DOI number: 10.5281/zenodo.5926395

Members of the working group:

Irene Adamski, Paul Ferris, Jessica Fritz, Karl Grün, Reka Hosszu, Tim Clement-Jones, Andreas Kaminski, Caroline Thomas

Chair: Sebastian Hallensleben

Disclaimer

Legal notice

About StandICT.eu

The StandICT.eu 2023 Coordination and Support Action project has received funding from the European Union's Horizon 2020 - Research and Innovation programme - under grant agreement no. 951972. The project is coordinated by Trust-IT Srl (IT), supported by its partners from the **Dublin City University** (IE) and AUSTRALO (ES). The content of the present report does not represent the opinion of the European Union, and the European Union is not responsible for any use that might be made of such content.

Table of Contents

Executive summary 3
Foreword 4
A new challenge for Europe 5
- The new challenge of AI-fabricated content and "people" 5
- Lag in European regulation and standardisation 5
- Capabilities of currently available fabrication technology 6
- Relevance for European standardisation and policy-making 10

Academic perspective: trust in the digital space 12
- Introduction 12
- Testimony, information, and trust 12
- Why does trust play a greater role in modern societies and in the digital space? 14
- The challenges of trust in the digital space 15
- What is a reasonable practice of trust? 15
- Summary: Fundamental strategies to combat misinformation and disinformation 19
- Bibliography for chapter 2 20

Elements of a sustainable solution 21
- Limits of the detection arms race 21

Benefits and necessity of considering identity 22
- Introduction 22
- Concept 22
- Centralised vs. federated vs. decentralized 22
- SSI – Principles, roles and modules 24
- Authentic pseudonymous identities as a solution class 25

Landscape and gap analysis 28
- Standardisation landscape 28
- Gaps and open questions 32

Conclusion and recommended next steps 34

Executive summary

With recent AI tools such as Deepfakes and OpenAI/GPTx it is possible to automate the fabrication of complex digital content on a large scale and to create large numbers of bots that can convincingly mimic human behaviour, e.g. posting product reviews, engaging in political and societal discourse, promoting stocks, acting as influencers and interacting with followers. Such tools could even mimic specific humans, e.g. making a head of state appear to deliver provocative statements or even declare war. The effects are difficult to undo, even if they are later shown to be based on a fabrication.

The consequences are a fundamental challenge to European society and business. While the issue of disinformation is already being addressed on multiple levels (including an EU HLEG in 2017/18 and various content marking standardisation efforts), the additional dimension of an automation of the fabrication of disinformation has not yet been acted on.

Most countermeasures require standards, e.g. standards for tracing information back to its source or creator, standards for bot-resistant pseudonymous identities, standards or protocols for assigning and tracking trust etc.

This report provides an overview of the recent and imminent technological capabilities as well as their impact on democracy, business and legal systems. It explores sustainable countermeasures, with a main conclusion being that the detection tools for fabricated persons and content are not sufficient and need to be supplemented by standards for handling the identity and trustworthiness of sources of information without sacrificing privacy. The report concludes with concrete recommendations to European standardisers, policy makers and other stakeholders.

Foreword

AI technology brings numerous benefits across European society. However, the rapid emergence of certain types of AI tools, their widespread availability and relative ease of deployment at low cost, pose a number of challenges in terms of risks in particular on Online Platform when it comes to Trusted Information.

Specifically, AI tools for fabricating arbitrary media, such as Deepfakes, and arbitrary text, such as OpenAI/GPTx, enable the automated fabrication of complex digital content on a large scale and the creation of large numbers of bots that can convincingly mimic human behaviour, or even specific humans. This can be considered a form of "inauthentic behaviour", and the resulting risks led the Commission to include specific risk assessments related to such manipulation in its proposal for the Digital Service Act and Artificial Intelligence Act regulation[1].

Next to a wide range of innocuous use cases, this new technology enables a potentially huge range of misuse, from posting fake product endorsements, to engaging in political and societal discourse, to creating fake profiles to engage and influence followers, or even emulating a head of state to deliver incendiary and destabilising statements. Even if later shown to be pure fabrication, the damage created by the propagation of false information and the generalised mistrust that this creates is difficult to undo. This poses a fundamental challenge to European society and business.

While many impacts of automation of the fabrication of disinformation have yet to be appreciated, the European Union has been active in addressing the challenges posed by fakes and disinformation for some time, such as the High-level Expert Group on Fake News and Disinformation (2017/18), and the Action Plan against Disinformation, among others.

Regulation and standardisation are intrinsically interlinked and a successful synergy between the two has been emerging in Europe since the 1980's, now well established as the "New Approach" [2]or "New Legislative Framework"[3] which provides structures and processes for standards defined by standardisation organisations while defining a way for the European Commission to issue standardisation requests to SDOs in cases where suitable standards are still needed to underpin upcoming regulation. This is also why in the proposed Regulation on Artificial Intelligence and the proposed Digital Service Act, standardization efforts occupy a prominent place.

We look forward to stimulating further thought leadership around standardisation efforts for Trusted Information outlined in the report and to favouring a structured dialogue between the EC, Member States and Standardisation organisations with the ultimate objective of robustly reinforcing current detection tools with complementary standards for handling the identity and trustworthiness of sources of information without sacrificing privacy.

Prabhat Agarwal

Acting Head of Unit F2 Digital Services & Platforms

DG Connect of the European Commission

1 https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

2 The 'New Approach' (cen.eu)

3 New legislative framework (europa.eu)

A new challenge for Europe

The new challenge of AI-fabricated content and "people"

Since 2018, tools based on artificial intelligence (AI) such as Deep Fake[4],[5],[6] enable practically everyone to fabricate arbitrary video and audio recordings. With sufficient processing power this can even be achieved in real time. The potential impact of such fakes is significant and mpossible to undo, even if they are later shown to be fabrications.

A July 2021 study[7] for the European Parliament presents an up-to-date analysis of Deep Fake technology and its consequences.

Also since 2018/19, convincing „photographs" of non-existent people can be generated automatically and in large quantities[8]. Furthermore, AI tools can now create credible text[9], even in interactive conversation. This technical capability can be used e.g. to flood the internet with fake news and firehose public discourse, to pretend the existence of public opinion and majorities using automated bots, or to promote hate speech or health disinformation to a broad audience..

These new technical tools can be used to automate previous efforts to influence and disrupt democratic processes and constructive political dialogue. Campaigns such as the long-running activities of the Russian „Internet Research Agency" (IRA) [10] [11] or the efforts of Cambridge Analytica[12] will become much easier to conduct and harder to counteract, especially when combined with microtargeting.

In addition to the political impact, the new ease of faking convincing reviews for online activities (e.g. for products, services, content, service providers, users etc.) can be expected to have a negative impact not only on traditional commerce but also the sharing economy and thus a crucial element of a more sustainable economy and society.

Overall, it is potentially becoming harder for individual citizens to distinguish truth from lies, to form their own well-founded opinion and to engage in fact-based constructive dialogue. The necessary critical mass of reasonably well-informed citizens as the foundation for sustainable and stable democratic systems becomes more difficult to sustain. An environment of "alternative facts", wide-spread distrust and cynicism, dominance of extreme positions as well as an inability to reach consensual solutions for pressing societal challenges becomes more likely, providing a more fertile breeding ground for authoritarian and populist leaders and ideas.

Lag in European regulation and standardisation

The European Union has addressed challenges around fakes and disinformation for some time, e.g. with a High-level Expert Group on Fake News and Disinformation[13] in 2017/18, and it has reacted with an Action Plan against Disinformation[14], amongst others. amongst others. However, both analysis and action narrowly predate the broad availability of tools for the automation and thus arbitrary scalability of the generation of disinformation (in particular Deep Fake and GPTx), and therefore could not take into account these more recent challenges. Regulatory solutions are hard to design as they are caught between the fight against fakes on the one hand and the indispensable protection of

4 https://gizmodo.com/researchers-come-out-with-yet-another-unnerving-new-de-1828977488
5 https://www.theregister.co.uk/2018/09/11/ai_fake_videos/
6 https://www.youtube.com/watch?v=gLoI9hAX9dw
7 https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2021)690039
8 https://www.theregister.co.uk/2018/12/14/ai_created_photos/ (man beachte auch das eingebettete Video im Artikel)
9 https://techcrunch.com/2019/02/17/openai-text-generator-dangerous/
10 https://www.nytimes.com/2015/06/07/magazine/the-agency.html
11 https://www.economist.com/briefing/2018/02/22/inside-the-internet-research-agencys-lie-machine
12 https://www.bbc.com/news/av/world-43472347/cambridge-analytica-planted-fake-news
13 https://digital-strategy.ec.europa.eu/en/library/final-report-high-level-expert-group-fake-news-and-online-disinformation
14 https://digital-strategy.ec.europa.eu/en/policies/online-disinformation

freedom of expression on the other, all in the context of the global scale, data resources and speed of large technology platforms. The Digital Markets Act and the Digital Services Act go some way in holding large platforms accountabe for specific kinds of disinformation but do not directly address the automation of fakes and bots.

In standardisation, the situation is similar: So far, very few standardisation activities related to fakes and disinformation have been launched at the European level, and none of them address the challenges of automated and scalable generation of disinformation.

Regulation and standardisation are entwined: Starting in the 1980s, a successful synergy between regulation and standardisation has emerged in Europe and is now well established as the "New Approach"[15] or "New Legislative Framework"[16]. It provides structures and processes for standards defined by three European standardisation organisations (CEN, CENELEC and ETSI, collectively known as ESOs) to gain a special "harmonised" status that regulation can rely on. It also defines a way for the European Commission to issue standardisation requests to ESOs in cases where suitable standards do not yet exist but are needed to underpin upcoming regulation.

With regulation and standardisation being so well connected at the European level, and given the political and regulatory significance of the issues discussed here, this report does not attempt to draw any sharp distinction between a standards development and a policy making perspective.

Capabilities of currently available fabrication technology

This section uses a range of examples to illustrate the capabilities of AI-based fabrication technology. The examples below also show that there are numerous positive use cases for the AI-based generation of content and even "people". As for many other technologies, the underlying tools are dual-use in the sense of not favouring benign over nefarious applications or vice versa.

These faces show "people" that do not exist. They are taken from an ongoing demonstration by AI hardware maker Nvidia at www.thispersondoesnotexist.com . Each reload of the page generates a new face. Most of the generated images are without any artefacts that would indicate that these "people" are not real.

15 https://boss.cen.eu/reference-material/guidancedoc/pages/newapproach/

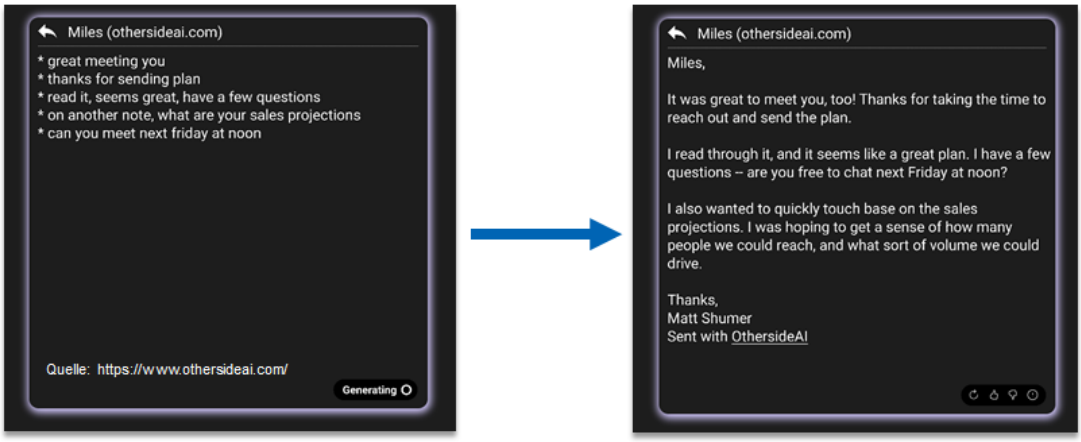
16 https://ec.europa.eu/growth/single-market/goods/new-legislative-framework_en

The tools for generating faces of "people" have been advancing rapidly, as shown e.g. in this sequence of AI-generated images of "people" that, in fact, do not exist[17]:

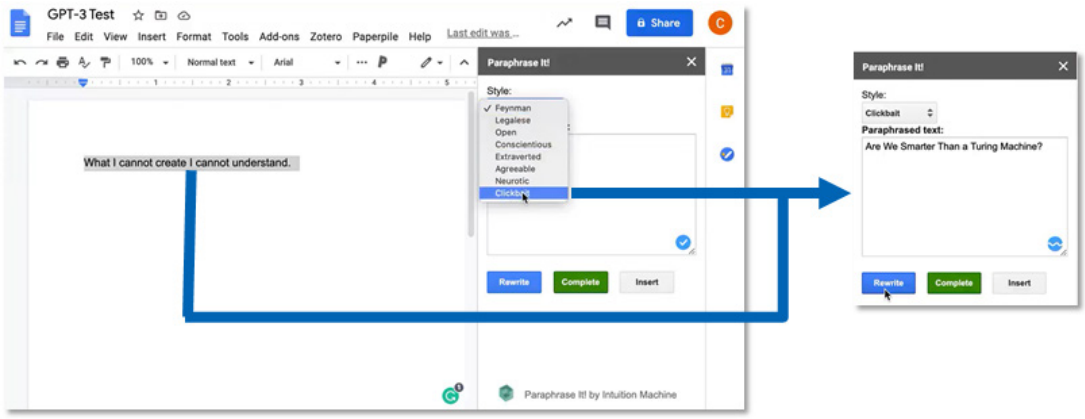
Deep Fake for generating video and associated audio has received substantial media coverage since 2019. Shown here is an example use case where the facial expression of a person (on the left) is imprinted onto another face[18], leading to the possibility of turning anyone effectively into a digital puppet:

Moving from image and video to the generation of text, this demonstration of GPT3 shows how a list of bullet points is automatically turned into a well-formed business email. This would work in the same way for a list of product characteristics being turned into a product review or, with some randomness added, into a whole set of similar product reviews.

