Hybrid Ethics for Generative AI: Some Philosophical Inquiries on GANs

Antonio Carnevale[*]
antonio.carnevale@ntnu.no

Claudia Falchi Delgado[♦]
claudia.falchidelgado@dexai.eu

Piercosma Bisconti [◊]
piercosma.bisconti@dexai.eu

ABSTRACT

Until now, the mass spread of fake news and its negative consequences have implied mainly textual content towards a loss of citizens′ trust in institutions. Recently, a new type of machine learning framework has arisen, *Generative Adversarial Networks* (GANs) – a class of deep neural network models capable of creating multimedia content (photos, videos, audio) that simulate accurate content with extreme precision. While there are several areas of worthwhile application of GANs – e.g., in the field of audio-visual production, human-computer interactions, satire, and artistic creativity – their deceptive uses, at least as currently foreseeable, are just as numerous and worrying. The main concern is linked to the so-called "deepfakes", fake images or videos that simulate real events with extreme precision. When trained on a human face, GANs can make the face assume hyper-realistic movements, expressions and (verbal and non-verbal) communication abilities. This technology poses an urgent threat to the governance of democratic processes concerning the production of public opinions and political discourses, with significant potential for reality-altering and disinformation. After a short introduction of their current technical state-of-the-art, in this paper, we want to enquire about the GANs` socio-technical system alongside different and intertwined philosophical accounts. Firstly, we will argue about the conditions that make perceived GANs-generated content trustworthy, arguing also about the general effects GANs might have on the perceived trustworthiness of individuals. Thereafter, we will discuss about the inadequacy to approach GANs only as perception-altering technology. Against this backdrop, we will propose a theoretical turn that considers the human-machine relationships of trustworthiness as elements of broader hybrid socio-technical systems. This turn comes up with political repercussions that we will discuss in the last part of the paper.

[*] Norwegian University of Science and Technology, NTNU, DEXAI – Etica artificiale.
[♦] Maastrich University, Department of Technology & Society Studies, DEXAI – Etica artificiale.
[◊] Sant'Anna School of Advanced Studies, DEXAI – Etica artificiale.

HUMANA.MENTE *Journal of Philosophical Studies,* 2023, 44, 33-56 ISSN: 1972-1293

1. Introduction

In the last twenty years, our digital societies have been profoundly impacted by the development of AI technologies, as for example the automation of textual content generation through GPT3, and profiling algorithms. These technologies enable novel circulation mechanisms for information, and therefore have unprecedented impacts on people′s attitudes and trust toward the official information channels, ultimately changing our social systems. This war of information and narratives takes place on social networks due to the new role of platforms as gatekeepers for online communication, leading to the rapid dissemination of content and an intense polarisation of public opinion. These events reaffirm what the COVID-19 pandemic showed: new technologies have facilitated the emergence of *infodemiology*. As the WHO states

> An infodemic is too much information including false or misleading information in digital and physical environments during a disease outbreak. It causes confusion and risk-taking behaviours that can harm health. It also leads to mistrust in health authorities and undermines the public health response. An infodemic can intensify or lengthen outbreaks when people are unsure about what they need to do to protect their health and the health of people around them. With growing digitization – an expansion of social media and internet use – information can spread more rapidly. This can help to more quickly fill information voids but can also amplify harmful messages[1].

This concept summarises an increasingly important phenomenon in the landscape of western liberal and social democracies: the continuous and uncontrollable emergence of fake news circulating within social networks, negatively affecting citizens′ trust in institutions. This distortion of information is combined with the loss of confidence and authoritativeness in the traditional means of information, contributing to a dangerous trend of erosion of the democratic basis of the legitimization of political power. Therefore, the circulation of so-called fake news is a problem of primary importance for the well-being of EU democracies, and so it should be swiftly tackled. Until now, fake news and its negative consequences have implied mainly textual content. However, new technologies constantly raise new challenges. *Generative Adversarial Networks* (GANs), part of the Generative Artificial Intelligence family (Zant, Kouw & Schomaker, 2013; Seow et al., 2022), are a class of deep neural networks model developed

[1] More details here: https://www.who.int/health-topics/infodemic#tab=tab_1

in 2016 by Ian Goodfellow capable of creating multimedia content (photos, videos, audio), run by freely available software and simulating media contents with extreme precision (Goodfellow et al., 2020). For example, if trained on a face, GANs can make it move and speak in a way hardly distinguishable from an actual video. This powerful technology can produce a "self-reenactment" video that reconstructs a speaker′s facial expressions in real-time (Rössler et al., 2018). These techniques then lead to the so-called deepfakes (Witness, 2018; Dagar & Vishwakarma, 2022).

Deepfakes are born mainly as pornographic content, but they have quickly moved into politics. Imagine a video depicting the Israeli prime minister in private conversation with a colleague, seemingly revealing a plan to carry out a series of political assassinations in Tehran. Or an audio clip of Iranian officials planning a covert operation to kill Sunni leaders in a particular province of Iraq. Or a video showing an American general in Afghanistan burning a Koran. Similar contents were produced, for example, with deep fakes of Barack Obama or Donald Trump (Seow et al., 2022). Moreover, during the war in Ukraine, a deepfake of President Zelensky surrendering was released. This time, quality was too low to be perceived as trustworthy, yet it is only a matter of time before the technology improves and becomes more deceiving.

GANs are a very recent technology – the first paper proposing this approach is from 2014 (Mirza & Osindero 2014) – but are evolving very rapidly, and we should expect that, in the coming years, deepfakes will be massively disseminated. The implications can be serious for the well-being of global democracies. In fact, until a few years ago, visual media, such as videos, were still considered reliable evidence, and most of the population is unaware of the existence of GANs. Nevertheless, soon social networks could be invaded by deepfakes indistinguishable from reality. It could be that i) fact-checking approaches might not be able to verify such a large amount of content and ii) the presence of GANs could allow individuals to deny actions they took, claiming to be victims of a deepfake in front of public opinion.

