase the answer is different. For 1 hypothesis we might say that is is highly unlikely that the hypothesis is true,

   We never give a definitive conclusion, the decision as to whether a certain piece of evidence is relevant in a court case is decided in the court. The NFI only gives their expertise and insight into a piece of evidence, but does not draw any conclusions as to whether someone is guilty or not. This is decided by the court using our explanation

   We are also expected to give a detailed explanation of our answer so that the people in the courtroom can understand how we came to the decision. This usually involves a description of the clues we found that gave us information as to how

6. How often does the case involve manipulated videos?

   R: The question into authenticity of video material is still rare, on average maybe 1 per year.

7. How many hours do you estimate go into the entire process?

   R: It's hard to say since there are so many different types of cases. Some cases may take 40-50 hours, while some more extensive cases might require upwards of 100-150 hours. A simple case such as identifying a suspect in video material, especially when he has very distinct face tattoos or other distinct bodily features might take around 40 hours. While for example a case where we needed to figure out whether a car was edited into footage of traffic took around 150-200 hours. This is an extreme case but still illustrates the huge differences between cases.

8. How are deepfake detection algorithms used in the process?

   R: Detection algorithms might be used once if we get a case like that, but whether we do anything with the result is another question.

9. Do you see a possibility for deepfake detection algorithms to aid in this process?

   R: It might be useful if we know the conditions that the neural network was trained in and whether this reflects the scenario which we have in the evidence. Often times the data that is used to train and test a deepfake algorithm is not representative of the evidence we have to investigate. The evidence we have to investigate are often very different from the videos that are used to train a deepfake detection algorithm. So if an algorithm has an accuracy of 99% on their train and test sample, we still have no clue what the accuracy of the algorithm is on our real-world examples. Another issue is that most deepfake detection algorithms offer no real insight into what clues the algorithm used to base their decision on. Being able to justify how we came to a decision is one of the requirements that we have for the total analysis of a case.

10. If a deepfake detection algorithm was run and pre-sorted possible clues of manipulation, would this cut down on the time needed to analyze a piece of evidence?

    R: This would probably help if all of the previously discussed conditions were met. Although the main part of the work would still be the full analysis by 3 experts.

11. Do you have any information that you feel I might have skipped over during the interview?

    R: Something that is also difficult in determining the authenticity of video material is the fact that we don't really know what the current capabilities of deepfake generation software is and the amount of technical knowledge that I needed to create a deepfake. This is very relevant since it is part of the likelihood calculations.

A.1.2 *Image Researcher #2 - NFI*

This respondent is an image researcher at the NFI and has several other functions in Digital forensics related jobs and organizations. He is also the chief R&D at the digital forensics department at the NFI. His perspective is very valuable to the research as the current capabilities of the NFI can be measured through this interview. As well as the current state of research on deepfakes at the NFI.

Interview Notes

1. Can you tell me a bit about yourself and what you do?

   R: I work at the NFI on several different divisions. I've promoted on searching image databases in 2003. I've been working at the department image research as image researcher since 1997. I've also worked in this department as research and development coordinator. One day a week I work as an endowed professor at the University of Amsterdam on the topic of Forensic Data Science. Additionally I am the chairman of the forensic IT work group at the European Network of Forensic Science institutes(ENFSI).

2. What is your role in the analysis of digital evidence?

   R: My role is on the one-side case research, so doing something with the video evidence we receive, like for example facial-comparison or investigating another type of claim. Additionally, research and development I the area of digital evidence, which includes but is not limited to finding new areas of research, participating in EU-research projects, and guiding interns that do research for us.

3. How often do you currently receive digital evidence for analysis for any purpose?

   R: Not very often. Most of the time we get these questions from the International courts like the ICC. And very rarely from the criminal court. In the 90s we got the question very often regarding child-pornography, because the creation of virtual child pornography was not a punishable offense. But at some point we were completely flooded with claims so they changed the laws surrounding this. Sometimes we get a criminal case where the defense claims that the evidence is manipulated, most of the time we find no evidence to back these claims up

4. Are there any protocols that you can follow for the analysis of manipulation claims?

   R: It is basically all custom work, we have some guidelines of how the process works in our system. Those are drawn from the best practice guide of the ENFSI, so that is in our quality assurance system.

5. What is the expected throughput time for a single piece of evidence?

   R: Usually from the moment of request until the final response to the customer, it's about 2 months. But this differs from case to case.

6. How do you receive the evidence from a customer?

R: The evidence gets delivered through a CD, DVD or over the internet.

7. What information about the evidence do you usually have when it comes from a court case?

   R: The video file and a paper request form, that's usually all. Sometimes we get some information about where the evidence comes from, but they often don't want us to have a bias so they leave out information that might affect our decision making. Whether or not we get the meta data associated with the files depends on the requester.

8. Which parties are able to submit requests to the NFI?

   R: There are various parties that could submit a request. Basically any government agency is able to submit a request. Below a small list of possible customers.

   - The police
   - Criminal Justice Courts
   - The public ministery(OM)
   - Public prosecutor(Officer of Justice)
   - International courts (Ex.: International Criminal Court of The Hague)
   - The social recherche
   - AIVD
   - Investigative service of financial and fiscal crime(FIOD)

9. How do you utilize Bayesian statistics to formulate your answer to the court?

   R: We setup two hypotheses, one says that it is manipulated and one that says it is not. And then we describe the likelihood ratios of each hypothesis being true. This is accompanied by a description as to why we think that these likelihood ratios represent the hypotheses. If these are not clear in our final report, we can be called to the court to explain the results.

10. How do you determine the priori probabilities when calculating the likelihood ratio?

    R: We usually decide this through discussion among the image research experts, and then an estimation is made. We will write in the report why we think that these prior probabilities are correct, but it is very difficult to determine it accurately, since there are many variables. In the case of deepfakes for example, we need to know whether someone has the knowledge, software, hardware and foresight to manipulate the evidence. This type of information is often not present.

11. How many actual man-hours go into the analysis of a single piece of evidence?

    R: We have a capacity model where this is described in detail, it is not something I know from memory. Usually for the preliminary research it's about 60 hours. If you want to know the specifics you can request the capacity model.

12. How do you determine whether the possibility of manipulation is present in a case?

    R: You have to look at the chain of evidence, whether someone has the capabilities to make deepfakes or even had the possibility of manipulating the evidence. If someone has had no possibility of interacting with the evidence then it is highly unlikely that it could have been manipulated. You have best practice guides for this from ENFSI.

13. Would deepfake detection methods be able to make the process more efficient?

    R: I see a possibility for it. We could use it for getting clues that might indicate manipulation.

14. What would a deepfake detection software have to be able to do to be used in the analysis?

    R: It would have to have a component of explainable AI in it. Without this component, it won't be of use since we can't use the results of a deepfake detection technology directly. We have to be able to explain our results.

15. Does the NFI have a general idea of what is currently possible with deepfake technology?

    R: We have a general idea since we have already done some research into it. So we know most of the current capabilities of deepfake technology.

A.1.3 *Legal professional - Cybercrime Knowledge Centre / Court of the Hague*

This interview is with a legal professional working for the High court in The Hague and the Cyber-crime knowledge center. The reason for the interview with this respondent, is since they can provide the perspective of judges that have to deal with a case involving deepfakes, and can also tell me about the perspective of lawyers. This is largely a case of figuring out how big the problem is for the legal community

Interview Notes

1. Can you tell me something about yourself and what you do?

R: I work for the Cyber Crime knowledge center. I am usually working on research in cybercrime and the law, I'm also a legal professional the Court of The Hague.

2. What is goal of the cyber crime knowledge center?

   R: The goal is to increase the knowledge on the topic of cybercrime within the rule of law. At the moment there is very little knowledge under judges and lawyers in this area. With the exception of a few judges and lawyers who have already dealt with these type of cases. Sometimes we get asked what telegram or signal is for example, that should fall under the umbrella of basic knowledge. If we look at the expectations in the increase in cases in which cybercrime is the main subject then the knowledge is way too low at the moment.

3. What is your specific area of expertise?

   R: I have mostly been busy with analyzing crypto transactions. Money laundring cases are about money, but in the future it will mostly involve cryptocurrencies, since these are hard to track and control.

4. What is you vision on judging the authenticity of digital evidence?

   R: This is a very interesting problem which we are currently working on with a project group. There are several ways to tackle this problem with legislature.

   There are legally 2 classifications on which you can classify such a defense; An alternative scenario or a reliability defence.

5. What definition of deepfakes do you work with?

   R: It is pretty difficult to give a specific definition to this, since there are many different and valid definitions. We use the audiovisual material that has been manipulated using artificial intelligence. If you just include video material manipulated using AI then you are selling the technology short, since as you know there are also audio and image deepfakes. .....

