ORIGINALARBEIT Elektrotechnik & Informationstechnik (2021) 138/7: 480–484. https://doi.org/10.1007/s00502-021-00929-7

The defalsif-AI project: protecting critical infrastructures against disinformation and fake news

D. Schreiber, C. Picus, D. Fischinger, M. Boyer

In this paper, we describe the concept and ongoing work of the project defalsif-AI, which addresses the protection of critical infrastructures against disinformation and fake news. Defalsif-AI deals particularly with the protection of the main democratic processes and the public trust in democracy and its institutions against engineered social media attacks, which, for example, attempt to manipulate the electoral process. Federal ministries and media institutions require new methods and tools to evaluate the ever increasing amount of digital media in terms of identification, verification, and correction of sources. Based on these requirements, the project focuses on research on audio-visual media forensics, text analysis, and multimodal fusion with the support of artificial intelligence (AI) and machine learning methods. One main focus of this research is to make the results more comprehensible and interpretable for non-experts in the forensic/technical field. The primary project outcome is a proof of concept of a multimodal detection platform, which can operate with a variety of sources, including the surface web and social media. Additional research carried out within the project focuses on providing and generating multimodal data necessary to train and test machine learning models. Finally, an analysis and assessment concerning the law and social science are carried out as well.

Keywords: critical infrastructures; fake news; disinformation; multimodal detection of fake news; multimodal fusion

Das Projekt defalsif-AI: Schutz kritischer Infrastrukturen vor Desinformation und Fake News.

In diesem Beitrag beschreiben wir das Konzept und die laufenden Arbeiten des Projekts defalsif-AI, das sich mit dem Schutz kritischer Infrastrukturen vor Desinformation und Fake News beschäftigt. Insbesondere geht es bei defalsif-AI um den Schutz demokratischer Kern-Prozesse und des öffentlichen Vertrauens in die Demokratie und ihre Institutionen vor konstruierten Social-Media-Angriffen, welche beispielsweise versuchen, den Wahlprozess zu manipulieren. Die öffentliche Verwaltung, aber auch Medienhäuser, benötigen neue Methoden und Werkzeuge, um die immer größer werdenden Mengen digitaler Medien hinsichtlich der Identifizierung, Verifizierung und Korrektur von Quellen auszuwerten. Ausgehend von diesen Anforderungen konzentriert sich das Projekt auf die Forschung in den Bereichen audiovisuelle Medienforensik, Textanalyse und multimodale Fusion unter Zuhilfenahme von Methoden der künstlichen Intelligenz (KI) und des maschinellen Lernens. Ein Hauptaugenmerk dieser Forschung liegt auf der Verbesserung der Verständlichkeit und Interpretierbarkeit der Ergebnisse für Laien im forensischen/technischen Bereich. Das primäre Projektergebnis ist ein Proof-of-Concept einer multimodalen Detektionsplattform, welche mit einer Vielzahl von Quellen arbeiten kann, wie etwa dem Surface Web und sozialen Medien. Weitere Forschung innerhalb des Projekts konzentriert sich auf die Bereitstellung und Generierung multimodaler Daten, die zum Trainieren und Testen von Modellen des maschinellen Lernens erforderlich sind. Schließlich werden auch eine rechts- und sozialwissenschaftliche Analyse und Bewertung durchgeführt.

Schlüsselwörter: kritische Infrastrukturen; Fake News; Desinformation; multimodale Detektion von Fake News; multimodale Fusion

Received June 8, 2021, accepted September 6, 2021, published online September 17, 2021

1. Introduction and motivation - the defalsif-AI project

Social media has made it possible to manipulate the masses via disinformation and fake news at an unprecedented scale [1]. One particularly sensitive target that is vulnerable to behavioral manipulation is critical infrastructure, as a direct attack on such infrastructures may have drastic implications nationwide. For example, the study of [1] considers an attack on the power grid in which an adversary attempts to manipulate the behavior of energy consumers by sending fake discount notifications encouraging them to shift their consumption into the peak-demand period. Furthermore, it then evaluates the impact of such an attack considering the distribution network of the Greater London area as a case study. Another example is the spread of a 5G coronavirus conspiracy across Europe. For example, as reported last year [2], in the UK, where the attacks began, almost 60 masts have been set ablaze, while two towers were vandalized in Ireland, and another in Cyprus. While in the Netherlands, there have been 11 recorded attempts.