17 Source: https://twitter.com/goodfellow_ian/status/1084973596236144640
18 https://youtu.be/qc5P2bvfl44

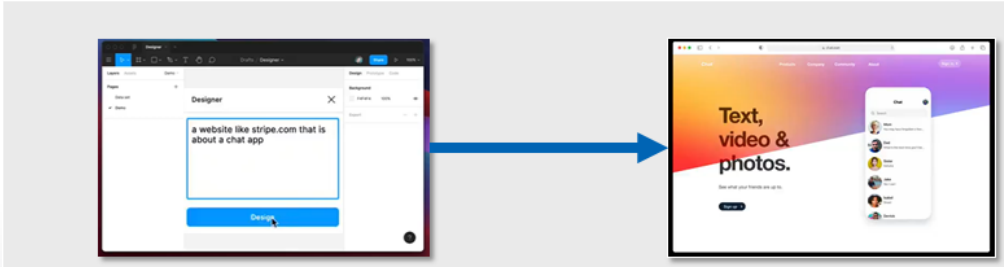


In contrast to its predecessor, GPT2, the current version GPT3 is also able to generate multiple speaking styles, as illustrated here:



https://twitter.com/IntuitMachine/status/1287050253103968257

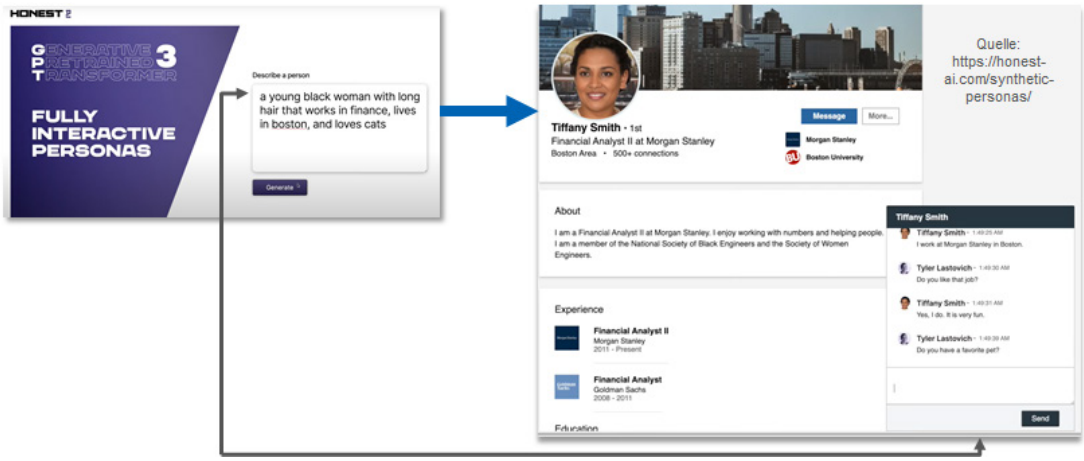
GPT3 is not only able to generate plain text but also markup and even code. Thus, simple prompts like the three examples shown here are sufficient to fabricate convincing websites and documents:



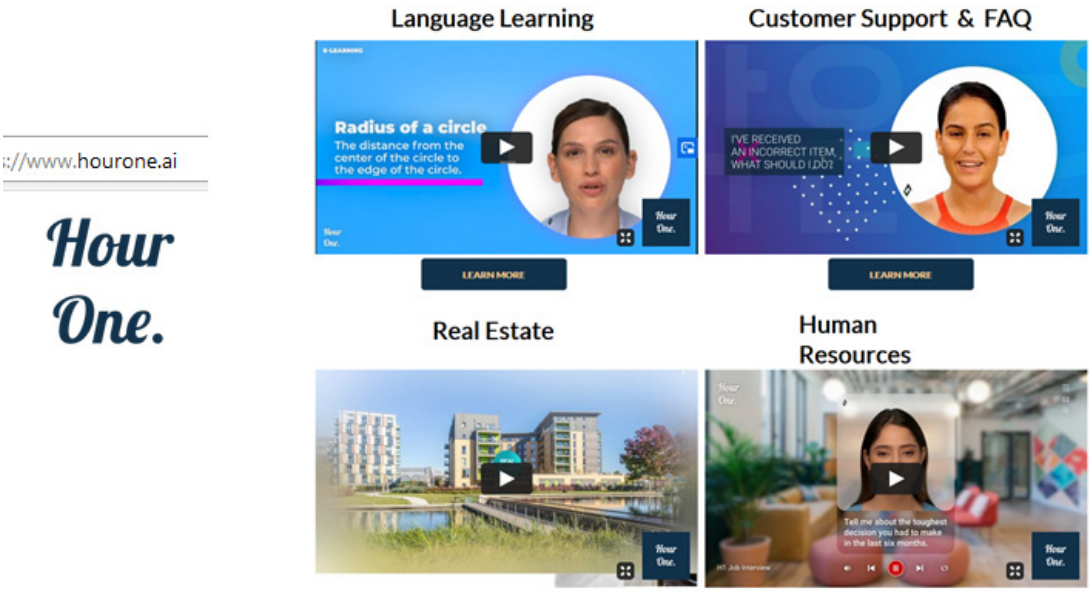
„A diploma from Harvard University in the name of John Doe with a top grade"

„A clone of the website of a rating agency that is modified to advise investing in company X"

Combining some of the above technologies leads to the following demonstration of generating a LinkedIn-like profile of a "person" from just a brief prompt, including basic chat functionality. This can be seen as the tipping point leading from individual fabrication technologies to the fabrication of complete human beings.



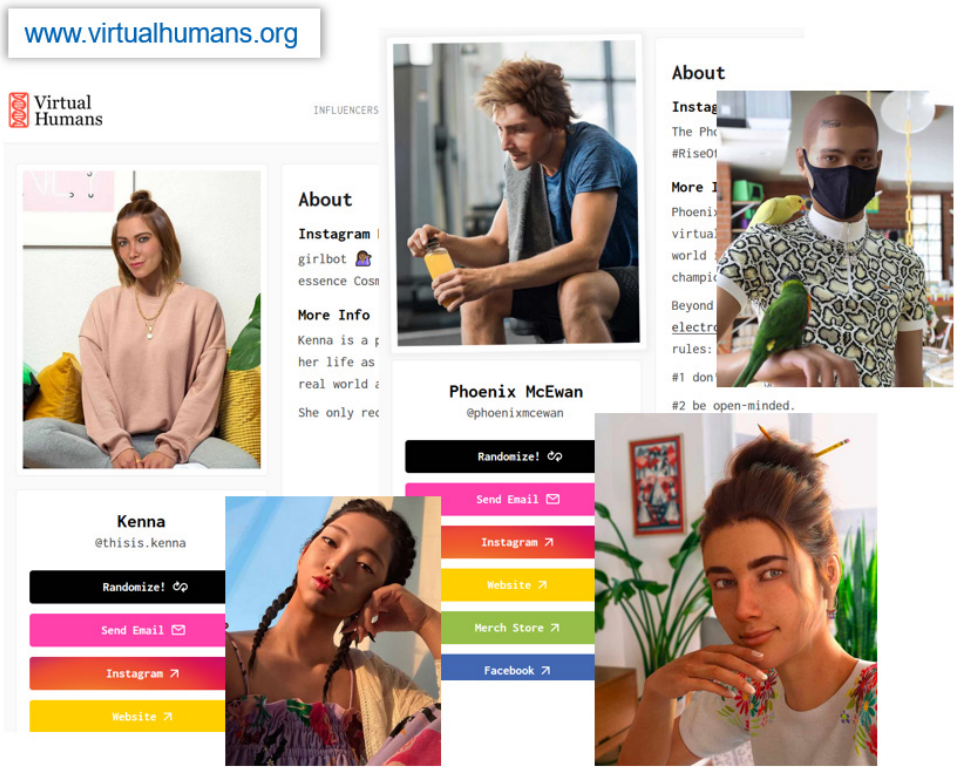
Some of the first commercial applications of AI-based generation of "persons" have been launched by US startup HourOne.ai. Here, uploading 7 minutes of video of a real human is sufficient to generate an AI Clone that is subsequently able to e.g. present arbitrary learning content, provide customer support, offer virtual tours or even conduct recruitment interviews:



Finally, an increasing number of "virtual humans" is being created as influencers with the promise of a scandal-free life and tireless promotion of brands and products. In some cases, their creators have not immediately revealed that the "persons" are, in fact, AI-generated[19].

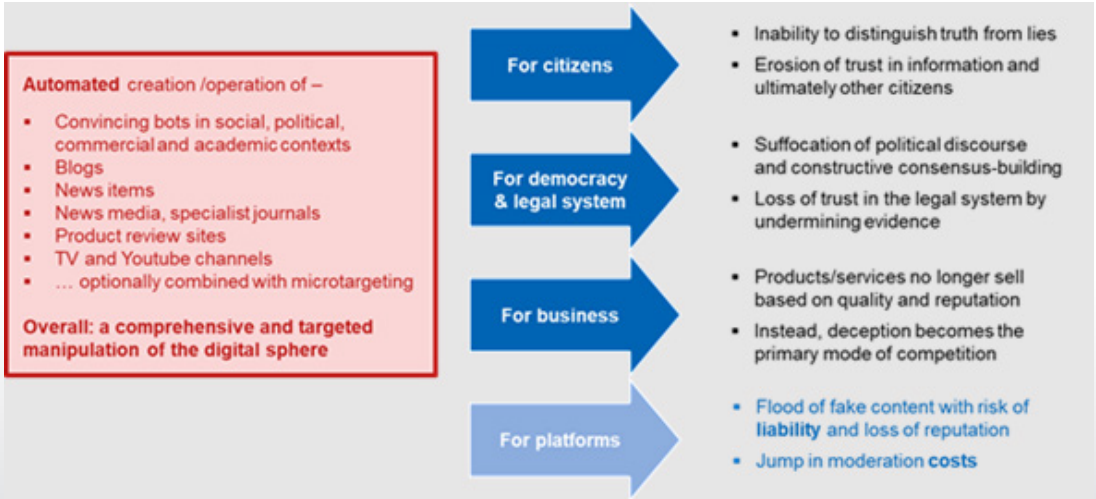
19 https://www.businessinsider.in/tech/news/meet-rozy-south-korean-influencer-who-doesnt-age-and-is-predicted-to-earn-almost-1-million-from-sponsorships-this-year/articleshow/86192294.cms

While audiences have become used to seeing avatars on chat rooms, and understand they are not real people, these can now be replaced by 'persons' without any clarifying disclaimer.



Relevance for European standardisation and policy-making

The following diagram summarises the significant negative consequences of the availability of tools for the mass fabrication of content and "people".



Imagine the current wave of Covid disinformation being further amplified and targeted, leading to an even higher pressure on public health infrastructures in future pandemics as well as an even higher number of avoidable deaths.

It is sometimes argued that, since those consequences have not yet materialised on a broad scale, no action is needed at this point. However, given that the means undeniably exist, and actors with sufficient (commercial and/or political and/or geostrategic) motivation undeniably exist, it does not seem appropriate to spend several years on observation and analysis rather than starting to develop countermeasures right now.

The July 2021 study20 "Tackling deepfakes in European policy", commissioned by STOA for the European Parliament, also reaches the conclusion that action should be considered without further delay, albeit only in the narrower context of deepfakes.

A further consideration is the ongoing European activity on creating a European digital identity based on self-sovereign identity (SSI) technology and the eIDAS standards and regulation. The pseudonymous identity solution proposed later in this document should be considered in the context of these ongoing activities, i.e. now rather than later. This is particularly true since any new ways of handling digital identity require broad changes in citizen/user behaviour which, again, will take time.

As will be explained in more detail further down in this report, most countermeasures to address this fundamental challenge require standards, e.g. standards for tracing information back to its source or creator, standards for bot-resistant pseudonymous identities, standards or protocols for assigning and tracking trust etc. Standardisation alsoe requires time – typically three to four years for the creation of a standard – and therefore should be started now rather than later. In the past, European regulation and standardisation have been the international vanguard in issues such as data protection, leading by example and setting the goal posts on a global level. There is a similar need and potential in the case of automated fabrication of content and "people", too.

