Team Report 1.1 Full Research and Innovation Project Proposal

DeepFakeChain

Due date	15/8/2022
Submission date	14/8/2022
Version	1.0
Authors	Nikos Giatsoglou (CERTH) Symeon Papadopoulos (CERTH) Dora Kallipolitou (Zelus) Stella Markopoulou (Zelus)

Grant Agreement No.: 957228
Call: H2020-ICT-2018-2020
Topic: ICT-54-2020
Type of action: RIA

Document Revision History

Version	Date	Description of change	List of contributor(s)
v0.1	15/7/2021	ToC version circulated by the TruBlo team.	TruBlo team
v0.2	22/7/2022	Defined subsection structure and content of TR1.1	CERTH & Zelus
v0.3	26/7/2022	Completed Scientific Background subsection on AI-based Deepfake Detection	CERTH
v0.4	1/8/2022	Completed Technical Implementation section	Zelus
v0.5	2/8/2022	Completed Related Work & the remaining Scientific Background section	CERTH
v0.6	5/8/2022	Completed Results & Impact section	Zelus
v0.7	8/8/2022	Completed Introduction, Executive Summary & Conclusions	CERTH
v0.8	11/8/2022	Completed revisions	CERTH & Zelus
v1.0	14/8/2022	Version ready for final submission.	CERTH & Zelus

Disclaimer

Project co-funded by the European Commission in the H2020 Programme	
Nature of the deliverable:	R: Document, report
Dissemination Level:	CO: Confidential to TruBlo project and Commission Services

Funded by the Horizon 2020 Framework Programme of the European Union

EXECUTIVE SUMMARY

This document is the first deliverable of the research project DeepFakeChain, selected by the cascade-funding project NGI TruBlo (GA 957228), which aims to enhance trust in online content through blockchain technology. The deliverable describes: (i) the motivation behind DeepFakeChain and the position of the project in NGI TruBlo, (ii) the societal and business needs that the project addresses, based on the identified gaps in the state-of-the-art and the market, (iii), an initial description of the technical implementation of DeepFakeChain along with the related scientific background, and (iv) the impact and scientific breakthrough expected by the project along with related KPIs.

In brief, DeepFakeChain is a collaborative cross-organisational platform for deepfake detection, targeted at journalists and fact-checkers. The platform offers a hybrid human-machine solution to deepfake detection, integrating advanced artificial intelligence (AI) algorithms and human feedback for the task, while providing ways to reach consensus in a transparent and trustworthy way. A key feature of the platform is that all annotations and decisions are protected by design from tampering and are stored in an external blockchain network, with an initial decision to rely on Alastria. The usage of blockchain technology is crucial to boost confidence in the platform's output, which will be accessible to third parties for auditing and research. The platform is designed to be extensible to future algorithms and needs; in particular, the project's partners consider the collaborative moderation of harmful content (e.g. NSFW, disturbing) as an attractive and economically sustainable sector for the developed platform.

Funded by the Horizon 2020 Framework Programme of the European Union

TABLE OF CONTENTS

Funded by the Horizon 2020 Framework Programme of the European Union

LIST OF FIGURES

Funded by the Horizon 2020 Framework Programme of the European Union

LIST OF TABLES

Funded by the Horizon 2020
Framework Programme of the European Union

ABBREVIATIONS

Abbreviations	Definitions
AI	Artificial Intelligence
API	Application Programming Interface
C2PA	Coalition for Content Provenance and Authenticity
CGI	Computer Generated Imagery
DLT	Distributed Ledger Technology
EDMO	European Digital Media Observatory
EEA	European Economic Area
EU	European Union
GAN	Generative Adversarial Network
IoT	Internet of Things
KPI	Key Performance Indicator
MAS	Multi-Sgent System
MVP	Minimum Viable Product
ML	Machine Learning
PoC	Proof of Concept
R&I	Research and Innovation
REST	REpresentational State Transfer
SUS	System Usability Score
US	United States

Funded by the Horizon 2020 Framework Programme of the European Union

1 PROJECT DESCRIPTION

Project name	DeepFakeChain
Link to project on TruBlo website	https://www.trublo.eu/deepfakechain/
Primary contact	Dr Symeon Papadopoulos, papadop@iti.gr
Project members	Mr Nikolaos Giatsoglou, ngiatsog@iti.gr Dr George Kordopatis, georgekordopatis@iti.gr Ms Stella Markopoulou, s.markopoulou@zelus.gr
Organisation(s)	CERTH, Zelus
Organisation's website	https://www.certh.gr, https://www.zelus.gr
Short project summary	
What is the focus of your project?	The development of a collaborative cross-organisational platform for deepfake detection with enhanced trustworthiness and transparency guarantees based on blockchain. The platform integrates algorithmic and human feedback on deepfakes, and offers advanced reputation and truth discovery methods to reach consensus. It is designed to be extensible with future algorithms, possibly targeting more general harmful content like hate speech, misinformation, and disturbing content, paving the way to be used as a collaborative content moderation tool.
Why is a new/better solution needed?	The generation of deepfake synthetic media that are exceedingly difficult to detect is a worrying trend with potentially devastating impact on society. Currently, no solution exists for perfect automatic detection, nor one is expected in the near future due to the ongoing refinement of generated deepfakes. DeepFakeChain addresses this issue by combining algorithmic solutions with human expert opinion, which can discern context and, in our opinion, is direly needed for the task.
How will your solution be better?	By bringing together expert journalists from multiple organisations and reaching trustworthy decisions through advanced reputation and truth discovery mechanisms.

Funded by the Horizon 2020 Framework Programme of the European Union

	By closing the gap between the users and the technical nature of state-of-the-art AI algorithms through a user-friendly design. By creating trustworthy and auditable annotations for deepfake detection that can be used by researchers.
Extra: How does this project contribute to "trustable content on future blockchains"	By using blockchain to guarantee the trustworthiness of media annotations and metadata, and make them available to third parties for auditing and research.
Type of project	(X) Scientific/research () Commercial, potential startup () Open source, non-commercial () Other, pls add 1-4 words if selected
Technologies used	Permissioned blockchain networks, consensus and truth-discovery algorithms, AI-based deepfake detection algorithms, reverse image and video search
Use of Alastria resources	Yes

Funded by the Horizon 2020 Framework Programme of the European Union

2 INTRODUCTION

New technologies bring new challenges. This is especially true for the Internet, a technology that has succeeded in removing barriers to knowledge and human communication. In the process, it has also shown potential for positive social transformation, contributing in such grassroots movements as the Arab Spring [1], but devious intentions have emerged as well. These include the dissemination of harmful content such as propaganda, hate speech, and disturbing videos depicting terrorist atrocities, child pornography, and revenge porn [2, 3]. While this content arguably has existed from the beginning of the Internet, its impact has become more problematic due to the global scale and influence of online platforms, the proliferation of user-generated content, and the increased proficiency of bad actors in cyberattacks. Due to the above, the EU commission has shifted its focus on online harmful content since at least 2017, and has been seeking ways to address it in consultation with the industry and the civil society[1]. The effort has culminated in the European Digital Services Act, ratified by Europe's policymakers in April 2022, which aims to modernise the EU's e-commerce Directive and force the big Internet companies to actively moderate the harmful content circulating in their platforms[2].

In this context, a special type of harmful content has emerged in recent years with devastating potential for societal harm: *deepfakes*. Deepfakes are synthetic images and videos, created from the manipulation of authentic content or fully synthetically, which are extremely difficult to detect by a human viewer [4]. Of course, synthetic media are not new, as image manipulation and CGI software have been around for many years, but deepfakes have achieved alarming degrees of photorealism, are resilient to traditional forensics, and can be created at a low cost and effort by everyone. This technology has become possible through advancements in AI technology and, up to now, has heavily targeted face manipulations. Initially, deepfakes were limited to humorous applications, for example, software to include the face of Nicolas Cage in movie clips[3] and the mobile applications FaceApp[4] and FakeApp[5], but the technology has taken a darker turn with the emergence of celebrity and other types of fake porn. Experts and journalists are also uneasy over the political repercussions of the technology and a future infection of the political sphere with fake media.

To address the threat of deepfakes, researchers and technology companies have come up with tools to automatically detect them, which are also based on AI technology [5]. These tools work by detecting artefacts from the generation process of the content that are undetectable to the human eye. Their accuracy however is not perfect and drops, most often when the data used for their training are not sufficiently representative of the target cases, or even in the presence of simple processing such as recompression. In addition, there have been efforts to securely embed provenance information to media from capture up to all possible manipulations[6] but the adoption of these technologies requires compliant capturing devices and is at a preliminary stage. Meanwhile, deepfake generation algorithms keep improving and in the near future it is conceivable that we will not be able to detect deepfake media with purely technical means.

---

[1] https://digital-strategy.ec.europa.eu/en/policies/illegal-content-online-platforms

[2] https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

[3] https://www.reddit.com/r/deepcage/

[4] https://www.faceapp.com/

[5] https://www.fakeapp.com/

[6] https://c2pa.org/

Funded by the Horizon 2020 Framework Programme of the European Union

Based on the above, we believe that deepfakes cannot be addressed effectively by machines alone but that human opinion is also needed. While humans do not have the processing capabilities of machines, they are better at understanding context and generalising so they can greatly contribute in deepfake detection. The task is best fitted to journalists and fact checkers, considering their professional expertise in verification and the potential for cross-organisational collaboration, as dictated by the growing trend of *collaborative journalism* [6]. In fact, collaborative practices have been applied extensively in recent years by the fact-checking community to verify the claims of politicians and protect against disinformation attacks during recent elections such as the 2016 US[7] the 2017 French Presidential Elections[8].

Of course, algorithmic tools should not be ignored as they can provide valuable feedback on a class of deepfakes that are undetectable even to the professional eye. Unfortunately, the highly technical nature and poor explainability of these algorithms hinder their widespread adoption by journalists. Currently, there are a few tools available for deepfake detection either by the research community or the industry. Tools by the research community like *DeepFake-o-meter*[9] tend to be open and extensible but have limited funding, which may jeopardise their sustainability and maintenance efforts. This concern is critical for deepfakes due to the rapid progress in generation and detection algorithms. In contrast, tools by the industry like *RealityDefender*[10] are well maintained by their companies but tend to be opaque and monolithic. Consequently, a market need exists for a user-friendly platform that is both extensible and economically sustainable, and connects state-of-the-art AI algorithms with their end users.