In this paper we describe concept and work in progress within the project defalsif-AI (Detection of Disinformation via Artificial Intelligence) [3], which addresses the safety and security of critical infrastructures against disinformation and fake news. In particular, defalsif-AI focuses on politically motivated disinformation and fake news, which weaken or threaten political as well as state institutions, e.g., protection against engineered social media attacks that

Schreiber, David, Austrian Institute of Technology (AIT) GmbH, Giefinggasse 4, 1210 Vienna, Austria (E-mail: david.schreiber@ait.ac.at); **Picus, Cristina,** Austrian Institute of Technology (AIT) GmbH, Vienna, Austria; **Fischinger, David,** Austrian Institute of Technology (AIT) GmbH, Vienna, Austria; **Boyer, Martin,** Austrian Institute of Technology (AIT) GmbH, Vienna, Austria

attempt to manipulate the electoral processes. Analysts in Federal ministries require new approaches for identifying and dealing with large-scale disinformation campaigns at the operative, strategic and political level, while media institutions are concerned about enabling democracy through its information and opinion-forming function. These organizations, therefore, require improved methods and tools for evaluating ever-increasing volumes of digital media in terms of identification, verification and correction of sources.

Based on these requirements, defalsif-AI focuses on research in the areas of audio-visual media forensics, text analysis and the multimodal fusion of all these with the support of artificial intelligence and machine learning methods. The primary project outcome is a proof-of-concept implementation that can operate on a variety of sources, including the Surface Web (e.g., news sites) and social media (e.g., Twitter). The planned proof-of-concept media forensics platform should allow for the upgrade of existing detection tools or integration of future tools into a scalable system so that end users can take advantage of new and updated tamper detection techniques in the future. An additional focus of this research is in enhancing the comprehensibility and interpretability of the results for non-experts in the forensic/technical field. Screening and monitoring tools are included to identify topics, trends, accumulations or anomalies in the dissemination of information. The detection platform should be able to ingest either individual media objects (e.g., video clip, image, audio clip, text) or eventually entire web pages – analyzing images, audio, video, and text material – providing the basis for recommended actions.

Additional research within the project focuses on providing and generating multi-modal data necessary to train and test machine learning models (in particular, deep neural networks). Consequently, an additional outcome of defalsif-AI will be a publicly available dataset to enable further research in the field. Furthermore, the aim is to provide a media forensic platform that is equally interpretable and explainable – that is, a tool that can be understood and used by non-technicians and non-experts - to a broad user base. A legal and social sciences analysis and assessment is to be carried out, recommending guidelines as well as the application-oriented derivation of technical and organizational measures for future compliant implementation and execution of disinformation analysis platforms. Finally, the project will deliver an exploitation plan detailing how the resulting proof-of-concept could be further developed and deployed as a productive system in both governmental and private agencies.

The rest of this paper is organized as follow. In Sect. 2, the state-of-the-art is outlined for visual-, audio-, text-based and multi-modal media forensics, followed by state-of-the-art in interpretation and presentation of machine learning results, as well as legal and social-sciences aspects. Section 3 describes the proof-of-concept media forensics platform, which is at the heart of the project. Section 4 contains our conclusions.

2. Previous work

With the rapid progress in recent years of machine learning methods based on deep neural networks, techniques that generate and manipulate visual content can now provide a very advanced level of realism (*deepfakes*). Deepfakes are synthetic media in which a person in an existing image or video is replaced with someone else's likeness. The main deep learning methods used to create deepfakes involve training generative neural network architectures, such as generative adversarial networks (GANs) [4]. For a recent review paper presenting an analysis of the methods for the detection of manipulated images and videos see [5], where special emphasize is placed on deepfakes and on modern data-driven forensic methods to fight them. In fact, [6] showed that both conventional and deep-learning detectors achieve up to 95% detection accuracy on *currently known* deepfakes. Unfortunately, as the counterfeiting industry will always be one step ahead of fake detection methods, algorithms must be constantly developed and updated for new types of deepfakes.

Similarly, a "cloned" voice that is potentially indistinguishable from the real person's is used to produce a deepfake audio. A broad and up-to-date overview of approaches in audio forensics is given in [7]. Currently, audio deepfakes are not as widespread and convincing as video deepfakes, due to the fact that details such as intonation, inflection, and pacing of the human voice are particularly difficult to model. However, human brains have a hard time detecting the differences between real and artificial voices; it's easier for humans to pick up on a fake image than to recognize an artificial voice. Therefore, there is a rising concern that in the future, audio deepfakes will be an even greater problem than visual deepfakes. A possible advancement in this direction is Adobe's Project Voco [8] where the aim is to develop what is essentially a Photoshop of soundwaves.