In a broader context, policy making and standardisation have a tendency to fulfil their responsibilities in a reactive manner but in this case proactiveness in shaping future developments is highly desirable. The issues raised by automated fakes and "people" require societal discourse and a broad consensus on what digital society in Europe should look like. How and whom do we trust? How can we maintain sovereignty in our capabilities to assess the trustworthiness of information and its sources? What are the key regulatory/standardisation, societal and commercial pillars that we need to develop?

Fundamentally, the question is how real people can trust each other in the digital space despite the availability of fabrication technologies. The following chapter puts this question into the context of academic research on trust and structures possible bases for trust. This forms the foundation for the subsequent discussion of concrete solutions.

20 https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2021)690039

Academic perspective: trust in the digital space

Introduction

In our beliefs, we are fundamentally dependent on what others tell us (climate change, viruses, elections, economy, pensions, AI). This is called "testimonial beliefs". The importance of testimonial beliefs is steadily increasing due to certain trends in modern societies. Consequently, instead of checking information ourselves, we often trust or mistrust information and its sources. Trust as well as mistrust can be justified or unjustified. Therefore, trust can be a reasonable (or unreasonable) practice. Disinformation exploits trust and intends that we base our trust and mistrust on (the) wrong (kind) of reasons. There are different approaches in the theory history of trust as to what a reasonable practice of trust is. The question then is, how can a theory of what characterizes a reasonable practice of trust help us make the real practice of trust more reasonable?

These thought steps explain the structure of the section: The first part of this section addresses the connection between testimony, trust, and information. The second part analyzes why trust plays a greater role in modern societies and in the digital space. The third characterizes the challenges of trust in the digital space, while part four deals with the answers of what constitutes a reasonable practice of trust. In the fifth and last part, we derive from this the solution space for trust in the digital space.

Testimony, information, and trust

How do you know that Japan is a group of islands? That nothing can move faster than light? That there are bacteria and viruses, some of which make us sick? Who won the last election? How do you know when you were born and where? Most of what we are convinced to know, we know through others. (Faulkner 2015; Coady 1992) The situation in which A tells B that something is so and so is called testimony in philosophy. Testimonies are therefore not merely the cases in which a witness testifies in court to what happened at a certain event, but also the many everyday situations in which someone tells us what is the case. Testimonies therefore have a *wide range*: they can be a *trivial* occurrence of everyday life: A tells B, for example, where the next bus leaves. But they can also be of *existential* importance – for example, when A assures B that a vaccine is safe and protects against a disease, or when a person who survived a genocide tries to find the language to give an account of what has happened.

Whether trivial or existential, testimony appears to be a *fundamental* source of our beliefs about the world (and ourselves). We are too limited spatially, temporally, and cognitively with respect to what we can know about the world. We are here in this place and therefore cannot see what is happening there; we are alive now and therefore do not know what occurred before we were born; we know something about this or that area – but there are so many more fields of knowledge to which we are (almost) completely ignorant. To put it positively: We are specialized. Negatively: we know very little from our own experience. In short: We are extremely dependent on what others tell us. True, we can go to a local council meeting to find out on the spot what is being discussed there – and who is advocating for what. But if we do that, we cannot at the same time learn what is being negotiated at the county level. We can study medicine and do a drug study to find out about the effects and side effects of this drug; but then we do not know firsthand how other drugs work. And when our car won't start, we often do not know what is wrong; we then seek out an expert at the garage to give us a diagnosis.

In fact, in most areas, even if we work directly in a field and are knowledgeable about it, we remain largely dependent on what others contribute. No one person can develop a computer from scratch on their own, no one can conduct a climate study on their own. Again, we remain dependent on what others tell us – even if the line between what we experience ourselves and what we know through others is very tangled. [21]

21 Coady explains this with the beautiful example of how presuppositional, for example, our own experience of 'I see the Queen' is. Even when we see her with our own eyes, our own experience is usually based on photographs and descriptions (e.g., how a state visit goes, that bodyguards are near Queen, etc.) that we have experienced

We refer to this fundamental dependence of our *own* beliefs on the beliefs of *others as* "testimonial belief". Two questions emerge with the importance of testimonial belief:

1. Can we *know* anything in this way? That we gain the belief p because others tell us that p is, as I said, a trivial occurrence of everyday life that happens almost non-stop. The question is: Can we know something on this basis? Or have we thereby merely acquired an opinion – at best a supposed knowledge? On the one hand, it seems strange to say that we merely believe that this or that country exists, but we do not know it because we have not been there ourselves; it also seems strange to say, without an appropriate context, that we think that this or that party won the elections. We do not know because we have not counted all the votes ourselves. On the other hand, to say that we know something merely *because* others claim it seems to remove the demarcation between *knowlege* and *opinion*. A classical proposal of how to conceive knowledge as distinct from opinion is: knowledge is a (i) true, (ii) justified (iii) belief. These three features are regarded as minimal conditions. They are *necessary*, though not *sufficient,* for someone to know something. While it should be understandable that knowledge presupposes a belief that is true, it should not be so clear what the second characteristic is about. Why is it essential that the belief be grounded? The answer is simple: someone can have a belief that happens to be true. For example, a person might be convinced that a drug used to treat malaria also provides protection against COVID-19 disease, or that the consequences of current climate change are less than expected – and this might be true, although the person just *happened to be* right. Similarly, we might believe someone who claims such things; but could we therefore also claim to know? Knowledge, therefore, requires more than just being "right" by chance. We must have good reasons for our belief.
2. So, what reasons do we have for our beliefs in general? This question arises even more with regard to testimonial beliefs: If others claim that something is so and so, when is it reasonable to believe them and when not? Others may be mistaken, or they may want to deceive us. We will take up and intensively discuss this question in 1.4.4. Therefore, a few introductory remarks will suffice here. Traditionally, it has been assumed that there are two sources of knowledge: one's own experience and one's own logical reasoning. These sources represent at the same time reasons par excellence to be convinced of something. 'Why do you think that p is the case' can usually be justified *either* by a reference to our experience and that means then sensual perception: 'I have seen it' or 'I have heard it', meaning: myself, with my *own* eyes or ears. *Or* it can be justified by a logical conclusion (here is meant again that we convince ourselves of the logical evidence).

Within the framework of these classical ideas, then, there are two sources of knowledge and thus two ways of justifying our knowledge: our own sensory experience and our own logical reasoning. How does witnessing fit in here? In principle, there are three possibilities:

- Testimony can be reduced to our own experience or our own logical reasoning; thus, we justify whether we believe a witness or not, for instance, by our own experience (this answer is given by the so-called *evidential view*, see 1.4.4.).
- Testimony constitutes a source of knowledge sui generis and it therefore requires its own kind of reasons (this answer is given by the so-called *assurance view*; but not only by it, as we will see).
- Whether we believe testimonies or not is not based on reasons. But this would also mean that we could not know anything that we experience and learn from others – and that would itself be very puzzling. Indeed, it would lead in a more radical skepticism than it might seem. For as we had seen before: it is difficult to draw a line at all and to delimit a sphere of knowledge which we gain purely and exclusively through ourselves.

The answer to the question of reasons is usually seen in research as an answer to what reasons we have for trusting others. In testimony, someone tells us something that we cannot directly verify ourselves. We trust or distrust the speaker and, accordingly, we believe or disbelieve him. Trust in others can be epistemically based, namely, when we *recognize* something about them that makes them trustworthy (or else not); this corresponds to answer (1). Or we believe them for normative or moral reasons (answer (2)). If trust were beyond reasons, this would correspond to answer (3); but then there could be no reasonable practice of trust either.

from others. Cf. Coady 1992, S. 167.

Why does trust play a greater role in modern societies and in the digital space?

Testimony, as we have characterized it above, seem to be either relatively timeless or a phenomenon that plays a greater role rather in pre-modern than modern societies. It appears to be rather *timeless* insofar as humans seem to be quite consistently dependent on what others know. It appears to rather of pre-modern importance, because processes of scientification, mechanization, in short rationalization seem to lead to the fact that we are less dependent on believing others.

It is one of the most important thinkers of the process of modernization and rationalization, namely the German sociologist Max Weber, who holds a different position. According to Weber, trust plays a greater role in modern societies. The reason for this, according to Weber's explanation, is precisely what characterizes modernity: Social differentiation, scientification, mechanization. In pre-modern societies, according to Weber, individuals know the basics of their lives much better from their own experience. The way in which "streetcar or elevator or money or court [...] or medicine" function, however, is beyond the grasp of most of them (Weber 1988, p. 449 – own translation). They have to trust that things will work out reasonably. Weber's contemporary, the sociologist Georg Simmel seconds: In modernity, persons depend to a much greater extent on other persons and systems; at the same time, their insight into them is far less than in pre-modern societies. In modern societies

> *"life stands on a thousand presuppositions which the individual cannot pursue and verify at all to their bottom, but which he must accept in good faith. To a much greater extent than one tends to realize, our modern existence - from the economy, which is increasingly becoming a credit economy, to the scientific enterprise, in which the majority of researchers must use innumerable results of others that are not at all verifiable to them - rests on faith in the honesty of others. We build our most important decisions on a complicated system of ideas, the majority of which presuppose the confidence that we are not deceived." (Simmel 1992, p. 389 – own translation)*

According to these early founding figures of sociology, trust consequently gains a greater relevance in modern societies - this also corresponds to the fact that trust only became a major research topic in the 20th century and, in essence, only towards the end of the 20th century. We can identify at least three reasons why trust has become more significant. All three have to do with our dependence on others:

1. Social differentiation: Modern societies are increasingly differentiated (different systems, division of labor). Thus, it is only possible to a very limited extent for individuals to directly verify the information they receive. This makes us more dependent on what others claim to be true. Consider, for example, medicine or politics. These areas appear to most of us as closed systems on which we are nevertheless dependent in many ways. Similarly, the modern world of work is so specialized that in many areas we are dependent on the willingness of others to cooperate.

2. Complexity of scientific methods: A special case of differentiation is the epistemic division of labor. Scientific methods have become so complex that it is difficult to understand them. Whether in an experimental setup, in a statistical analysis, or with regard to computer simulations and AI models, people are hardly able to comprehend these complex methods themselves. Nevertheless, many decisions are based on what they believe about them. The methodology of a computer simulation may already be comprehensible to many simulation scientists only with a lot of time. Laymen are faced with difficulties that can hardly be mastered. Nevertheless, the results of simulation are relevant to their lives - for example, when it comes to forecasting the utilization of ICU beds during a pandemic or climate change.

3. Information and information technology: Information technology significantly increases the amount of generated information, its communication and reception. Here, a trivial property plays an important role: Information can be sent from one place to another by means of information technology. In our context, however, this means: The number of testimonies increases dramatically. And these are testimonies that we are generally unable to verify ourselves. Increasingly, we receive information about events in other places, at distant times, or in contexts that are difficult to understand. Nevertheless, we are supposed to act on them, for example because our political decisions are affected by them. We read what is happening on the Turkish-Syrian border, which virus variants are appearing in country X, or how batteries used in this country lead to environmental problems in other countries. Information technology decouples situation and information, so that we are hardly in a position to decide on the basis of our own experience whether what we are told is true.

The challenges of trust in the digital space

Information technology does not only create more situations in which to commit crimes. Information technology also poses a challenge to our trustworthiness because it opens up new kinds of possibilities for deception. Here are some examples:

- Imitations: Social bots that pretend to be people; Generative Adversarial Networks can synthetically generate faces, speeches, or actions (deep fakes) which look like real photos or videos.
- Ruses: Incentives (games, free apps) that enable the collection of data (including geofencing) to create user profiles (e.g., for mass personalization); the feint is based on drawing attention to something else.
- Camouflages: Social networks in which the selection of posts is performed by user profiles based on data analytics and adaptive processes, without this preselection being visible.
- Potemkin networks: Posts generated by bots as well as user-adaptively segregated posts are made in social networks that appear to represent social majorities that then appear underrepresented outside the network, resulting in the impression of bias in what is then called mainstream media.

The art of digital deception can have two general purposes. (1) Deception can have the purpose of causing a person or group of people to adopt beliefs that are untrue because this provides a strategic advantage to the deceiving party. However, deception can also be (2) for the purpose of generalized mistrust of information. In the first case we speak of *misinformation*, in the second case of disinformation. Thus, *disinformation* is not successful only when a person or group accepts a certain (false) belief, but already when they 'no longer know what else to believe.'

What is a reasonable practice of trust?

Modern societies thus have a higher need for trust; at the same time, however, the (technological) possibilities of deception are increasing. Not to fall into a modern reflex right away and to look for technical solutions immediately, we first want to develop an orientation, which answers could be given. We therefore ask: What strategies can modern societies develop to deal with this difficult situation? All possible answers lead to the question: What is a reasonable practice of trust? (The technical solution is also *one* answer to that question.)