Finally, due to their centralised nature, online platforms face issues of transparency and trust, which have become all the more sensitive after the Cambridge Analytica scandal in 2018 [7]. This incident has sparked discussions around the Decentralised Web and Web 3.0, with blockchain anticipated to be a key component [8]. Blockchain has entered the mainstream agenda thanks to its usage in cryptocurrencies like Bitcoin[11] but, at its heart, it is a technology for distributed storage. The distinction from other distributed storage solutions is blockchain's focus on security and data consistency guarantees, achieved through powerful consensus algorithms. In addition to storage, simple commands can be implemented on top of blockchain in the form of *smart contracts*, which can be executed without the supervision of a central entity [9]. Powered by these features, a great wealth of blockchain applications have been proposed that are unrelated to cryptocurrencies [10]. Critically, blockchain networks can be used for the notarisation of online transactions, which are immutable and accessible to all concerned parties. This property can be capitalised to increase the transparency of online platforms and make their actions accessible and auditable by pertinent third parties.

Based on the above, we propose *DeepFakeChain*, a collaborative platform for journalists that will offer valuable tools for deepfake detection. On one hand, our platform will give access to state-of-the-art algorithms for automatic detection, which will be easily updatable to match the progress of deepfake generation. The interface to these algorithms will be non-technical and user-friendly, to encourage end user adoption. On the other hand, the platform will enable journalists from different organisations to offer feedback on uploaded media on the prospect of being a deepfake, in the form of *annotations*. After gathering sufficient annotations, a decision will be made based on the consensus of human annotators. We note that although

[7] https://www.propublica.org/electionland/

[8] https://crosscheck.firstdraftnews.org/france-fr/

[9] http://zinc.cse.buffalo.edu/ubmdfl/deep-o-meter/

[10] https://www.realitydefender.ai/

[11] https://www.bitcoin.com/

Funded by the Horizon 2020 Framework Programme of the European Union

the platform primarily focuses on deepfakes, it can be extended to other types of harmful content such as hate speech and disturbing content. Therefore, we envision the application of our platform in the general area of collaborative content moderation.

The DeepFakeChain design integrates a large number of security features that aim to enhance the trustworthiness of the platform's decisions. *Firstly*, proof of all uploaded media and annotations will be uploaded in an independent blockchain network to offer readily available evidence of possible tampering. *Secondly*, all annotators will be screened before registration to verify their affiliation to an accredited organisation and establish a minimum level of trust among collaborators. *Thirdly*, an advanced reputation and truth discovery protocol will be implemented that will protect the platform's decisions from erratic or malicious annotators. These decisions will be reached through smart contracts without the platform's intervention to ensure their independence and further enhance their trustworthiness. Lastly, the decisions and related metadata will be stored in the blockchain network and will remain available to pertinent third parties for auditing or research purposes.

The remaining document is structured as follows. Section 3 presents existing platforms, features of which are similar to DeepFakeChain. Section 4 describes the scientific background that is relevant to our solution. Section 5 illustrates the technical implementation of our platform. Section 6 elaborates on the anticipated impact of our solution. Section 7 presents key performance indicators (KPIs) and concludes the document.

Funded by the Horizon 2020
Framework Programme of the European Union

3 RELATED WORK

Our solution lies at the intersection of human and automatic verification methods, therefore, in this section, we present related tools from both categories. In particular, we first describe platforms for collaborative annotation to understand their evolution and current status. Then, we narrow our focus on collaborative verification tools that are used in journalism and are closest to the functionality of our platform. Finally, we present existing online tools and software for deepfake media detection and state our conclusions.

3.1 PLATFORMS FOR COLLABORATIVE ANNOTATION

In general, annotation is present in every user-generated online activity. Mailing lists, forums, and social media, are all implicit ways of crowdsourcing user opinion, without framing annotation as their primary purpose. Comments, as present in news, e-commerce, and recommendation sites are a more explicit form of annotation that have been with us for a while and allow people to express their opinion on published content. Taken to the extreme, the *Hypothes.is* platform[12] allows its users to annotate every possible webpage through a browser plugin.

These simple annotation methods face significant reliability issues, related to both the annotator and the annotation itself. For example, trolls, paid commenters, human disinformation agents and bots or simply misinformed users degrade the quality of the annotation. To resolve these issues, simple trustworthiness mechanisms have been proposed and implemented, including upvoting / downvoting and flagging comments as inappropriate or illegal. This feedback can be subsequently used for moderating, downranking or simply indicating the annotation as untrustworthy. Similarly, users gain reputation points by the quantity of upvotes or awards / badges that they receive. *Reddit*[13] is a representative example, where comments are ranked according to various criteria such as freshness, hotness, or trendiness. From their side, Redditors receive Karma points when their submissions contribute positively to the community[14]. Still, these methods have their limitations, for example, strict moderation can receive backlash from the users and lead into inflammatory discourse.

Question answering sites such as *Quora*[15] and *StackOverflow*[16] are a more deliberate way of crowdsourcing, if we consider the answers as annotations. In this case, we have the added feature that a decision is reached by the consensus of the user base, namely, the best answer provided. *Wikipedia*[17] is a representative example of this category, as it allows both unregistered and registered users, called Wikipedians in the latter case, to edit the encyclopaedia. In particular, for each Wikipedia page, a Talk page exists that documents and publicly displays all discussions around an edit. Considering that every user can contribute publicly, to ensure the quality of the final content, these platforms utilise elaborate reputation and consensus methods based on user hierarchies with different rights and privileges. For example, although the comments of all Wikipedians carry the same weight, older and more active members have less limitations in editing the encyclopaedia. Unfortunately, the

[12] https://web.hypothes.is/

[13] https://www.reddit.com/

[14] https://reddit.zendesk.com/hc/en-us/articles/204511829-What-is-karma-

[15] https://www.quora.com/

[16] https://stackoverflow.com/

[17] https://www.wikipedia.org/

Funded by the Horizon 2020 Framework Programme of the European Union

reputation systems of these platforms have an unintended consequence. As user contribution is voluntary, users are motivated to flaunt their opinions and even use their reputation scores for self-promotion in job-seeking platforms. A product of this behaviour is impoliteness and the platforms need to state elaborate codes of conduct to mitigate this problem[18].

Even more specialised platforms are Amazon's *Mechanical Turk*[19] and *Appen*[20], which allow crowdsourcing annotations from human workers for a small fee. In this case, human work is preferred because the annotation tasks are more expensive and difficult to automate, or to generate training data for AI algorithms. These platforms are popular, even among researchers, but have been criticised for evading labour legislation to pay the annotators below the minimum wage, along other ethical issues. Ethics aside, the annotations are frequently of low quality due to poor task description, users' dissatisfaction with their payment, or lack of expertise of the annotators, all of which are common issues of crowdsourcing.

3.2 PLATFORMS FOR COLLABORATIVE VERIFICATION

In recent years, due to the rise of false online information, collaborative platforms have become increasingly popular for fact-checking and verification in journalism, following the trends of collaborative and citizen journalism. We note here that collaborative journalism refers to the collaboration among professional journalists, while citizen journalism refers to the collaboration of amateur reporters, possibly under the guidance of professional staff.

WikiTribune[21] was a news wiki site, founded as a for-profit organisation by Wikipedia founder, Jimmy Wales, that operated between October 2017 and November 2019. Its goal was to recruit volunteers to write and curate articles about highly publicised news, following professional journalistic practices (fact-checking, citing sources, proof-reading, etc.) and guided by established reporters and editors. After one year of operation, the company laid off its professional staff and relied solely on volunteers, however, the project was not successful and eventually merged with the social network *WT.social*.

TruthSquad[22] was an initiative for training citizens (*truth squads*) on fact checking by reporting and researching controversial reported statements by politicians. The first pilot was launched in August 2020. Civilian fact checkers were guided by professional journalists who provided expert feedback on the users' input until each truth squad reached a verdict. By gamifying the experience, the TruthSquad initiative enjoyed high participation but keeping the users motivated and receiving quality feedback was limited, placing the hardest work on the shoulders of the professionals. This showed that TruthSquad worked better as an educational tool while highlighting the limitations of citizen journalism.

Meedan's *Check* platform[23], powered by Facebook and WhatsApp, intends to streamline collaborative verification efforts between professional journalists and simple users. In particular, users can submit claims and reports to the platform to be verified by the journalists. Interestingly, to scale the platform to large numbers of users, the heavy lifting of guiding users how and what to upload was done by WhatsApp bots. This ensured a reduction in number and

[18] https://en.wikipedia.org/wiki/Wikipedia:Essay_directory#Wikipedia's_code_of_conduct

[19] https://www.mturk.com/

[20] https://appen.com/

[21] https://web.archive.org/web/20191123162712/https://www.wikitribune.com/

[22] http://mediashift.org/2010/11/crowdsourced-fact-checking-what-we-learned-from-truthsquad320/

[23] https://meedan.com/check

Funded by the Horizon 2020 Framework Programme of the European Union

improvement in quality of the original submissions, while saving time-consuming manual labour. The journalists then worked to verify the uploaded claims, leaving a trail of annotations, ultimately reaching a verification label.

Meedan's Check has already powered various collaborative reporting and fact-checking projects in the political sphere such as Propublica's *Electionland*[24], reporting on voting issues during the 2016 US elections, and FirstDraft's *CrossCheck*[25], addressing misinformation in the 2017 French Presidential Election. Both of these initiatives were awarded by the Online News Association in 2017 and spurred similar activities[26]. Meedan also targets other type of harmful content such as hate speech, with Propublica's *DocumentingHate*[27] and the *Co-Insights*[28] project, as well as health misinformation[29].

FactcheckEU[30] was an initiative launched by the International Fact-Checking Network for fact-checking during the 2019 European parliamentary elections, gathering 19 European media outlets from 13 countries. Fact-checks from every outlet were gathered under a common portal and most of them were translated in many European languages.

Birdwatch is an initiative by Twitter to provide user comments and annotations on misleading Tweets. The goal is for Tweets to become flagged as misleading when there is sufficient justification and consensus among users that the Tweet is misleading. The service is currently experimental and raters are reviewed by the Associated Press and Reuters. Integration however is slow. The project was announced in October 2020 and the first pilot operated in January 2021. It operates on a different site from Twitter, which is open only to US users, and the results do not influence Twitter's ranking algorithm. The results of Birdwarch only started appearing in Twitter in March 2022 to a small subset of its users, after complaints of delay during the Ukrainian war. Interestingly, Twitter commits to not label, address, delete or modify user comments, unless they violate their terms of service.