Both consequently enable another shift toward a *post-truth* world (Harsin, 2018). Although GANs are extensively studied in computer science and all relevant technical aspects (Goodfellow et al., 2020), their impact in media contexts and democratic processes has not yet been examined from a social sciences and philosophical perspective. This lack of socio-technical studies on GANs reflects a more widespread lack of preparedness in steering their ethical and political impacts, a delay that involves Information Technology (IT) experts as well as common users, who tend to believe, trust, and share social media contents.

In this paper, first, we give a brief technical overview of how GANs work and how they can be currently applied (Section 2). Then, we discuss – from a philosophical perspective – three core challenges essential to understanding the ethical and political implications of GANs-generated content on our societal shapes.

A fundamental philosophical order of questions concerns the *system of perception* (Section 3): are realities synthetically generated by GANs hyperreal, so much as to confuse the image with their real reference? Or are they distortions that frighten our familiarity with the world of technological things in which we are immersed?

We cannot answer these questions without dealing in this article with a second problematic order: that of *information system* (Section 4). Whether the perception of the truthfulness of generative reality will be deceptive for us or not, will depend on how these technologies manage to disrupt the information system, with particular attention to the dynamics of building trust and certainty. Therefore, the question we raise is how these "synthetic social actors" might modify social systems, and how we can enquire these effects. In this section we propose a theoretical turn from enquiring the perceived trustworthiness of AI-generated contents, to analysing how these contents affects the trustworthiness of the relations between the agents in hybrid system.

Lastly, in (section 5), we discuss the *political implications* of generative AI and reflect on their impact on digital democracy and their possibility of enhancing or diminishing political trust (i.e, on social media). The emerging blockchain technology seems the most suitable, if not the only one up to date, to fulfil these requirements due to its embedded secured nature.

2. Overview of GANs Reality, hyperreality, or uncanny valley?

In the last two years, much attention has been devoted by the AI community to Generative Artificial Intelligence (GAI), namely a type of AI systems able to generate new data such as images, videos, audio, and texts (Zant, Kouw & Schomaker, 2013). Beyond medical applications, for example data augmentation, generative AI has been widely used for artistic purposes: just with a short paragraph of text, it assists you with creating art and pictures. One outstanding example, recently released, is Open AI's DALL-E system, which enables the creation of hyper-realistic images from text descriptions. Other examples include videos generation (Meta's text-to-video AI platform), AI avatars (Synthesia's text-to-speech synthesis which generates a professional video with an AI avatar reciting your text), music (Jukebox neural net) and much more. The magic behind most of this generative reality stands within GPT-3 – a language model creating human-like text that enables the creation of hyper-realistic images. Simply speaking, we can understand it as "text in – text out" – the model processes the text given and produces what it 'believes' should come next based on the probabilities of sequences of particular words in a sentence.

Among these AI systems, we will focus on GANs as AI-based systems of simulation / synthetic generation of reality, known with the hasty formula "deepfakes".

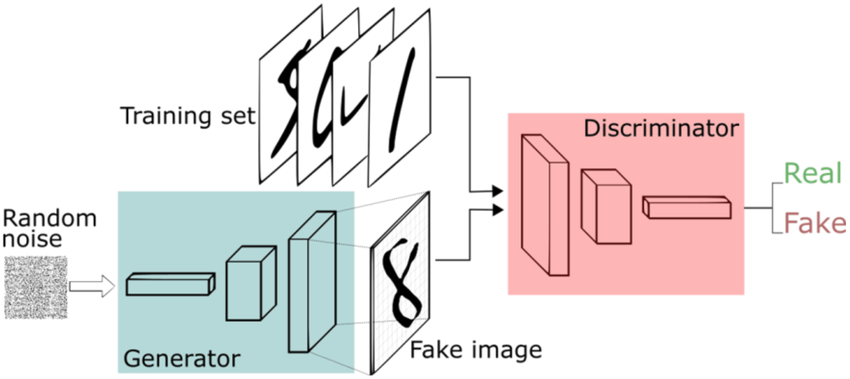

Figure 1 – Thalles Blog (Jun 7, 2017) [2]

GANs are an emerging technology for both semi-supervised and unsupervised learning and are composed of the i) generator and ii) the discriminator.

The discriminator has access to a set of training images depicting real content. The discriminator aims to discriminate between "real" images from the training set and "fake" images generated by the generator. On the other hand, the generator generates images as similar as possible to the training set. The

[2] https://sthalles.github.io/intro-to-gans/

generator starts by generating random images and receives a signal from the discriminator whether the discriminator finds them real or fake. At equilibrium, the discriminator should not be able to tell the difference between the images generated by the generator and the actual images in the training set; hence, the generator generates images indistinguishable from the training set (Elgammal et al., 2017).

We can think about GAN, in a more metaphorical and intuitive way, as a synthetically-generated game between a forger (generator) and a policeman (discriminator) (Creswell et al., 2017). The counterfeiters produce fake money, while the police attempt to arrest the counterfeiters, but doing this, in the meantime, will allow the production and circulation of counterfeit money. Thus, competition between counterfeiters and the police will make the production of counterfeit money increasingly realistic, as counterfeiters will have time and a way to "train" themselves to improve production. At this point, counterfeiters, until they are caught, will be able to produce perfect counterfeits, so that the police will no longer be able to distinguish between real and counterfeit money.

A further complication to this metaphor: it is as if the generator learned through the gradient of the discriminator, that is, as if the counterfeiters had a "mole" between the police to report the specific methods used by the police to detect forgeries (again, allowing you to improve the production of fake money by making it more and more likely to the original and therefore misleading, difficult to distinguish).

In the next sections, we discuss interesting philosophical aspects questioning the way we perceive and interact with synthetically-generated agents, and what are the implications for social systems.

