ORIGINAL RESEARCH

The future of online trust (and why Deepfake is advancing it)

Hubert Etienne[1,2,3]

Received: 21 December 2020 / Accepted: 13 June 2021 / Published online: 28 June 2021

Abstract
Trust has become a first-order concept in AI, urging experts to call for measures ensuring AI is 'trustworthy'. The danger of untrustworthy AI often culminates with Deepfake, perceived as unprecedented threat for democracies and online trust, through its potential to back sophisticated disinformation campaigns. Little work has, however, been dedicated to the examination of the concept of trust, what undermines the arguments supporting such initiatives. By investigating the concept of trust and its evolutions, this paper ultimately defends a non-intuitive position: Deepfake is not only incapable of contributing to such an end, but also offers a unique opportunity to transition towards a framework of social trust better suited for the challenges entailed by the digital age. Discussing the dilemmas traditional societies had to overcome to establish social trust and the evolution of their solution across modernity, I come to reject rational choice theories to model trust and to distinguish an 'instrumental rationality' and a 'social rationality'. This allows me to refute the argument which holds Deepfake to be a threat to online trust. In contrast, I argue that Deepfake may even support a transition from instrumental to social rationality, better suited for making decisions in the digital age.

Keywords Trust · Deepfake · Disinformation · Fake news · AI ethics

1 Introduction

Trust has become a particularly trendy concept in AI. Nowadays, most major technology companies claim their commitment to building a 'trustworthy AI' while social media and governments worry about ensuring trust in online information. The European Commission even formed an expert group to write 'ethics guidelines for trustworthy AI' [22] to establish an initial framework for regulating the development of AI in the EU. The danger of untrustworthy AI culminates with Deepfake, often presented as the gravedigger of online trust. This solution, which notably permits to create synthetic videos of existing people, is widely perceived as a deadly threat to democracies, given its potential of serving sophisticated disinformation campaigns [10, 41, 43], manipulate elections [36] and annihilate any trust in online information [40, 42], thereby paving the way for a nihilist post-truth world [13]. But what actually is trust? Usually left aside, this question happens to be trickier than it seems, and its complexity is testament of a rich evolution in its theory and manifestation across societies.

In this paper, I mobilise anthropological and philosophical theories of trust to defend an unconventional position: not only is Deepfake not a threat for online trust, but it could even represent the critical ally we need to promote trust in the digital age. The first section lays out the original dilemma of building trust, presents the solution found by traditional societies, and how trust evolved across political systems up to modern theories thereof—leading me to formulate three conclusions on trust. The second section criticises the modern rational theories of trust, presenting three main arguments against the suitability of the rational choice theory to model trust and prompting me to consider an opposition between two types of rationality. The third section breaks down the argument justifying Deepfake as a unique threat to online trust, and individually refutes its three components. It then provides reasons for switching from instrumental to social rationality when making decisions in the digital age, and explains how Deepfake supports such a transition.

✉ Hubert Etienne
hubert.etienne@sciencespo.fr

[1] Facebook AI Research, Paris, France

[2] Department of Philosophy, Ecole Normale Supérieure, Paris, France

[3] Sorbonne Université, LIP6, Paris, France

2 The social dilemma of trust: from the enabling of positive circularity to the management of risks

2.1 Solving the social dilemma of trust to enable positive circularity

The apparent familiarity we cultivate with the concept of trust justifies the awkwardness experienced when it comes to defining it, a question we might be tempted to answer in the Augustinian way: 'If nobody asks me, I know: but if I were desirous to explain it to one that should ask me, plainly I know not' ([5], 239). We may then try to approach the question through other angles: Whom do we trust? Someone 'trustworthy'. When do we trust them? When trust is needed. However, and as intuitive and circular as these replies can be, not only are they useless in grasping a better understanding of trust, but also can be widely refuted by the reality of social interactions. Obviously, I do not mean the same thing when saying 'I trust people not to murder me when I walk in the street', or 'I trust the news published by the Guardian' or 'I trust my friend to keep my secret'. While I have no idea whether people on the street are actually trustworthy, in many cases, I also have no need to confide a secret. The etymology of the French word for trust (*confiance* deriving from the Latin *confidentia*) then permits to rationalise the semantic explosions of expressions around one original word, *fidere*, from which derives *se fier à quelqu'un* (trust someone), or *confier quelque chose à quelqu'un* (entrust someone with something) who is *fidèle* (faithful) or *digne de confiance* (trustworthy). More importantly, it reveals a key connection between *fidere*—which has also provided *avoir foi en* (have faith in)—and *credere*—from which derives *croire* (believe), *donner credit à* (give credit to) or *crédible* (credible)—as trust somewhat involves the mediation of a transcendent order.