The *European Digital Media Observatory* (EDMO) has brought together a community of more than 30 fact-checking organisations[31] representing all EU countries and working together to carry out investigations and fact-checking briefs that are published under the EDMO portal[32]. EDMO hubs are also contributing to this effort providing fact-checks and investigations on a national level[33].

---

[24] https://www.propublica.org/electionland/

[25] https://crosscheck.firstdraftnews.org/france-fr/

[26] https://wan-ifra.org/2019/11/verificado-2018-fighting-misinformation-collaboratively/

[27] https://projects.propublica.org/graphics/hatecrimes

[28] https://meedan.com/project/co-insights

[29] https://meedan.com/programs/digital-health-lab

[30] https://www.facebook.com/factcheckeu

[31] https://edmo.eu/fact-checking-community/

[32] https://edmo.eu/

[33] https://edmo.eu/edmo-hubs

Funded by the Horizon 2020 Framework Programme of the European Union

3.3 PLATFORMS FOR DEEPFAKE DETECTION

While deepfake generation and detection are still hotly investigated, progress has been rapid and some online tools are already available, offered both by academic institutions and the industry. The tools are typically limited to face manipulation, which is currently thought to pose the greatest threat.

Deepware[34], a Bosnian company working on cybersecurity, created the deepfake detection software *DeepWare Scanner* in 2019. This software is offered for free as a web service, an API, and a mobile application, and is even open source. On the technical side, Deepware's algorithm is based on the EfficientNet B7 convolutional network, which works on isolated frames. Since the algorithm is compatible with the winning model of the DFDC challenge, the developers include the option to parameterize it with the weights of this model, offering some diversity in detection. The developers have also committed to improving their algorithm, for example, taking into account temporal information and voice manipulation, and keeping the software open source. Interestingly, there is an option for requesting human expert validation by the company.

DeepFake-o-meter[35] is a free web service and deepfake detection library, implemented by the University of Buffalo's Media Forensics Lab. The library integrates 12 detection algorithms from the literature and is developed with extensibility in mind, even though its Github code has not been updated since October 2020 at the time of writing. The service allows users to upload a video, select their preferred detection algorithms, and provide an email to obtain the results.

RealityDefender[36] is a proprietary software for deepfake detection intended for big companies and governments. It has been developed by a San Francisco start-up in collaboration with Microsoft. In particular, it is based on Microsoft's Video Authenticator tool, which was released just before the 2020 US elections. The software can be tried for free for a limited duration.

DuckDuckGoose[37] is a Dutch company, offering their deepfake detection software DeepDetector. The software is closed source, claims to have an accuracy of 93% and to offer explainable results in the form of an activation map, showing the detected synthetic areas. It targets a wide area of applications, from digital ID verification, forensics, videoconferences, to journalism. The company also offers the browser plugin DeepfakeProof, which checks visited websites for deepfakes, and the deepfake generation software Replicant, which is intended for penetration testing of biometric authentication systems.

Sensity[38] is a Dutch company, founded in 2018, which offers secure authentication services such as ID verification, face recognition, and fraudulent document detection. Considering the threat of deepfakes to identity verification, the company has developed a detection software to complement its liveness detection module, which is available as an API or on-premises installation.

---

[34] https://deepware.ai/

[35] http://zinc.cse.buffalo.edu/ubmdfl/deep-o-meter/

[36] https://www.realitydefender.ai/

[37] https://www.duckduckgoose.ai/

[38] https://sensity.ai/

Funded by the Horizon 2020 Framework Programme of the European Union

BioID[39] is a German company, working on proprietary ID verification software. In 2021, it began participating in the German state-funded program FAKE-ID, which aims to develop new deepfake detection methods in collaboration with German universities. Consequently, the company has started integrating deepfake detection features in its liveness detection software.

Although not meant as a tool for identifying deepfakes, the *Coalition for Content Provenance and Authenticity* (C2PA)[40] addresses the prevalence of misleading information online through the development of technical standards for certifying the source and history (or provenance) of media content. C2PA is a Joint Development Foundation project, formed through an alliance between Adobe, Arm, Intel, Microsoft and Truepic.

3.4 CONCLUSIONS

The above show that collaborative editing/annotation tools have a long history in the Internet's history with important successes (Wikipedia, StackOverflow, Mechanical Turk). In journalism, this model trend is very much alive with the trends of collaborative and citizen journalism. Besides, the identified platforms for media verification have a strong focus on fact-checking, leaving a gap for the more specialised task of deepfake or more generally harmful content detection. We noted that the citizen journalism initiatives tended to produce mixed results and place a lot of work to the shoulders of the supervising professionals. Therefore, considering the specialised nature of deepfake detection, we opted against opening the platform to amateur reporters in this initial stage, which helps establish a minimum of trust among our platform's users.

Regarding the algorithmic tools for deepfake detection, we note that tools from academic institutions tend to be open-source and extensible but typically lack a business model, which threatens their economic sustainability. On the other hand, industrial tools monetize their software by targeting mainly identity verification and forensic applications, and are typically closed source and opaque. The only exception is DeepWare's software, which is open source, free and available in a variety of forms (web service, API, mobile app). We also note a trend of deepfake detection being offered as a feature instead of a standalone software. Compared with these platforms, DeepFakeChain includes the feature of collaboration, addressing deepfake detection much more comprehensively. In addition, DeepFakeChain recognizes the trustworthiness issues of human decision making and exploits blockchain technology and reputation models to output highly reliable trustworthiness scores, available to parties external to the platform. The scores are also complemented with comments protected from tampering, which support the made decisions and aid in auditability.

[39] https://www.bioid.com/

[40] https://c2pa.org/

Funded by the Horizon 2020 Framework Programme of the European Union

4 SCIENTIFIC BACKGROUND

Our solution brings together various important technologies: trust and reputation schemes, blockchain, and AI algorithms for deepfake detection. The necessary background on these technologies is provided in the following sections. In particular, we begin by describing trust and reputation models, specifically, why they are needed, in which networks they are applied and the contribution of blockchain. We then continue with AI algorithms on deepfake detection.

4.1 TRUST AND REPUTATION

To facilitate our presentation and position our project within the existing literature, it is worthwhile to distinguish three representative cases of networks on which trust and reputation schemes have been applied. These are depicted in Figure 1. Specifically:

- Figure 1.a depicts a *centralised network* where users interact with a server. The server may represent a seller, a root certificate authority, or simply a data aggregator.
- Figure 1.b depicts a *bipartite network* that represents a buyer-seller or provider-consumer scenario. This is the typical model of an e-commerce market with many sellers.
- Figure 1.c depicts a *decentralised network* where every user can interact with everyone else. Decentralised networks have been used in many paradigms such as peer-to-peer, ad hoc, wireless sensor and Internet-of-Things networks, as well as multi-agent systems (MAS)[41] [11].

DeepFakeChain makes use of both the centralised and the decentralised paradigm. On one hand, the platform's services will be deployed centrally for the convenience of the users, who will not have to install additional software. On the other hand, the platform's operations will be committed and notarised in an external decentralised network, based on blockchain, which will guarantee their transparency and trustworthiness. Key to this design is that the users will have independent access to the blockchain network and the stored data, which are immutable.

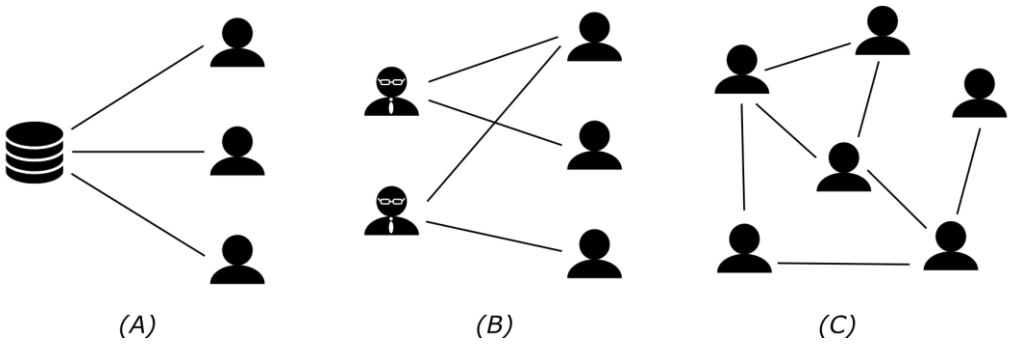

FIGURE 1: REPRESENTATIVE NETWORKS (A) CENTRALISED, (B) BIPARTITE, (C) DECENTRALISED

[41] The term MAS originated before 2000, replaced the older term *distributed AI,* and used to describe systems of intelligent agents interacting among each other autonomously. Today, the agents could refer to robots. MAS have offered a fertile ground and contributed greatly to the research of trust and reputation systems.

Funded by the Horizon 2020 Framework Programme of the European Union

4.1.1 TRUST AND REPUTATION

The notion of trust and reputation has been extensively studied in the literature, both from the lens of humanities and computer science as *computational trust* [12]. Yet, a common definition has not been reached, which has been attributed to the abstract and complex nature of trust. The most commonly cited definition is the one given by Gambetta [13], which defines trust as the subjective probability that an agent[42] will perform an expected action, assumed beneficial to us, in other words, a *service*. This definition is popular because it suggests the interpretation of trust as probability, although other models exist as well. Trust is also commonly accepted to be *dynamic* (changes with time), *contextual* (people are not trustworthy for all requests), and *multidimensional* (trust has many sources and aspects), and these characteristics are sometimes considered in computational models.

At this point, it is worth commenting on the distinction between trust and reputation. Like trust, reputation is difficult to define precisely although it involves the general perception of an agent by others. While most researchers agree that they are distinct concepts, in practice they are frequently used interchangeably [12]. A further complication is that reputation can be viewed both as a contributor to trust and the aggregation of the trust of others towards an agent. In our opinion, trust is tied to an interaction and reputation is a quality of an agent. Both quantities feed to each other as an agent's reputation influences the trust of others over an interaction while the validation of the interaction influences the agent's reputation.

Considering a set of agents, trust can be modelled computationally as:

1. a binary or discrete variable, e.g., from [-1, 0, +1] representing untrustworthy, neutral, and trustworthy states respectively.
2. a continuous variable or trust score, representing more granular notions of trust and uncertainty.
3. a tuple of values representing different dimensions of trust in different contexts.