A reasonable practice is one that is grounded in reasons. Testimony raises the question of when we are justified in believing others. Others may overestimate their cognitive competence, that is, deceive us by deceiving themselves, or deliberately aim to deceive us, especially in the digital space. In other words, for us to be justified in believing the testimonies of others, they must exhibit competence in some area and be of good will. But how to decide when this is the case? This is the trust problem. It is the question of the reasons we have for believing others. When is the practice of trust justified and thus reasonable?

In philosophy, a reasonable practice is classically spoken of when reasons can be given. The question with which we started can therefore now be reformulated:

- How can what we learn through others be justified as our knowledge?
- How can trust be justified? Put another way: When is trust reasonable?

Historically and systematically, there have been three different answers to the question of how to solve the trust problem. Knowing them is important because it promises to expand the sociotechnical solution space. This is because most projects use only the first approach. We briefly review the three attempted solutions to the trust problem, their merits and their shortcomings. The three responses differ in how they conceive of the relationship between reasons and trust and what *role* reasons play.[22]

(A) The epistemic approach

The *epistemic* solution considers the trust problem to be solvable by cognition, namely *one's own* cognition.[23] This means that the reasons for trust can be traced back either to one's own experience or to one's own logical reasoning (Figure 1).

Let us begin with the reduction to our own experience. This solution has been attempted in two ways. First, B can create a track record of A; B then puts the number of true testimonies of A in a ratio to the

22 Cf. Kaminski 2017.

23 Representatives of this theory are the so-called Evidential View of Hume, furthermore especially the rational choice and game theorists. Cf. Hume 1999, 2012; Coleman 1990; Hardin 2002, 2006; Gambetta 1988a, 1988b.

total number of testimonies of A. The ratio indicates how trustworthy A is. For example, B can count how many times A made a promise and how many times A kept it. Here A does not have to be an individual, B could also develop a typology here: Scientists, Politicians or Young People, etc. Second, B can use an indicator that relates trustworthiness to an *observable* sign (called an observable), e.g., A's voice, gestures, specific situation, etc. The frequency with which the observable is observed when A tries to deceive or speaks truthfully indicates the degree of trustworthiness. Which types (if any) or which indicators are reliable is a matter of experience. Only it can show this. That, after all, is the approach.

The problem in both variants is that it is not a *general* applicable solution strategy of the trust problem. This becomes clear when we ask *how B knows which* statements are true. That knowledge is mandatory in both variants. Certainly, B can create a track record for A's promise. But the possibilities to do so are limited. How, for example, can a person know, based on *his own experience* mind you, how often medical professionals or politicians testify true? If anything, the problem is exacerbated when looking at the second option. How often an observed property occurs when testifying true can be researched psychologically – and has been researched psychologically. But even then, it is not a universally applicable strategy. For it to be, all persons would have to become psychologists and, moreover, conduct their research alone.

The epistemic attempt at a solution can therefore only succeed in individual cases, but not in general. It presupposes that B knows in a sufficiently large number of cases on the *basis of his own experience* when A speaks truthfully. B must not fall back on the experiences (evidence) of others, for example C to decide whether A is truthful. Otherwise, there is a circle or infinite regress.[24] B would then have to know when C speaks truthfully and so on. B may therefore judge only on the basis of his own knowledge when another person is truthful. For this to be a generally applicable solution strategy, however, B would have to know not only for A, but for any number of other speakers C, D, E, etc., when they speak truthfully and when they do not. Such comprehensive knowledge cannot be the rule. This is also shown by the fact that if this strategy were generally feasible, it would moreover make testimony superfluous. After all, we are interested in the knowledge of others precisely because it greatly expands the realm of what we can know on our own. The extensive field research that would be required for the epistemic approach would result in witnessing precisely no longer being (as) relevant.[25]

The objection that the epistemic justification strategy is not universally feasible may evoke a sense that there must be something wrong with the criticism of this approach. After all, don't we proceed in the same way? We look at how often a politician or doctor says something – and how often this is true? But this response does not take the problem at hand seriously enough. After all, do we evaluate (a) whether what a politician says is true based on *our* own experience or (b) with the help of *others* such as journalists, neighbors, peers in social networks, etc.? The systematic point is that our own expertise and knowledge is far too limited to evaluate, say, how often the car mechanic who tells us what needs to be done to the car speaks truthfully. We can certainly check this on a point-by-point basis: "He had said the engine needed to be replaced, but it was just a faulty spark plug!" But this point-by-point verification is not sufficient for a track record. A track record actually requires a statistical analysis, i.e., big data.

However, the epistemic approach has other fundamental problems:

1. *The relationship between epistemic reasons and trust:* The epistemic approach leads to the fact that the better epistemic reasons B has that A will not disappoint trust, the higher B's trust in A will be. This gives rise to fundamental paradoxes. For example, behavior that is a breach of trust can increase trust. Let us imagine that B and A are in a relationship. A promises B to be absolutely faithful. However, B wants to increase the epistemic reasons for trusting A, so he secretly reads A's diary. In it, B finds a lot of evidence that A is indeed behaving perfectly faithfully. From an epistemic point of view, this would lead to B now having greater trust in A, although B's behavior is in practical contradiction to trust. Overall, the conceptual relation seems to be misconceived. This is because the more and better epistemic reasons there are for the probable course of a process (the more certain the prediction), the more superfluous trust seems to become.[26]
2. *The form of the relationship*: Another set of problems concerns the form of the relationship that A and B have to each other from the point of view of epistemic theories. In principle, it is a prediction combined with a decision under uncertainty.[27] Therefore, the epistemic approach

24 This is shown in particular by Coady (1992).
25 Coady 1992.
26 Cf. Kaminski 2013.
27 The founder of the rational choice theory of trust, Coleman, also explicitly introduces trust as a special case of a

can be applied to things as well as to persons. We can make a track-record of how reliably our car ignites (number of successful start attempts/number of start attempts); in the same way we grasp the trustworthiness of a friend.

3. *Dealing with disappointed trust:* A final point concerns how to deal with disappointed trust. Epistemic approaches allow only one reaction when A disappoints B: surprise. This is because, in the end, it is a prediction that does not occur. However, the real way of dealing with disappointed trust is also moral disappointment: "Why did you do that? You shouldn't have done that!" This moral disappointment, the accompanying reproaches, claims, the hurt do not find any theoretical space in the epistemic approach - because ultimately B does *not* trust *A at* all in this perspective, but only his *own cognitive faculty*.

As can be seen, the epistemic approach attempts to solve the trust problem entirely on the shoulders of the individual. This does not succeed. The social problem of knowledge can obviously only be solved socially (networks of trust).

The epistemic approach is so important because it is the basis of the vast majority of technical attempts to model and assess trustworthiness.[28] Even if technical modeling works without an explicit reference to a trust theory, they predominantly draw on the assumptions of epistemic theory, for example in connection with psychological research.[29] Track records and indicators are either implemented explicitly (based on psychological theories) or an algorithm is supposed to independently develop a model of trustworthy behavior based on machine learning.

By realizing the epistemic strategy, the technical approach inherits its problem. Thus, a person using such a tool to evaluate the trustworthiness of messages or people must trust the tool. She can check and evaluate the tool by other persons or by means of other tools. However, this results in the circular or infinite regress presented above. The technical approach, however, exacerbates this problem. The boundary between fake news and trustworthy news, for example, is the source of controversy, which itself operates with the distinction between fake news and trustworthy information. That is, the claim that a news item is disinformation may itself be labeled disinformation.

(B) The normative approach

The epistemic approach leads to consequences that contradict trust. A person who is planning a burglary and therefore logs at what times the police patrols may recognize a pattern (every three hours) and then say, quite along the lines of epistemic thought: I trust the police not to come by now.[30]

The *normative* approaches in trust theory, which go by title words such as "assurance view" or "affective trust", have reacted to these and other problematic situations.[31] In particular, they try to correct a fundamental problem: In the epistemic-technical approach, one does not trust the other person at all, but one's own or technical abilities. One ultimately trusts to correctly discern who can be believed. In the normative approaches, on the other hand, a categorically *different reason is* asserted why one trusts another person. This consists not in something one recognizes about him or her, but in the fact that the other person vouches for the truth of what he or she says; he or she assumes a *responsibility* and thus offers a (non-epistemic) assurance. The *value-based* relationship between two is thus ultimately the ground of trust. Thus, the answer to why B believes A when A asserts that p is: because I trust A. From the point of view of epistemic approaches, this is not an answer. We will also see in a moment for good reason this can not be the answer. However, it is first important to understand that there is a change of perspective involved in the approach, which remains incomprehensible as long as we consider and evaluate it from the point of view of epistemic theory. If we go along with this change of perspective, the normative view indeed offers us something that found no place or remained incomprehensible in the epistemic approaches:

- the special importance of the relationship between two.
- That I trust you and not something I recognized in you.
- that you will morally disappoint or even hurt me and not only cognitively surprise me if you break my trust.[32]

decision under uncertainty and uses the model of the bet.

28 Cf. Kaminski 2019a; Lampe und Kaminski 2019.

29 Cf. Kaminski 2019b.

30 This is an example by Bernd Lahno from a workshop. For his critique of game theory cf. Lahno 2002.

31 Cf. Moran 2006; Faulkner 2007; Lagerspetz 1998.

32 The normative approaches often take place under title words such as "affective trust" (as opposed to "predictive trust") or trust as emotion. Since they try to distinguish themselves from rational choice theory, the impression could be created here that it is supposed to be about a contrast between rationality and emotionality. However,

You further take over voluntarily a responsibility for that communicated by you to me. You stand up for the truth.

As good as normative approaches are at capturing the moral violation of disappointed trust, they too have a fundamental problem. It results from one of their merits. In their perspective, one does not have (epistemic) reasons for trust, but trust *is* the reason based on which one accepts another person's testimony. However, this leads into a peculiar dogmatism: B trusts A because B trusts A.

At this point, the normative view knows nothing further to say. That is all that is sufficient to render it speechless, to render it dull. The reason for this is simple: it leaves no room for the fact that there are epistemic reasons for trusting A. The assurance view does an excellent job of teaching us what it means to trust someone (and not merely to predict their behavior). But it can say nothing about how we can *tell* whether a person is trustworthy. The fundamental problem with this approach, then, is that it does not allow us to specify criteria for a person's trustworthiness. This is because, from its point of view, all criteria would take the form of indicators, thus reducing the normative approach in turn to the epistemic one.

It is foreseeable that the two perspectives on trust must be combined. We must *have reasons for* trust and the other person in his relationship to us must *be* this reason. We must be able to respond to disappointed trust in a morally wounded way *and* learn to grant trust appropriately (that is, not naively and blindly, but also not only after a no-cause screening of our counterpart). Simple mediation in the form of a layered model (an epistemic + a normative layer) does not work because the two approaches are mutually exclusive in this present form. But this form of mediation is not the only possible solution, there is an alternative.

(C) The virtue ethics approach

Virtue ethics offers a mediation between the epistemic and normative approach.[33] It starts with the simple consideration of when we are justified in ascribing a virtue to a person. In thinking through the answer, we find that attributing a virtue requires both epistemic and shared normative grounds. A virtue can be justifiably attributed to a person only if there is something about her behavior that justifies attributing the particular virtue to her. Thus, A can be described as just or kind if and only if B recognizes something in the latter's behavior that justifies this description. However, this is not sufficient. Both A and B must see a value (a virtue) in acting in this way. Wanting to be kind or just must be the normative reason for their action, that is, it must motivate and justify for them to act in this way; otherwise, their action only coincidentally bears a resemblance to a kind action.

The proposal is now to consider trust and trustworthiness as virtues. This brings four conditions into view that must be fulfilled.