The relationship between inter-human trust and faith in transcendent divinities was found by anthropologists, when investigating the first dilemma traditional societies had to overcome to exist—trust referring to both an absolute necessity and a practical impossibility. The existence of a social system is conditioned by the development of non-destructive interactions between different communities, including the exchange of goods through the *gifts and counter-gifts* logic [32] or family members through the *alliance theory* [29]. These interactions do not only result from individual decisions but are mainly enabled and driven by wider circular dynamics at the social system's scale, which enjoys a certain autonomy over that of its members. These dynamics can then be either positive as illustrated by the gift and counter-gift theory (the ontological debt received by someone when accepting a gift triggers a whole dynamic of positive indirect reciprocity), or negative as exemplified by the autogenerative logic of vengeance (the duty to murder the one who has murdered creates a new murderer, calling for a new vengeance and resulting in an endless negative reciprocity of violence) and the development of society is conditioned by its success in finding means to defuse negative reciprocities and enable positive ones [4]. In both cases, this implies a passage to the transcendental level and an authentic belief.

Sacrifices of goods (*potlatch*) and people (sacrifices) allows containment of the effects of the 'mimetic desire' [19], within and between tribes according to René Girard, resulting in the production of divinities with whom the group can establish a relationship through cathartic rituals, thus preventing its own destruction. For Mark Anspach, the power a group acquires over itself to counter the dynamic of revenge is given by the reification of vengeance itself, and the possibility of pre-empting the sacrifice: killing an innocent person instead of 'the person who had killed', allows for an exit of the vicious circle of vengeance as illustrated by the story of *Sulkana* in Pays Moussey [17]. Both interpretations converge on the subterfuge developed by traditional tribes to keep violence in check by reifying it as a third party and establishing a ritualised relationship with it, based on a genuine belief, to live together more harmoniously.

Once negative reciprocity is defeated, the establishment of a positive reciprocity for a group to prosper comes by the enabling of a gift economy. Gifts are in many ways like sacrifices—or rather 'auto-sacrifices' as 'we give ourselves in giving' says Mauss ([30], 125),[1] remarking the ontological dimension implied by the material gift which binds people together at the transcendent level—but implies a reversed temporality. Whereas vengeance comes in response to an anterior action (a murder) in the name of justice, the gift anticipates reciprocity and triggers it: it calls for a reaction. This latter cannot be a direct counter-gift to the donor, but instead take the form of a gift to another person, as part of a wider indirect reciprocity scheme at the social group level. 'We do not give to receive in return: we give so that the other also gives' ([25], 1415),[2] that is to say to establish a relationship, which would otherwise be closed by the reception of a direct counterparty.

Here comes the dilemma: if, by definition, the gift has to be spontaneous (i.e. purely disinterested), how can the giver know that it will effectively lead to an indirect counter-gift, and then initiate a virtuous circle of positive reciprocity? From the receivers' view, the spontaneity of the gift seems to convey a double bind: on the one hand an obligation to give something back, and on the other hand the impossibility

[1] My translation of 'on se donne en donnant'.

[2] My translation of 'On ne donne pas pour recevoir: on donne pour que l'autre donne'.

of accepting this message without denaturing the gift itself. The receiver then faces two contradictory signals: a message saying, ‘I present you with a gift’ and a meta-message saying ‘you need to give something back’. This dilemma is then to be overcome by the introduction of a third party, the *Hau* (the spirit of the gift) for the Maori, which ensures the reciprocity of the gift while maintaining its spontaneity, by dissociating the sender of the two messages: the donor sends the message and the third party sends the meta-message [4]. More than just reflecting the social interaction on the meta-level, the third party that emerges from the interaction of the social group transforms it through a mechanism of ‘auto-transcendence’, which enables trust within society as long as they keep faith in the transcendent entity.