Frequently, trust is modelled as a continuous variable in [0,1], which facilitates its interpretation as a probability. More general formulations exist, which interpret trust scores as beliefs and define special rules for manipulation (for example, see Dempster-Shafer theory [14]). Other mathematical tools that are frequently employed are *network analysis*, *Bayesian inference*, and *game theory*. Trust scores are typically inferred from:

- **direct information**, which stems from our direct interactions with an agent. It assumes a method to rate the success of that interaction, e.g., with a service level agreement (SLA), and the recording of past interactions. This is the most dependable source, as it relies on our own experience.
- **indirect information**, which stems from the information that other users provide on an agent. It is also called *witness information*, *recommendation*, *referral*, or *endorsement*. This source provides much richer information on agents, which is especially valuable for new users with few interactions. Unfortunately, it also contains uncertainty as the recommending users may be unreliable or malicious, hence their trustworthiness must be considered.

[42] The term *agent* comes from the MAS literature which has contributed a lot to computational trust. Here, it is used interchangeably with *person*, *node*, or *user*.

Funded by the Horizon 2020 Framework Programme of the European Union

- **pre-existing knowledge**, which is typically acquired from real life. This information is not always available but if it exists, it can guide users in new interactions with agents.
- **sociological or socio-cognitive information**, which takes into account social characteristics of the users. For example, these models may factor the frequency and consistency of interactions between two users or try to estimate the social roles of the users with social network analysis.
- **prejudices and bias**, which take into account the negative stereotypes that users have for each other.

Most computational trust models consider direct and indirect information since they are the most straightforward and cheap to compute, as well as pre-existing knowledge used for initialisation if it is available. The other sources may lead to more nuanced notions of trust but are also more complex and expensive to compute. Their usefulness depends on the application.

Computational trust received intense research after 2000 due to two crucial Internet applications: e-commerce and decentralised networks. In e-commerce, buyers needed to estimate the reliability of sellers to avoid scams, considering that word-of-mouth information for far away companies was scarce. In decentralised networks, users needed to a way to trust their peers, which were anonymous and potentially malicious. With reference to the networks of Figure 1:

- The centralised network of Figure 1.a conveniently aggregates the transactions of all users, performs the reputation computations centrally, and makes the scores and comments available to all users. Thanks to its global view, the server can monitor all operations, detect malicious users, and mitigate their activity. On the other hand, the centralised model introduces a big issue: the trustworthiness of the server.
- The bipartite network of Figure 1.b is a generalisation of the centralised network. In this case, buyers need to decide which seller to trust, although both sellers and buyers can be malicious. For example, sellers may not deliver the bought service or artificially raise their reputation with fake transactions (*ballot stuffing attack*), while buyers may not return accurate ratings or slander the seller even for services that they have not bought (*bad-mouthing attack*). Reputation scores must be diffused on all parties for the network to operate correctly.
- The decentralised network of Figure 1.c allows users to interact directly with their neighbours, and potentially everyone else. A big part of the literature is concerned with this case, which foregoes the central servers but opens new challenges in the process. In the decentralised case, the users must estimate the trustworthiness of everyone else but due to the absence of global information, asynchronous communication, and different latencies, they are the most vulnerable to attacks. An infamous attack is the *Byzantine generals attack* [15], in which nodes can present conflicting information to other nodes in order to jeopardise consensus.

4.1.2 BLOCKCHAIN-BASED TRUST

Blockchain is a special type of distributed ledger technology (DLT). A distributed ledger is a distributed database where transactions can be stored and verified by all nodes, and validated through consensus. While there many possible data models and technologies for DLTs, they rest on three common pillars: (i) *public-key cryptography*, to guarantee who initiated a transaction, (ii) *decentralised networks*, to broadcast transactions to all nodes without a central authority, and (iii) a *consensus mechanism,* for nodes to agree on a transaction [16]. In blockchain in particular, transactions are bundled and validated in the form of blocks, where each block points to the previous one. The most well-known consensus algorithm for

Funded by the Horizon 2020 Framework Programme of the European Union

blockchain is *proof-of-work*, popularised by Bitcoin, which relies on a competition of artificially hard computational puzzles to validate a block of transactions. Since proof-of-work has received criticism over its excessive energy consumption, alternative consensus mechanisms have been proposed and analysed over their security properties such as *proof-of-stake*, *Paxos-based* and *Byzantine fault tolerance*-based consensus [17].

Blockchain has revolutionised decentralised networks due to its ability to create a global state among distrustful nodes. The technology is also very resistant to data changes as, to tamper with the blockchain data, an attacker would have to compromise the majority of the network's nodes. This effectively ensures the *immutability* of the stored data. Other properties of blockchain storage include *transparency*, *auditability*, and *fault tolerance*, as all nodes store a copy of the blockchain and have access to its records. In the canonical vision of blockchain, all users are members of the decentralised network and register their transactions but the blockchain network can also be used as a secure independent storage for notarisation purposes.

With respect to trust and reputation systems, blockchain has had a big impact because it can make all past transactions and reputation values available to all users [16]. Figure 2 depicts the evolution of the networks of Figure 1 with access to blockchain storage. In the decentralised network case (Figure 2.c), this solves the difficult *Byzantine attack* as users cannot present conflicting information to different users. In the bipartite network (Figure 2.b), seller ratings can be recorded in the ledger with possible measures to blind the identity of the buyer. Even in the centralised case (Figure 1.a), the blockchain can help the users trust the server since all transactions are recorded transparently. Despite its benefits, blockchain cannot guarantee the trustworthiness of the stored data or if they were inserted by an unreliable or malicious user. Therefore, reputation systems are still needed to ensure data quality and protect from attacks.

The establishment of a common global state by blockchain and the centralised implementation of DeepFakeChain obviates the need for computing trust in a decentralised way, which has been investigated extensively in the literature. In contrast, the reputations of the users can be evaluated by everyone through the sources already described, e.g., direct, indirect, social information etc. Another attractive approach is *truth discovery* [17], which proposes techniques to estimate the true value of an object based on the noisy ratings of various users. Truth discovery approaches are based on voting and they allow calculating source reliability scores along with the desired truth in an unsupervised way. In fact, the source reliability of the truth discovery literature can be considered as a special form of trust score and can complement traditional techniques that estimate the reputation of users from past transactions. In DeepFakeChain, we intend to use a combination of these techniques to establish the presence of a deepfake from the ratings of human experts and use these results to form the reputation of the experts.

Funded by the Horizon 2020 Framework Programme of the European Union

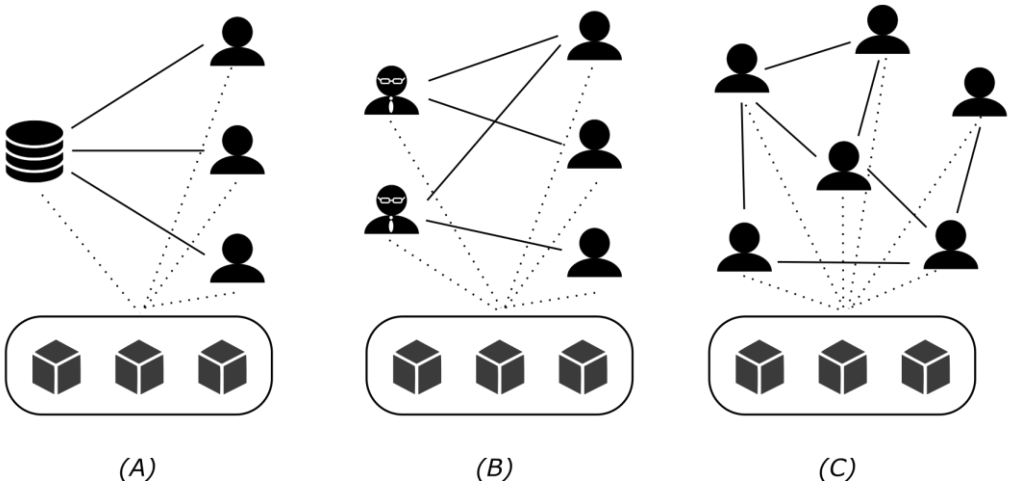

FIGURE 2: REPRESENTATIVE NETWORKS WITH BLOCKCHAIN STORAGE (A) CENTRALISED, (B) BIPARTITE, (C) DECENTRALISED

4.2 AI-BASED DEEPFAKE DETECTION

Deepfake generation algorithms are special cases of generative AI models. Generative models can generate new data samples (e.g., new images) from a given collection of samples that in an abstract sense specify a data distribution (e.g., a given collection of images) [18]. While some traditional machine learning (ML) models, such as fully visible belief networks, variational autoencoders, and restricted Bolzman machines, had some limited success in this task, a breakthrough came in 2014 when Ian Goodwill proposed generative adversarial networks (GANs) [19]. GANs contain two models: i) a *generator*, which creates random samples, and ii) a *discriminator*, which classifies samples as real or fake. During training, the generator creates new samples and the discriminator is presented with both real and fake data. By parameterizing both the generator and the discriminator, the ground truth labels (real or fake) can be used to improve both models, which have competing goals. As both models improve, the training continues until the generator is powerful enough so that the discriminator can no longer distinguish fake data from real. This process can be seen as a zero-sum game that plays out until the model-players reach a Nash equilibrium, a radically different paradigm from conventional supervised learning.

Since 2018, research on GANs and generative models has increased exponentially resulting in new tasks and impressive photorealism; see for example OpenAI's DALL-E 2[43] and Google's IMAGEN[44] diffusion models. Some of these tasks include:

- **image to image translation**, translating a scene representation to another, for example, transforming a horse into a zebra or changing day to night.
- **image in-painting**, inserting a new image, for example a face or an object, in an existing image or video.
- **image harmonisation**, removing objects from images and videos and replacing them with photorealistic content that matches the scene.
- **text to image synthesis**, generating an image that matches a given textual description.
- **resolution upscaling**, increasing the resolution of an image by adding plausible pixels.

[43] https://openai.com/dall-e-2/

[44] https://imagen.research.google/

Funded by the Horizon 2020 Framework Programme of the European Union

It is worth mentioning that a large part of the literature is concerned with face manipulation in photos and videos due to the significance of the human face in communication. Face manipulation is further classified as [20]:

- **entire face synthesis**, creation of an entirely synthetic face.
- **identity swap**, replacing a face with another person's face.
- **attribute manipulation**, modifying parts of a face, for example, a nose.
- **expression swap**, modifying the expression of a face.