Visual and audio fake detection are currently evaluated on popular datasets such as the DFDC (DeepFake Detection Challenge) [9]. However, German-language datasets are hardly available - the majority of datasets are in English. Similarly, the vast majority of datasets are unimodal (usually images or videos) - multimodal manipulation is currently hardly considered, although combining and fusing audio and visual clues will likely improve fake detection. For example, the work in [10] evaluated fake video detection based on inconsistencies between lip movements and audio speech. Similarly, [11] proposed a method for representing the temporal relationship between the auditory and visual streams. Moreover, even less effort has been invested to combine audio, images, videos, and text for deepfake detection.

Text-based fake news refers to information content that is false, misleading or whose source cannot be verified. This content may be generated to intentionally damage reputations, deceive, influence or incite. The main two approaches for text-based fake news detection are traditional machine learning (classification) methods, where Natural Language Processing (NLP) may play a role in extracting features from data, and deep learning approaches. A comprehensive review of detecting fake news on social media can be found in [12].

Although tools for generating data manipulations are easily accessible to a broad basis of users, there are hardly any tools to detect such manipulations for non-experts. Moreover, there are hardly any commercial platforms that offer a broad collection of tools for end users, nor providing explainable as well as interpretable results for non-expert users. Despite their widespread adoption and success, machine learning models remain mostly black boxes. Understanding the reasons behind a specific prediction is, however, quite important in assessing trust, which is fundamental if one plans to take action based on a prediction. Only a limited effort has been made to represent and interpret the results of machine learning models in general, and of neural networks in particular, and hardly any has been made for the non-expert users. In [13], the "Lime" explanation system was introduced, that explains the predictions of any classifier, by learning an interpretable model locally around the prediction. The "TELEGAM" [14] is a prototype system, that demonstrates how visualizations and verbalizations can collectively support interactive exploration of machine learning models, while "NeuralVis" [15] is designed for visualizing the internal structure of deep neural networks models.

Legal aspects of fake news have been addressed and researched by numerous academic disciplines and political institutions in recent

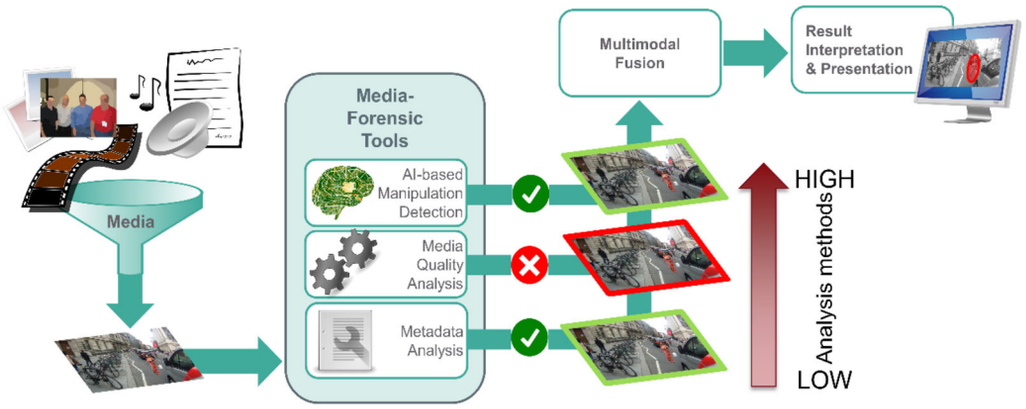


Fig. 1. The generic approach followed in defalsif-AI for detecting disinformation and fake news using multi-modal media forensic tools and three-level hierarchy of algorithms

years, such as [16]. From a juridical perspective the debate around fake news is primarily about the distinction between the fundamental human rights of freedom of expression and free circulation of information from various other criminal-, civil and media law provisions for the protection of personal rights.

In addition, further research is needed regarding the concept of "truth." From social sciences aspect, the major challenge is that false news, even after being corrected, has become the majority opinion and thus cannot be reversed. The state of current research shows the need to further take a closer look at a society's handling of misinformation as well as its effects. Furthermore, in the development of artificial intelligences, the supposed objectivity of detection should be questioned, as the developers of technical systems inscribe their subjective worldviews into the technology, which can thus contribute to the reinforcement of inequality [17].