6. Is the cyber crime knowledge center aware of the current capabilities of deepfake technology?

   R: We are keeping track of which applications there are to generate deepfakes. We are mostly aware of the the capabilities of deepfakes detection technologies. We don't think it is very likely that the public ministry enters a deepfake as incriminating evidence into a case. We have to assume that the material that the public ministry brings in as evidence is authentic, unless there is a significant reason to doubt that fact.

7. Is the manipulation of evidence something that is getting more attention in the law community right now?

   R: Since November deepfakes have become a larger topic. Before that it was already a topic but the level of research surrounding it was low. Knowledge in the general law community is very low, a few judges do have a lot of knowledge, but most of them have a low knowledge about cyber crime in general.

8. How many cases have you seen where the claim was made that deepfakes were used to call the authenticity of evidence into question?

   R: Next to none, I work for the Court in The Hague, which basically only handles cases that are appealed. Which might be why I haven't come across any of these cases. I am aware of a case that is currently busy, in which the claim was made that access was given to a banking system with the use of AI generated images, while this shouldn't have been possible. Whether this claim has any grounds to support it or whether this defense will lead anywhere is currently unclear. Sometime around June there will be a decision so then you can contact me about this case and I can tell you how it went.

9. Do you expect the amount of claims to increase?

   R: It is a real possibility that the amount of defenses in which deepfakes are claimed increases. It is kind of surprising that it's not a big subject already. It might be relevant in phishing cases. It is kind of an interesting method to use while phishing.

10. How does the collaboration between the the court and the NFI work when analysing evidence?

    R: The public ministry has the possibility to request an investigation before the case goes to court. Very often the evidence has already been analysed before the case goes to court, so the defense can't claim that the evidence gets manipulated.

11. How would the response of the NFI be built up to make it presentable in court?

    R: The meaning of the question is not clear to me. The NFI gives a likelihood ratio about whether a digital trace originates from a certain source(sourcelevel) or whether it can be explained using a certain factual event(event level). The NFI determines how they build up this likelihood ratio, the law doesn't put any requirements to the way that the NFI has to present the evidence or how to explain the results. They do have to be able to explain the level of expertise of a NFI-professional, through for example the quality requirements of the NRGD(Dutch Register of Law experts). The judge decides the value of the probability judgement that the NFI made in the context of the case during the hearing.

12. Do you think something is missing in the support that is given by the NFI in the context of evidence analysis?

    R: That question is better asked to the police or to the public ministry, since the NFI supports them in analyzing evidence, not the court.

13. Do you see any other methods of authenticating digital evidence without having the NFI do all the work?

    R: Well the NFI is not the only organisation of experts that can handle the analysis of such evidence. There are more forensic experts linked to the NRGD that can analyse evidence for cases.

14. Do you have any information that I maybe have left out during questioning that you think is important for me to know?

    R: I think the most important part of this issue now is to create a clear division between when a claim from a suspect should be taken seriously and when this is not the case. Not every claim needs to be analysed at the same level of precision, some claims can already be refuted before it is analysed just because of the context information of the case or the way the evidence is structured. There needs to be a clear framework for how to handle cases, similarly to the hack-defense that was used often in child pornography cases.

A.1.4 *Strategic Digital Innovation manager - Dutch National Police*

This respondent is the strategic digital innovation Manager at the Dutch National police. As part of her work she looks into new technologies and the impact they might have on police work. They can provide the perspective of the police who is also dealing with deepfakes and will be the first line which handles new evidence. They can give an insight in how the collaboration with the NFI works. Unfortunately, the audio recording stopped three-quarters through the interview, so some questions and answers have been lost.

Interview Notes

1. Can you tell me a bit about yourself and what you do at the Dutch police?

   R: I'm the strategic digital specialist at the Dutch National Police. Which means that I look at new technologies and what they mean for society and with the police work. This includes integrating the work with our mission to be vigilant and of service to the values of the rule of law. So really looking into the pro's and con's of technology, especially their long term effects.

Sometimes technology can be of value in police research, but what are their long term effects. For example, think of facial recognition which can be really useful but also has a lot of bias.

2. What is your role at the Dutch police?

   R: Currently I'm looking into 3 big themes, connected society, which is about the internet of everything. Everything is being fit with sensors, what happens behind the curtains with that data. Which assumptions, algorithms, values are associated with the use of that data. And what does that mean for the autonomy of people in society. So what does that mean for the rule of law. Where are the opportunities, the threats for the police work.

   The second theme is synthetic media and deepfakes. Of course I'm also working with deepfakes, looking and the opportunities and the risks. But the big consideration for the police is how do we arrange a process in such a way, with the right checks and balances. That we can give meaning to digital information in a certain context, so that decisions can be made based on that information. Because in essence we can no longer rely on our own perception, because everything is becoming fake.

   The third theme is about the metaverse, NFTs, blockchain which is irrelevant for this interview.

3. What definition of deepfakes do you work with?

   R: Well there are two definitions, the traditional definition, which is kind of obsolete. Which was a kind of deep learning specifically focused on faceswapping with GANs. With which you could swap someones face or synthesize completely new faces. The definition for deepfakes which is used now is all synthetic or manipulated media generated by AI. The original definition also had a negative sound too it, in which it was mostly relevant in a criminal context or misinformation. While most applications for deepfakes are actually positive!

   Q: I don't agree with that statement, but continue.

   R: Maybe if you just count the amount of cases of sextortion for example, then yes the volume of negative applications is large, but if you look outside of those example then most applications are positive. Such as for example Val Kilmer who had throat cancer and they were able to clone her voice, or allowing people with ALS to retain their original voice. Or on website which you can include your own face to allow for more inclusivity. But what for me and the police is most relevant is that the experts say that within 5 years 90% of content online is synthetic in some way. For example video conferencing might include deepfake technology to change the facial expressions of people to always look at the screen or keep eye contact with the camera, while inherently there is nothing wrong with this technology. In the case of the police talking to a victim or a doctor who talks to a patient remotely, they might rely on subtle facial expressions to derive what is really going on with the victim/patient. If the information that they receive is synthetic in some way, they might draw the wrong conclusions. So it is important to be aware that this is happening. The essence for me is that when 90% of content online is synthetic, then does it even matter whether images have been manipulated or not, what is important is the value of the information that is still present in the image. How do we construct the argument of which value you can give some information in the context in which you want to apply it.

We started in September with a movement from the police since this is coming our way, to work together with academia and the NFI to tackle this problem. And we came to the conclusion that we needed to create a process of sufficient quality, with the right checks and balances, by asking the right questions, so that you can sufficiently support a certain decision.

Q: So you want to avoid having the NFI analyse evidence when it is not relevant to the case?

R: Yes, in essence that is what we would like to do. Since if someone says that an image is manipulated, they are likely to be right, since the software that created the image might have manipulated it.

Q: Well the suspect can't just say that the image has been manipulated, he has to point to a specific part of the image that was manipulated before the NFI will analyze it. They have to at least back up their claim and say which parts of the image was manipulated before such a claim will be investigated.

R: Well yeah, but often there is much more data available, for example meta-data of the video that includes location data or the IP-address that it was uploaded from. They have several other things surrounding it, so this information also has to be taken into account when investigating such a claim.

For example, say someone uploads a video to the police that shows his neighbour committing some criminal offense. In the past we might have reacted to that video immediately, nowadays we first have to analyse the evidence carefully and determine, who is the suspect here actually, the neighbour or the one uploading the video? This is why we need a process for determining the value of the information that is present in evidence.

4. Which projects are currently running at the police on the topic of deepfakes?

R:There is one big project within it there are several projects, so with the judicial powers, academia, the public ministry and lawyers. These projects are to look at how we can better construct a process so that the judge knows what questions to ask and not blindly follow all defenses. For example, in the past we had people who used the defense: "My computer was hacked". Well he's probably right since there is a high probability that there are some traces of malware on every system, but what does this fact change about the value of the information of the evidence in it's current context. Because in 99% of cases its not relevant to the case if here is malware on the device. So the judge had to learn to understand that and to ask the right questions. In the beginning we had to show that the computer was not hacked.

In the case of deepfakes we are trying to look at the front of the chain, what does this development mean for our working process. Looking at where the weaknesses are in our processes. For example, we use passport photo's for the request process of drivers licences, ID cards and passports. But the current request process is insecure, because everyone can supply their own photos, and with deepfake technology it is possible to create a photo that for a human looks like person A, but for facial recognition on person B. So we have to look with the ministries at how we can make the total process more secure, so we don't have to deal with the aftermath of a synthetic photo at the end of the chain.

The primary focus of our projects is awareness and training surrounding the subject, and discovering together what these technologies mean for us.