Due to the increased realism of deepfakes that can evade the human eye, researchers have turned to automatic methods of detection. Interestingly, the adversarial training of GANs produces a discriminator that is optimised in the same training set as the generator but it is not as useful for detection for two reasons: i) in practice, the GAN architecture used in the deepfake generation is unknown, and ii) the training of the GAN aims at optimising the generator. Consequently, new algorithms for detection have been proposed [4]. Some of them try to detect artefacts that reveal the synthetic generation of deepfakes. These can be inconsistent local artefacts (eyes, teeth, head poses), artefacts in physiological signals (eye blinking, heart beats), artefacts in the frequency domain, inconsistencies between images and audio, etc. Other techniques forego the explainability of specific signals and train ML models treating detection as a pure classification problem. Irrespective of the approach, videos offer more opportunities than static images for detection due to the existence of temporal and audio information. Yet, this comes at the cost of more intensive processing.

While some proposed methods work well for specific datasets, there are important challenges:

- the compression of videos degrades the deepfake artefacts and the accuracy of detection, and is frequently not considered in the literature.
- the proposed methods tend to overfit the training datasets and have lower performance in other datasets.
- the increased effort placed on face detection may render detection models less effective for other types of synthetic content.
- the improvement of generative models can conceivably lead to undetectable fake media in the future.

Due to the above, our platform can greatly help in the fight against harmful deepfakes by i) including in the loop humans who are (currently) more adequate to evaluate contextual information than machines, and ii) by creating a pool of resources that can be used in training more powerful deepfake detectors. Our platform is explicitly designed to be extensible so that new detection models can be added as deepfake generation algorithms improve. In addition, deepfake detection algorithms can greatly benefit from signals from multiple models that consider different features. There is currently scarce knowledge on how well models combine, and our platform will help in this direction.

Funded by the Horizon 2020 Framework Programme of the European Union

5 TECHNICAL IMPLEMENTATION

This section describes our technical implementation design. We begin by outlining the user requirements of our platform that helped guide our design. We then illustrate the implementation design with an architectural diagram and elaborate on the role of each component. We conclude with a summary of the tasks required to implement our platform. Note that several of the presented requirements and design decisions are tentative and may be revisited during the project.

5.1 USER REQUIREMENTS

The users of DeepFakeChain belong to the following categories:

- **Platform staff** who manage the platform and ensure its correct operation.
- **Journalists** from professional media companies and organisations, who use the platform for collaborative annotation and integrate it to their everyday workflow. Depending on their role in their company, they can have different levels of access to the DeepFakeChain platform.
- **Stand-alone verification experts** who are not affiliated with a professional media company but have the credentials to collaborate in the verification process.
- **Simple users** who can monitor the platform for verified deepfakes and request the examination of uploaded media.
- **Third-parties** such as researchers, auditors, and other unanticipated categories who can use the platform for specialised goals with specialised access rights and privileges. For example, researchers can access the media annotations to use as training data, and auditors can access annotations and user profiles to evaluate the platform's operations.

Before we describe the user roles that cover the above categories, we need to define the concept of a **channel**, which enables flexible cross-organisational collaboration among journalists in our platform. A channel is a task group within DeepFakeChain with well-defined rules targeting a specific category/topic of media, for example, related to the unfolding Ukrainian War. Journalists from different organisations can subscribe and collaborate in any channel subject to the rules set forth by the channel's administrators. These rules concern civil conduct within the channel, its level of visibility, and the criteria for reaching consensus, we stress though that the channel administrators cannot tamper with the annotations of annotators nor the decisions reached by the annotators' consensus. Regarding visibility, channels can be *open*, *public,* or *private,* which determines the authorisation required to view and annotate contents in the channel. In particular:

- **open channels** are open for users to view and annotators to collaborate without explicit approval by the channels' administrators.
- **public channels** allow users to view their uploaded media but annotators require the explicit approval by the channels' administrators to begin collaboration.
- **private channels** are searchable by users but subscription and approval is required to view their contents. Annotators also require permission to collaborate in these channels.

Based on the above, we define the following user roles:

- **Public viewers** who can view uploaded media and annotations from public channels. Public viewers can upload media anonymously to be evaluated by our AI algorithms but further actions require registration.

Funded by the Horizon 2020 Framework Programme of the European Union

- **Registered viewers** who can create a simple user profile and subscribe to channels inside DeepFakeChain. Registered users can access the contact information of reviewers if the latter have revealed it and can submit media for human annotation by a channel's reviewers. This submission can be accepted or rejected based on the reviewers' discretion.
- **Reviewers** are the annotators of the platform. They are registered on the platform only via invitation by the Editors, which represent credible media organisations and companies. The consortium will consider additional registration procedures that cover the case of unaffiliated annotators.
- **Editors** represent media organisations and companies and have the rights to invite new reviewers to the platform. To register in the platform themselves, they need to contact and be approved by the Global Administrators of the platform, who will perform the necessary screening.
- **Channel administrators** are users responsible for a channel. By default, the creator of a channel becomes its first administrator who can invite and manage subscription requests by reviewers, and define additional administrators. The creators of a channel can be Editors and, possibly, simple reviewers to increase flexibility. A channel administrator can set the rules of operation of a channel but cannot influence the comments and consensus of its reviewers.
- **Global administrators** are DeepFakeChain's staff, who are responsible for the smooth operation of the platform. Beyond their technical assistance and maintenance of the platform, they are responsible for screening and approving the requests of Editors, thus setting in motion the operation of the platform. Global administrators have global rights over the platform but even they cannot tamper with the reviewers' annotations and decisions. As the platform grows, we anticipate this role to divide between maintenance staff, a research unit, which will upgrade and add new algorithms to the platform, and an administrative unit, which will be responsible for approving new editors in the platform.

From the above, we highlight the top-down registration process of editors and reviewers, which shifts the administrative burden of registering reviewers to the editors. In this design, the global administrators are responsible for screening only the editors, while the latter are held accountable for their own reviewers. On the other hand, after their registration, the reviewers have equal rights in the platform and are free to participate or oversee any channel in the platform, irrespective of their editors' actions. We believe that this design strikes a good balance between flexibility and accountability, and promotes trust among the users, a key ingredient of every successful collaborative platform.

To clarify the actions and rights of each user role, we present them formally as user requirements in the following tables.

TABLE 1: USER REQUIREMENTS FOR PUBLIC VIEWERS

ABBR	REQUIREMENT
PV-1	Public viewers shall search via keyword for channels and media from public channels based on their stored metadata (title, tags, description, etc.).
PV-2	Public viewers shall playback searchable media.

Funded by the Horizon 2020 Framework Programme of the European Union

PV-3	Public viewers shall view the annotations and metadata with public visibility of searchable media. The visibility will be determined by the rules of the channel and reviewer configuration, set forth by the channel administrators.
PV-4	Public viewers shall have access to reviewers' personal profiles subject to the visibility constraints established by the reviewers' privacy configurations.
PV-5	Public viewers shall submit videos anonymously to be evaluated by the platform's deepfake detection algorithms. The platform will maintain an indication of the processing status and will load the results immediately once they are read, depending on the available computational resources.
PV-6	Public viewers shall register to the platform via a public online form to achieve the status of registered reviewer. Registration of a viewer will require only the verification of the user's email.

TABLE 2: USER REQUIREMENTS FOR REGISTERED VIEWERS

ABBR	REQUIREMENT
RV-1	Registered viewers shall retain all the privileges of public viewers.
RV-2	Registered viewers shall create and manage a user profile in the platform with basic information. At a minimum, this information will contain the user's email and password.
RV-3	Registered viewers shall close their user accounts. All personal data will be removed from the platform within a maximum retention period defined by the applicable law.
RV-4	Registered viewers shall subscribe to channels to facilitate monitoring their operations. In particular, the registered viewers will have direct access to their subscribed channels through their account page. Subscription to private channels requires the approval of one channel administrator.
RV-5	Registered viewers shall select to be notified by new media uploads in their subscribed channels. They may also be notified by new comments and reached decisions in selected media.
RV-6	Registered viewers shall view metadata with higher visibility restrictions than public viewers.
RV-7	Registered viewers shall view reviewers' profiles with higher visibility restrictions than public viewers, which will be established by the reviewers' privacy configurations.

Funded by the Horizon 2020 Framework Programme of the European Union

RV-8	Registered viewers shall search for similar videos from their subscribed and public channels via the reverse search functionality.
RV-9	Registered viewers may request the human annotation of uploaded media that are submitted to a subscribed channel. They will be notified via email about the approval or rejection of their request, which will be subject to the discretion of the channel's administrator and reviewers.

TABLE 3: USER REQUIREMENTS FOR REVIEWERS

ABBR	REQUIREMENT
RE-1	Reviewers shall retain all the privileges of registered viewers.
RE-2	Reviewers shall register on the platform via invitation by editors. They may also register via other means that do not require an affiliation to an editor, for example, invitation by channel and global administrators.
RE-3	Reviewers shall create and manage a user profile with more information than a registered viewer. This information will include a name, a photo, a self-description, affiliation and contact information. The platform may implement more advanced identity verification procedures.
RE-4	Reviewers shall register to open channels or request registration from public and private channels to work as annotators. Registration to public and private channels will require the approval of the channels' administrators. Reviewers shall also register to channels via invitations sent by their administrators.
RE-5	Reviewers may view lists of content pending annotation, which are assigned to them by the administrators of their registered channels. The deadlines for completing these annotations will be according to the rules set forth by the channels' administrators.
RE-6	Reviewers shall annotate media with annotations that are general or linked to a specific spatiotemporal context. The annotations will be simple statements or contain a judgement.
RE-7	Reviewers shall control the visibility of their annotations, judgements, and comments to public and registered viewers if permitted by the rules set forth by the channel administrators. Visibility will be always full for peer reviewers in the same channel, the reviewers' editors, and the administrators of the channel.
RE-8	Reviewers shall communicate indirectly via comments in the pages of media. The comments may have a tree structure to allow reviewers to reply to each other and may contain reputation features such as awards and votes to promote helpful comments. The reviewers may also communicate via direct means.

Funded by the Horizon 2020 Framework Programme of the European Union

RE-9	Reviewers shall be notified of activity in their assigned media such as new annotations, comments, and reached decisions.
RE-10	Reviewers will receive invitations to become channel administrators extended by existing channel administrators. They may create channels and become their first channel administrators. This is a typical editor privilege but may be extended to reviewers to increase flexibility.

TABLE 4: USER REQUIREMENTS FOR EDITORS

ABBR	REQUIREMENT
ED-1	Editors shall retain all the privileges of reviewers.
ED-2	Editors shall request registration to the platform through a public online form. The form will provide contact information, a description of the editor's organisation, and the intended use of the platform. The editor's request will be approved or rejected by the platform's global administrators possibly after contacting them for further screening.
ED-3	Editors shall create and manage a user profile that contains additional information about the represented organisation. This may include the organisation's address, websites, and VAT number.
ED-4	Editors shall invite reviewers to create accounts on the platform. The editors will be responsible and held accountable for the actions of their reviewers.
ED-5	Editors may monitor the activities of their reviewers through a special panel in their account page.
ED-6	Editors shall deactivate or close the accounts of reviewers who misbehave.
ED-7	Editors shall create channels and become their first channel administrators.