3. The proof-of-concept media forensics platform

The project defalsif-AI addresses the safety and security of critical infrastructures against disinformation and fake news. Consequently, the main outcome of the project is to demonstrate a proof-of-concept state-of-the-art media forensics platform, integrating detection tools for fake text, image, video and audio modalities. Figure 1 illustrates the generic approach followed in the project for detecting disinformation and fake news: the defalsif-AI platform receives either individual media objects (e.g., video clip, image, audio clip, text) or an entire web page; the system then analyses each media object via the complete tool chain corresponding to the specific modality of the media object. Each tool chain is structured according to a three-level hierarchy of algorithmic sophistication; each level provides an evaluation regarding the authenticity of the media object. Finally, the three per level predictions are fused via machine learning tools to provide a final verdict, where the individual and final predictions, as well as their confidence estimates, are presented to the non-expert user in an interpretable way. This approach aims to make the processes behind the system more transparent, and the results of the analysis better explainable to and thus trusted by the user.

The media forensics platform is hierarchically structured, with algorithmic sophistication ranging from low to high, where the low and intermediate levels comprise traditional computer vision, audio and text analysis methods, while the high level is based on recent state-of-the-art deep learning techniques, employing deep neural networks. An example of a low-level method is the metadata analysis of an image, namely determining whether the image was created at the same date as is claimed in a social media post, and if time of day and location match with the content of the image. The medium-level methods are concerned with the quality analysis of the content, such as whether the noise characteristics of a sound recording matches with the known noise characteristics of the specific recording device that was being used. Finally, the highest-level methods deal in particular with content-specific characteristics of the various digital artifacts, e.g., hateful text content, manipulated audio content and most notably deepfake photos and videos.

Deepfakes are altered photos and videos that use deep learning models and have become increasingly realistic in recent years, making it harder to detect the real from the fake with just the naked eye. A key milestone of deflasif-AI is delivering the first version of the proof-of-concept forensics platform, at the end of the project's first year, with an emphasis on visual deepfakes detection. The first version already integrates the following state-of-the-art deepfake detection techniques: (i) A fake face detector for images based on the StyleGAN network [18], which offers a novel alternative architecture for generative adversarial networks (GANS); (ii) A deepfake (face swaps) detector for videos, using the solution of the NTech Kaggle Deepfake Detection Challenge [19], which is based on the EfficientNet-B7 network model; and (iii) A general forgery analysis tool for images, based on the ManTraNet [20], a fully convolutional network. Additionally, the first platform version includes a reverse image search engine as well (currently integrated through an API to the Google Reverse Image Search Engine). As part of the on-going research within defalsif-AI, it is planned to modify the design of the ManTraNet network, as well as to re-train the modified network on defalsif-AI's own acquired dataset. Design modifications and or re-training for the other networks within the forensic platform are under consideration as well. As deep learning models tend to increase their accuracy with the increasing amount of training data, obtaining the best possible abstract representation of the input data, it is expected that with modified and re-trained models, the performance of the proof-of-concept platform will be enhanced.

Figure 2 shows an example use case, namely applying the image forgery tools and the corresponding interface within the platform. In the manipulated picture on the left, a splicing operation was executed to add an additional person from another image. The picture on the right shows the prediction mask output, identifying the potentially manipulated image region. In addition, the platform provides metadata information like the image size and the modification

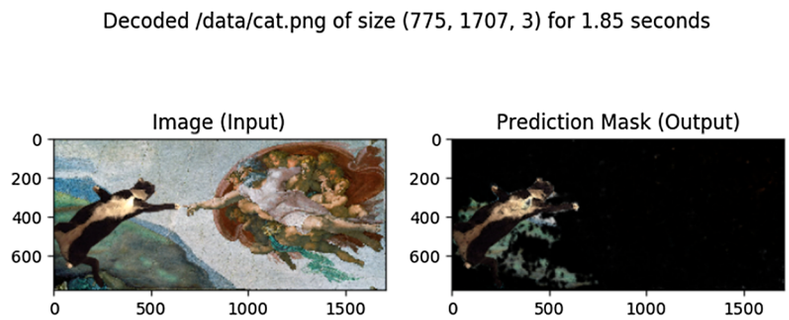


Bild-Metadatei

Eigenschaft	Wert
File:FileName	cat.png
PNG:BlueX	0.15
PNG:ProfileName	Photoshop ICC profile
File:FileType	PNG

Fig. 2. An example of an image forgery analysis by defalsif-AI's proof-of-concept platform, and the corresponding interface: detection of a splicing manipulation and analysis of metadata

date. For the depicted example, it also provides the information that the PNG file was created with the program Photoshop, which is a valuable indication of potential manipulation.