5. How does the police currently handle digital evidence?

   R: We do some research, if we get video evidence, we also look into whether it might be manipulated or not. But not in the context of AI generated manipulations. There is no good tooling available for that, and there will never be one.

   Q:There will never be good tooling? Duckduckgoose says they have one.

   R: No, they're probably doing their best, but this problem won't be solved with tooling. If you look at the current best tool, that was tested in a big international challenge, it only detects 65% of the deepfakes in the wild. Then it would be very special if they rise above that and the question remains, who can confirm their claim aside from themselves.

   Another thing that often isn't mentioned is the amount of false positives. You can imagine that in policework it is very important that we don't say that something is fake when in reality its real. So tooling is not the answer, because this is about traditional deepfake technology. Deepfakes that are created with new technology are often not detected by older detection methods.

Something in detection that can be useful is a browser extension that keeps track of older known deepfakes and alerts users on them. For example, a lot of people use pictures from thispersondoesnotexist.com, all detection methods are now able to detect these pictures.

The important thing to remember is that everything will be synthetic media in the future, deepfakes are nothing different than synthetic media. So how are you going to make a tool that separates the different contexts from each other, which piece of media was made in the context of cybercrime or disinformation and which was made for entertainment.

6. Do you have a clear picture of what the possibilities are with current deepfake technology?

   R: Yes we have a clear picture. The several projects related to awareness of the subject has contributed a lot to the knowledge surrounding deepfakes. We are not only looking at negative applications but positive ones too, so we can integrate synthetic media into our own processes. For example, we are looking into using virtual police officers to interact with the public. But we have learned from New-Zealand where they have a similar virtual agents, that it currently does not work due too technological limitations.

   Another application is one from the police in Australia, Victoria. They have a colleague who committed suicide, so they created a deepfake from him in the context of suicide prevention.

7. In your current deepfake strategy, have you looked into prevention methods such as proof of authenticity systems?

   R: Yes, but the blockchain won't help with this. Blockchain has one vulnerability, everyone can put everything onto it. So you still have to trust someone. For example, there was this initiative in the Netherlands with putting free-range eggs on the blockchain. So how can I know for sure, that when the stamp is put on the egg and it is put on the blockchain, that it was actually a free range egg before it was put on the blockchain?

   Q: But what about the moment you record a video?

   R: Yes you can add a hash value to the metadata but no blockchain is needed for that. We do that with bodycams for example. Same thing when we confiscate a computer, we make an image and at that moment a type of hash value is created for the image. So that you always know that it is the original image you are working with. Within the police we have to look at the bigger picture at how we can authenticate our own information, so that we can remain a trusted party. Which is separate from deepfakes and more about trust in a digitalized society.

But it is very difficult to do with deepfakes, someone might send us a video, but if we create a hash at that moment there is no point to the hash, it doesn't tell us if it was manipulated before the video came to us.

There is a content authenticity initiative and witness.org is one of the most important trackers. It is aimed at journalists, because you have a lot of activist journalists in authoritarian regimes. For example they want that their information can be used as evidence for war crimes. So how do you add a sort of authenticity stamp when they create this evidence on their phone which includes crucial information about the evidence, without revealing too much about the creator of the evidence and potentially putting them in danger. There are big initiatives and it's very complex, so it's not something we as the Netherlands or the police have to think of on our own. I suggest that we join one of those content authenticity initiative. And use those findings in our own environment.

8. Does the police ever have to send digital evidence to the NFI? If so, what are the expectations on delivering this type of evidence?

   R: In the context of the deepfakes we have had no cases. Sometimes there are possibly deepfakes in a case, but whether they are deepfakes is then not relevant to the case.

   Q: What do you mean?

   R: Well for example, if you have a case where someone is being extorted over a private porn video, then its not relevant to the case whether this video is actually real or a deepfake. It could be a deepfake, but it's not relevant for the case. The crime that is being committed is extortion.

A.1.5 *Duckduckgoose - Deepfake Detection startup*

This interview is with two of the founders of the deepfake detection startup Duckduckgoose. The company aims to create a deepfake detection software that can be used by consumers and organizations like the NFI to detect deepfakes quickly and effectively. Since there are 2 persons in the interview the CEO will be marked as R1 and the CTO as R2. *Interview Notes*

1. Can you tell me a bit about yourself and your background?

   R1: Well we started during a minor from the TU Delft called tech-based entrepreneurship. In which all kinds of students from different study programs participated. The aim was to look for a socio-technical problem in society and find a solution to that problem including finding a market with potential customers. Quickly we thought of deepfakes as a possibility, because at that time, two years ago, they were even more new than now. It was really an emerging technology. So we thought, if we tackle this then we have a new societal problem with a lot of impact, including a solution, which does not happen often.

   So we started out by making a prototype and talking to people who were also dealing with this problem such as Zeno and people from the police. And after 6 months we decided that we would start a real start-up. That's how we started, and now we are here and we now focus on detecting deepfake faces in videos and images. And implementing this with explainable AI techniques.

2. Can you tell me a bit about duckduckgoose?

   R1: Our solution is used by Forensic institutes in the Netherlands and in foreign countries. And aside from that we are looking into whether we can provide a solution to the digital-id verification world. For example, banks with who you can open a bank account with a selfie and a passport photo. Since selfies are very deepfake fraud sensitive, a solution to this can be very valuable. Someone can make a deepfake video of me and open a bank account in my name.

   Q: Since all of these processes are automated?

   Yes, the banks have automated this process. It is cheaper and more efficient to automate this stuff. Especially after the pandemic all of these things have been automated.

3. How did you get interested in Deepfake detection?

   R1: We saw that everywhere where you are dealing with image and sound there is a danger for deepfakes being misused. So we wanted to supply all markets with the right tooling for detecting deepfakes.

4. What kind of method does your deepfake detection software use?

   R2: That is a very good question. We use artificial intelligence for our deepfake detection. So it's a bit of an AI versus AI problem for the creation and detection of deepfake material. And we use several neural networks for that which can supplement eachother to create a classification of whether something is a deepfake. The question of what does such a neural network look for to decide to classify it as a deepfake, is not really feasible. You could write a very good thesis about that.

Q: But you do know what type of data you feed the algorithm, either frames or consecutive frames?

R2: We do frame based classification, but we are able to see the differences between consecutive frames as well. There are multiple different ways to feed your algorithm, and we aren't convinced that one method is the best. So we experiment with multiple different methods to what best fits our use case

R1: It is a very ongoing process, so you have to keep improving, keep trying out new things. One of the things we looked at in the short term is using multiple modalities. So not just looking at the frames or the differences between them, but taking audio into account as well.

5. How do you ensure the deepfake software is capable of handling the latest deepfakes?

   R2: We do this by keeping up with the latest scientific papers and deepfakes on the internet. As well as making our own training material to better understand how deepfakes are made and to ensure that our algorithm is not only trained on existing datasets. Which might cause an overfit of the data to the model.

6. What kind of accuracy do you reach with your model on deepfakes in the wild?

   R2: That is a very good question. We have numbers of how are model performs on the test set of our training dataset, which is at around 93% accurate. About deepfakes in the wild we don't really have a measurement, but if you can hand us a dataset as part of your thesis we would be very interested in testing it out for you.

   Q: Some papers will come out soon probably with new datasets.

   R2: Of course, and when they come out we will add them to our repertoire of training datasets.

   R1: One of the reasons for our partnership with Zeno is that the NFI will test it with data that we don't have in our trainings datasets. And then they give us feedback about how well the model performed, which helps us a lot in improving our model and further developing our training material.

7. Do you think you will ever be able to tell a company with certainty that your tool will detect a certain high percentage of deepfakes?

R1: yes that is a very real possibility. When we talk about certain types of deepfakes, for example the deepfakes of non-existant persons, the styleGANs, we can detect them with a very high accuracy. But if we are talking about new types of deepfakes that get created in the future, by criminals or scientists, then it's more difficult. At the moment we can't really ensure that we detect them with a high confidence. But we are busy with research to solve it, so even when it is a deepfake type that we don't know, that we can't detect, that we can communicate this with the client and end-user and give assurances that way.

R2: What I think we can offer is the guarantee that we keep up with recent development, so we can add new types of deepfakes fast. Just like an anti-virus software never 100% is able to detect all viruses, it's only safe until the next exploit is found or a genius hacker thinks of some new virus. Then it won't work anymore, but the anti-virus companies deal with this by providing a quick update, I'm convinced we can do a similar thing.

8. How do you ensure that your algorithm is up-to-date?

   R2: we keep up with the most recent scientific developments. We see that the innovation in the deepfake world mostly comes from the academic world. Not as much the criminal corner as is more common in the virus story. So we keep up to date with the publications, and our tech team adapts based on the findings in those papers.

   R1: and by making our own deepfakes we understand deepfakes better. We also create training material that is not in any of the datasets.