TABLE 5: USER REQUIREMENTS FOR CHANNEL ADMINISTRATORS

ABBR	REQUIREMENT
CA-1	Channel administrators shall set the rules governing the channel's operations, including its general visibility (open, public, private), the visibility options of the registered reviewers, the rules of conduct, and the criteria for reaching consensus.

Funded by the Horizon 2020 Framework Programme of the European Union

CA-2	Channel administrators shall participate in the annotation process of their own channels.
CA-3	Channel administrators shall approve requests of reviewers to collaborate in their channel if the channel's visibility is set to public or private.
CA-4	Channel administrators shall invite reviewers and editors to become peer administrators in their channel.
CA-5	Channel administrators may monitor the reviewers and registered users subscribed in their channel through a special panel in their account page.
CA-6	Channel administrators shall unsubscribe or deactivate the activity of reviewers who misbehave in their channel.
CA-7	Channel administrators shall accept or reject requests by registered users to evaluate their uploaded media.

TABLE 6: USER REQUIREMENTS FOR GLOBAL ADMINISTRATORS

ABBR	REQUIREMENT
GA-1	Global administrators shall have full access to all data stored in the platform. They will not be able to remove data that has been stored in the external blockchain storage.
GA-2	Global administrators shall evaluate the requests of media organisations to create editor accounts in the platform.

5.2 SYSTEM ARCHITECTURE

The system architecture consists of four main layers that will interact with each other to realise the project's vision. Namely:

- The User Layer, where the user will interact with the system.
- The AI Layer, where the core AI-based tasks are performed.
- The Data & Content Layer, to gather and maintain all the media content, accompanying metadata and the results from the processing capabilities of the DeepFakeChain platform.
- The Blockchain Layer, which contains the actual ledger and all the blockchain related functionalities and technologies.

Funded by the Horizon 2020 Framework Programme of the European Union

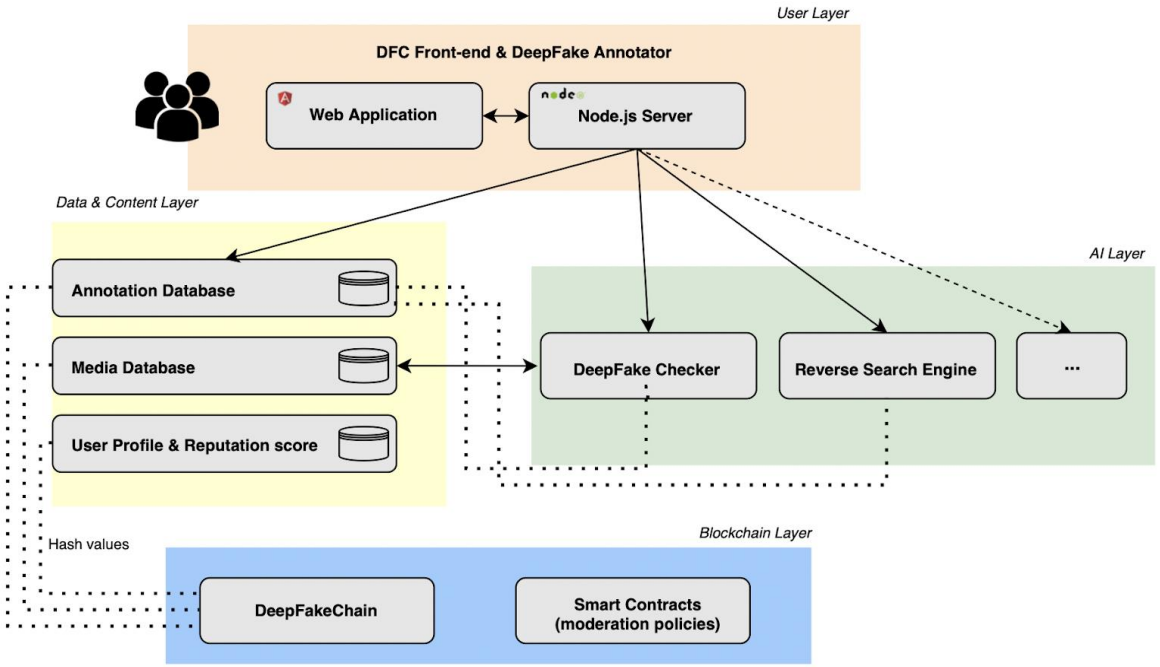

FIGURE 3: ARCHITECTURE OF DEEPFAKECHAIN

Each layer plays an important role for the realisation of the project's vision and all necessary precautions regarding the confidentiality and integrity of the stored data will be taken. All the technologies that will be used are state-of-the art solutions that are suitable for each part and functionality of the project in order to guarantee the security of the system's data and communications. Following is a more detailed description of each layer.

5.2.1 USER LAYER

The user layer consists of two main components: i) the user interface and ii) a dedicated server. For the user interface, we intend to extend the base implementation of the existing platform DFDLab[45], provided by CERTH. DFDLab offers a simple interface for users to upload and annotate videos, and we intend to extend it by implementing the functionalities of DeepFakeChain's user roles such as all possible registering methods, searching by text and similar multimedia, voting for deepfake judgement etc. In addition, we are considering the SmartViz toolkit provided by Zelus to enhance the user experience with visually attractive dashboards. These dashboards will contain intuitive visualisations and tables tailored to each user role and/or business need. The users will also be able to customise them by adding and removing widgets from the rich widget pool of the SmartViz toolkit.

For the dedicated server, a Node.js express server can be used as a middleware of all the backend communications and services of the DeepFakeChain solution. This server will provide a REST API, through which all the communications from the user interface to the actual AI-driven services will be realised. The server can also be described as the orchestrator of the entire architecture, since it will organise and provide to the user all the available services and

[45] https://dfdlab.mever.gr/, not publicly accessible

Funded by the Horizon 2020 Framework Programme of the European Union

functionalities from the AI layer. The orchestrator will also handle all the communications with the system's persistent storage and the blockchain network.

5.2.2 DATA & CONTENT LAYER

The data & content layer is the persistent storage of the DeepFakeChain solution. It sits between the user layer and the AI layer, and stores both media- and user-related data. Regarding the media=related data, it stores the media files along with their metadata and URL links, the annotations from users and the AI algorithms, and the decisions reached by the consensus mechanisms. Regarding the user-related data, it stores the users' profiles, information about their actions in the platform, and their reputation scores.

We highlight that the persistent storage of DeepFakeChain is off-chain and will be synchronised with the ledger of the blockchain ledger. This design was chosen because storing and indexing large files such as media data is more efficient in a conventional database than a blockchain network. Our storage solution thus combines the efficiency of conventional databases with the robustness of blockchain. Regarding the storage technologies, we consider MongoDB for the platform's metadata in MongoDB, extended from DFDLab, and Elastic Search for the searchable metadata.that need to be indexed. Elastic Search is an attractive choice because it offers quick indexing and fetching of large amounts of data, state-of-the-art security, multiple APIs, and good interoperability with JavaScript applications.

5.2.3 AI LAYER

The AI layer contains the core functionalities of the DeepFakeChain, specifically, state-of-the-art ML services for processing multimedia. Initially, the AI layer will contain two important services that have been developed by the CERTH team: i) the Deepfake Checker [21], which outputs the possibility of a video of being a deepfake, and ii) the Reverse Search Engine [22], which allows uses to search for similar content via a multimedia query. Both services will be available to the end users through appropriate interfaces at the user layer. In addition, the AI layer will be extensible to future additions of services and algorithms. This will be enabled through the Node.js server that will orchestrate all AI-assisted data pipelines.

5.2.4 BLOCKCHAIN LAYER

The blockchain layer contains the blockchain network of the DeepFakeChain solution, acting as a ledger of record trails of all media files, human annotations, and results from the AI layer. The ledger will permanently store evidence of the DeepFakeChain's operations, which can be used as proof of trustworthiness to external parties and auditors. In particular, the ledger will only hold hashes of the records that are stored in the data & content layer. Thus, the blocks will have to be carefully organised to store hashes of the actual media files along with annotations. In parallel, the user's reputation score should be also stored in the blockchain since the profiling of a user-content creator or uploader is of high importance in drawing conclusions. Therefore, a hash that describes a record of the reputation score and other user's metrics should also be part of each block-transaction.

For the governance of the blockchain infrastructure, we will define smart contracts that will interact with the network. Specifically, smart contracts will be used to query and insert new entries to the ledger in the agreed format, as well as reach consensus among human annotators without the intervention of the DeepFakeChain platform or its users. In this initial design, we foresee the usage of Alastria's blockchain network by registering DeepFakeChain's server in the network. In addition, we require independent access to the Alastria network through an auditing interface to guarantee the users' trust in the platform. The exact implementation of this interface is under investigation at the moment.

Funded by the Horizon 2020 Framework Programme of the European Union

5.3 TASKS & MILESTONES

Below, we describe the tasks and milestones of the DeepFakeChain project.

Task 1: DeepFakeChain Business Analysis and Specifications (M1-M3; Lead: Zelus)
Includes the first sprint (M1-M3) of Phase 1, resulting in the detailed description of the business plan of DeepFakeChain in D1.2, including its scope, drivers, and user requirements. In M2, an infographic of the proposed solution will be designed and delivered for dissemination purposes (MS1).

Task 2: DeepFakeChain Proof of Concept Implementation (M4-M9; Lead: CERTH)
Includes the second (M4-M6) and third (M7-M9) sprint of Phase 1. The second sprint is concerned with the deployment of the individual components of DeepFakeChain and their initial integration with TruBlo's ecosystem, culminating in video presentations of the individual components (MS2) and a detailed technical report (D1.3). The third sprint is concerned with the release of DeepFakeChain's first PoC, culminating in a demo presentation of the platform (MS3) and the submission of a scientific paper to a peer-reviewed Conference or Journal (D1.4). This paper is expected to be accepted and published by M14, fulfilling MS4 of Phase 2.