1. In order to ascribe such a virtue to a person, something must be evident about A's behavior that justifies calling her trustworthy. But - and this is the fundamental difference with epistemic approaches - what she shows is not a casual sign, but an expression of her will and desire to be trustworthy. A, because it is a value for her to be trustworthy, must act trustworthily. In the epistemic approach, for example, a modulation of voice (detected by a technical device) or a microexpression on the face could be an indication of how the person will act. In the virtue ethics perspective, it is not about signs that appear unintentionally on the person, but about the reasons that the person gives me voluntarily (therefore, looking at the diary is not an acceptable reason either).
2. A must also appear so because being trustworthy is a value for A that motivates and directs A's actions. If A does not conceive of being trustworthy as a virtue, then his actions could not be meaningfully described in this way. The point here is not the word, but that A acts in such and such a way for the reason that being trustworthy represents a value for A. We therefore see how both must share a common value.
3. But it must also be a value for B to be trustworthy. Only then can B also attribute this virtue to A. After all, let us imagine that B would not consider it a value to be trustworthy. Then B cannot regard A as virtuous either. Perhaps B would then even be more inclined to describe A as easy to see through, easily predictable, or naive. But for B to consider A's behavior virtuous, trustworthiness must be a virtue for B himself.
4. However, in order for both of them to have a relationship that they understand to be one of trust,

this would mean that the rational choice approaches would throw in their towel on the concept of reason and claim it exclusively for themselves; however, this is not theoretically compelling. Moreover, the choice of the title "affective trust" goes back to another thought:

33 Kaminski (2020) developed such an approach.

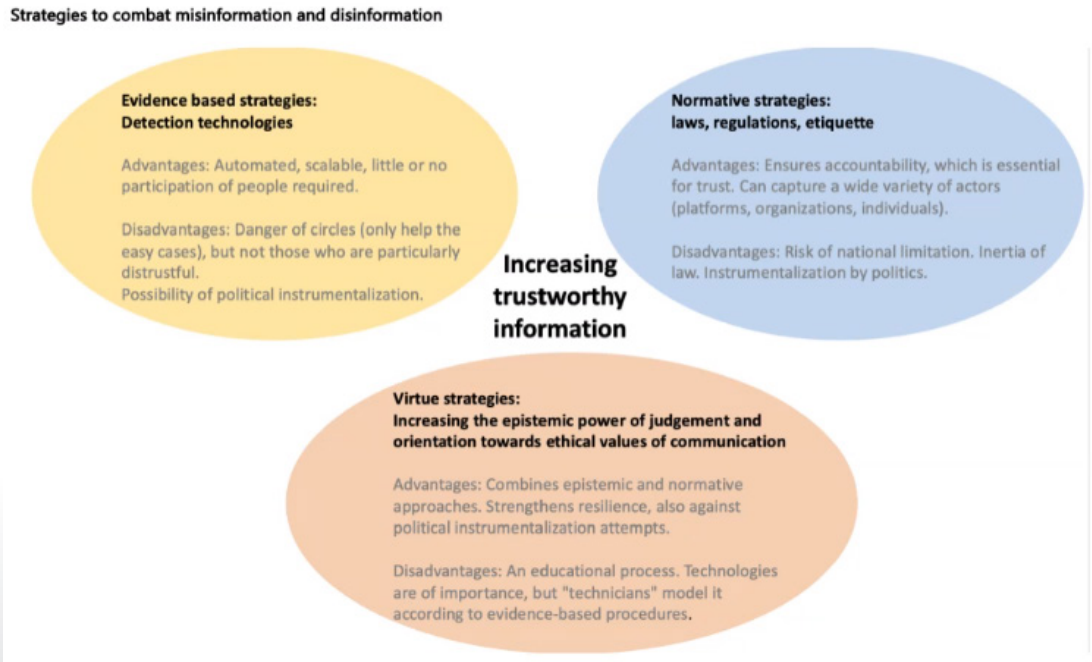
this value must not only be shared by them in general, but they must share it in *their relationship* with each other.

What is gained by this? The first thing to notice is how difficulties we had before dissolve. It no longer causes problems that we have very good epistemic reasons *and* normative claims. Trust is no longer rendered superfluous even by superior epistemic reasons. Furthermore, both forms of response to disappointment are united: we can *learn to* trust appropriately *and* we are morally disappointed when someone does not prove to be trustworthy. Finally, the relationship between the two is now addressed. They share a common value - their relationship and therefore want to be trustworthy for each other.

There is one last gain we can take away from the virtue-theoretical perspective: An answer to when this practice is reasonable. Virtue theory starts from the assumption that virtuous action is appropriate in view of the situation. That is, it is precisely *not virtuous to* trust blindly; or to always be trustworthy, regardless of how others behave. It is virtuous to be trusting or trustworthy when it is appropriate to the situation. This sounds like a circular description. And there is nothing wrong with this idea. If virtuous behavior is *appropriate* with respect to the *situation*, then it depends crucially on the situation. Period. Just as equitable distribution does not always translate into everyone being awarded the same amount. Thus, the community actually plays a more important role than it might seem: it captures what matters in the situation. Virtue is therefore currently also seen as a sensitive knowledge (similar to a perceptual ability to be sensitive to the crucial circumstances of a situation that cannot be listed a priori).

Summary: Fundamental strategies to combat misinformation and disinformation

The diagram below summarises the three strategies described above. In the following chapter, we will build on this analysis to consider the implications for addressing the challenges of AI-automated fakes and bots. This ensures an overarching perspective on the solution space, although, given time and resource constraints of this Technical Working Group, this report cannot be an exhaustive discussion.



Bibliography for chapter 2

Coady, Cecil (1992): Testimony. A philosophical study. Oxford, New York: Clarendon Press; Oxford University Press.

Coleman, James S. (1990): Foundations of Social theory. 1. Aufl. Cambridge, Mass., London: Belknap Press of Harvard Univ. Press.

Faulkner, Paul (2007): On Telling and Trusting. In: Mind 116 (464), S. 875–902. DOI: 10.1093/mind/fzm875.

Faulkner, Paul (2015): Knowledge on Trust. Reprint: Oxford University Press.

Gambetta, Diego (1988a): Can We Trust Trust? In: Diego Gambetta (Hg.): Trust. Making and breaking cooperative relations. New York, N.Y, Oxford: Blackwell, S. 213–237.

Gambetta, Diego (Hg.) (1988b): Trust. Making and breaking cooperative relations. New York, N.Y, Oxford: Blackwell.

Hardin, Russell (2002): Trust and trustworthiness. New York: Russell Sage Foundation (The Russell Sage Foundation series on trust, v. 4).

Hardin, Russell (2006): Trust. Cambridge, U.K: Polity (Key concepts in the social sciences).

Hume, David (1999): An enquiry concerning human understanding. 1. publ. Oxford: Oxford University Press (Oxford philosophical texts).

Hume, David (2012): Eine Untersuchung über die Prinzipien der Moral. Hg. v. Gerhard Streminger. Stuttgart: Reclam (Reclams Universal-Bibliothek, 18993).

Kaminski, Andreas (2013): Die Form vom Vertrauen und ihre verwickelte Praxis. In: Alfred Hirsch, Petar Bojanić und Željko Radinković (Hg.): Vertrauen und Transparenz - für ein neues Europa. Belgrad: Inst. für Philisophie und Gesellschaftstheorie (Conferentia), S. 163–183.

Kaminski, Andreas (2017): Hat Vertrauen Gründe oder ist Vertrauen ein Grund? Eine (dialektische) Tugendtheorie von Vertrauen und Vertrauenswürdigkeit. In: Jens Kertscher und Jan Müller (Hg.): Praxis und 'zweite Natur' - Begründungsfiguren normativer Wirklichkeit in der Diskussion. Münster: Mentis, 121–139.

Kaminski, Andreas (2019a): Begriffe in Modellen. Die Modellierung von Vertrauen in Computersimulation und maschinellem Lernen im Spiegel der Theoriegeschichte von Vertrauen. In: Nicole J. Saam, Michael Resch und Andreas Kaminski (Hg.): Simulieren und Entscheiden. Entscheidungsmodellierung, Modellierungsentscheidungen, Entscheidungsunterstützung. 1. Auflage 2019. Wiesbaden: Springer Fachmedien Wiesbaden GmbH; Springer VS (Sozialwissenschaftliche Simulationen und die Soziologie der Simulation), 167-192.

Kaminski, Andreas (2019b): Gründe geben. Maschinelles Lernen als Problem der Moralfähigkeit von Entscheidungen. In: Klaus Wiegerling, Michael Nerurkar und Christian Wadephul (Hg.): Ethische Herausforderungen von Big-Data. Bielefeld: Springer, 151-174.

Kaminski, Andreas (2020): Die verwickelte Einfachheit von Vertrauen. und seine spekulative Struktur. Marburg: [Habilitationsschrift, in Vorbereitung].

Lagerspetz, Olli (1998): Trust: The Tacit Demand. Dordrecht: Springer Netherlands (1).

Lahno, Bernd (2002): Der Begriff des Vertrauens. Paderborn: Mentis.

Lampe, Hildrun; Kaminski, Andreas (2019): Verlässlichkeit und Vertrauenswürdigkeit von Computersimulationen. In: Kevin Liggieri und Oliver Müller (Hg.): Mensch-Maschine-Interaktion. Handbuch zu Geschichte – Kultur – Ethik. 1. Auflage 2019. Stuttgart, [in print].

Moran, Richard (2006): Getting Told and Being Believed. In: Jennifer Lackey und Ernest Sosa (Hg.): The Epistemology of Testimony: Oxford University Press, S. 272–306.

Simmel, Georg (1992): Soziologie. Untersuchungen über die Formen der Vergesellschaftung. 1. Aufl. Frankfurt am Main: Suhrkamp (Suhrkamp-Taschenbuch Wissenschaft, 811).

Weber, Max (1988): Ueber einige Kategorien der verstehenden Soziologie. In: Max Weber: Gesammelte Aufsätze zur Wissenschaftslehre. 7. Aufl. Hg. v. Johannes Winckelmann. Tübingen: Mohr (UTB, 1492), S. 427–474.

Elements of a sustainable solution

Limits of the detection arms race

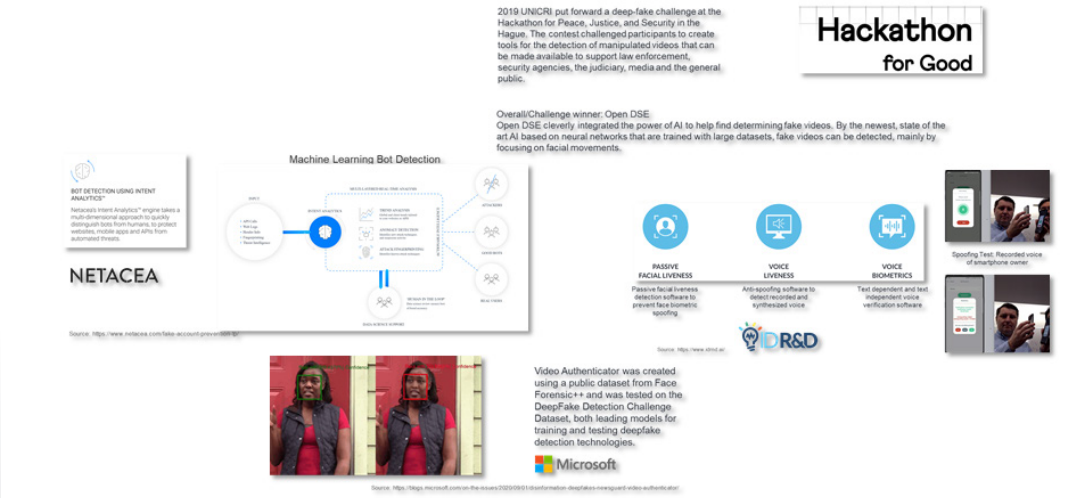
Most existing measures against fakes and bots such as fact-checking services, AI-based analysis of images or the detection of activity patterns inside large social networks find themselves in an unwinnable arms race against better and better fabrication tools. The very nature of AI plays into this arms race, on the one hand it can be trained to get better and better at detecting fakes, but at the same time AI will also learn to be able to better blur the telltale clues in the creation itself. At present, AI-automated bots are particularly good at imitating quick, fleeting human interaction but struggle to mimic deeper, more sustained interaction – but this rarely happens in a fast-paced digital life today. Moreover, we can expect to see such bots improve to become capable of longer-term interaction (even grooming) as well.

The numerous efforts in detecting fakes and bots, both by established corporations and recent startups as well as by non-profit organisations (cf. the examples in the illustration below and the comprehensive list here[34]) are definitely valuable and worthwhile but can only play a supporting role for a broader solution. In some cases, especially fact-checking, detection methods also require considerable resources, and hence they do not scale well when faced with large volumes of AI-automated fakes.

Similar limitations apply to a related approach that is often presented as a countermeasure against fakes and bots: education. Internet users are admonished to "check their sources". Banks are providing their customers with information and official warnings to train them to spot fraud. Policy makers call for better digital education in schools etc. Such approaches shift the burden of responsibility to individuals who have neither the resources nor the time nor necessarily the will to spot fakes and bots, especially when they encounter them on a massive scale and when sharing them to al their friends is only one click away.

A comprehensive discussion of the strengths and weaknesses of specific tools or methods to detect fakes and bots is beyond the scope of this document. For the present purposes, the important message is the fact that they are valuable and necessary but not at all sufficient to address the challenges of their AI-based automation on a massive scale

Within the three approaches to establishing trust described in chapter 2, all such detection efforts fall firmly into the first, "epistemic", category and thus ignore the potential of the two other categories, i.e. "normative" and "virtue ethics".



34 https://www.rand.org/research/projects/truth-decay/fighting-disinformation/search.html

Benefits and necessity of considering identity

Introduction

To address the full solution space, all three approaches to trust described in chapter 2 (epistemic, normative, virtue ethics) need to be considered. While the three approaches are quite distinct they do have one important commonality: the need to establish the identities of the persons involved. There can be no trust without identity (although, as we will see, this does not necessarily involve a person's real name).

Consequently, we need to turn our attention to a discussion of digital identity which will then lead us to a specific type, namely authentic pseudonymous identities, as the foundation for sustainable countermeasures against AI-automated fakes and bots.