2.2 The modern conception of trust grounded in rational choice theory

According to Girard and Anspach, the forms of exchanges and the types of third parties have evolved across ages, but the structure of social trust remains unchanged. Originally established on the belief in spirits, it became faith in a unique God and, by delegation, in its terrestrial lieutenant, the monarch of divine right. With the end of political theology, the advent of Modernity led to a major shift in the perception of social order and the approach of the future. From natural and divine, the social order is now perceived as a human institution resulting from the interactions of free agents with unpredictable behaviours, thus calling for a need to reposition trust on a new basis [12]. The future becomes all the more synonymous with uncertainty that it is no longer ruled by tradition, and that the mode of social interaction progressively switches from ‘familiarity’ to ‘anonymity’ [29]. This is when the contract theoreticians acknowledge the role of the State as a third party to set up the conditions for trust to make social life possible in a territory ruled by law. This exigency first relates to the confidence that any attempt against one’s life [23] or predation against one’s goods [28, 34] would be severely punished, then more generally to the sacralisation of all contracts made legally between individuals in a society.

It is worth noting here that despite the great difference between Thomas Hobbes and Jean-Jacques Rousseau’s anthropology, the mechanisms playing in the construction of such trust are fundamentally rational. With the State being charged to enforce the consequences of individuals’ actions under all circumstances in an impersonal way—sanctioning an assassin is not a personal act of vengeance, but a collective reply to crime against all—the target goal is that the certainty and the severity of the sanctions operate as an *ex ante* regulatory mechanism to discourage attempts to break the law. Rational expectations are at the core of Hobbes’ conception of the State, promising negative incentives to extinguish opportunities to free ride. By doing so, he tends to substitute systematic general distrust (war of all against all) with a systematic common trust (the impossibility of a war of all against all) through a promise of mutually assured destruction, which remains relevant today in nuclear dissuasion doctrine. Rousseau goes even further towards modern economic rational thinking by explicitly presenting his *pactus associatis* under a costs and benefits scheme: ‘What man loses with the social contract is natural freedom and […] what he earns is civil liberty and the property of everything he owns’ ([34], 38).[3]

Nevertheless, Hobbes and Rousseau’s conception of the State as a third party does not relate to trust itself (Hobbes remains suspicious of the State and devises an exit clause in the case it would turn against himself)—but to the conditions for the development of trust between individuals. It also rests on the assumptions that the State has both the right intentions (is not corrupted) and the effective capacity (power) to find contract breakers and to sanction them accordingly. Not only do people lack the same trust in their political systems nowadays, but all betrayals are also not illegal, as falling under the State’s jurisdiction. The need to refine the theory of trust to decentralise it from the orbit of the State, and to extend the scope of social interactions it can account for, found in the rational choice theory a promising pathway.

Anthony Giddens notes that ‘the first situation calling for a need to trust is not the absence of power, but the insufficiency of information’ ([18] 1990, 40),[4] or rather a situation of imperfect information between full omniscience and perfect ignorance, as ‘who knows everything does not need to trust, who does not know anything cannot reasonably trust’ adds Georg Simmel ([37], 356).[5] In addition to being a poor substitute to good information, trust would be entirely finalised, characterising relationships between rational agent who only trust each other when they have an interest to, expect some benefits for themselves [16], and anticipate a rational interest from others to be trustworthy in the right way, at the right moment [21]. The influence of rational choice theory has been so important that trustworthiness nowadays seems to be associated with a simple absence of rational antagonist interests, like when situations call for the arbitration of a third party, supposed ‘trustworthy’ on the sole basis it has no a priori direct interest at stake.

[3] My translation of ‘Ce que l’homme perd par le contrat social, c’est sa liberté naturelle & […] ce qu’il gagne, c’est la liberté civile & la propriété de tout ce qu’il possède’.

[4] My translation of ‘la première situation exigeant un besoin de confiance n’est pas l’absence de pouvoir, mais l’insuffisance d’information’.

[5] My translation of ‘celui qui sait tout n’a pas besoin de faire confiance, celui qui ne sait rien ne peut raisonnablement même pas faire confiance’.

From a precious social good, trust would have become a blemish associated with a situation of involuntary vulnerability in a context of poor information to be avoided at all costs. As a mechanism to control risks and uncertainty, we would then mainly reach to trust in case of *strict necessity*, in a situation of *alignment of rational interests* (e.g. when walking safely on the street without expecting anyone to assault me, or when a creditor lends money to a debtor), or *by convenience*, as a mixture of both of these two: when I trust a doctor to perform a medical surgery, assuming that filling the competency gap to do it myself would be much too costly. This is also the case when I trust a newspaper to convey news that is properly verified by its columnists, supposing the business interest of the company in only sharing good quality information, and assuming that fact-checking everything myself would have a higher cost than the value of the information itself.