Task 3: DeepFakeChain final solution delivery (M10-M15; Lead: CERTH)
Assuming that DeepFakeChain enters Phase 2 of TruBlo OC3, this task will include the activities of the final two sprints for the development and delivery of the close-to-market DeepFakeChain solution. The fourth sprint (M10-M12) will create the detailed development plan (D2.1) and result in the draft MVP of the integrated DeepFakeChain (D2.2), based on the PoC from phase 1 (D1.3). The final sprint (M13-M15) will produce the final MVP of DeepFakeChain to be tested by the end-users (will be reported in D2.5); Adjustments and patches based on the issues and comments extracted from the validation phase (Task 4). All this will be reported in D2.4.

Task 4: DeepFakeChain Demonstration, End-User Validation and Business applicability (M10-M15; Lead: Zelus)
Assuming that DeepFakeChain enters the second phase, this task will include two parallel sprints for the validation and business analysis of the developed solution. The first sprint (M10-M12) will develop a detailed validation plan including technical KPIs, time-frame and test cycles with end-users (D2.3). A direct link will be set with Task 3 in order to timely deliver the user's comments and validation results, to be incorporated in the final platform. The demo of the completed platform along with the details of the end-user demonstrations (one online webinar, two hand-on workshops) will be reported in D2.5. All the technical details and the results of the validation will be reported in a technical white paper (D2.6) that might also be submitted to an appropriate venue. In parallel the initial business analysis of phase 1 (D1.2) will be revised and enhanced with a market analysis, questionnaires to validate the business applicability and sustainability. The final Business Plan will be reported in D2.7.

Funded by the Horizon 2020 Framework Programme of the European Union

TABLE 7: DELIVERABLES AND MILESTONES FOR THE 1ST PHASE OF TRUBLO OC3

Nº	Deliverable or milestone name	Description	Type	Delivery Month
D1.1	Full Research and Innovation Project Proposal	The present report.	Report	M1
MS1	DeepFakeChain Objectives/ Impact/Vision	An infographic providing the DeepFakeChain vision, objectives, offerings and impact.	Diss. Info	M2
D1.2	Project Solution Design and Business Applicability	A report detailing the business aspects of DeepFakeChain, including the project's scope, market drivers, and planned functionalities.	Report	M3
MS2	DeepFakeChain Components	A video demo of individual DeepFakeChain's components.	Demo	M4
D1.3	Technical Report	A report detailing the technical architecture of DeepFakeChain's initial PoC.	Report.	M6
MS3	DeepFakeChain PoC Demo	A video demo of DeepFakeChain's initial PoC.	Demo	M7
D1.4	Scientific Publication	A submission of a research paper presenting the results of D1.3 and MS3 to a peer-reviewed Journal or Conference.	Publ.	M9

TABLE 8: DELIVERABLES AND MILESTONES FOR THE 2ND PHASE OF TRUBLO OC3

Nº	Deliverable or milestone name	Description	Type	Delivery Month
MS4	Accepted Scientific Publication	The acceptance of the work presented in D1.4 at a peer-reviewed Journal or Conference.	Publ.	M14
D2.1	Detailed Development Plan	A report detailing the development plan for the phase 2 of the project.	Report	M10

Funded by the Horizon 2020 Framework Programme of the European Union

D2.2	Detailed Development Plan Monitoring	A report detailing the development progress of DeepFakeChain.	Report	M12
D2.3	Validation Plan	A report detailing the validation of DeepFakeChain.	Report	M12
MS5	User Feedback	The initial user perception on DeepFakeChain.	Report	M14
D2.4	Report on DeepFakeChain's MVP	A report detailing the final form of DeepFakeChain solution.	Report	M15
D2.5	Demonstration of Developed Solution with Users	The demo of the developed solution to end users and feedback collection.	Demo, Report	M15
D2.6	White Paper	A white paper detailing DeepFakeChain.	Diss. Info	M15
D2.7	Business Plan	DeepFakeChain's short and long term Business plan.	Report	M15
MS6	Final solution and business readiness assessment		Report	M15

Funded by the Horizon 2020 Framework Programme of the European Union

6 RESULTS & IMPACT

We anticipate that the DeepFakeChain platform and the project's results will have a notable impact on collaborative annotation and verification of harmful content. The following sections describe the expected impact on the scientific, technical, societal, and economic front.

6.1 SCIENTIFIC IMPACT

DeepFakeChain is a new solution in the scientific area of blockchain-based trust and reputation systems and is expected to have a notable impact there. Users will be able to verify the integrity and provenance of the uploaded media, annotations, and reputation scores of the platform, even in the presence of centralised services such as the web portal and the services of the AI layer. Considering that the annotations on public media are useful beyond the confines of our platform, the data uploaded in the data will be designed to have significance to outside observers. This will contribute to interoperability, which is a frequently overlooked aspect of trust and reputation schemes.

In addition, our solution will combine conventional trust and reputation techniques with techniques from the scientific field of truth discovery to derive the trustworthiness of both annotations and annotators in the absence of ground truth labels (if the content is indeed deepfake or not). We intend to characterise the security guarantees of our solution and investigate its attack surface, considering attacks from both the users and the administrators of the platform and its constituent services. We highlight that the public notarisation of the platform's decisions in the blockchain storage and the decentralised execution of the smart contracts will play a key role in mitigating potential attacks from malicious actors. The investigation of the above is expected to fuel a submission to a relevant peer-reviewed venue.

Another important aspect of our solution is the crowdsourcing of training data on deepfake media and its possible use in training new AI algorithms for detection. Considering the scarcity of public datasets on the detection of deepfakes and other types of harmful content, by opening our platform to third party researchers, the contribution to the scientific community will be high. In addition, the generated data from our platform will include trustworthiness scores, which will greatly raise their value, mitigating a key issue of conventional crowdsourcing approaches. This opens up attractive research avenues of how these trust scores can be fully capitalised during the training of new algorithms. Finally, as our platform is designed to bring together multiple AI services for deepfake detection, it can provide a testing ground for the effectiveness of their combination and how complementary they are. This is a largely unexplored part of the literature on deepfake detection, which our solution hopes to illuminate.

6.2 TECHNICAL IMPACT

From the technical point of view, the DeepFakeChain solution will push the modern technologies and techniques to its limits. The blockchain technology has been advancing since 2008 and it is now used in many different fields, proving to be an important problem solver that transfers the trust from the users to the system. Our approach uses an advanced blockchain network, which will be controlled by smart contracts, designed to exploit all the capabilities of this technology. The advanced security features of blockchain for the notarisation of media data and metadata will be complemented by our system's logic, which orchestrates the AI services and the off-chain data storage. This orchestrator will have full control over the system and adhere to all the modern security mechanisms in order to safeguard the user data and the system's data and metrics.

Funded by the Horizon 2020 Framework Programme of the European Union

DeepFakeChain will also offer access to advanced and specialised AI algorithms for automatic media verification, as well as algorithms for search similar audiovisual content. Further algorithms can also be included owing to the extensible design of the platform. This sophisticated technological substrate must be presented to the user in the most user friendly way possible, and our platform aspires to create a user interface that is simultaneously clean, i.e., highlighting only the most important features to the user, and comprehensive, i.e., offering the full details and digital trail around any selected media item.

6.3 SOCIETAL IMPACT

A recent study published by the European Parliamentary Research Service =[23]=[23] concluded that deepfakes give rise to a wide range of societal harm, for example, to the financial, judicial, and political systems. To the individual, deepfakes enable many kinds of fraud such as increased risk of defamation, intimidation, and extortion. Figure 4 presents an overview of different categories of risks associated with deepfakes, which are not expected to be resolved in the coming years.

Psychological harm	Financial harm	Societal harm
• (S)extortion • Defamation • Intimidation • Bullying • Undermining trust	• Extortion • Identity theft • Fraud (e.g. insurance/payment) • Stock-price manipulation • Brand damage • Reputational damage	• News media manipulation • Damage to economic stability • Damage to the justice system • Damage to the scientific system • Erosion of trust • Damage to democracy • Manipulation of elections • Damage to international relations • Damage to national security

FIGURE 4: RISKS ASSOCIATED WITH DEEPFAKES

DeepFakeChain proposes a practical solution to identify and expose untrustworthy content by building a trusted repository of deepfake multimedia, which will be made available to professionals and non-expert citizens alike. For professionals, the platform will help streamline their work and increase their productivity through the feature of channels and tools for similar multimedia search. The reputation system will also mitigate erratic or even malicious judgements, enhancing the trustworthiness of the final decisions. For non-experts, our platform will provide trustworthy judgements that can be backed with sufficient justification, for example, what was the consensus of judging a media content deepfake and which reviewers participated in the judgement. This will enhance the credibility of the platform's output to citizens and help prevent individual harm.

Our platform also aspires in alerting citizens on the presence of deepfake media in the online world and educating them on how to detect them. The transparency of our platform will play a key role in this direction by showcasing all the steps taken to verify whether a piece of content is manipulated and highlighting the visual areas suspected of manipulation. In addition, we plan to deliver education sessions for the different types of users, in order to explain to them how the platform works, lessons learned throughout the development process as well as its benefit for them.

6.4 ECONOMIC IMPACT

We plan to make DeepFakeChain available to the EDMO communities of fact-checkers and researchers to collaborate on deepfake detection and have access on previously identified

Funded by the Horizon 2020 Framework Programme of the European Union

cases of media manipulation. EDMO includes national and multinational hubs with specific knowledge of local information environments so as to strengthen the detection and analysis of disinformation campaigns, improve public awareness, and design effective responses for national audiences. Currently, the EDMO hubs cover Ireland, Belgium, Czech Republic, Denmark, Finland, France, Italy, Luxemburg, the Netherlands, Poland, Portugal, Slovakia, Spain, Sweden, as well as Norway in the EEA, making these countries the first target markets for the commercialization of DeepFakeChain.

Within the e-publishing and user generated content markets, verification and fact-checking are of paramount importance. Our main target audience is media organizations, independent reporters, bloggers, public speakers in general, who want to protect their credibility and mitigate the risk of civil compensation for defamation[46] which ranges from 100€ to 1€ million depending on the country. In addition, DeepFakeChain can offer valuable services to a large number of professions where fake media content can have an adverse effect. Examples include:

- *Public relations and crisis management consultants*, considering the high potential impact of deepfakes on public opinion. For example, in the economic sector, a crisis may result in massive drops of the stock price of a business and negative publicity for their products/services and/or owners/executive boards. In the political sphere, deepfakes can irreparably damage the reputation of political parties and politicians and disorient the public discourse.
- *Insurance companies*, considering the potential use of deepfakes for insurance fraud. In Europe alone, the cost of insurance claim frauds has a current annual cost of EUR13 billion[47].
- *Legal and investigation firms*, considering the use of hard to detect tampered evidence inside the courthouse. Lawyers will need proof of the trustworthiness of the presented evidence and education to understand the implication of deepfake technology in their profession.
- *Social media platforms*, considering the pressure that the upcoming Digital Services Act regulation places on big online platform to address harmful content within tight time constraints. DeepFakeChain can be used for the collaborative annotation of uploaded multimedia content by the staff (or contracted third parties) of these platforms. In addition, DeepFakeChain can be extended to include more general types of harmful content to widen its market scope.