4. Conclusions

The project defalsif-AI addresses the safety and security of critical infrastructures against disinformation and fake news, spread by social media and manipulating the behavior of the masses, with a focus on the example of political processes and institutions. At the heart of the defalsif-AI project is the AI-based proof-of-concept platform for analyzing media information with regard to possible targeted manipulation of audio-visual as well as text-based content. Consequently, the benefits and outcomes of the defalsif-AI project are not restricted to the political or journalistic cases which are the focus of the project, but rather to threats of any kind to critical infrastructures. The approach followed in the defalsif-AI project, regarding tools development and practical applications will result in a considerable increase in knowledge about the detection and handling of disinformation and fake news, both for the consortium and for subsequent users in the media industry.

The innovative aspects of defalsif-AI are manyfold: First, providing collections of media forensic tools across different modalities integrated into a single platform. Furthermore, allowing to update and integrate future tools into the system, enabling end-users to leverage new and updated fake detection techniques in the future. Second, providing a multimodal dataset comprising images, videos, audio and text, and including German language (and possibly other end-user relevant languages), and making the dataset publicly available for other research purposes on a non-profit basis. Third, the platform will provide interpretation and explanation of the detection results that can be understood, trusted and used by non-technical or non-expert users. Fourth, providing epistemological sharpening and operational formalization of the fake news phenomenon in terms of legal informatics and developing an interdisciplinary valid typology and legally sound operationalization of the fake news concept for the technological development of the media forensic tool. Fifth, enabling for the first time to incorporate a single integrated tool into the current journalistic workflow. Finally, from a sociological point of view, the project will include a comprehensive analysis of the requirements and needs in the design of the media forensic tool, as well as an assessment and discussion of potential risks and effects on the public-media debate and the socio-political fabric of society.

Of course, such envisioned proof-of-concept forensic platform would never be able to provide 100% reliable decisions. Rather, it is intended as a "pre-screening" method that can, at any rate, automatically detect the tampering generated by current counterfeiting tools. Unfortunately, the counterfeiting industry will always be one step ahead of the detection methods for fakes. On the one hand, this means that an assessment by experts will still be necessary if the security requirements are very high, and on the other hand, the tool must be constantly developed and updated - similar to anti-virus protections for PCs.

Funding information The work described in this paper was funded in the context of the defalsif-AI project (FFG project number 879670, funded by the Austrian security research program KIRAS of the Federal Ministry of Agriculture, Regions and Tourism BMLRT).

Data availability Not applicable.

Code availability Not applicable.

Conflicts of interest/Competing interests Not applicable.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

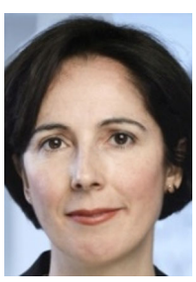
References

1. Gururaghav, R., AlShebli, B., Waniek, B., Rahwan, T., Peng, J. C.-H. (2020): How weaponizing disinformation can bring down a city's power grid. PLoS ONE, 15, e0236517.