Our perceptual system conveys two worlds that have always conflicted with each other and that many religious, philosophical, and scientific theories have tried to resolve in the course of human history: the world of things and the world of their representation. That all perceptions are acts of interpretation is the take-home-message of all the constructivist approach to human cognition, first and foremost represented by the contribution of Humberto Maturana and Francisco Varela in their seminal book *Autopoiesis and Cognition* (1991). Without entering the details, their theory of the *perceptive apparatus* is influenced by the internal organisation (the *autopoiesis*) of the perceiving agent. Particularly representative in this sense is the paper on the relationship, in the frog, between vision and the cognitive apparatus (Lettvin et al., 1959). This paper claims that the frog's perceptual system, even before hyletic signals are processed at the cognitive level, "constructs" the frog's reality in a specific way that is useful for the frog's organisation. In particular, because of the frog's interest in flies, the frog's perceptual system is more "interested" in very fast-moving objects than in slow ones. The key point in this argument is that the frog's perceptual apparatus itself already organises the world in a way that is functional for its autopoietic system. The same gnoseological constructivism is also to be found in the very definition of the autopoietic system: objective reality, the real as such, is inaccessible to subjects. In the same way, the neuroscientist Anil Seth argues that perceptions are acts of informed guesswork that the brain applies when it encounters sensory data. Perception is a processor of active construction, a "controlled hallucination" (Seth & Bayne, 2022). It seems as if the world out there has all these properties like redness, shape, and temperature and that we detect these through our senses and something in our brain reads out this information from the outside world. In this view, perception is a bottom-up outside-in process.

> "You can build systems that perceive things this way, and that's not entirely wrong per-say, but it's missing the central part of the story which is the context of what we perceive. When I perceive something as being red, it's not just sensitive to an externally existing redness. Redness is coming from within my brain, as a way of predicting how certain patterns of light appear, how surfaces reflect patterns of light. Sensory data by itself is not red, it's not anything. It's just energy. Sensory signals don't come with labels attached. Everything we perceive is a kind of inference, a burst guess about what's out there. The question for me is how we use these ideas to explain not just what we perceive 'oh, I perceive a cup because a cup is somehow there...' but why the experience of cupness is the way it is. What is it about the predictions that my brain is making that makes the experience of cupness different from the experience of tomatoness, jealousy or something very different?" (Seth & Bayne, 2022)

Our perceptions are therefore complex constructs continuously in the balance between determining the sense of reality but also controlling hallucinations, distortions, alienations, and reifications. Precisely for this reason, the idea of being deceived becomes strongly paralysing, the idea of being victims of some unconscious illusion. This is the case of the hyperrealities synthetically generated by GANs: what happens when an image no longer resembles its object but identifies itself with it on a perceptual level? What happens, then, when *similarity* vanishes and becomes indistinguishable from the identity?

However, being indistinguishable does not imply that it can deceive a human. As the literature on social robots suggests (Li et al. 2011), humans easily perceive the mismatch between verbal and non-verbal cues or the mismatch between the different metacommunicative contents conveyed by different communicative registers.

This kind of mismatch, as Kätsyri et al. (2019) suggested, can be at the root of the "uncanny valley" (UV) effect, one of the most-known phenomena in human-robot perception studies (Bisconti & Carnevale, 2022). The effect is not just relevant to robots but to any form of a human-like object, including dolls, masks, facial caricatures, virtual reality avatars, and characters in computer-generated movies (Seyama & Nagayama, 2007). This effect is referred to the sensation of unfamiliarity while perceiving artificial agents that seem not quite human (Laakasuo et al., 2021). UV studies have suggested different outcomes. Laakasuo et al. research revealed a moral uncanny valley effect. People evaluated moral choices by human-looking robots as less ethical than the same choices made by a non-human or a non-uncanny robot.

Similarly, Kätsyri et al. (2019) results showed a linear relationship with a slight upward curvature between human likeness and affinity. In other words, less realistic faces triggered greater eeriness in an accelerating manner. Thus, a weaker UV effect for Computer Generated (CG) faces. Further studies indicated that UV has a real influence on humans' perceptions of robots as social partners, robustly influencing not only humans' conscious assessments of their reactions but also able "to penetrate more deeply to modify their actual trust-related social behaviour with robot counterparts:" (Mathur & Reichling, 2016, p. 31).

BuzzFeed, a news and entertainment company, created its video using increasingly common techniques known as synthetic media (Witness, 2018) or deepfakes. However, generating credible deepfakes is still challenging lip-synching, coherent facial mimicry, and smooth interaction between the human subject of the deepfake and its environment are still open issues. This is because often, deepfake lacks credibility in the small details of the image, such as blurry outlines. This might be a fact-checking consequence that might not enable the verification of such a large amount of content. However, notwithstanding all these technical limitations, some studies suggest that GANs-generated content, which often relies on visual and audio contents (as opposed to text), can seem very realistic and hence may be perceived as more trustworthy compared to other

forms of simulation. For instance, UV research has shown that when sets of photos depicting faces were given to participants, the ones generated by GANs obtained more perceived trustworthiness by its users. Yet, this perception may vary according to psychological content varies such due to objectivity, subjective, and both internal individual and external-environmental characteristics.
All these studies challenge an exciting philosophical question: based on what, therefore, does the perception of trustworthiness for the contents of reality synthetically generated by the GANs vary?

Perceived trustworthiness can be measured as confidence in the information validity regarding sincerity and objectivity (Hovland & Weiss, 1951). General trustworthiness of GANs-generated contents, which determines whether individuals will believe the message they are exposed to, can be viewed as a broad evaluation that consists of several judgments of the source and the message. For example, previous research has highlighted the importance of perceived source credibility, source vividness or perceived salience (Lee & Shin, 2021), information believability, and overall persuasiveness (Hwang et al., 2021), when making conclusions regarding deepfakes and other kinds of information. This first challenge also stands behind the lack of current literature addresses this topic partially (Vaccari & Chadwick, 2020; Etienne, 2021; Langguth et al., 2021) and thus extensive work must follow to understand the feature of synthetic agents increasing or decreasing perceived trustworthiness. Studies on perceived trustworthiness are lacking in the case of GANs, and in general for what concerns generative AI. Thus, this raises an issue not only related to individual relationships with synthetic social agents but also to the psychological aspects of human-AI relationships.