9. How do you ensure that when deepfakes are no longer able to be seen as fake by humans, that you algorithm will be able to detect them?

   R1: We think that when this is the case then especially algorithms will have to be used to detect them. Since human experts can only see so much, while models can see what we can't see and base their decision around that.

   Q: But if we can't see the difference, how will you be able to trust that the algorithm gets it right?

   R1: that is a very good question. We highlight the area's in the image on which the algorithm has based its classification. It's what we call explainable AI and we show what the network has found in the image. So we can give the person looking at the results an idea of what could be wrong about the picture. What would happen if the artifacts that the result is based on is no longer visible to the human eye? Well that is a question that we have to think about. We have some ideas about it, but nothing concrete yet. It is one of the challenges which we have to solve for the future.

10. I read a paper that stated that 90% of the content on the internet within 5 years is synthetic in some form, such as filters from TikTok, Snapchat, Instagram, etc. How do you ensure that your model differentiates between this type of synthetic content versus more harmful deepfakes?

    R1: That is something we are working on. We asked ourselves the question, well what if we became an analytics tool in Instagram or TikTok. On those platforms you can also find deepfakes for disinformation, but also kids who make themselves look different with filters. How are we going to detect the difference, are we only going to detect the bad ones from the harmless ones. How do you handle all of the context information and those parameters? These are challenges that we have to work out.

    Q: Well if it is created and uploaded directly onto the app then you would have the metadata information of that video so you could know what kind of filter was used and how this affected the image?

    R1: Yes we could do that.

11. Are you the only deepfake detection start-up that offers a tool? Because as far as I know in the academic world it's purely theoretical in how deepfakes could be detected and no tools are available.

    R1: No, as far as we know there are a few more start-ups that offer deepfake detection like us. They say that they offer about the same as us in terms of classification, but whether they actually offer a product similar to ours I'm not sure. Everywhere we went we were the first one offering something like this, so that's what we do know. Something we are at least unique in is the explainable AI component of our product.

    Q: That(explainable AI) is also very important in the judicial system.

    R1: at the moment it is. In a few years it will be mandatory though. The European commission is currently working on a law for regulating AI applications. In those laws they make a separation between AI systems that play a role in someone's life or society. And the more important your AI system is, the more requirements you have to follow, in area's such as transparency and insight into the workings of the AI system.

12. Are you working on other types of applications for deepfakes outside of detection, such as prevention systems?

R2: We give presentations to spread awareness of the issue, so making people more aware of what they see online. Because a lot of people just assume that something is true if they see images without really thinking about it. So we want to stimulate thinking a bit more critically about what you see online.

On a more technical level, we are also focusing more on the multiple modalities such as combining audio and video. But in the short term that's it. We have looked into AI generated text but we don't see many applications for this so it's not worth investing in at the moment.

R1: We have also thought about using blockchain to capture authenticity at the moment the video is made. But that was a bit too complex. Another reason that we stayed with detection is the way it would be used. If we partnered with youtube for example and they used our algorithm at the start of the upload process. Then you can prevent deepfakes from being uploaded altogether just by having good detection methods.

13. What applications do you see for your deepfake detection algorithm?

    R1: There are a lot of applications. For example, video conferencing is now very prevalent which is also deepfake sensitive. Deepfakelive is a new deepfake technology which allows for the creation of real-time deepfakes. Another application are dating sites, which apparently also have a lot of deepfakes on them. There has been a case where someone was scammed for $300.000 because they saw a deepfake of an admin. Aside from that there are a lot of deepfakes in adult content, which is the most used application of deepfakes. There is currently no regulation for this, but a lot of people are victim to this type of deepfake.

    In addition, there are applications in news and media, because journalists who work online have to know whether the sources they have and the people they talked to are actually real. In the end we want to go to a full consumer product, a detector on every system. Because you could get a video in an email from a know person or another video, and that could also be fake.

    Q: So you really believe that technology is the solution?

    R1: We believe in a mix of technology and the human capability to discern things. That is why we have the explainable AI component. Technology is of course one of the better ways to tackle these kinds of problems, combined with awareness in the general population.

    Q: When you say this you are assuming that your algorithm is good enough to classify new deepfakes in the wild, because at the deepfake detection challenge, the best algorithm only reached a 65% accuracy on deepfakes in the wild.

R2: Well what we offer is that we are an aid in avoiding the negative effects of deepfakes. We are there to help you get insight into the classification of deepfakes.

R1: One of the key challenges is detecting as many types of deepfakes as possible. I believe that we can build a model which can do that as much as possible. It is dependent on many different factors.

R2: I think we can draw a parallel with the anti-virus world, at some point they probably also thought: "okay, viruses are getting better and better, how are we going to keep up with the firewall?". While in actuality the anti-virus software is pretty good at keeping up with recent developments.

14. Do you have a vision for the future of deepfake detection?

    R1: We want to create a digital environment in which we can believe what we perceive. So we want to give individuals and company's the tools to keep seeing the difference between real and fake wherever necessary.

15. And what are the costs of a license of your software?

    R2: It depends on what kind of solution you are getting, whether you have an enterprise license or and individual license. So I can't give you a singular answer to that question.

    R1: Well as mentioned earlier it depends on several factors, but if we had to say a standard price it would be 6000 euros a year in the case of forensic applications. That is for 1 license, so one computer that can use it.

16. Do you think there are deepfakes out there that cannot be detected with your algorithm?

    R1: Well we actually experimented with that, what we saw was that with a few deepfakes the classification said is was real. But the algorithm showed some highlights, so some activation map outputs. If an image researcher would look at those highlights, then they would find some inconsistencies.

17. Since I forgot to ask in the beginning: What definition for deepfakes do you use?

    R2: The definition we use is video material that is manipulated using AI or fully synthesized by AI technology. Within that definition, we only detect the face swap aspect of those deepfakes, but in the long-term we want to expand that to cover the full definition.

A.1.6 *Public Prosecutor - Public ministry*

This respondent is a public prosecutor for the public ministry. It is important to get his perspective of the process to further increase the knowledge into the needs of customers of the NFI.

1. Can you tell me a bit about yourself and your background?

   R: I've been the National Officer digital investigations for the public ministry for a couple of year. Which is a new function within the public ministry and is meant to increase the knowledge of digital investigations with all public prosecutors. Things like how to safeguard digital evidence, how do you analyze it, and what is the forensic value of digital evidence? Before that I was the national public prosecutor cybercrime, at which I mostly researched international cybercrime. I've been doing this for about 8 years, and before that I was a Fraud investigation officer in The Hague.

   Q: So you are still a public prosecutor?

   R: Yes I am still a public prosecutor, but I don't have many cases at the moment. I will add, I haven't had any cases where the defense has claimed that the video evidence was manipulated. I also haven't had any colleagues that have come across it and came to me for help.

   Q: Is image authentication something that you are working on within your job or is it not yet a subject since the question has not been asked yet?

   R: It is more on the background right now, because we do expect that the defense is coming in the future. So the NFI is also working with us, what do we have to do if these questions are coming in. How can we handle this defense? What kind of software we should use, that sort of thing.

2. What definition of deepfakes are you familiar with?

   R: That is a difficult question. I am not familiar with a specific definition for deepfakes. We had a masterclass deepfakes with the NFI. If you ask me then deepfakes are mostly manipulated media, which is mostly relevant in Criminal law within the sexual offences and identity theft. But it is not limited to impersonating another person(faceswaps).

3. How do you see deepfakes being used in the context of cybercrime?

   R: I see them mostly being used in cases of identity theft and revenge porn. Something I think we are underestimating is the use of deepfakes to extort or scam someone, where someone might be suspected of this crime while in reality not having anything to do with the case. This will undoubtedly be used as a defense in some cases, but to verify this we have to determine at an earlier point in the investigation whether the suspect is actually the suspect. That the suspect was not set-up through the use of a manipulated video.

Q: So in that case most of the investigation into the manipulation of digital evidence will have been done before the case ever comes to court?

R: Yes, and in that case most of the evidence is digital and is the criminal offence committed digitally. So in that case you have to investigate carefully whether the evidence you have against a suspect is reliable and credible. While in other cases it will be used more as a defense of suspects that you have already identified. Such as violence on the street, or statements from well-known individuals that may be manipulated.

Q: Well in that case you would have to figure out whether the manipulation defense is actually a valid concern. Because similarly to the Hack-defense, which I am sure you are aware of, whether or not something is manipulated is might not be relevant to the the burden of proof of the evidence. Someone can shout that a piece of media is manipulated, while this may be true, it doesn't change anything about the evidence that actually matters in the context of the crime that was committed.

R: Well that is an interesting one, the hack-defense we get very often. Where someone says I want my computer to be investigated by the NFI or the police since it has been hacked. Which is a difficult request since is highly likely since most computers contain some kind of malware, especially the computers of those involved in shady activities online.