3 The instrumental rationality of reliance and the social rationality of trust

All these approaches grounded in rational choice theory, however, should be rejected for three reasons. First, they are based on a confusion between trust and expectation. Second, they are invalidated by the reality of social interactions. Third, they fail to recognise trust as an objective in itself.

3.1 The independence of trust with the degree of information

There is no harm in recalling that rational agents are only a radical simplification of human's decision-making process. The theory is based on a tripod, including a recursive metacognitive knowledge (the agent makes decisions based upon certain principles and is aware of this cognitive process), a projective metacognitive knowledge (all other agents are also supposedly rational, thus making decisions based on the same principles) and information about the evolution of the system (deriving from the observation of the environment and other agents' behaviours). The financial markets, in theory, relatively suit this description and this is why rational choice theory can be helpful here to model economic behaviour. Such an environment is said to be relatively efficient because all agents are supposed to access the exact same information, process it in a similar way and aim for the same unique objective: profitability. However, there is no trust playing in the market, only rational decisions made on the basis of available information proceeding from more or less long-term strategies and more or less risk-aversion. Irrational perturbations are said to come from non-professional investors (the famous fear of the trading housewife), human mistakes or psychology (fat finger, aversion loss biases) and market abuses (rumours, inside trading), that is to say from human factors, justifying their replacement by algorithmic trading. In real life, people are only partially rational as illustrated by the extensive literature on cognitive biases (e.g. [24]), the extent of their desires largely exceeds that of their economic interests and the dynamics at stake in social interactions are much more complex than the macroeconomic laws of the market. This is what prompts Jean-Pierre Dupuy to conclude that "the concept of 'equilibrium' imported from rational mechanics by the market theory is not suitable to characterise the 'attractors' of mimetic dynamics" ([15], 71), playing at the heart of social systems.

A rational agent is, by definition, purely rational. It makes decisions based on available information, which is processed through calculation rules, aiming for expected consequences that maximise its objectives. Its cognitive process does not vary with the degree of information available, so that a lack of information would automatically make it switch to another decision mode, that of trust. Would we want to integrate trust relationships between agents in a simulation, it would be represented by a variable attributing different weights to each agent, modifying the probability distribution for each one to be expected to become adversarial under specific circumstances, or the credibility of their announcements. Such variables would, however, be deemed to remain an externality which the agent cannot access by itself, nor modify, but only receive and integrate in its calculations. In other words, it would modify the agent's rational expectations, not replace them, and if Ludovic does not trust Laurence in general circumstances, he will not start trusting him in a critical situation, where information is lacking. This is why we cannot talk about trust in a situation of *strict necessity*—when an agent's fate is completely dependent on another's will—because there is no choice. We can call this uncertainty and Ludovic may hope that Laurence takes the decision that would be favourable to him, but there is no trust at play here.

Likewise, it would be erroneous to invoke trust in situations where agents perceive they have *aligned interests*. Here again, what is at stake is nothing else than rational expectations because of the metacognitive assumption of rational choice theory: Ludovic predicts Laurence's behaviour because he assumes Laurence is rational, has access to the same information, and also assumes that Ludovic is rational himself. Only the metacognitive assumption enables both agents to realise they have an interest in collaborating to maximize their chances to reach their objective. Hardin and Gambetta's trust then is no more than rational expectations leading to a behavioural 'synchronisation', rather than a trust relationship. One may argue here that a true alliance can exist between agents as 'objective allies', when objectives are sufficiently far away, so that were Ludovic to be temporarily vulnerable, Laurence would refrain from taking advantage of such a situation though he could. However, here again, Laurence would not refrain from benefiting from Ludovic's situation for the sake of loyalty, because he is trustworthy, but solely because the optimum scenario for him satisfying his objectives requires Laurence not to take advantage of it. We cannot even talk about an alliance here—for which Laurence would sacrifice his short-term interests to keep Ludovic as an ally in the long run to increase his chances of reaching a higher gain—because there is no such thing as an alliance or retaliation for purely rational agents, but only synchronisation. In fact, Laurence could be entirely opportunistic, taking advantage of Ludovic's weakness if he had interest in that. This would not change the so-called cooperating strategy in the future. Such reasoning is that of the efficient breach of contract theory defended by the judge Richard Posner [33], as part of the Law and Economics doctrine, which is entirely grounded in the rational choice theory. Situations change, interests aligned yesterday are not necessarily still aligned, and there should be no hard feelings in breaking former engagements, at the cost of potential penalties, would this allow the agent to reach a higher level of utility.