The issue of trustworthiness implies social systems at large when we consider that social agents are the main driver of information. The massive spread of deepfakes in information channels might worsen an increasingly important phenomenon in the political and democratic landscape of Europe: the continuous and uncontrollable emergence of fake news circulating within social networks, negatively affecting citizens′ trust in institutions. This excess of information is combined with the loss of trust and authoritativeness in the classic means of information, contributing to a dangerous trend of erosion of the democratic values. The introduction of generative technologies in the system of information generation and circulation might become a disruptive element, lowering the perceived trustworthiness of information by human social actors. In fact, the criteria for the trustworthiness of information content profoundly change with generative AI.

The main issue of not being able to recognize fake contents from reliable ones depends on the increase of uncertainty for the human agents, with respect to any information that the social system provides. The relationship between these elements is, in our opinion, the most important issue pertaining to GANs. Scholars have examined the relationship between trust and uncertainty. Even if a deepfake does not always deceive its viewers, they may become uncertain whether their content is true or false. Uncertainty is experienced when insufficient information is available to make a choice, and thus it can be overcome by introducing new information (Alvarez & Brehm, 1997). Thus, if deepfakes, among other methods of disinformation, succeed in increasing uncertainty, one of the main implications may be a reduction of epistemic certainty of the public opinion in the information channels. Tsfati and Cappella (2003) revealed that for trust to be relevant, there must be a degree of uncertainty on the side of the trustor, and this is implied inherently in the notion of trust. On the other hand, trusting others may become more complex when uncertainty increases. Concerning GANs, increased uncertainty may explain why deepfakes may cultivate the assumption among citizens that fundamental and rooted truth cannot be established.

Therefore, we claim that the synthetically-generated contents of GANS may deceive people distinctively for (a) the objective characteristics of the content, (b) individuals' subjective perception of the content and (c) the (internal) individual characteristics of subjects who are exposed to the content. Since no methodologies have been formalised to assess the features enhancing the perceived trustworthiness of GANs, it becomes crucial to understand individual differences (i.e., demographic, ideological, and cognitive) related to their propensity to trust GANs contents.

At this level of analysis, the analysis of GANs effects relies only on the analysis of the human-AI relationship, which must specify which characteristics of GANs are implied in increasing or reducing the perceived trustworthiness of the GANs-generated content. Namely, GANs account only for a troubling noise of information and social systems. We claim that this standpoint cannot capture the whole positive and negative potential owned by generative AI with respect to social systems. AI-generated social agents entail a qualitative shift from a social paradigm composed of only human actors to hybrid social systems. Hybrid social systems, composed of human and non-human actors, cannot be assessed from the perceptual point of view only since this perspective can only be based on an anthropocentric standpoint. And that is why we must introduce another level of challenge that the generative reality of GANs poses to philosophical reasoning. *From the perceptual to a system approach.* This will allow us to introduce a novel standpoint on the implications of synthetic social agents for social systems.

3. GANs, social agency and system trustworthiness

In this chapter, we briefly sketch a systemic approach to hybrid systems. We will discuss three main issues:

I. We should move from an anthropocentric theoretical framework of social systems, where only human agents are proper social agents.

II. The analysis of the systemic implications of synthetic social agents on social systems brings us to the notion of hybrid social systems.

III. We claim that analysing the individual perceived trustworthiness of GANs cannot capture the complexity of hybrid systems. Therefore, we propose a theoretical turn where trustworthiness is considered a proxy property of the socio-technical system.

To frame the capacity for the social agency of synthetic actors, we start with the Actor-Network Theory (ANT). Its best-known theorizer is Bruno Latour, also in his joint work with Stephen Woolgar (see e.g., Latour & Woolgar 1987; Latour 2005), but the contributions of Michel Callon and John Law have been equally pioneering and foundational (Callon et al., 1986; Law & Hassard, 1999). ANT allows us to approach the question of the social agency of GANS because, in its theoretical framework, actors of the social are both humans and non-humans, and their *agency* is measured considering their capacity to mediate the relations between actors. This enables a *symmetrical* treatment of human and non-human actors, with all the actors of the social system equally able to modify the relations between actors. Technical objects, artificial agents such as GAN and, their generated contents can be therefore considered, under ANT, equally significant for the circulation of contents, beliefs, and information inside the social system. In revealing the complexities of socio-technical environments in their social, material, cultural, or political dimensions, ANT puts all the emphasis on the relations between the actors.

The most important distinction is therefore the difference between mediator and intermediator. A mediator modifies the information and the relations between the actors of the social system. Instead, an intermediary is a messenger that "transports meaning or force without transformation" (Latour, p. 39, 2005), providing information and connections but no more. In relational terms, they may facilitate or enact introductions between parties. The mediation process is furtherly specified in the theory of translation. In practice, applying ANT means identifying and spelling out the "translation process", namely the process by which an actor acquires knowledge of a system and begins to be an "active" actant in that system. Networks of actors, in Latour's theory, create so-called "panoramas": this concept identifies narratives and beliefs that are shared inside a network of actors, informing their actions and their future associations, and therefore shaping the socio-political system. While with the methodology of translation, we describe how artificial actors can be involved and become spokespersons of a network, with the concept of panoramas we analyse the process of belief change among actors in the system. It is thus at the level of panoramas that we should enquire how synthetic social agents might modify the equilibrium of social systems. Narratives and beliefs, that are semantically conveyed mainly through verbal and visual content, are the direct result of GANs-generated content.