Q: I would expect that since this defense is so common that there is a standard way of working to handle the situation and most public prosecutors are aware of this. Which is also what I expect would be useful in the case of deepfakes.

R: Actually I still get colleagues asking me about how to handle such cases. But this problem has been scoped in the court and there are standards available on how to handle these cases.

4. Do you do your own research into evidence at the public ministry or do you sent everything to the NFI?

   R: Most of the investigation work is done by the police and the team digital support. So should such a defense come, then firstly the police will try to figure out whether they can investigate this claim, if they have the knowledge to do so. And for most hack-defenses the technical detective will determine whether the defense has any bearing on whether a criminal offence was committed by the suspect. So most of the evidence is analysed without going to the NFI.

5. So you work with the police to setup a case against a suspect, with the evidence that the police has gathered and analyzed?

   R: Yes, so the role of the public ministry and the public prosecutor is to lead the investigation. Which means that if you have a criminal offence, that we work together with the police to figure out how to set-up a case against the suspect and how we will collect and present the evidence. If a suspect then uses the hack-defense or the manipulation defense, then the public prosecutor will determine what to do with these claims and maybe start an investigation into the claims.

   Q: So the public prosecutor is also authorized to say which investigation methods are warranted in a case?

   R: Yes that is typically the role of the public prosecutor. The public prosecutor determines what kind of investigation methods are used and under which conditions they can be used, for example how far you will go with breaking the right of privacy.

   Q: Is it possible then that a public prosecutor decides to use methods which are not lawful in the context of the case, meaning that the evidence that is collected is not usable for the case?

   R: That is a possibility, a public prosecutor in collaboration with the police might make a mistake and use unauthorized methods, but that doesn't necessarily mean that the evidence cannot be used. This is a very big discussion in criminal law, but let me just say that there are very few cases in which evidence obtained in such a manner cannot be used.

6. How do you decide to send evidence to the NFI?

   R: We do this in case we think we don't have the knowledge at the police to judge a claim made by a suspect, and we need the knowledge of an expert to look at the evidence more carefully.

7. Lets say there is a sextortion case in which someone has used revenge porn to extort someone, is it in that case relevant whether the video was real or manipulated?

   R: That depends on what the offense is that the public prosecutor decides to pursue, but you would have to talk to a public prosecutor that is specialized in sexual offences. You could say that by publishing a fake video with the goal to shame someone, you commit an act of libel against that person. That is a different offense and different punishment than when it is categorized as having committed lewd acts or have published revenge porn.

Q: For me the most important part of this issue is when it is relevant to a case whether deepfakes were used.

R: For that case the most important part is what kind of offense we charge the suspect with. Whether this is relevant in a case is dependant on the type of offense that has been committed. So to give a clear answer as to when it is relevant and when it is not is difficult.

8. Are you aware of the detection methods that can be used to detect deepfakes, are there any plans to implement these in your investigations?

   R: I am aware that they exist, we at the public ministry are not planning on using this, but the police and the NFI who actually execute the investigations are planning on using this technology as far as I know. And I would expect them to use this technology a lot more in the future as the software gets better.

   Q: Well that is actually a contentious issue since not only the detection technology will improve, but the deepfake technology will improve with the detection techniques. Which makes it difficult to determine whether the detection methods can keep up with the creation technology.

9. Is there anything you feel I might have missed during the interview?

   R: I feel like you should be very aware of the type of cases that we are likely to get. I think that high-profile libel cases are most likely to contain the manipulation defense in which celebrities make statements that they say they haven't made. Or that they have been put in a bad light in order to attack their personality. This is also since there is a lot more material of these people available to create these deepfakes so the creation of them should be easier.

   Q: Yeah that is something that could happen, like an influencer promoting some cryptocurrency on instagram that turns out to be a rugpull in which the value of the cryptocurrency drops to next to nothing in minutes. The influencer can then say that they didn't say that and that the video in which they promoted this was manipulated. And since there is so much video material available from them that would be a very real possibility.

   R: I expect those type of cases to be the first that will use this defense. I do not expect cases such as nightlife violence in which a lot of video material is available to be featured in court. Since it is highly unlikely that multiple different videos from different sources have been deepfaked.

A.1.7 *Deepfake Detection Company CEO - Sensity AI*

1. Can you tell me a bit about yourself and your background?

   My background is in academia, I have a PHD in machine learning and application to computer vision that I got during studies in Australia. I have multiple post-doc, one that was related to machine learning and encrypted data. And another one at the University of Amsterdam on generative models. Co-founded Sensity 4 years ago, which focuses on deepfake detection and creative technologies. I am currently still leading the company, but I still work a lot in research.

2. Can you tell me a bit about Sensity AI?

   Sensity started out in creative technology, specifically anti-deepfakes, but has evolved to incorporate more creative technologies. We really do the entire track, from the research and development to the commercial application of the technology. We bet on deepfakes in the beginning, but have created more versatile technologies since the applications of deepfake technology were limited, and more could be done with machine learning and facial recognition than just deepfake detection. We use AI, computer vision and digital forensics for anti-fraud applications. So essentially we do a lot of biometrics, like liveness detection, scanning passports for fraud, face matching, we also do checks on pdfs of financial documents, and a lot more. We are essentially targeting security organizations, government organizations, but mostly fintech and banking clients. So deepfakes have become a part of many different security technology on faces.

3. Which definition of deepfakes do you work with?

   What we care about when talking about deepfakes is digital manipulation made with AI. And also we usually use the term specifically in relation to face swapping. So while the general definition is still relevant and correct, it's less relevant for us since most of the applications of deepfakes that can be harmful involve face swapping.

   Such as for example in biometric identification companies. We will release a rapport in which we did penetration testing using a deepfake on biometric identification of companies that can be viewed as our competitors, and we found that most of the systems were not able to detect the deepfake. So we are releasing the software open-source that can be used to detect these types on github.

4. What kind of detection techniques are used in the deepfake detection network of Sensity?

In practice, most of the time with AI the simpler approaches are more effective because there are less assumptions that it's based on. So I think in most cases the latest ideas are just a bit too complex to make it work in practice. We still use ideas of a few years ago that we mix with our own experience and data we collect on the internet. So you can say we use a frame-based approach while also being capable of using the video information as input.

We use deep-learning, so we don't need to tell it what to look at, you can tell it to look at faces, but you don't need to tell it what kind of artifact to extract or to look at.

In a lot of older technologies that don't use deep learning an expert defines some kind of filter of what features to extract from an image and then build the machine learning model on top of it. So the artifacts are extracted manually, like for example jpeg-compression. Deep learning has automated some of this, meaning that if you have enough data that says this is good and this is bad, it can figure out these features on its own. And nowadays you have approaches that combine these two methods, where you apply the deep learning principle but also create a network that extracts specific types of data, like for example looking at certain frequencies on Fourier's base or looking at edges. So I think ours is more of a convergence of the two ideas.

5. Do you think deepfake detection can be applied in forensic situations? If so, how?

Clients do find it useful, so I think they can be applied. It really depends on the type of application, what the scale is and the type of data that has to be analyzed. It is very challenging to detect whether a random video on the internet that someone has spent hours to produce is a deepfake, however if we have the context information surrounding the video or force attackers to use our system with the right checks in place.

I'll give you another situation, in the forensic situation I don't think you want an automated decision, you need a human in the loop. But in other cases where you can constrain the input video in a way that you decide and not the users decide, like for example liveness detection, you can force the user to use your software to record a video. Then the kind of stuff they can do to inject deepfakes, which is possible, is very limited. Therefore a deepfake detector can be very effective and even automated.

Forensic is more difficult. When you do forensic modeling and you want to know how difficult it is for an attacker to do something, which are the things that they can do in a specific situation. If you force them to use your system, the things they can do is very limited and can be automated against. In the forensic analysis this is much more difficult since you don't know how the video was made, someone could have worked for 2 months on a certain video and you wouldn't know.

6. What kind of accuracy does your deepfake detection model currently reach on deepfakes in the wild?

   I prefer not to comment on this.

7. How do you ensure that your model can detect the latest deepfakes?

   We keep up with the latest academic literature. You can draw a comparison between deepfake detection and anti-virus software. No anti-virus company will make the claim that they can defend against all viruses. It is basically the same thing with deepfake detection, we can guarantee that we will keep up with recent developments and that novel deepfakes will be detected as soon as we are aware of them. However, it is impossible to guarantee that we will detect all deepfakes. We can guarantee that the use of the software will lower the probability of being affected by deepfakes.

8. In your opinion, will the detection technology ever reach a high enough accuracy on new deepfakes to warrant it's use of commercial purposes?