Finally, the observation of social interactions reveals a greater complexity for trust relationships than what these theories could describe—also suggesting we have a limited power over our relationship to trust. Some people have a capacity to trust easily while others are more mistrustful, some naturally inspire more trust than others, and these distinctions cannot be attributed to a variation in rationality. We also tend to offer some people our trust on the basis of very little knowledge, for reasons which do not even seem rationally grounded, often in an involuntarily and even unconscious way [6]. There are many examples of situations where we give our trust, although it is not in our interest to do so—e.g. when telling a friend a terrible secret they could use against us with no apparent benefit in telling them.

3.2 The fundamental distinction between trust and reliance

To Simmel's argument that we cannot reasonably trust when we do not know anything, some have argued that we, however, tend to trust a doctor we just met for non-trivial decisions. Relying on a doctor's prescription does not, nevertheless, mean that we completely abandon ourselves to their goodwill [31] and this is why the term 'reliance' is often preferred, considered as a weak degree of trust [8]. I reject the idea that such reliance would be of the same order as trust, only differing in degree. The reliance is here based on perfectly rational information (white blouse, people in the waiting room, doctor listed on the official register, etc.), so here again we are facing rational expectations made on limited information. Just like I do not trust the barriers to cross the railway safely when they are open, but only process it as a signal which leads me to expect no train should arrive immediately, I do not trust, but only expect, someone who looks like a doctor in a place which looks like a medical office to be one. In such situations of *convenience*, we do comply with the paradigm of rational choice theory, and our decision to abide by the doctor's advices does not proceed from trust, but from rational expectations. Doing so, we do not so much *rely* on the doctor, rather than *rely* on our conception of the world, just like I expect it to be more likely to be assaulted by a gang member and to have a preppy-looking young person bring me back my lost wallet, than the contrary. Were the opposite to happen, I would certainly be surprised because my assumptions would have been proven wrong, but not feel betrayed as no trust was involved.

Ultimately, Simmel is, however, right in saying that we cannot 'reasonably trust', because trust is beyond reason, or more precisely beyond this rationality. This is particularly clear when considering his second argument, according to which someone who knows everything does not need to trust. In fact, not only is trust disconnected to the level of information, but it often competes with it. Only someone with all evidence of a crime against them would tell their friends 'I am innocent, you have to trust me'. People choosing to trust their partner again, although these latter were proven untrustworthy by cheating on them several times, clearly does not reflect a rational behaviour, but a wish to repair a relationship. This is because trust is not a matter of reason and is even most spectacularly exhibited when one puts someone else's words above all other contradictory information they may have, to make a decision against all rational expectations.

Trust reflects an alternative mode of decision-making, resulting from both a choice to put oneself in a situation of voluntary vulnerability (as trust always comes at the cost of the possibility of being betrayed) by putting someone's words above any other information, and a desire to build a relationship with this person. This is precisely because humans are only partially rational agents that they are capable of trust, which permits them to transcend rationality to make decisions towards a greater goal, which is ultimately social, not purely individual. Trust abides by a mode of decision-making which may seem irrational when considering particular decisions such as short-term transactional relationships where the incentive to betray can be high and the cost small. It however becomes perfectly rational, as soon as trust is not considered anymore as only a means to an end, but also as an end in itself, recognizing the building of relationships as an objective. Let us refer to these two rationalities as instrumental rationality and social rationality. The former refers to the mode of reasoning of the rational agent as previously defined, whose cognitive process is entirely directed towards the making of decisions for the purpose of maximising private interests. This is the mode of reliance and rational expectations. The latter is a mode of reasoning closer to a vertigo of reason. It plays a role in decision-making, principally in competition with the outputs of instrumental rationality, but its end is not to be exhausted in an action theory or a theory of knowledge. While the instrumental rationality gathers data to produce knowledge (which can be an end itself for the individual) or to make better decisions against others, the social rationality is to be satisfied by the sole existence of trust relationships, *id est* by the simple fact to be relating to others in a certain way, conceiving social integration as an end itself. Whereas it is certainly true that the decline of familiarity coming with modernity led to the need to rethink our relationship with trust, it however, did not change its principle. From this perspective, modernity may rather have brought a need for new ways to develop meaningful relationships in an environment of unfamiliarity, rather than to preserve oneself against risks. Instrumental rationality is what allows humans to survive in the Hobbesian state of nature. Social rationality is what enables them to flourish in society.