This leads us to a consideration that moves beyond ANT. In fact, the relationship between actors and networks is analysed by ANT without considering the semantic dimension (Law & Hassard, 1999), and therefore the different ways semantic modifications can happen inside the relations between actors. Actors, for ANT, mediate all in the same way, whatever action they perform, a drawback of the ANT manifesto on "flattening the social". On the other hand, synthetic social actors are a class of non-human agents that enter the social sphere with all the instruments to semantically mediate narratives and beliefs, namely they could convey meaningful verbal and visual content. This concept can be summarised under the umbrella term, usually applied to social robotics, of "quasi-other" (Ihde, 1990). This peculiarity of synthetically generated agents breaks in an unprecedented way the distinction between human and non-human agents. If Latour` s aim, and that of ANT in general, is to show that objects are actors even if they do not speak, in this case, the object of our research can speak and is difficult to distinguish from a human speaker. ANT allows us to consider human and non-human actors on the same level. Nevertheless, we must go further to fully grasp the complexity introduced by synthetic social actors. Generative AI, in fact, will not simply redundantly reproduce the narratives of current social systems but will generate new semantic patterns of interpretation of the social facts, becoming a proper *social* and *active* agent of the social system. This is the fundamental difference with "usual" technical objects: generative AI is a learning system that produces highly unpredictable semantic outputs. These outputs, being socially conveyed by autonomously generated verbal and visual contents, have a qualitatively different power of modifying social systems with respect to other technological objects. For this reason, we consider these socio-technical systems as structurally hybrid: non-human agents have reached the same (or very similar) ability as human agents in generating and vehiculating social narratives. Given this, the theoretical turn we envision is to analyse how generative AI can modify the processes regulating the trustworthiness of the relationships between social actors when conveying verbal content within information systems.

Namely, the open challenge is to connect the analysis of the individual's perceived trustworthiness with a holistic approach encompassing the trustworthiness of the relationships between all the actors and processes that are part of the socio-technical system. The human-perceived trustworthiness of AI is still detached from its systemic implications. We can move beyond this paradigm by analysing the concept of trustworthiness as a property of the relations between the (human and non-human) actors of the system. This moves us forward from a framework where trustworthiness is a property of the AI agent with respect to the human agent.

We claim that, in hybrid systems, we should understand what kind of AI-generated content are able to increase or decrease the trustworthiness of the relationships between the actors of the system. We make an example to clarify the relevance of this theoretical turn: let's imagine a deepfake depicting the ministry of economics in Germany stating that the economy will drop next year. The interesting issue to analyse is not how trustworthy the deepfake itself is, but what kind of implications this specific deepfake (with its peculiar feature) will have on the relationship between the other actors of the system, i.e., the stock market investors. This example, while trivial, shifts the focus of the analysis from the AI perceived trustworthiness directly to the AI effects on the trustworthiness of the hybrid system.

4. The political implications of synthetic social agents: a matter of trust

Such a necessity of a holistic consideration of the trust/trustworthiness of AI, however, does not yet find an effective and systematic resonance in the debate on the psychological, social, and ethical aspects of AI. As we shall soon see later in this section, there is no scarcity of such types of studies. Rather, the shortcoming lies in the observation that such approaches mostly try to find methods for measuring and assessing the social acceptability of AI (Occhipinti et al., 2022). We think that, here, there are *political issues* at stake that go beyond the acceptability-base challenge.

Technology acceptance research has shifted in the later years from acceptability, in terms of technological usability, to incorporating social factors into specific technology-oriented assessments (Malhotra & Galletta, 1999). Such a holistic consideration of the AI effects on the trustworthiness of the hybrid system does not yet find an effective and systematic resonance in the debate on the psychological, social, and ethical aspects of AI. As we shall soon see later in this section, there is no scarcity of such types of studies. Instead, the shortcoming lies in the observation that such approaches mostly try to find methods for measuring and assessing the social acceptability of technology. The approach cannot answer the question of accepting new technologies only by looking at individual concerns; it also involves analyzing the activities, decisions, and consequences of any new digital technologies within which those concerns persist (Kantar & Bynum, 2021). We can further relate this holistic conception to Gilles Deleuze's notion of "societies of control" expressed in his 1990 essay *Post-scriptum sur les sociétés de contrôle.* According to Deleuze, the overall objective of societies of control is no longer simply to govern abnormal behaviour in closed environments (e.g., psychiatric hospitals and prisons) but to ensure a regime of unrelentless surveillance in the open spaces of our communities. Imaging a city where one could leave one′s apartment, street, and neighbourhoods, thanks to one′s (individual) electronic card that raises a given barrier. Still, the card could just as quickly be rejected on a given day or between certain hours. What counts in this situation is not the barrier but the computer that tracks each person′s position – licit or illicit – and effects a universal modulation. Who decides and controls which technologies function as mediators of human-world relations? (Ihde, 1990; Verbeek, 2016). We think that, here, there are political issues at stake that go beyond the acceptability-base challenge.

Political deepfakes are an essential product of the Internet′s visual turn, providing the first evidence of the risk of the deceptiveness of deepfakes. Anecdotal evidence suggests that the prospect of mass production and diffusion of deepfakes by malicious actors could present the most serious challenge yet to the authenticity of online political discourse (Vaccari & Chadwick, 2020). Additionally, citizens have frail defences against this form of visual deception because visuals have a more significant persuasive effect than plain text (Newman et al., 2015; Stenberg, 2006). This could, however, vary from each digital culture and from the demographic size of societies (Roozenbeek et al. 2020). Research focusing on demographic variables studies the link between essential individual characteristics (such as gender and age) and susceptibility to misinformation. For example, higher exposure to misinformation is generally associated with older age (Grinberg et al., 2019). Nevertheless, this association might differ based on the specific topic of misinformation; for example, a recent study by Roozenbeek et al. (2020) showed that older individuals were less susceptible to misinformation regarding COVID in four of the five countries (Ireland, Spain, UK, and USA). Although these factors, especially age, are important, they are often investigated in combination with either motivational or cognitive characteristics.