   There are definitely commercial purposes in which deepfake detection can reach a high enough accuracy to be viable. It's all dependent on the type of constraints you can apply to the input and what the specific application requires of the video. We already have systems that do fraud detection on passport images and software that can detect deepfakes in live settings. Such as video conferencing and liveliness checks. But in these cases the deepfakes aren't that good and aren't made to fool humans, but they will fool the computer systems if the right checks aren't in place.

9. I read a paper that stated that 90% of the content on the internet within 5 years is synthetic in some form, such as filters from TikTok, Snapchat, Instagram, etc. How do you ensure that your model differentiates between this type of synthetic content versus more harmful deepfakes?

   In most cases this is not feasible and it is also not necessary. I don't think you can train an algorithm to detect whether a manipulation is malicious or not, that is just not what technology can do.

   What we do is that we make the attacker use our system, which is constrained in ways which makes it feasible to detect deepfakes. In cases such as a random video on the internet on which you have no background information, it is very hard to do such an analysis.

10. How do you keep up with recent developments in deepfake technology?

    We keep up with the development by following the literature and by actively participating in it. Tomorrow we are releasing a rapport on the automatic biometric detection systems of some of our competitors, in which we researched whether we could get into their systems using deepfakes for biometric identification. In this research we found that most systems did not have sufficient deepfake detection to prevent our attacks from happening.

11. Are you working on other types of applications for deepfakes outside of detection, such as prevention systems?

    Sensity is working on many different applications which use facial recognition and deep learning to detect attempted fraud or biometric security breaches. We are working on prevention on the commercial level, to prevent deepfakes from being used to infiltrate a system or commit identity theft/fraud. We don't do anything with any social media platforms yet to prevent deepfakes from being uploaded, nor can we control what is uploaded on the internet so we can't create a system that outright prevents deepfakes from being distributed.

12. How do you see image forensics experts using your tool?

    They can use it as a guide during the forensic process, they just can't fully trust on the results of the algorithm, but it can be used to give an indication or to support the analysis of a specific video. The type of information a deepfake detector can still be valuable to the forensic image expert even if he can't trust the result outright. It is a signal that you can use in combination with other information. Its usefulness is also dependent on context information you have about the video. It can be easier to use the deepfake detector if you know the constraints in which the video was taken.

13. What kind of explainable AI features does your software have?

    Yeah, its part of it. It is very difficult to explain a deepfake. Especially in the case where there are no visible artifacts, it's difficult for the model to show exactly what the result was based on. You can explain why the algorithm took a decision, but if you plot it on a video humans might still not be able to see it. So unless there are visible artifacts, or you have the original video, it can be very hard to explain to a human why the model decided on a result.

14. Do you have any other information I might have skipped over in the interview that is relevant to the research?

A.1.8 *Image researcher #3 - NFI*

1. Can you tell me a bit about yourself and your background?

   Originally I was a biologist, I've worked for years on biometric research. Classification of images was something I was always interested in, so it is at the basis of my interests. Around 20 years ago I started working at the NFI in biometric facial comparison.

2. What do you do at the NFI?

   When I started working at the NFI 20 years ago, I started on facial comparison with biometrics, which is something that started then and wasn't very advanced. It was very limited, mostly on controlled images. And continually develop the capabilities on the area of biometrics. For me the automatic analysis is overtaking the humans, definitely with the neural networks. Within certain constraints of course, because just like for humans, the training set that you are used to is crucial to how accurate you are. Although humans are less stable than most automated systems, automated systems give the same result when you input the same image.

   We used to have a lot of new techniques to analyze video material and different techniques for different videos. We used to use a tool that could compare certain features and where you could specify what you were looking for.10 years ago most of the innovation in facial comparisons kind of stalled, so when neural networks became a thing 5 years ago suddenly a lot of innovation in this field was added. In which you don't really specify what the neural network should look for, but it does get the results you want.

3. Can you tell me how the current process of analyzing evidence goes step-by-step?

   The request comes in with the NFI then it comes to us. Then we look at what material comes in and what the question about the material is, which isn't always a match. We will look at the material and whether we can answer the question that the requester asked about the material. If we think, we can't answer this certain question, but we can answer it if it is phrased in a different way or they could ask us something related to the original question. Then we will suggest that after the preliminary research.

Usually the question is, is the person in the video the same person as the suspect. Then the first question we have about the images is whether we have the original video, because we always need the most original image. Then we need to make sure that this is the complete video, or whether any other evidence is related to this evidence. And if this is everything, can we do something with these images. For facial comparisons we would like to record reference images with the suspect in the same circumstances as the original evidence.

We send a letter after the preliminary research in which we say how much time and organization the analysis is going to take. Then we wait for a response and if they want us to analyse it then we will. Sometimes we have to say, with these images we can't do anything, but then he customer can still say that they want us to analyze it. This happens often when a case goes is appealed. And sometimes it turns out that we can do more than we expected.

Then if the custom request comes in we will collect all of the available material, do the analysis and report about it. Sometimes we have to explain the results in court.

4. Do you have a moment of feedback in that process with the customer somewhere?

   Yes, that is possible between the moment of the preliminary research and the custom research. And it happens that there are informal calls between us and the customer. It is only when it's needed, there is not a predetermined moment for feedback.

5. What information do you receive from a court case?

   That depends on the case. Most of the time we receive the video material without context information. Sometimes, like in a case that we are currently handling, we only receive a few frames and questions about those frames, well in that case we ask for the original video, and that can take a while. We usually want the full chain of evidence, that is every piece of information about the evidence, where does it come from, who has already worked on it, and are we working with the original material?

6. How many cases can you handle at once?

   Currently we are capable of handling around 5-10 cases simultaneously. This is with a team of around 6 people who work on various disciplines in the image research. So some people might be better at doing certain types of research.

7. How often do you redefine the existing processes?

Yes the current processes are registered in our quality control system Inception. In that system all of the documents describing our processes, procedures, research methods, etc. are registered.

This is because we have to describe what we do, and do what we describe, so that it can be verified that the process is sound by outside parties. But you have to watch out that you don't start documenting too much. This will get in the way of the actual work and might make the process less efficient.

The cases are also documented in the same system, if we do something that differs from the standard we also document this. And the communication between parties is also documented. We also document how we came to a certain decision. There is a case log in which it is visible who had which communication with whom, who did what and when the rapport was written. We always let a 2nd person follow the rapport to ensure that mistakes are caught. Concept versions and version history are saved as well. Different types of research are also stored in our quality system.

8. Do you keep up with new techniques for your research methods?

   We continually try to refine our current processes. This we do through the extensive R&D efforts that we do through thesis assignments for students, and research done by our own image researchers. We usually let students do pre-liminary research that we feel we don't have time for, and if the results are positive we might continue developing upon that. We also keep up with scientific publications for the development of new techniques. If we find something that is interesting and might work for us, we test it out on older cases and see if it works for us.

9. Do you look into previous cases for how you solved certain cases in the past?

   We sometimes do go back and look at old cases. For example to test new research techniques. To see if we would have gotten the same answer if we used the new technique. So in that case it is part of the validation process of new research methods. If we get similar answers that can be an indication that the new technique is a valid way of working. If we get a different answer then maybe the new technique isn't working properly or we originally came to the wrong conclusion.

A.1.9 *Digital Forensic Specialist - Dutch Police*

1. Can you tell me something about yourself and your background?

I work at the digital investigations unit of the Dutch police since 2008. Before that I worked as a asset officer at Eneco for 27 years. I completed various different engineering studies.

2. What are your tasks at the dutch police?

   The job description is digital forensic detective. I mostly work on safeguarding data, extracting data from different sources, such as mobile phones, security camera's, computers, etc. And making sure that the data is backed-up, ready for analysis, in its original state and authentic. This data is collected into evidence files and saved in the system where the practical detective can access them and analyse them. I then categorize the data, basically I change the 0's and 1's into insightful categorized items. Basically I make sure that the data is able to be searched and that it remains available in it's original state. If you ask me in 10 years, what did you do with the data of this case? Is it still available in it's original state? Then I would be able to say yes, since everything is logged and backed up in it's original state. In addition I do analyses of video material to see if the video is in its original state, or whether the video was manipulated. We also do enhancing of video material to answer questions about it, such as for example extracting car license plates from video where this is not visible due to motion blur.

3. What kind of analyses on video material do you do?

   We start at the beginning, starting with finding out whether we have the original file format. Very often the systems have a native video format in which they are recorded. We have to make sure that the person extracting the video has extracted it in the original format, since a lot of information in the video can be lost during compression or conversion to a different format.

   If the video is compressed then information is lost that can in no way be recovered. So it is very important that the evidence in the case is in it's proprietary format.