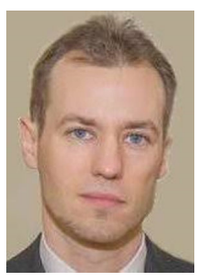
2. Financial Times [Online]. Available: https://www.ft.com/content/1eeedb71-d9dc-4b13-9b45-fcb7898ae9e1.
3. defalsif-AI project (10.2020 - 9.2022) [Online]. Available: https://defalsifai.at/.
4. Charleer, S. (November 2019, 17 May 2019): Family fun with deepfakes. Or how I got my wife into the Tonight Show Medium. Retrieved 8 November 2019, 17 May 2019 [Online]. Available: http://svencharleer.com/2018/02/02/family-fun-with-deepfakes-or-how-i-got-my-wife-onto-the-tonight-show/.
5. Verdoliva, L. (2020): Media forensics and DeepFakes: an overview. IEEE J. Sel. Top. Signal Process., 14(5), 910–932.
6. Marra, F., Gragnaniello, D., Cozzolino, D., Verdoliva, L. (2018): Detection of GAN-generated fake images over social networks. In IEEE conference on multimedia information processing and retrieval (MIPR) (pp. 384–389).
7. Zakariah, K. M., Khan, M. K., Malik, H. (2018): Digital multimedia audio forensics: past, present and future. Multimed. Tools Appl., 77(1), 1009–1040.
8. [Online]. Available: https://www.youtube.com/watch?v=I3l4XLZ59iw.
9. Deepfake Detection Challenge (DFDC) [Online]. Available: https://www.kaggle.com/c/deepfake-detection-challenge.
10. Korshunov, P., Marcel, S. (2018): DeepFakes: a new threat to face recognition? Assessment and detection arXiv:1812.08685 [cs.CV].
11. Le, N., Odobez, J. M. (2016): Learning multimodal temporal representation for dubbing detection in broadcast media. In Proceedings of the 24th ACM international conference on multimedia.
12. Shu, K., Wang, S., Tang, J., Liu, H. (2017): Fake news detection on social media: a data mining perspective. In ACM SIKGDD explorations newsletter (pp. 22–36).
13. Singh, S., Ribeiro, M. T., Guestrin, C. (2016): "Why should I trust you?": explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.
14. Hohman, F., Srinivasan, A., Drucker, S. (2019): TeleGam: combining visualization and verbalization for interpretable machine learning. In IEEE visualization conference.
15. Zhang, X., Yin, Z., Feng, Y., Shi, Q., Liu, J., Chen, Z. (2019): NeuralVis: visualizing and interpreting deep learning models. In Proceedings of the 34th IEEE/ACM international conference on automated software engineering.
16. EU Code of Practice on Disinformation (2018): [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation.
17. Benjamin, R. (2019): Engineering inequity. Are robots racist?. In Race after technology. Abolitionist tools for the new Jim code (S. 49–76). Cambridge: Polity.
18. Karras, T., Laine, S., Aila, T. (2019): A style-based generator architecture for generative adversarial networks. In IEEE/CVF conference on computer vision and pattern recognition (CVPR).
19. NTech-Lab (2021): Kaggle DeepFake Detection Challenge (DFDC). [Online], Available: https://github.com/NTech-Lab/deepfake-detection-challenge.
20. Wu, Y., AbdAlmageed, W., Natarajan, P. (2019): Mantra-net: manipulation tracing network for detection and localization of image forgeries with anomalous features. In IEEE/CVF conference on computer vision and pattern recognition (CVPR).

Authors

David Schreiber
is a scientist working at the Austrian Institute of Technology since 2001, in the field of computer vision. He has received B.Sc. in Physics, and an M.Sc. and a Ph.D. in theoretical High-Energy Physics. Before joining AIT, he worked within several industrial companies, in the areas of computational geometry, image processing and computer vision.



Cristina Picus
is a scientist of the Austrian Institute of Technology, working in the research area of sensing and vision solutions. She has been involved in several research projects developing real-time image processing and machine learning algorithms for applications of visual surveillance, e.g. human tracking, multi-camera surveillance. She studied physics at the University of Cagliari (Italy) and received 2004 her Ph.D. degree in theoretical Physics at the University of Heidelberg (Germany). She worked subsequently as researcher in the field of Computer Vision, 2004 in the company Advance Computer Vision and since 2007 at AIT.



David Fischinger
is a Research Engineer at the Austrian Institute of Technology. In 2005 he graduated in Technical Mathematics at the Vienna University of Technology. In 2007 he graduated in Computational Intelligence (CS) and CS Management. 2014 he finished his PhD in robotics. Since 2020 he has been a member of the Sensing & Vision Solutions Team of the AIT. Special areas of research include Computer Vision, Machine Learning and Service Robotics.

Martin Boyer
is a Senior Research Engineer at AIT Austrian Institute of Technology GmbH – working as a project manager and research engineer in the field of Sensing and Vision Solutions since 2011. His research interests lie in the areas of video analysis frameworks, SW architecture, code generation and VideoAnalytics-as-a-scalable-Service. Currently Martin is coordinating the Austrian national security research project defalsif-AI "Detection of Disinformation via Artificial Intelligence".