Concept

The concept of a 'digital identity' aims to solve the puzzle of enabling trust between two essentially unknown entities. Humans are not able to verify whether they are interacting with a person or a machine, and whether the person or machine is trustworthy. As a remedy, we exchange tokens of trust to establish a relationship before the actually desired interaction or transaction occurs. This usually happens vie the presentation of data that represents identity claims, credentials or attributes – on a simple level this can be the knowledge of an accounts name and password, with 2-factor authentication both log-in information as well as device-access is demonstrated, while the most complex processes require knowledge of information, access to devices, as well as proof of prior interaction with a trusted authority (e.g. access files received from a tax authority, an employer or a client to access sensitive information).

In essence, a digital identity creates a trail of verifiable information that links our identity back to a trusted source of information. In the progressing Digital Age the question(s) are and continue to be:

- How transparent and/or verifiable is this trail of information?
- How secure, durable and authentic is the link between identity information and its source?
- How reliable and trustworthy is the source of the information?

There are different technological and governance approaches to solve these questions. They are commonly categorized in three groups: centralized identity systems, federated identity systems or decentralized identity systems. The history of the current century has shown that the first two options have reliable traits and benefits but also proven and reoccurring flaws and drawbacks. The decentralized identity approach has been discussed in various ways in theory, but only in the past decade has the technology matured to a point where digital decentralized identity systems can be implemented.

Centralised vs. federated vs. decentralized

As with all digital processes the major questions in digital identity systems revolve around the handling of data. Data must be acquired, curated, made accessible and guarded at the same time. The process of acquiring and curating data (verifying the validity of information, updating when necessary, categorizing, etc.) is time and resource intensive process – and as such lends itself to the creation of mono- or oligopolies. This is what we have seen in the beginning of the Internet Age. Centralized systems centralize the cost of data acquisition and curation under one roof, for example via in-house IT-departments in large organisations. This has however led to the creation of data silos and made it difficult to share data outside of the silos. This approach also creates large 'honey pots' for malevolent actors, thereby increasing the cost of securing data. At the same time, the lack of data

portability decreases efficiency and effectiveness in digital solutions due to the fragmentation of data among third parties. The limited interoperability further reduces innovation.

Outsourcing issues of accessibility, interoperability and adherent questions of convenience to new service providers has led to the advent of the federated identity system. With this approach so-called Identity Providers take over the tasks of data curation, accessibility and protection, while data acquisition is handled by users and organisations via subscription. The benefit is dilution of the 'honey pot' and thus the threat of being targeted. It also allows for specialisation of service providers on aspects of usability and curation efficiency. The drawback is that such an oligopoly of providers (e.g. Applie-ID, Facebook log-in, Google account log-in) and particularly the business models connected to them incentivise profit maximisation via data maximisation and cost efficiency via bulk processing of data. This severely reduces the ability of the individual to make individual choices about the handling of their data and reduces control and sovereignty of the identity subject. A 'one size fits all' approach also greatly increases the risk of successful hacks and information going into unauthorised hands.

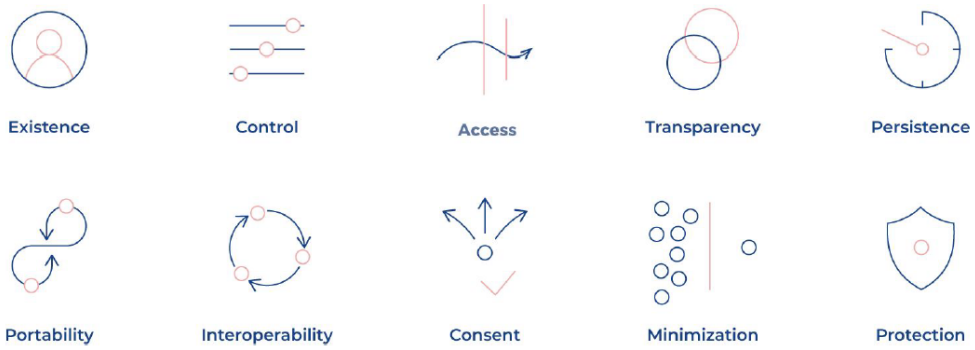
In summary, in centralised identity systems organisations can benefit from tailored data acquisition but bear the cost of data curation and protection. Data accessibility is severely limited. In federated identity systems, organisations and users can benefit from more efficient data curation and data accessibility but pay for this with bulk data acquisition and the ensuing risk of data theft. In neither scenario does the individual have sovereign control over how their data is being handled.

With regards to the questions of trusted information raised above, the centralized approach offers easy answers to all three questions but is limited in providing trusted information services outside the specific silo. Also, once the system is compromised the trustworthiness of all information is in jeopardy. A federated identity system can provide secure links between points of information but can have gaps in the transparency of the information trail, making it difficult for third parties to verify the trustworthiness of the information and its source. One has to trust the intermediary first and foremost. This provides for user convenience and puts a large burden on the service provider to handle the verification of information and correctly assess the trustworthiness of a data source. It also creates existential dependencies and power imbalances of otherwise sovereign identity subjects within the digital space.

The alternative to the centralised or federated systems is a decentralized identity approach. In this option the resource intensive cost of data acquisition and curation is diminished via individualisation. If an identity subject is enabled to curate and share their data sets by themselves across secure channels, then the number of tasks dependent on third parties diminishes to such a degree that the balance of power between service providers and users reaches a new equilibrium. The most common name for this is that of Self-sovereign (digital) identity.

SSI – Principles, roles and modules[35]

The term was coined in 2016 by Christopher Allen[36], who identified ten criteria that should be fulfilled by a sustainable identity system that returned control over private data to the individual. These criteria focus on the quality of identity data (existence, persistence, transparency), respect for the sensitivity of identity data (minimisation, protection, consent) and a sovereign handling of personal identity data (control, access, portability, interoperability).



Design by Iryna Nezhynska from Jolocom GmbH based on a concept by Christopher Allen

Since then, the concept of SSI has seen a lot of support and growth. There is a global community of developers, political proponents and legal or civil society experts working on further maturing the technology, standardising the existing work and adjusting the concept to fit real world applications.

While the technical standardisation work is ongoing and can best be monitored in specialised or technical fora (e.g. W3C, DIF, ToIP), the central tenets of the SSI concept are a division of work among three separate roles: an Identity Holder, an Identity Issuer and an Identity Verifier.

- **Identity Holder** – is the subject engaging in or initiating the process of identification or authentication. The existence of identity data is centred on them. They can selectively choose to share or receive identifying information, and where and how this information is stored.
- **Identity Issuer** – is a trusted source of information, often an authority that can attest specific identity data (e.g. a public administrative organisation, a private business, an institute of education). They must be able to provide an identity claim, credential or attribute to an identity holder upon request, and also able to revoke such data if necessary.
- **Identity Verifier** – is a third party that provides the service of confirming whether a presented identity information is true and reliable. The provision of the verification service enables trusted interactions between two unknown actors within the digital space.

The innovation of this approach lies in the fact that only the Identity Holder actually holds all the information. An Identity Issuer will only know what information was issued to whom, but not how or then this information is being used. An Identity Verifier will only know whether data provided is untampered and valid, thereby classifying an Identity Holder as 'trustworthy' or 'not trustworthy' without needing full access to the full information that might be connected to a data set. The classic example is verifying whether a potential buyer is of age within the jurisdiction of an online shop. For this step, all that the seller needs to know is whether the buyer on the other side has reached of surpassed a certain age. The exact number of years, date of birth, place of birth, current nationality etc. is irrelevant for this interaction. In order to establish trust within an SSI model, the buyer (as the Identity Holder) will only have to offer the identity attribute 'of age' to the seller. The seller can then verify the validity of this attribute via an Identity Verifier. No data aside from this needs to be shared, meaning the buyer can be assured that their private information is secure and the seller is spared data protection measures and the cost of proper data storage.

Technical components or modules necessary for such interactions are so-called Identity Wallets, often smart phone applications, that store our identity information decentralised on privately owned

35 Further Reading: Allen, Christopher. Life with Allacrity.com. (2016) http://www.lifewithalacrity.com/2016/04/the-path-to-self-soverereign-identity.html
Decentralized Identity Foundation. INATBA. WG Identity. (2020) "Position Paper on SSI"
W3C. Verifiable Credentials WG.

36 Allen, Christopher. http://www.lifewithalacrity.com/2016/04/the-path-to-self-soverereign-identity.html

devices. The pandemic situation has provided a lot people around the globe with a rudimentary example of this, where test results or vaccination proofs are stored locally on a smartphone app to be presented upon request. On the software level, all three roles (Holder, Issuer, Verifier) interact via agents based on role-specific technology stacks and their respective SDKs. On the legal level, the interactions between all three roles are governed via trustframeworks that need to be provided either directly by legal authorities (i.e. governments) or indirectly by subject-matter authorities (e.g. FINDY, ToIP Foundation) authorised by responsible legal authorities and with appropriate privacy and consent mechanisms.

Authentic pseudonymous identities as a solution class

Authentic pseudonyms as the missing quadrant

The analysis so far can be summarised in the quadrant diagram below. One might argue that the top left quadrant is already a solution to the challenge of AI-automated fakes and bots: An individual can verify their real name using traditional documentation (as is requested e.g. by Twitter to obtain the "Verified" tick) or SSI, and this of course proves that they are human. In terms of trustworthiness, it is cleary a step up from the bottom left quadrant, i.e. the use of unverified names that could be associated with a real human being but might just as well denote a bot that just pretends to be human.

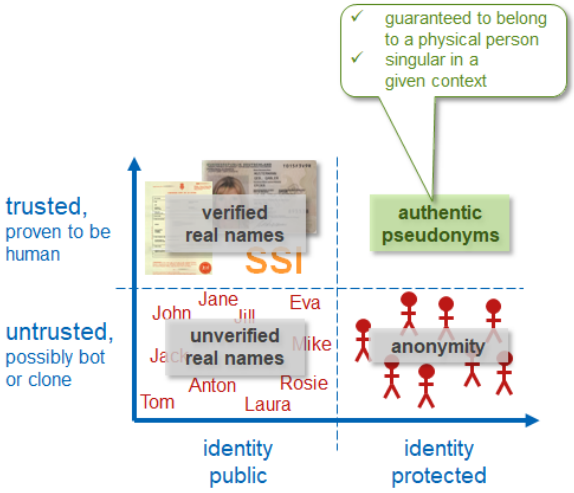
In both of the quadrants on the left, one's real-name identity becomes public. This can be problematic and undesirable, though. There are numerous situations where revealing one's real name should be avoided, e.g. in a whistleblowing event, in the context of authoritarian regimes, in sensitive commercial contexts, or in cases where people want or need to keep professional and private opinions separate. Furthermore, anonymity can be regarded as a psychological requirement for sustainable social interactions and free societies.

Of course anonymity in the digital space is possible today, at least to a large degree. However, even if successful, anonymity keeps an individual in the bottom half of this diagram where they compete for attention with bots. They have a voice, and indeed a protected voice, but might be drowned out by bots.

What we need is the top right quadrant for which we have coined the term "authentic pseudonyms". These are digital identities that -

- are guaranteed to belong to a physical person but are not traceable to a specific person
- are singular in a given context, i.e. it is possible to have different authentic pseudonyms on different platforms or for a, but within the same platform a person can only have one such pseudonym.

AI-automated bots are unable to acquire authentic pseudonyms[37] whereas human beings can. Thus, the digital space can be reclaimed, and a fundamental requirement for trust is safeguarded.



37 At least not in a significant number of cases. Credentials can be stolen and abused but this is not sufficient to supply mass-fabricated bots.

Note that the concept of authentic pseudonyms as presented above allows people to assume different personas in different contexts given that such pseudonyms can be different on each platform. It needs to be impossible (ideally at a cryptographic level) to find out that two pseudonyms on two platforms belong to the same person since such an aggregation of behavioural observations would make it too easy to identify the real person behind the pseudonyms or at least facilitate undesirable profiling.

Relationship between pseudonymous identities and accountability

Authentic pseudonyms provide a crucial side benefit that is hard to realise with conventional usernames: Suppose a person uses their account on a social media platform to call for mass murder o. Such a person is likely to be see their account blocked fairly quickly. However, they can then easily create a new account and call for mass murder again, get blocked again etc.

With accounts backed by authentic pseudonyms, this cat-and-mouse games no longer happens. The person can still call for mass murder for the first time which will lead to a block. However, they then cannot establish a new account since they are not able to generate a different authentic pseudonym for this platform. Thus, sustainable blocking becomes possible which helps platforms fulfil their regulatory obligations far more easily. This approach also strikes a balance between freedom of expression of one hand, and enforcing bans against hate speech on the other hand.