Drawing from various theories and phenomena, such as motivated reasoning, which refers to biases that lead to decisions based on their desirability rather than an accurate reflection of the evidence (Kunda, 1990), researchers have already identified a few specific individual variables that may motivate the person to believe misinformation she/he is exposed to. These include, but are not limited to, minority status, belief in conspiracy theories, low trust in science, media, the government, and politics (Roozenbeek et al., 2020). Additionally, previous research has also explored the role of cognitive abilities and other related variables (Sirota & Juanchich, 2018). These studies have revealed that particularly education, reflective thinking – measured for instance with the *Cognitive Reflection Test* (Sirota & Juanchich, 2018) – and the so-called "bullshit receptivity" are consistently associated with the processing of misinformation. Research specifically focusing on the association between individual differences and propensity to trust GANs-generated content is much less common and thus just starting to emerge (Ahmed, 2021).

If this technology is mainly trained on the raw data scraped from the Internet, does it mean that it can reinforce social stereotypes and harmful points of view that are an inherent part of the Internet? Even though it is possible to try to filter undesirable content like pornography, inappropriate language, or racist comments, the amount of data on the Internet is so large that it is not unlikely that some of the malicious content will slip through. Although there is no proof that fraudulent political deepfakes misled participants, we contend that many of them were left questioning the veracity of their claims. Because of this ambiguity, people may have less trust in the news on social media. An approach to deal with raw data scraped from the Internet is to consider other types of technologies that have been used to make increasingly important decisions, raising the question of whether they can be trusted to act fairly and transparently or, more generally, in the interests of their users (Pasquale, 2015).

The studies we have mentioned highlight a point that we think is essential and that we would highlight in this final part of the paper. In their own way, GANs represent, at the level of computational gain, what the social realities shared and signified by human beings (and among these, a place of honour obviously goes to language) represent at the level of complex cognitive acquisition. Provocatively, we could say that artificial agents such as GANs and human agents are not at odds with each other as "artificial agents vs. natural agents ", but rather *as agents differently able* (less complex vs. more complex) *to synthesize reality in a discourse that makes sense*. And here lies the new dilemma of trust and technology.

Trust is a complex feeling, perhaps not even a feeling at all. From first observation, one thing is clear: it is *exclusive* (Pettit, 1995). There is or there isn′t. Better to say, when it exists and is lost, it is difficult to regain it. Second evidence also tells us that it is the basis of many human relationships, even very different from each other: love, friendship, business, politics, relationships with institutions. It is commonly believed that trust has to do with having certainties, with being sure of something – "having faith in one′s means", "I am confident that things will go well". Differently, its underlying reason is quite the opposite: *trust is a controlled uncertainty over time*. It is the antidote that human relationships have developed to mitigate the indeterminacy of the future and the impossibility of fully knowing the others we face or the situations that happen to us in life. Because of this primeval connection with the unknown, *trusting is risky*. Trust and risk form a primary axis of fiduciary relationships, starting with self-confidence, a founding characteristic of the balance of the human personality marked by the success of emotional stability to be able to believe in one′s life plans (Luhmann, 1979). In addition to the Self, trusting and entrusting oneself are however important dispositions for entering into a relationship with the

Other, up to its extreme, entrusting oneself totally, abandoning oneself into the hands of the other, from which similar motives arise of faith and fidelity ("blind trust").

Returning to our discourse, technologies more able to generate synthetic realities in some way "social" are of concern not so much because they make it more articulated and difficult to answer the classic questions on the political implications of technology (for example, "How can we trust these technologies?"; "How to design them so that they are reliable and human-centric from their conception?"). That′s the least of the problems. They question the link between risk and trust. By securing risk-free trust nodes inside our social world, will we still be able to build trust relationships?

A very topical and discussed case of technology, which is inscribed in this register of issues, is the one of blockchain.

5. Technology-mediated system trustworthiness: the Blockchain example

Blockchain (BC) is an outstanding example of a technology that should produce what we named "system trustworthiness", namely the developed technology is not trustworthy itself, but ensure trustworthiness among the actors of the system. We claim that the systemic perspective on trustworthiness that pertains to BC technologies should be adopted also in designing the interaction between AI-generated agents and social systems. BC, by replacing seemingly untrustworthy intermediaries with a technological system designed to minimise the need for trust (BC is also said to be "trustless"), aims to reduce the trust requirement by restricting user behaviour and eliminating the option of non-compliance through its design (Werbach, 2018). Similarly, to frame the capacity for social agency of synthetic actors, Bitcoin can be seen as a significant non-human actor (blockchain and wallets), including human and non-human factors that affect the process of division as embodied in the focus actors (e.g., code, algorithms, electricity), institutions (regulators, central banks, European institutions), and ideology (e.g., an inclusive and democratic global payment system) (Islam et al., 2019). These play an *equally* important role in reshaping the translation process and becoming active game changers in the network. In fact, what BC does is to ensure that the relationship between two (human) actors is trustworthy "by-design", and therefore it changes the nature of the relationships among actors inside the system. Certainly, from a system perspective, BC is a mediator, since it enables new trustworthy relational configurations between (human) actors. On the other hand, BC does not generate by itself new semantic patterns: it adds a property – trustworthiness – to previous relational configurations (exchanging money on a stock market).

On the other hand, generative AI goes far beyond Distributed Ledger Technologies (DLTs) such as blockchain, as *they do not simply verify the narratives of current social systems but will generate new semantic patterns of interpretation of social facts, becoming an active social agent of the social system itself.* This is the fundamental difference between the "usual" technical objects (the blockchain) that have been socially accepted due to their ability to achieve decentralised trustworthiness on the Internet of Things (IoT) and the "not usual" synthetic and hybrid systems (GANs), which have not yet been socially accepted due to their novelty, *uncanny valley* (UV) effect and therefore, unsure impact on their perceived trustworthiness.