4. How do you decide when to send evidence to the NFI for analysis?

   This rarely happens, if we really can't find what we are looking for or the analysis has failed, then we will work together with the NFI to figure out whether we are missing something. Sometimes they have better tooling than us and can help us figure something, but most of the time we have similar capabilities and they won't find anything extra. We do most of the analysis ourselves. If we really can't figure something out then often we will look at it together, but often we have the same results. We have very advanced tools nowadays so most of the time we can use these software tools and get basically the same answer as the NFI. The NFI is mostly responsible for improving the investigation methods we have, which they do through developing new methods to solve specific problems.

Q: But the image you are creating for me is weird, since you say that almost no evidence is sent to the NFI. Yet I know that they work on cases often.

I don't know where else they get questions from, but it's not from us. We do most of the analysis ourselves.

5. Do you send the original case information to the NFI?

   Yes, the original case files are sent to the NFI when we require them to do an investigation.

6. Have you had any experience with cases in which deepfakes were a factor?

   No not yet. We had one case in which someone said that the video was manipulated, but in that case the video material came straight from his own security camera, which didn't make it possible for the video to be manipulated. This made the claim very unlikely. What you have to keep in mind is that most of the video evidence we receive comes straight from security camera's, CCTV, doorbell cameras, etc. When we get the evidence straight from these sources in their original format, it's basically impossible that they are manipulated.

7. At what point in the evidence chain are you analysing the evidence? Where does the evidence come from, what do you do with it, and where does it go after you are done with it

   The tactical detective actually analyses the evidence, and if something is not clear about the picture or has a question about a face or a logo in a video then he comes to us to ask questions about this.

8. How do you obtain the evidence, do you collect the evidence yourself or do you get it from another source?

   The evidence is collected by the police when they arrest a suspect and directly given to us. When a suspect is arrested, we usually get the evidence for indexing and safekeeping the same day. We are mostly responsible for collecting and indexing the data, and we do some specific analyses or transformations is there are parts of the video that are unclear.

9. Are you planning on having more capabilities for dealing with deepfakes in the future?

   Well I am part of a working group in which we had a pilot about deepfakes. In this working group we worked with Duckduckgoose and tried out their detection method on some deepfakes.

> We also had to make our own deepfake so we know what the process is like. And we got instructed as to what to look for in an image to spot a deepfake. So we are looking at implementing new tools to deal with deepfakes.

A.2 EXAMPLE EMAIL STAKEHOLDERS

Dear <stakeholder>,

My name is Quinten Riphagen, I am currently doing my graduate thesis of my Master degree in the Business Information technology program from the University of Twente at the NFI. My research is on deepfakes and their effects on the forensic analysis of digital evidence. Specifically I am researching the current process of the submitting of video evidence to the NFI for forensic analysis and every step until the verdict is returned to the court. Which will be redesigned in order to accommodate for an expected increase in questions from the court regarding the authenticity of video evidence.

This is a qualitative research in which I will do interviews with stakeholders to get their perspective on the current process and what is needed when a new process is designed. I am reaching out to you since you might have information about the process, or incentives in a new approach.

If you are interested in helping me design the new process, let me know. We would have at least 2 interviews of about 30-45 minutes about your part in this subject and what can be done at the NFI to improve this process. One interview will focus on the current process and one interview will be feedback on the initial model I create. A possible third interview could be planned if feedback on the final model is needed. These interviews can be held online, but I would prefer to travel to your location to execute them if possible.

Please consider participating in this research as your inputs and opinions on this would be very valuable to me and might improve the situation at the NFI. If you reply to this email, we can set a date for an appointment.

Kind regards,
Quinten Riphagen

APPENDIX B

B.1 ASSESSMENT MATRIX

Level	Item	Source
Dimension: People		
2	**PEO2a:** Digital forensics researchers are formally examined?	Interviews
	PEO2b: The research is exclusively executed by digital forensics researchers?	Interviews Literature
3	**PEO3a:** Is collaboration between digital forensics investigators at the police and Forensic institute encouraged?	Interviews
	PEO3b: Formal examination specifically for image researchers?	Interviews
	PEO3c: Are the roles in the process formally defined? - Image researchers - Technology experts - Digital forensics experts	Interviews
4	**PEO4a:** Is image authentication is only done by image research experts?	Interviews
	PEO4b: Is collaboration with key partners (police, public prosecutor) standard practice?	Interviews
5	**PEO5a:** Are most of the cases handled by the key partners?	Interviews
	PEO5b: Is a constant improvement of image researchers through novel image research projects achieved?	Literature
Dimension: Process		
2	**PRO2a:** Are the standard image authentication processes defined? Define different request categories.	Literature
	PRO2b: Are cases being documented during the investigation?	Interviews & Literature
	PRO2c: Are historic case solutions documented and added to standard techniques?	Interviews

3	**PRO3a:** Are the image authentication processes are followed as they are defined and exceptions to the rule documented?	Literature
	PRO3b: Are feedback mechanisms with the client integrated into the process?	Literature Interviews
	PRO3c: Is the reporting on cases standardized?	Literature
	PRO3d: Implemented preparation process for estimating value of evidence?	Literature
4	**PRO4a:** Are process monitors implemented into the process? (examples:) - Average throughput time - Context analysis success rate - # of cases requiring additional explanations in court - % of cases which have verified manipulation	Literature
	PRO4b: Are there biannual meetings to improve the process and process definitions?	Literature
	PRO4c: Is context analysis the main part of the image authentication process?	Interviews
5	**PRO5a:** Is the continuous improvement of the IA process facilitated?	Literature
	PRO5b: Are the process outcomes predictable?	Literature
Dimension: Technology		
2	**TEC2a:** Is a documentation system in place?	Interviews Literature
3	**TEC3a:** Is there a DFaaS platform implanted within the organization and key partners? - Hansken	Literature DFaaS
	TEC3b: Deepfake detection is used to test the results of the case?	Interviews
4	**TEC4a:** Is deepfake detection implemented to initially get information about the evidence? Are the explainable results used for local analysis?	Interviews
	TEC4b: Are the possibilities of deepfake technology well known and continuously updated?	Interviews
	TEC4c: Are tools implemented for specialized auxiliary data analysis?	Literature
5	**TEC5a:** Is deepfake detection integrated into the DFaaS platform so key partners can use it as well?	Interviews
	TEC5b: Is the deepfake detection continuously developed in collaboration with the deepfake detection provider?	Interviews Literature
Dimension: Policy		

2	**POL2a:** Does current policy ensure the creation of processes?	Literature
	POL2b: Does current policy dictate the use of a documentation system?	Literature
	POL2c: Does current policy facilitate training of personnel?	Literature
3	**POL3a:** Does the policy dictate that IA is done by formally examined image researchers?	Literature
	POL3b: Does policy ensure that the process documentation is followed?	Literature
	POL3c: Does current policy dictate the use of the DFaaS platform?	Literature
4	**POL4a:** Does the current policy ensure the image researchers are experts in IA?	Literature
	POL4b: Does the current policy facilitate bi-annual improvement of processes?	Literature
	POL4c: Does the current policy allow for the use of deepfake detection?	Interviews
5	**POL5a:** Does current policy achieve continuous improvement of the process?	Literature
	POL5b: Does the current policy achieve continuous development of deepfake detection capabilities?	Interviews
	POL5c: Does current policy facilitate full collaboration between the key partners(customers, deepfake detection companies, police)	Interviews

BIBLIOGRAPHY

[1] Tobias Mettler. "Maturity assessment models: a design science research approach." In: *International Journal of Society Systems Science* 3.1-2 (2011), pp. 81–98.

[2] Yisroel Mirsky and Wenke Lee. "The creation and detection of deepfakes: A survey." In: *ACM Computing Surveys (CSUR)* 54.1 (2021), pp. 1–41.

[3] Thea van der Geest. *'Ik ben onschuldig, want ik ben gehackt!'* 2018. URL: https://magazines.openbaarministerie.nl/opportuun/2018/02/ik-ben-gehackt.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." In: *Advances in neural information processing systems* 27 (2014).

[5] Jason Brownlee. *A Gentle Introduction to Generative Adversarial Networks (GANs)*. 2019. URL: https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/.

[6] S. Tariq, S. Lee, and S. Woo. "One detector to rule them all: Towards a general deepfake attack detection framework." In: *The Web Conference 2021 - Proceedings of the World Wide Web Conference, WWW 2021*. 2021, pp. 3625–3637. DOI: 10.1145/3442381.3449809.

[7] H. Khalid and S.S. Woo. "OC-FakeDect: Classifying deepfakes using one-class variational autoencoder." In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. Vol. 2020-June. 2020, pp. 2794–2803. DOI: 10.1109/CVPRW50498.2020.00336.

[8] N. Carlini and H. Farid. "Evading deepfake-image detectors with white-and black-box attacks." In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. Vol. 2020-June. 2020, pp. 2804–2813. DOI: 10.1109/CVPRW50498.2020.00337.