7 PROJECT KEY PERFORMANCE INDICATORS (KPIS)

The tables below summarise the initial KPIs of the DeepFakeChain project covering both Phase 1 & 2 of TruBlo OC3, organised in innovation, scientific, technical, business, and dissemination categories.

[46] http://legaldb.freemedia.at/2017/06/09/trends-in-civil-compensation-for-defamation-in-europe/

[47] Many of Insurance Europe's member associations collect precise data on successfully detected fraud. An example, in Italy, is the IVASS AIA database (https://www.ivass.it/media/avviso/new-phase-aia). Collective number taken from Insurance Europe's white paper: Insurance fraud: not a victimless crime (https://www.insuranceeurope.eu/mediaitem/2bf88e16-0fe2-4476-8512-7492f5007f3c/Insurance%20fraud%20-%20not%20a%20victimless%20crime.pdf)

Funded by the Horizon 2020 Framework Programme of the European Union

TABLE 9: INNOVATION KPIS

ID	Description	Measure	Deadline
1	Development of a collaborative platform for human-machine deepfake media verification.	Implement selected user requirements in the DeepFakeChain's PoC.	M7
2	Trustworthiness of media annotations through blockchain technology.	Integration blockchain storage to DeepFakeChain's PoC.	M9
3	The increase in the perceived trust of multimedia annotations among content reviewers.	> 3.8/5 SUS score by external test reviewers.	M14
4	The increase in the perceived trust of multimedia annotations among viewers.	> 3.8/5 SUS score by external test reviewers.	M14

TABLE 10: SCIENTIFIC KPIS

ID	Description	Measure	Deadline
1	The proposal of a novel blockchain-based trust and reputation scheme with truth discovery features.	Submit 1 scientific paper to a peer-reviewed journal or conference.	M9
2	The extension of DeepFakeChain's AI layer for deepfake detection in Phase 1.	Deploy 3 algorithms (total) for automatic deepfake detection.	M9
3	The publication of a novel blockchain-based trust and reputation scheme with truth discovery features.	Publish 1 scientific paper to a peer-reviewed journal or conference.	M14

Funded by the Horizon 2020 Framework Programme of the European Union

4	The extension of DeepFakeChain's AI layer for deepfake detection in Phase 2.	Deploy 6 algorithms (total) for automatic deepfake detection.	M15

TABLE 11: TECHNICAL KPIS

ID	Description	Measure	Deadline
1	Implementation of all user roles in DeepFakeChain.	Create one user account for each user role.	M6
2	Timely uploading of multimedia in DeepFakeChain.	Upload videos of < 10 mins within 10 secs.	M6
3	Timely search of multimedia via keyword search.	Return videos to query within 5 secs.	M6
4	Timely search of multimedia via reverse search.	Return videos to query within 15 secs.	M6
5	Timely automatic detection of deepfakes by DeepFakeChain.	Check videos of < 10 mins within 20 secs.	M6
6	Integration of blockchain to DeepFakeChain.	Register at least 1 node in the blockchain network.	M7
7	The timely notarisation of annotations in the blockchain storage	Notarise a media's annotations in the platform within 1 sec.	M8
8	Independent auditing of the blockchain storage.	Build an independent interface to the blockchain storage to verify annotations.	M8
9	Internal testing of DeepFakeChain.	Test with >10 internal users by CERTH & Zelus.	M9

Funded by the Horizon 2020 Framework Programme of the European Union

10	External testing of DeepFakeChain.	Test with >20 external users.	M15

TABLE 12: BUSINESS KPIS

ID	Description	Measure	Deadline
1	The positioning of DeepFakeChain in the market.	Present a sustainable business plan in D1.2.	M3
2	The growth of the stored content in DeepFakeChain in Phase 1.	Ingest >100 media files (total).	M9
3	The growth of the stored content in DeepFakeChain in Phase 2.	Ingest >200 media files (total).	M15
4	The growth of the users in the DeepFakeChain platform.	Create user accounts by > 20 annotators and >80 citizens.	M15
5	The application of DeepFakeChain to different markets.	Detail distinct business plans for the identified markets of the Economic Impact section in the DeepFakeChain's WhitePaper (D2.6).	M15

TABLE 13: DISSEMINATION KPIS

ID	Description	Measure	Deadline
1	The dissemination of DeepFakeChain's vision.	Disseminate DeepFakeChain's infographic (MS1) through CERTH's & Zelus' social media.	M2
2	Public engagement in Phase 1.	Present DeepFakeChain in at least 1 public venue (academic open day,	M9

Funded by the Horizon 2020 Framework Programme of the European Union

		newspaper interview, Meetup etc).	
3	Public engagement in Phase 2.	Present DeepFakeChain in at least 1 public venue (academic open day, newspaper interview, Meetup etc).	M15
4	The dissemination of DeepFakeChain's results.	Disseminate DeepFakeChain's White Paper (D2.6) through CERTH's & Zelus' social media.	M15

Funded by the Horizon 2020 Framework Programme of the European Union

8 CONCLUSIONS

In this document, we have described our vision for DeepFakeChain, a collaborative platform for deepfake detection powered by both human evaluators and AI algorithms alike. Based on our review of the market and existing software, we believe that our solution fills an important gap and has the potential to disrupt the practice of deepfake detection. Key to our solution is the capitalisation of blockchain technology to enhance the trustworthiness and transparency of both the media annotation and the decision making process of deepfake detection. Through the openness of the blockchain storage, we also aspire to create valuable metadata and trust scores that can be used by third parties to assess the genuineness of online content, as well as researchers on their work on deepfake technology. This is especially important considering the common interoperability issues of conventional trust management systems, namely, the produced trust scores are relevant only within the considered platform. Finally, we wish to highlight that our platform can form part of a more general solution on collaborative harmful content detection, thus playing a significant role in content moderation, which is a promising market for online platforms.

Funded by the Horizon 2020
Framework Programme of the European Union

REFERENCES

[1] H. H. Khondker, "Role of the new media in the arab spring," *Globalizations*, vol. 8, no. 5, pp. 675–679, 2011.

[2] T. Keipi, M. Näsi, A. Oksanen, and P. Räsänen, *Online hate and harmful content: Cross-national perspectives*. Taylor & Francis, 2016.

[3] E. Kapantai, A. Christopoulou, C. Berberidis, and V. Peristeras, "A systematic literature review on disinformation: Toward a unified taxonomical framework," *New media & society*, vol. 23, no. 5, pp. 1301–1326, 2021.

[4] L. Verdoliva, "Media forensics and deepfakes: an overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

[5] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: a systematic literature review," *IEEE Access*, 2022.

[6] S. Stonbely, *Comparing models of collaborative journalism*. Center for Cooperative Media, Montclair State University, 2017.

[7] J. Isaak and M. J. Hanna, "User data privacy: Facebook, cambridge analytica, and privacy protection," *Computer*, vol. 51, no. 8, pp. 56–59, 2018.

[8] J. Zarrin, H. Wen Phang, L. Babu Saheer, and B. Zarrin, "Blockchain for decentralization of internet: prospects, trends, and challenges," *Cluster Computing*, vol. 24, no. 4, pp. 2841–2866, 2021.

[9] B. K. Mohanta, S. S. Panda, and D. Jena, "An overview of smart contract and use cases in blockchain technology," in *2018 9th international conference on computing, communication and networking technologies (ICCCNT)*. IEEE, 2018, pp. 1–4.

[10] J. Abou Jaoude and R. G. Saade, "Blockchain applications–usage in different domains," *IEEE Access*, vol. 7, pp. 45360–45381, 2019.

[11] V. R. Lesser, "Multiagent systems: An emerging subdiscipline of ai," *ACM Computing Surveys (CSUR)*, vol. 27, no. 3, pp. 340–342, 1995.

[12] D. D. S. Braga, M. Niemann, B. Hellingrath, and F. B. D. L. Neto, "Survey on computational trust and reputation models," *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, pp. 1–40, 2018.

[13] D. Gambetta *et al.*, "Can we trust trust," *Trust: Making and breaking cooperative relations*, vol. 13, no. 1, pp. 213–237, 2000.

[14] K. Sentz and S. Ferson, "Combination of evidence in dempster-shafer theory," 2002.

[15] L. Lamport, R. Shostak, and M. Pease, "The byzantine generals problem," in *Concurrency: the works of leslie lamport*, 2019, pp. 203–226.

[16] E. Bellini, Y. Iraqi, and E. Damiani, "Blockchain-based distributed trust and reputation management systems: A survey," *IEEE Access*, vol. 8, pp. 21127–21151, 2020.

[17] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han, "A survey on truth discovery," *ACM Sigkdd Explorations Newsletter*, vol. 17, no. 2, pp. 1–16, 2016.

Funded by the Horizon 2020 Framework Programme of the European Union

[18] G. Harshvardhan, M. K. Gourisaria, M. Pandey, and S. S. Rautaray, “A comprehensive survey and analysis of generative models in machine learning,” *Computer Science Review*, vol. 38, p. 100285, 2020.

[19] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, “Generative adversarial networks: An overview,” *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[20] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “Deepfakes and beyond: A survey of face manipulation and fake detection,” *Information Fusion*, vol. 64, pp. 131–148, 2020.

[21] S. Baxevanakis, G. Kordopatis-Zilos, P. Galopoulos, L. Apostolidis, K. Levacher, I. Baris Schlicht, D. Teyssou, I. Kompatsiaris, and S. Papadopoulos, “The mever deepfake detection service: Lessons learnt from developing and deploying in the wild,” in *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*, 2022, pp. 59–68.

[22] G. Kordopatis-Zilos, C. Tzelepis, S. Papadopoulos, I. Kompatsiaris, and I. Patras, “Dns: Distill-and-select for efficient and accurate video indexing and retrieval,” *arXiv preprint arXiv:2106.13266*, 2021.

[23] M. van Huijstee, P. van Boheemen, and D. Das, “Tackling deepfakes in european policy,” 2021.

Funded by the Horizon 2020 Framework Programme of the European Union