As blockchain is in its nature verifiable, decentralized, distributed, immutable, transparent, and auditable (Chen et al., 2018), it will be less likely, to reinforce social stereotypes and harmful points of view that are an inherent part of the Internet sphere. In fact, it adds a new property – trustworthiness – to previously existing interactional configurations of the system. This differs from generative AI, able to generate novel semantic and relational patterns, so we should keep researching and anticipating the political implications of GANs. As focusing on the individual level of perceived trustworthiness is not enough, we suggest considering the perceived trustworthiness of the hybrid socio-technical system itself. Trustworthiness could help us measure the acceptance of decisions, with requirements such as consensus, economic models, and incentives for honesty, explainability, and robustness of predictors (Nassar et al., 2019). As many more infrastructure requirements are needed to secure political trusts, such as (security, privacy, reliability, usability, dependability, performance, and governance), emerging blockchain technology seems the most suitable, to fulfil these requirements. Still, many challenges must be tackled, and the most important ones are minimizing humans in the loop for validating explanations and actual timelines for specific applications.

This last section briefly sketched an example of technology where trustworthiness has been considered at the system level, instead of the technology level. As we discussed, Blockchain technologies and generative AI both have relevant implications regarding trustworthiness, if analysed systemically. Artificial actors (GAN) becoming active spokespersons of a network could mislead and likely harm online civic culture and reduce trust in online news (Vaccari & Chadwick, 2020). Furthermore, this implies the consequence of enhancing misinformation and damaging digital democracy. On the other hand, technology, such as blockchain, plays a similar role in reshaping the network. In this case, misinformation would be more controlled thanks to blockchain's trust-embedded characteristics.

In this paper, we discussed the implications of generative AI for social systems, particularly GAN-generated content, focusing on the concept of trustworthiness. First, we discussed the current technical state-of-the-art of GANs to show how these AI systems can generate a nearly perfect representation of reality. In the second section of this paper, we focused on the main implication currently addressed by the literature: at what condition a GANs-generated content is perceived trustworthy, and what general effects GANs might have on the perceived trustworthiness of individuals. Then, we discussed the limitation of an approach focusing only on the individual perceived trustworthiness, and we proposed a theoretical turn to consider the system trustworthiness in hybrid socio-technical systems. The fourth chapter discussed the possible political implications of GANs and how humans and non-humans can both play a role in shaping the socio-technical system. Conclusively, the fifth chapter discusses how the type of technology used (i.e., blockchain) plays a role in some intrinsic differences related to susceptibility to misinformation.

During our research, we encountered three main limitations:

I. Although GANs are extensively studied in informatics (Goodfellow et al., 2020; Seow et al., 2022), their impact in media contexts and democratic processes has not yet been examined from social sciences and philosophical perspectives.

II. This lack of socio-technical research on GANs reflects how we cannot yet precisely understand, in hybrid systems, what kind of AI-generated content can increase or decrease the trustworthiness of the relationships between the actors of the system.

III. It is not sufficient to focus on the individual level of perceived trustworthiness. Hence, we suggest considering as a whole the perceived trustworthiness of the hybrid socio-technical.

In future works we will focus on narrowing down a methodological approach for assessing the trustworthiness of hybrid systems. In this paper we outlined the importance of a theoretical turn from human-AI interactions to the study of socio-technical societies as hybrid systems.

ACKNOWLEDGMENTS

Piercosma Bisconti and Antonio Carnevale conceived the idea for the paper. Antonio Carnevale wrote the Introduction and Chapter 4. Claudia Falchi Delgado wrote Chapter 2 and 5. Piercosma Bisconti wrote Chapter 5. All authors contributed to the final manuscript.

REFERENCES

Ahmed, S. (2021). Fooled by the fakes: Cognitive differences in perceived claim accuracy and sharing intention of non-political deepfakes. *Personality and Individual Differences*, 182:111074. https://doi.org/10.1016/j.paid.2021.111074

Bisconti, P., & Carnevale, A. (2022). Alienation and Recognition: The Δ Phenomenology of the Human–Social Robot Interaction (HSRI). *Techné: Research in Philosophy and Technology*. https://doi.org/10.5840/techne202259157

Callon, M. et al. (1986). Mapping the dynamics of science and technology: sociology of science in the real world. Basingstoke: Macmillan.

Chen, W., Xu, Z., Shi, S., Zhao, Y., & Zhao, J. (2018). A survey of blockchain applications in different domains. *Proceedings of the 2018 International Conference on Blockchain Technology and Application – ICBTA 2018*. https://doi.org/10.1145/3301403.3301407

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2017). Generative Adversarial Networks: An Overview. *arXiv*. https://doi.org/10.1109/MSP.2017.2765202

Dagar, D., & Vishwakarma, D.K. (2022). A literature review and perspectives in deepfakes: generation, detection, and applications. *International Journal of Multimedia Information Retrieval*, 11, 219–289. https://doi.org/10.1007/s13735-022-00241-w

Deleuze, G. (1992). Postscript on the Societies of Control. October, 59, 3–7. http://www.jstor.org/stable/778828.

Elgammal, A., Liu, B., Elhoseiny, M., & Mazzone, M. (2017). CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms. *arXiv*. https://doi.org/https://arxiv.org/abs/1706.07068v1

Etienne, H. (2021). The future of online trust (and why Deepfake is advancing it). AI Ethics 1, 553–562. https://doi.org/10.1007/s43681-021-00072-1

European Commission. High Representative of the Union for Foreign Affairs and Security Policy. (2022). Joint communication to the European Parliament, the European Council, the council, the European economic and social committee and the committee of the regions on the defence investment gaps analysis and the way forward.