Possible implementations of "authentic pseudonyms"

There are several possible ways of implementing "authentic pseudonyms", each with different characteristics as well as strengths and weaknesses. They are outlined in the table below:

Implementation	Description/ Characteristics	Strengths	Weaknesses
TR-03110 "Restricted Identification"	Optional feature in the eIDAS standard: An ID card is able to calculate on demand a pseudonym token specific to both the person and the requesting entity (e.g. a platform). It is cryptographically impossible to link the token back to the person. The card always provides the same token for the same inputs.	Capability already implemented in ID cards in Germany (since 2011) No need for the person to trust any specific organisation (except for the card issuer) Mature cryptography Already standardised Same card for real-name and pseudonymous identification	Capability not yet implemented in ID cards of countries other than Germany User acceptance for using ID cards to access digital services still low in many EU countries High barrier of entry for developers
Pseudonyms on SSI[38]	Extend SSI to cover the placement of pseudonym tokens in SSI wallets, calculated from existing content of the wallet (typically real-name identity).	Compatible with plans for the SSI-based European Digital Identity Word-wide rollout possible, including in countries without trustworthy insitutions	Currently multiple competing approaches and implementations of SSI; standardisation landscape immature Trusted third party required in some cases

38 The trust framework determines the criteria that deem a source of information reliable and trustworthy, that mark the quality of an Identity Issuer. Questions about the secure link between information and source are answered by standardisation efforts within the SSI tech stacks. Efforts are on-going in different standardisation bodies and communities to ensure state-of-the-art, secure and interoperable solutions (e.g. Verifiable Credential WG at the W3C). Finally, current efforts are focused on open-standards and, to a comparably large part, open source solutions, enabling a transparent and easily auditable trail of information between identity subject and

Implementation	Description/ Characteristics	Strengths	Weaknesses
	Depending on the SSI implementation, the calculation might need to be performed by a trusted third party[39].	No need to rely on physical token Compatible with distributed and federated approaches to establishing identity	Meaningful practical rollout likely to take several years unless accelerated by institutional recognition
Commercial IDs plus identity trustee	Large corporations such as Apple, Google, Facebook and Amazon are able to establish real-name identities with a high degree of certainty. Pseudonym tokens can be derived from these by an identity trustee as an intermediary who is trusted by both the corporation and the individual.	Worldwide huge market penetration Convenience likely due to the powerful user interface development capability of corporations	Not compatible with European drive towards digital sovereignty and European data protection policies Unclear accountability in the absence of jurisdictional clarity and of a paying customer relationship Not compatible with data- and advertising driven business model of large corporations Non-trivial role of identity trustees; licencing/vetting infrastructure needed Standardisation likely to be extremely hard due to commercial interests of competing big players

the source of its digital identity markers. - Based on this foundation, SSI can provide the underlying technological approach to enable data and identity verification without diminishing individual control in private interactions. In order to enable such an identity system on a broader scale, it will take a high quality supply of SSI solutions, which will require efforts on the part of the tech community with regards to standardisation and on the side of law makers and civil society where trust frameworks are concerned.

39 Remote data assessment might provide an approach that avoids any direct private data sharing in which specific access to a one way algorithm is granted. The algorithm can then demonstrate a trustable match (which can include an assessment of certainty, for instance) between private data sets (for instance, between an individual and a verified source) without sharing the data set itself.

Landscape and gap analysis

Standardisation landscape

This section provides an overview of current standardisation activities as well as existing standards that have relevance to the discussion in the earlier chapters. The overview is primarily grouped by committee.

ISO/IEC JTC1

ISO/IEC JTC 1/SC 27, *Information security, cybersecurity and privacy protection,* is responsible for international IT security. The most relevant standards to electronic identification and trust services are developed by SC 27/WG 5 *Identity Management and Privacy Technologies*. After completion of foundational frameworks, specifically, the ISO/IEC 24760 series *A framework for identity management* and ISO/IEC 29100 for *Privacy framework*, priorities for WG 5 are related standards and Standing Documents on supporting technologies, models, and methodologies. WG 5's Projects include:

- A framework for identity management – Part 1: Terminology and concepts (ISO/IEC 24760-1, 2nd edition:2019)
- A framework for identity management – Part 2: Reference framework and requirements (ISO/IEC 24760-2, 1st edition:2015)
- A framework for identity management – Part 3: Reference framework and requirements (ISO/IEC 24760-3, 1st edition:2016)
- Privacy framework (ISO/IEC 29100, 1st edition:2011; Amendment 1:2018)
- Privacy architecture framework (ISO/IEC 29101, 2nd edition:2018)
- A framework for access management (ISO/IEC 29146, 1st edition:2016)
- Requirements for partially anonymous, partially unlinkable authentication (ISO/IEC 29191, 1st edition:2012)
- Privacy enhancing data de-identification terminology and classification of techniques (ISO/IEC 20889, 1st edition:2018)
- Privacy impact assessment – methodology (ISO/IEC 29134, 1st edition:2017)
- Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy management – Requirements and guidelines (ISO/IEC 27701, 1st edition:2019)
- WG 5 Standing Document 2 – "Privacy references list"
- WG 5 Standing Document 4 – "Standards Privacy Assessment"

ISO/IEC JTC 1 SC 27 is working in close collaboration with CEN/CLC/JTC 13 *'Cybersecurity and Data protection'* on eIDAS related standardization activity.

ISO/IEC JTC 1 SC 42 has a brief reference to fake news detection in the context of advanced language models in TR 24030, Information technology — Artificial intelligence (AI) — Use cases (https://standards.iso.org/iso-iec/tr/24030/ed-1/en/Use+cases-v05_electronic_attachment_022021.pdf)

ISO

Several Technical reports are in progress in ISO TC 307:

- TR 23644 Trust Anchors (in progress) https://www.iso.org/standard/81773.html?browse=tc
- TR 23244 Blockchain and distributed ledger technologies - Privacy and personally identifiable information protection considerations considers different levels of privacy and their mapping to blockchain environments https://www.iso.org/standard/75061.html?browse=tc
- DTR 3242 Blockchain and distributed ledger technologies - Use cases https://www.iso.org/standard/79543.html?browse=tc
- DTR 23249 Blockchain and distributed ledger technologies - Overview of existing DLT systems for identity management https://www.iso.org/standard/80805.html?browse=tc (reference rather than analysis document)

Published standards on privacy: https://www.iso.org/committee/6266604/x/catalogue/p/1/u/0/w/0/d/0

There is also early work on pseudonymity/multiple identity use cases, e.g. as a way of avoiding the direct use of biometrics, as a way of balancing police powers (e.g. when monitoring sporting events), or for covid test&trace

Further, there is a working group on identity aspects of blockchain and on SSI with several Technical Reports in preparation.

ISO/TC 307/JWG04 Road Map on Security standards in DLT/Blockchain is progressing in that the interest among the group's experts is intense. Fundamental work needs to be tackled by this group such as measuring how secure a DLT/Blockchain system is that is compatible with the existing standards for system security measurement in ISO/IEC 27001 and 27002 and common criteria agreed that is comparable with ISO 15408. This fundamental work has not yet been started.

ISO/TC 307 & TC 68 have formed a **Joint Advisory Group on Digital Currencies.** The work will rely on Trust, Identity, Privacy and Security foundational documents as it encompasses the reality that unlike paper cash, digital cash will be traceable and directly taxable, among many other effects. The present report is expected to become one of the foundational documents.

CEN-CENELEC

CEN/CLC/JTC 19/WG 1 - Decentralised Identity Management based on Blockchain and other Distributed Ledgers Technologies has been formed. A New Work Item Proposal has gone to Ballot 'Decentralised Identity Management Model based on Blockchain and other Distributed Ledgers Technologies' . If approved, work is likely to start in March 2022

ITU

ITU-T SG17 is responsible for the study and coordination of the work on security and identity management. It has approved the following recommendations:

- ITU-T X.1058 "Information technology - Security techniques - Code of practice for Personally Identifiable Information protection",
- ITU-T X.1148 "Framework of de-identification process for telecommunication service providers",
- ITU-T X.1212 "Design considerations for improved end-user perception of trustworthiness indicators",
- ITU-T X.1250 "Baseline capabilities for enhanced global identity management and interoperability",
- ITU-T X.1252 "Baseline identity management terms and definitions",
- ITU-T X.1403 "Security considerations for using distributed ledger technology data in identity management",
- ITU-T X.1451 "Risk identification to optimize authentication",
- ITU-T X.1363 "Technical framework of personally identifiable information (PII) handling system in IoT environment" and is developing six draft Recommendation in this domain: (X.5Gsec-t, X.sec-QKDN-tn, X.smsrc, X.scpa, X.sgos, X.rdda). More info: http://itu.int/ITU-T/go/tsg17

OASIS

The OASIS Security Services (SAML) TC maintains and extends the widely used Security Assertion Markup Language (SAML, also ITU-T Recommendation X.1141) standard. A profile of SAML is used for cross-border identification and authentication of citizens in the eIDAS nodes provided by the eID Building Block of the Connecting Europe Facility (CEF). SAML is also used at national level in Member States.

The OASIS Trust Elevation TC defines a set of standardized protocols that service providers may use to elevate the trust in an electronic identity credential presented to them for authentication.

The OASIS DSS-X TC defines standard Digital Signature Service Core Protocols, Elements, and Bindings. The latest version provides both JSON- and XML-based request/response protocols for signing and verifying, including updated timestamp formats, transport and security bindings and metadata discovery methods. This TC works in close liaison with the ETSI Electronic Signatures and Infrastructures (ESI) TC.

The OASIS ebXML Message TC maintains the OASIS ebMS3 (also ISO 15000-1) standard and the AS4 standard (also ISO 15000-2). AS4 is profiled as the message exchange protocol of the European

Commission's eDelivery Building Block. Several dozens policy domains use eDelivery for cross-border secure and reliable exchange of documents and data. AS4 is also used in the EESSI system for digitalisation in social security coordination.

The OASIS Business Document Exchange TC provides complementary eDelivery specifications for service location and capability lookup.

The **OASIS ebCore TC has delivered version 3 of the CPPA specification**. CPPA3 provides standard data definitions, and formats for electronic, XML-based protocol profiles and business collaboration agreements, as well as algorithms for formation, matching, discovery and registration. Version 3 is an evolution of work done in the joint ebXML project with UN/CEFACT. It complements other ebXML standards for messaging including **AS4**.

OIDF

Set of standards and related certification profiles addressing identity transactions over the internet.

Active working groups in this area include:

- OpenID Connect WG,
- AccountChooser WG,
- Native Applications WG,
- Mobile operator Discovery, Registration and Authentication WG (MODRNA),
- Health Related Data Sharing WG (HEART), and
- Risk and Incident Sharing and Coordination WG (RISC)

http://openid.net/wg/

IETF

The Web Authorization Protocol (OAUTH) WG developed a protocol suite that allows a user to grant a third-party Website or application access to the user's protected resources, without necessarily revealing their long-term credentials, or even their identity. It also developed security schemes for presenting authorisation tokens to access a protected resource.

The ongoing standardisation effort within the OAUTH Working Group is focusing on enhancing interoperability of OAUTH deployments.

The Public Notary Transparency (TRANS) WG develops a standards-track specification of the Certificate Transparency protocol (RFC6962) that allows detection of the mis-issuance of certificates issued by CAs or via ad-hoc mapping by maintaining cryptographically verifiable audit logs.

The Automated Certificate Management Environment (ACME) WG specifies conventions for automated X.509 certificate management, including validation of control over an identifier, certificate issuance, certificate renewal, and certificate revocation. The initial focus of the ACME WG is on domain name certificates (as used by web servers), but other uses of certificates can be considered as work progresses.

https://trac.ietf.org/trac/iab/wiki/Multi-Stake-Holder-Platform#eIdentity

W3C

Verifiable Credentials provide a mechanism to express credentials, e.g. driving licenses, on the Web in a way that is cryptographically secure, privacy respecting, and machine-verifiable. Currently, the following Specifications and Notes have already been issued:

- Verifiable Credentials Data Model 1.0 https://www.w3.org/TR/vc-data-model/
- Verifiable Credentials Implementation Guidelines 1.0 https://www.w3.org/TR/vc-imp-guide/
- Verifiable Credentials Use Cases https://www.w3.org/TR/vc-use-cases/

Decentralized Identifiers (DIDs) are a new type of identifier that enables verifiable, decentralized digital identity. A DID refers to any subject (e.g., a person, organization, thing, data model, abstract entity, etc.) as determined by the controller of the DID. In contrast to typical, federated identifiers, DIDs have been designed so that they may be decoupled from centralized registries, identity providers, and certificate authorities:

- Decentralized Identifiers (DIDs) v1.0 https://www.w3.org/TR/did-core/

- Use Cases and Requirements for Decentralized Identifiers https://www.w3.org/TR/did-use-cases/
- DID Implementation Guide v1.0 https://www.w3.org/TR/did-imp-guide/
- DID Specification Registries https://www.w3.org/TR/did-spec-registries/
- DID Method Rubric v1.0 https://www.w3.org/TR/did-rubric/

Web Authentication defines an API enabling the creation and use of strong, attested, scoped, public key-based credentials by web applications, for the purpose of strongly authenticating users. The current work is on Web Authentication: An API for accessing Public Key Credentials - Level 3 https://www.w3.org/TR/webauthn-3/

Web payments: An important goal of Secure Payment Confirmation (SPC) is to streamline strong customer authentication (SCA). One way to reduce friction is to allow many authentications for a given registration. In other words, ideally the user registers once and can then authenticate "everywhere" (consistent with the policies of the relying party; they have to opt-in). The following Specifications are relevant:

- Secure Payment Confirmation https://www.w3.org/TR/secure-payment-confirmation/
- Payment Request API http://www.w3.org/TR/payment-request/
- Payment Method IDs http://www.w3.org/TR/payment-method-id/
- Payment Handler API https://www.w3.org/TR/payment-handler/
- Payment Method Manifest https://www.w3.org/TR/payment-method-manifest/

Work on **Social Networking** includes identity schemes that can play a role:

- ActivityPub https://www.w3.org/TR/activitypub/
- Social Web Protocols https://www.w3.org/TR/social-web-protocols/
- IndieAuth https://www.w3.org/TR/indieauth/

The **Web Crypto API** describes a JavaScript API for performing basic cryptographic operations in web applications, such as hashing, signature generation and verification, and encryption and decryption: https://www.w3.org/TR/WebCryptoAPI/ See also the note on use cases: http://www.w3.org/TR/webcrypto-usecases/ narrowing the scope of the Web Crypto API.