Three conclusions then arise about trust. First, it is a choice which necessarily implies putting oneself in a situation of voluntary vulnerability for the purpose of social integration. While one can prove oneself trustworthy over one's past choices, this is only possible if they have also been given the possibility to deceive us. Trust can then only be won after it was given and accepting the possibility of being betrayed is necessary to enable the possibility of developing trust relationships. Second, trust cannot be captured by rational decision theory, and it is most strongly experienced precisely when it dictates a behaviour opposed to the recommendation of the decision-making process, based on rational expectations. It does not follow that trust is irrational, but rather that it abides by another type of rationality, as an alternative mode of reasoning dedicated to the building of strong social relationships, even at the cost of truth, efficiency or one's own life. Third, given that trust derives from social rationality and is necessarily associated to the possibility of betrayal, implying intentionality, trust can only characterise relationships between agents provided with a free will. This is why we cannot be betrayed by false news or a broken chair, but only by their personified source or manufacturer.

4 Deepfake promotes online trust instead of ruining it

4.1 Deepfake is not a threat for democracies

Deepfake is a computer vision technique, using deep learning methods to generate synthetic images for the purpose of reproducing the features of a source image in a target image. It was principally mediatised with the Face2Face project [39] and the Synthesizing Obama project [38]. It has since found applications in a wide range of domains from internet memes to art and the cinema industry. However, the applications which have caught the most attention were those related to political contents, such as the fake videos speeches of Boris Johnson and Jeremy Corbyn released by the think tank Future Advocacy in the context of the UK's 2019 general elections.[6] In 2020, the activist group Extinction Rebellion released a fake video of the Belgian Prime Minister, Sophie Wilmès, suggesting a possible link between deforestation and Covid-19.[7] These highly mediatised examples, together with the rapid improvement of Deepfakes' performance, fed a great concern within the AI ethics community: we may soon be incapable of distinguishing machine-generated content from real content, leaving us vulnerable to sophisticated disinformation campaigns for the purpose of elections manipulation. By preventing us from trusting anything online, Deepfake would thus bring disinformation techniques to their paroxysm and even pave the way to a post-truth world, characterised by an unprecedented relativism and a systematic distrust. Although this concern a priori seems legitimate, I am now to show that it has no solid ground. The argument can be broken down as such:

(1) Deepfake's performance represents an unprecedented potential for information manipulation
(2) The major issue deriving from Deepfake relates to disinformation and election manipulation
(3) Used as such, Deepfake could then definitely ruin online trust

With regards to the first part of the argument, we shall indeed concede that Deepfake techniques are improving rapidly and that it will certainly soon be impossible for a human being to discriminate between synthetic and non-synthetic content without computer support. However, Deepfake is neither the first, nor the most effective technique of information manipulation. Ancient Greece's rhetoricians were already using a vast range of sophistic techniques to convince or persuade an auditorium, and selling their art to the wealthy Athenian youth, preparing it for the practice of power in democracy. Since Plato, we tend to dissociate truth from eloquence in political discourses, and, however, rhetoric is still taught in political schools and shapes every public allocation. Dupuy [15] for instance explains how the argument of the reversal of the burden of proof is used to reject the application of the Precautionary principle theory in the innovation domain, while it is based on a *petitio*

[6] https://futureadvocacy.com/deepfakes/.

[7] https://www.extinctionrebellion.be/en/tell-the-truth.

principii. Considered as one of the founders of public relations, Edward Bernays [7] even considered 'propaganda' as a necessity for political systems following universal suffrage, 'to create order from chaos' (1928, 141),[8] and the manipulative strategies he developed notably permitted him to persuade American women to smoke, for the benefit of the American Tobacco Company.