[9] *What are deepfakes? Are They a Security Threat?* 2021. URL: https://www.tessian.com/blog/what-are-deepfakes/.

[10] Brenda Marie Rivers. *FBI issues guidance on identifying 'deepfake' content*. 2021. URL: https://executivegov.com/2021/03/fbi-issues-guidance-on-identifying-deepfake-content/.

[11] Jon Bateman. *Deepfakes and synthetic media in the financial system: Assessing threat scenarios*. Carnegie Endowment for International Peace., 2020.

[12] Catherine Stupp. *Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case*. 2019. URL: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

[13] Donie O'Sullivan. *House intel chair sounds alarm at Congress' first hearing on Deepfake videos*. 2019. URL: https://edition.cnn.com/2019/06/13/tech/deepfake-congress-hearing/index.html.

[14] Saifuddin Ahmed. "Fooled by the fakes: Cognitive differences in perceived claim accuracy and sharing intention of non-political deepfakes." In: *Personality and Individual Differences* 182 (2021), p. 111074. ISSN: 0191-8869. DOI: https://doi.org/10.1016/j.paid.2021.111074.

[15] *National Center for the middle market survey*. 2016. URL: https://www.middlemarketcenter.org/.

[16] David Bisson. *Small companies overconfident about their security posture, finds survey*. 2017. URL: https://www.tripwire.com/state-of-security/risk-based-security-for-executives/connecting-security-to-the-business/small-companies-overconfident-security-posture-finds-survey/.

[17] Elaine Lee. *Preparing for the next cybersecurity epidemic: Deepfakes*. 2022. URL: https://www.darkreading.com/operations/preparing-for-the-next-cybersecurity-epidemic-deepfakes.

[18] Stephen Viña. *Digital deception: Is your business ready for "deepfakes"?* 2020. URL: https://www.marshmclennan.com/insights/publications/2020/october/digital-deception--is-your-business-ready-for-deep-fakes-.html.

[19] Bobby Chesney and Danielle Citron. *Deep Fakes: A Looming Challenge for Privacy*. 2020. URL: https://doi.org/10.15779/Z38RV0D15J.

[20] *Our Product*. URL: https://www.duckduckgoose.ai/detector.

[21] Mary Ann Azevedo. *Truepic, which just raised $26M in a Microsoft-led round, aims to verify the authenticity of photos and videos*. 2021. URL: https://techcrunch.com/2021/09/14/microsofts-m12-leads-26m-investment-into-truepic/?guccounter=1.

[22] Leo Kelion. *Deepfake detection tool unveiled by Microsoft*. 2020. URL: https://www.bbc.com/news/technology-53984114.

[23] *About Us*. 2022. URL: https://deepware.ai/about/.

[24] URL: https://defudger.com/.

[25] Cooper Hood. *How Deepfake Technology Can Change The Movie Industry*. 2021. URL: https://screenrant.com/movies-deepfake-technology-change-hollywood-how/.

[26] Jacek Naruniec, Leonhard Helminger, Christopher Schroers, and Romann M Weber. "High-resolution neural face swapping for visual effects." In: *Computer Graphics Forum*. Vol. 39. 4. Wiley Online Library. 2020, pp. 173–184.

[27] James Vincent. *Deepfake dubs could help translate film and TV without losing an actor's original performance*. 2021. URL: https://www.theverge.com/2021/5/18/22430340/deepfake-dubs-dubbing-film-tv-flawless-startup.

[28] Paengsuda Panyatham. *Deepfake Technology in the Entertainment industry: Potential Limitations and Protections*. 2021. URL: https://amt-lab.org/blog/2020/3/deepfake-technology-in-the-entertainment-industry-potential-limitations-and-protections.

[29] Panda Security. *Deepfake Fraud: Security Threats Behind Artificial Faces*. 2021. URL: https://www.pandasecurity.com/en/mediacenter/technology/deepfake-fraud/.

[30] Ellen Daniel. *Phishing scams account for half of all fraud attacks, according to RSA*. 2019. URL: https://www.verdict.co.uk/phishing-rsa-report/.

[31] James Vincent. *A spy reportedly used an AI-generated profile picture to connect with sources on LinkedIn*. 2019. URL: https://www.theverge.com/2019/6/13/18677341/ai-generated-fake-faces-spy-linked-in-contacts-associated-press.

[32] Ashish Cauhan. *Ahmedabad: Deepfakes replace women on sextortion calls: Ahmedabad News - Times of India*. 2021. URL: https://timesofindia.indiatimes.com/city/ahmedabad/deepfakes-replace-women-on-sextortion-calls/articleshow/86020397.cms.

[33] Mohamed Thaver. *Instagram message, nude video calls, blackmail: Modus operandi of a new phishing crime wave*. 2021. URL: https://indianexpress.com/article/cities/mumbai/message-video-call-blackmail-modus-operandi-of-a-new-phishing-crime-wave-7493107/.

[34] *The Disinformation Dozen: Center for Countering Digital Hate*. 2021. URL: https://www.counterhate.com/disinformationdozen.

[35] T Hwang. "Deepfakes: A Grounded Threat Assessment." In: *Centre for Security and Emerging Technologies, Georgetown University* (2020).

[36] Roel J Wieringa. *Design science methodology for information systems and software engineering*. Springer, 2014.

[37] Ebrahim Hamad Al Hanaei and Awais Rashid. "DF-C2M2: a capability maturity model for digital forensics organisations." In: *2014 IEEE Security and Privacy Workshops*. IEEE. 2014, pp. 57–60.

[38] Ahmed Almarzooqi and Andrew Jones. “A framework for assessing the core capabilities of a digital forensic organization.” In: *IFIP international conference on digital forensics*. Springer. 2016, pp. 47–65.

[39] B. Kitchenham and S Charters. “Guidelines for performing Systematic Literature Reviews in Software Engineering.” In: (2007).

[40] Jennifer Rowley. “Conducting research interviews.” In: *Management research review* (2012).

[41] M Muswazi and Edmore Nhamo. “Note taking: A lesson for novice qualitative researchers.” In: *Journal of Research & Method in Education* 2.3 (2013), pp. 13–17.

[42] Hennie Boeije. *Analysis in qualitative research*. Sage publications, 2009.

[43] Mark C Paulk, Bill Curtis, Mary Beth Chrissis, and Charles V Weber. “Capability maturity model, version 1.1.” In: *IEEE software* 10.4 (1993), pp. 18–27.

[44] Diogo Proença and José Borbinha. “Maturity models for information systems-a state of the art.” In: *Procedia Computer Science* 100 (2016), pp. 1042–1049.

[45] Martin Kerrigan. “A capability maturity model for digital investigations.” In: *Digital Investigation* 10.1 (2013), pp. 19–33.

[46] Ronald Krutz. *Methodology for assessing the maturity and capability of an organization's computer forensics processes*. US Patent App. 10/952,537. 2006.

[47] RB Van Baar, Harm MA van Beek, and EJ Van Eijk. “Digital Forensics as a Service: A game changer.” In: *Digital Investigation* 11 (2014), S54–S62.

[48] H.M.A. Beek, J. Bos, A. Boztas, Erwin van Eijk, R. Schramp, and M. Ugen. “Digital forensics as a service: Stepping up the game.” In: *Forensic Science International: Digital Investigation* 35 (Dec. 2020), p. 301021. DOI: 10.1016/j.fsidi.2020.301021.

[49] Graeme Horsman. “Defining ‘service levels’ for digital forensic science organisations.” In: *Forensic Science International: Digital Investigation* 38 (2021), p. 301178.

[50] *Best practice manual digital image authentication*. 2021. URL: https://enfsi.eu/about-enfsi/structure/working-groups/documents-page/documents/best-practice-manuals/.

[51] Y. Mirsky and W. Lee. “The Creation and Detection of Deepfakes.” In: *ACM Computing Surveys* 54.1 (2021).

[52] Justin D Cochran and Stuart A Napshin. “Deepfakes: awareness, concerns, and platform accountability.” In: *Cyberpsychology, Behavior, and Social Networking* 24.3 (2021), pp. 164–172.

DECLARATION

I, Quinten Riphagen, hereby declare that this research is made by my own work. Any information used outside of this work has been properly referenced. This research was carried out at the Netherlands Forensic Institute in commissioned by my supervisor Prof. dr. ing. Zeno Geradts. In order to graduate the degree of Business Information Technology at the University of Twente. During this research I was supervised by Abhishta Abhishta and Jan-Willem Bullee from the University of Twente.
Zwolle, August 2022

Quinten Riphagen

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

`https://bitbucket.org/amiede/classicthesis/`

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

`http://postcards.miede.de/`

Final Version as of September 21, 2022 (`classicthesis` version 1.3).