European Union. (2021). EU annual reports on human rights and democracy. https://www.eeas.europa.eu/eeas/eu-annual-reports-human-rights-and-democracy_en.

Feng, S., Wang, X., Wang, Q., Fang, J., Wu, Y., & Li, Y. (2018). The uncanny valley effect in typically developing children and its absence in children with autism spectrum disorders. *PLoS One*, 13(11), e0206343

Goodfellow, I., et al. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144. https://doi.org/10.1145/3422622

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 US presidential election. *Science*, 363(6425), 374–378. https://doi.org/10.1126/science.aau2706

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., & Alahi, A. (2018). Social GAN: Socially acceptable trajectories with generative adversarial networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.2018.00240

Harsin, J. (2018). Post-Truth and Critical Communication Studies. *Oxford Research Encyclopaedia of Communication*. https://doi.org/10.1093/acrefore/9780190228613.013.757

High-Level Independent Group on Artificial Intelligence (AI HLEG). (2019a). Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

High-Level Independent Group on Artificial Intelligence (AI HLEG). (2019b). Policy and investment recommendations for trustworthy Artificial Intelligence. https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence.https://doi.org/10.1002/9781119471509.w5GRef225

Hwang, Y., Ryu, J. Y., & Jeong, S. H. (2021). Effects of disinformation using deepfake: The protective effect of media literacy education. *Cyberpsychology, Behavior, and Social Networking*, 24(3), 188–193. https://doi.org/10.1089/cyber.2020.0174

Ihde, D. (1990). Technology and the lifeworld: From garden to earth. Indiana University Press.

Islam, N., Mäntymäki, M., & Turunen, M. (2019). Understanding the role of actor heterogeneity in blockchain splits: An actor-network perspective of bitcoin forks. *Proceedings of the Annual Hawaii International Conference on System Sciences*. https://doi.org/10.24251/hicss.2019.556

Kätsyri, J., de Gelder, B., & Takala, T. (2019). Virtual Faces Evoke Only a Weak Uncanny Valley Effect: An Empirical Investigation with Controlled Virtual Face Images. *Perception*. https://doi.org/10.1177/0301006619869134

Laakasuo, M., Palomäki, J., & Köbis, N. (2021). Moral uncanny valley: A robot's appearance moderates how its decisions are judged. *International Journal of Social Robotics*, 13(7), 1679–1688. https://doi.org/10.1007/s12369-020-00738-6

Langguth, J., Pogorelov, K., Brenner, S., Filkuková, P., & Schroeder, D. T. (2021). Don′t Trust Your Eyes: Image Manipulation in the Age of Deepfakes. *Frontiers in Communication*, 6, 26

Latour, B. (2005). Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford: Oxford UP.

Latour, B., & Woolgar, S. (1987). Laboratory Life: The Construction of Scientific Facts. Princeton University Press.

Law, J., & Hassard, J. (eds) (1999). Actor Network Theory and After. Oxford and Keele: Blackwell and the Sociological Review.

Lee, J., & Shin, S. Y. (2021). Something that they never said: Multimodal disinformation and source vividness in understanding the power of AI-enabled Deepfake news. *Media Psychology*, 25(4), 531–546. https://doi.org/10.1080/15213269.2021.2007489

Li, H., John-John, C., & Tan, Y. K. (2011). Towards an effective design of social robots. *International Journal of Social Robotics*, *3*(4), 333–335.

Luhmann, N. (1979). Trust and Power. John Wiley.

Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition*, 146, 22–32. https://doi.org/10.1016/j.cognition.2015.09.008

Maturana, H. R., & Varela, F. J. (1991). *Autopoiesis and cognition: The realization of the living*. Springer Science & Business Media.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*. https://doi.org/10.48550/arXiv.1411.1784

Nassar, M., Salah, K., Ur Rehman, M. H., & Svetinovic, D. (2019). Blockchain for explainable and trustworthy artificial intelligence. *WIREs Data Mining and Knowledge Discovery*, 10(1). https://doi.org/10.1002/widm.1340

Occhipinti, C., Carnevale, A., Briguglio, L., Iannone, A. and Bisconti, P. (2022). SAT: a methodology to assess the social acceptance of innovative AI-based technologies. *Journal of Information, Communication and Ethics in Society*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/JICES-09-2021-0095

Pettit, P. (1995). The Cunning of Trust. *Philosophy and Public Affairs*, 24, 202–225. https://doi.org/10.1111/j.1088-4963.1995.tb00029.x

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information.* Harvard University Press.

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L., Recchia, G., ... & van Der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science*, 7(10), 1–15. https://doi.org/10.1098/rsos.201199

Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience*, 23, 439–452. https://doi.org/10.1038/s41583-022-00587-4

Seyama, J., & Nagayama, R. S. (2007). The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence: Teleoperators and Virtual Environments*, 16(4), 337–351. https://doi.org/10.1162/pres.16.4.337

Seow, J. W., Lim, M. K., Phan, R. C. W., & Liu, J. K. (2022). A comprehensive overview of Deepfake: generation, detection, datasets, and opportunities. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2022.09.135

Sirota, M., & Juanchich, M. (2018). Effect of response format on cognitive reflection: Validating a two-and four-option multiple choice question version of the Cognitive Reflection Test. *Behavior Research Methods*, 50(6), 2511–2522. https://doi.org/10.3758/s13428-018-1029-4

Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, 6(1). https://doi.org/10.1177/2056305120903408

Werbach, K. (2018). The blockchain and the new architecture of trust. The MIT Press. https://doi.org/10.7551/mitpress/11449.001.0001

Zant, T. V. D., Kouw, M., & Schomaker, L. (2013). Generative artificial intelligence. In *Philosophy and Theory of Artificial Intelligence*, 107–120. Springer, Berlin, Heidelberg.