Identity for WebRTC 1.0 defines a set of ECMAScript APIs in WebIDL to allow and application using WebRTC to assert an identity, and to mark media streams as only viewable by another identity. This specification is being developed in conjunction with a protocol specification developed by the IETF RTCWEB group. https://www.w3.org/TR/webrtc-identity/

IEEE

IEEE has standards and pre-standards activities relevant to Electronic Identification and Trust Services, including dealing with blockchain technology, authentication, and biometric identification. More information can be found at:

IEEE P2049.3 - Standard for Human Augmentation: Identity,

IEEE 2410-2019, IEEE Standard for Biometric Open Protocol,

IEEE P2733, Standard for Clinical Internet of Things (IoT) Data and Device Interoperability with TIPPSS - Trust, Identity, Privacy, Protection, Safety, Security,

IEEE P2790, Standard for Biometric Liveness Detection,

IEEE P2799, Standard for Confirming and Conveying Identity Over the Internet,

IEEE P2989 – Standard for Authentication in Multi-Server Environment, and

IEEE P3210 - Standard for Blockchain-based Digital Identity System Framework.

There are also several pre-standards activities looking at digital identity, including guidelines for the provision and use of digital identities for digital resilience. For more information, see: https://ieeesa.io/rp-eidentification .

ENISA

Relevant standards for digital identities in Europe are typically defined by the European Commission rather than by standards development organisations. A detailed report published in January 2022 by ENISA (European Union Agency for Cybersecurity) "critically assesses the current literature and reports on the current technological landscape of SSI and existing eID solutions, as well as the standards,

communities, and pilot projects that are presently developing in support of these solutions. This study takes a wide view of decentralised electronic identity, considers possible architectural elements and mechanisms of governance, and identifies security risks and opportunities presented by SSI in view of cross-border interoperability, mutual recognition, and technology neutrality as required by eIDAS. " .

Under the title "DIGITAL IDENTITY: Leveraging the Self-Sovereignty Identity (SSI) Concept to Build Trust" it is available at https://www.enisa.europa.eu/publications/digital-identity-leveraging-the-ssi-concept-to-build-trust

Others

- The Journalism Trust Initiative (JTI) aims at a healthier information space. It is developing and implementing indicators for trustworthiness of journalism and thus, promotes and rewards compliance with professional norms and ethics. The JTI was originally launched and is now operated by Reporters Without Borders (RSF) and was cofunded by the European Commission.https://www.journalismtrustinitiative.org/
- SOLID project for data and identity sovereignty led by Tim Berners-Lee

Gaps and open questions

The preceding chapters have made it clear that this working group sees authentic pseudonymous identities as a cornerstone for ensuring trust in the digital space, for ensuring fairness in the sense of "one person, one voice", and for sustainably addressing the challenge of AI-automated fakes and bots.

The conceptual thinking inside and outside this working group has reached a certain level of maturity, with discussions involving numerous experts and stakeholders going back to the end of 2018. Nevertheless, many open questions remain that relate to aspects of the concept itself, implementation issues and standardisation:

- Standardisation gaps: To what extent are the numerous relevant standardisation activities listed in the preceding section likely to support authentic pseudonymous identities? Where should the numerous groups and fora collaborate and how could these be achieved? How do we deal with competing standardisation activities (in particular outside of the ISO/IEC/ITU or CEN/CENELEC/ETSI pillars where processes exist to prevent competing standards)? Which gaps do remain, either because of a lack of activity or because of an unclear timeline for results? How can CEN-CENELEC and ETSI address these gaps at the European level?
- Evolution of technology: Which strategies are most suitable to handle the discrepancy between the rapid evolution of technical capabilities for generating fakes and bots on a large scale, and the much slower standardisation and policy making processes? Where are technology-agnostic incentive mechanisms potentially more relevant than standards or regulation? What needs to be open, what can be proprietary? What needs to be considered regarding the open source / open standards debate?
- European values and goals: Which normative standards might be useful? What are the privacy, freedom of expression and other civil liberty implications of connecting information to the (pseudonymous) identity of its source and/or distribution chain, and how can standards help to strike an acceptable balance? How can mass-produced information be categorised into different risk levels? How do pseudonymous identities need to be implemented to optimally support the vision of the EU for the digital society in specific sectors, e.g. logistics, media, education, health, communication, leisure etc.)?
- Security considerations: How can the potential for gaming pseudonymous identities be minimized? How can pseudonymous identities be made quantum-proof or at least quantum-resilient? How can the security of pseudonymous identities be measured and tested in both centralised and decentralised contexts? How can individuals recover from one or more of their pseudonymous identities being compromised? What needs to be considered in IT security standards development?
- Entities with identities: What would benefit from having authentic pseudonymous identities? Apart from real people, how much value is in considering them for other categories of entities?[40]

40 ISO TC 307 WG2 proposes that four categories of entities could be distinguished: (1) Real people, individual humans, subjects; (2) Organisations, often legally constituted entities (firms, corporations, governments) or socially understood entities (families, villages, groups); (3) IoT to include anything owned by the first two

- Role of time: How long should a given pseudonymous identity have been established, and can we use a track record to build trust? In cases where the establishment date of a pseudonymous identity can be fabricated, a resilient system could require a "breadcrumb trail" of events that can be verified – but what is the equivalent of real-name world breadcrumb events (e.g. travel on an airline, transit through an airport, renewal of a passport with personal attendance, school or college attendance, a tax payment, etc.) in the context of pseudonymous identities? How could we cover the significant portion of the world population where such breadcrumb events are unavailable or undocumented?
- Media and journalism: How do AI-automated fakes and bots impact different types of media, journalists, influencers, bloggers, reporters, professional reviewers etc.? How can standards cover the whole spectrum? To what extent are these standards technical, and to what extent do the refer e.g. to journalistic standards for accuracy, or quality standards for media literacy education?
- Feedback, learning and community: How can standardised feedback mechanisms for the reliability of one's own judgement be formulated? Can "critical thinking" ever be the subject of standardisation, and if so, how? Are there innovative concepts for setting communication standards, beyond codes of conducts, and ideally developed bottom-up by relevant communities? Can there be a competition of "good" communication standards?
- Stakeholders: What responsibility should the EU have for a pseudonymous digital identity infrastructure and what standardisation activities should the EU drive as a result (e.g. by means of standardisation requests to CEN, CENELEC and ETSI)? How can the need for trusted third parties in the case of SSI-based solutions be minimized? What is the right balance/synergy between public and private responsibilities? What role (if any) should large internet corporations play, and how can they be forced to adapt and to adhere to public goals? Where are potential synergies? Where does Europe need to assert/enforce sovereignty?
- Non-AI-automated disinformation: How do we deal with disinformation/misinformation that is spread by real people rather than automated bots? How can standards be formulated to cover both types?

catagories, and not necessarily connected to any network; (4) Processes; often excluded, but with the existence of smart contracts within the DLT world, the actual contract (or process, to include any externalities to the smart contract itself)

Conclusion and recommended next steps

Europe is vulnerable to the new challenge of AI-automated mass generation of fakes and bots. In this report, we have described the potential of AI-based fabrication technologies to erode trust through fabricating "persons" and content on a massive scale that are indistinguishable from real human beings and their online activity. This can have potentially dramatic consequences, both for the economy and the functioning of democratic systems. Europe is faced with the prospect that society – both as citizens and as consumers - no longer knows which information or person it can trust in the digital space and which "person" is indeed a natural person at all.

It would be negligent to delay dealing with this threat because it has not yet manifested on a large scale. Countermeasures take time to develop and roll out. Therefore, attention and resources must be put into them right now, even while further analysis is being conducted in parallel. Europe needs to build immunity against mass-fabricated disinformation, needs to become able to defend civil society against actors using fakes and bots on a massive scale, and needs to discourage further pollution of the digital space.

The protection of the European economy and democracy must be at the top of the todo list. Standards developers as well as policy makers need to take action urgently.

We suggest that the following recommendations are considered:

General:

1. **Recognize** that the AI-based a**utomated mass fabrication** of fakes and bots is a **new challenge** that is not covered by existing actions and initiatives and requires a dedicated policy and standardisation response.

2. **Accept** that **evidence-based countermeasures** (fact checking, image analysis, user education etc.) are in a "detection arms race" with fabrication tools and therefore **insufficient** on their own.

3. **Accept** that a **compulsory use of verified real names** throughout the digital space would be effective against bots but comes with a high price in loss of civil liberties and privacy and is therefore **not a viable approach** in general.

4.**Start the process of establishing authentic pseudonymous identities** as an effective way of protecting trust in the digital space without loss of civil liberties, and of ensuring **"One Person, One Voice"** per context.

European standards developers and organisations:

5. **Refine our review of existing standards** related to forgeries, bots, physical trust anchors and/or self assured identities and make it more accessible, e.g. through a graphical representation

6. Establish a **CEN-CENELEC Focus Group on pseudonymous digital identities**, expanding on already ongoing activities, including CEN-CENELEC JTC19/WG1

7. Develop standards for connecting information back to its source or originator

8. Develop standards for bot-resistant pseudonymous identities (e.g. protocols, interfaces) compatible with eIDAS regulations and standards

9. Develop a **roadmap** for further standard development at the European level

10. **Continue** the analyis in **StandICT EUOS TWG TRUSTI** to address some of the questions raised in chapter 4.

European policy makers:

11. **Recognize a pseudonymous identity infrastructure** as a **governmental responsibility**, similar to real name identity infrastructures

12. **Provide policy support and a permanent line of funding** for necessary innovation activities, including:

- pilot applications e.g. for product reviews, political discourse, consumer information, or journalism
- R&D, e.g. for different kinds of (physical or non-physical) trust anchors and for decentralised/ federated approaches
- methods for achieving broad user acceptance
- integration with the European Digital Identity and with SSI
- integration with other emerging identity frameworks

13. **Create safe spaces to discuss and analyse** the following questions:
    - How can people trust each other in the digital space despite fabrication technologies?
    - What are key regulatory, societal and commercial pillars that need to be developed?
    - How can sovereignty be preserved in terms of assessing the trustworthiness of information and its sources?
    - How can professional accountability be designed?
    - What are the considerations around the use of federated vs. distributed vs. decentralised identities?
14. **Establish a High-Level Expert Group "Digital Identity, Trust, Sovereignty"**
15. Connect these activities to the **Recovery & Resilience programme**
16. Introduce pseudonymous identity infrastructures into the **ongoing political processes** for eIDAS, the Digital Markets Act, the Digital Services Act and the AI Act
17. **Conduct public-sector pilots to establish the principle "One Person, One Voice"** e.g. for petitions or public consultations

Multiple stakeholders:

18. Develop a pair of scenarios in an EU-wide dialogue: "How would the European digital society look like ...
    - with a well-established privacy-protecting identity infrastructure?"
    - with unmitigated spread of mass fabrications and bots and/or authoritarian countermeasures?"
19. Develop educational curricula as well as teacher training to provide children to become sovereign digital citizens and to make use of ways to spot and resist fakes and bots
20. Build coalitions, e.g. with G7, Council of Europe, OECD, GPAI, ANEC ...

We believe that, like in many other areas, Europe has the potential to be a trailblazer in the world for ensuring trust and constructive commerce and discourse in the digital space through thinking about identity in a broader and innovative way and establishing the necessary infrastructure and standards. We hope that this paper triggers an intensive and action-oriented dialogue in Europe.

We propose a bold approach that ultimately leads to an **ecosystem of trust between people in a digital Europe and beyond**.

Imagine a digital future –

- where the reliability of a piece of information can be linked to well-founded trust in its **source**,
- where people routinely participate in online discourse using **pseudonym mechanisms** that protect their real identity and at the same time prevent the creation of multiple identities in a given context,
- where **decentralised tools and protocols** assist citizens in judging how much they can trust other online actors and their information, allowing for **varied assessments of trustworthiness** rather than an "objective" score,
- where popularity and influence is **earned by constructive discourse and consensus building** rather than polarisation, and.
- where the "rules of the game" both in society and in business are **stacked against** trolls, bot herders and manipulators and their AI-based automation tools.

StandICT.eu has received funding from the European Union's Horizon 2020 (H2020) research and innovation programme under the Grant Agreement no. GA 951972.