Claims may be subjective, discourses misleading and communication campaigns deceiving, but facts are facts one may say. However, facts are always captured within a certain context and conveyed in a certain way, which supports a particular vision of the world. Besides outright lies, common misleading techniques used on social media for disinformation purpose include real photos, videos and quotes either truncated, or shared outside their original context, suggesting that the battle for accurate information rages in the field of misleading suggestions, rather than that of factual accuracy. In another context, finance workers excel in the art of presenting univocal data in different ways, highlighting some aggregate instead of others or changing the scale of the graph to modify the shape of the curve, to support the story they aim to tell. Public administrations also demonstrate a great ingenuity in this domain when communicating on the performance of their actions to reduce the unemployment rate [35] or when soliciting polling institutes to build a public opinion suitable for the political measure they aim to enforce [11]. Finally, even a purely factual message will certainly not produce the same impact, whether I say 'George died yesterday at 3 pm' or 'Yesterday, a Black American citizen was murdered by a White police officer', which prompted Friedrich Nietzsche to claim that there are no facts, only interpretations. It results from what precedes that Deepfake should only be considered as one trick among others within the large spectrum of manipulation techniques. Some have come to consider it as an evolution rather than a revolution in the history of manipulation techniques [43] and I would add that deepfakes may be even easier to counter, as they relate to a question of factual accuracy (did X really pronounce this discourse or not?), rather than vicious misleading suggestions.

The second part of the argument states that Deepfake's greatest issue relates to disinformation and elections manipulation. It explains why the main efforts to address its potential negative impacts have so far not been focused on the regulation of its uses, but on the detection of synthetic content. This is illustrated by Facebook's Deepfake Detection Challenge,[9] Google's open-sourced dataset of Deepfake videos for training purposes[10] or Microsoft's Video Authenticator,[11] all initiated for the explicit purpose of supporting the fight against misinformation. Deeptrace's 2019 report states that of the 15,000 deepfakes found online and analysed by the researchers, 96% of these were actually pornographic, principally representing fake videos of celebrities used without consent [2]. The application DeepNude is already leveraging Deepfake techniques to monetise the undressing of women, offering on demand services to reconstruct a naked body from a given picture. In addition, a second report from Deeptrace (now Sensity) revealed in October 2020 the existence of a deepfake robot operating on Telegram, which has been used to strip c.700,000 women, with over one hundred thousand of them being publicly shared on the social media, warning against the dangers of such robots being 'weaponized for the purpose of public shaming or extortion-based attacks' ([1], 8). These two applications obviously raise major concerns for privacy, personal image property and reputation. In contrast with the fear of Deepfake's potential use for disinformation, they epitomise the reality of Deepfake's uses today, opposing confirmed dangers and present victims to hypothetical risks, and truly constitute a novel and unique threat for people's privacy.

The third component of the argument finally states that, were Deepfake's performance to become sufficiently advanced to make the detection of synthetic contents impossible and be used for the purpose of disseminating false news, it could then deal a fatal blow to online trust. As previously said, we never trust a piece of information, but always the moral person responsible for it (I would say here either an individual or a community of people), as the content itself cannot be granted a proper intentionality. The whole question then is that of the definition of *online* trust.

If what is meant by this is that believing the false news spread against their political representatives, people would end up losing faith in them and come to distrust political institutions as a whole, I would reply that such a state of general distrust already exists and is not to be attributed to Deepfake. It is not because of Deepfake that a great number of U.S. citizens have little trust in their political representatives, with 81% of them believing that members of Congress behave unethically some or most of the time [32], nor that three quarters of the French population considers members of their political representatives corrupted [26]. Arthur Goldhammer and Pierre Rosanvallon [20] wrote a whole book titled *Counter-Democracy: Politics in an Age of Distrust* in 2008 to investigate the reasons for the general

[8] My translation of "pour créer de l'ordre à partir du chaos".

[9] https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/.

[10] https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html.

[11] https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/.

level of distrust ten years before the first deepfakes. Disinformational deepfakes could then indeed take advantage of the generalised level of distrust which leaves people more vulnerable to anti-elite hoaxes, and transform what used to be perceived as impossible, now as improbable. Yet, technology should not serve as a scapegoat, taking the blame for a political issue, which first and foremost calls for political change. The confirmation bias regularly cited as a key vector of false news spread would hereby not play such a significant role if people were to consider their representative trustworthy.

4.2 The role of Deepfake in promoting trust for the digital age

On the other hand, if what is to be understood by 'online trust' is rather a sort of general 'reliance' in cyberspace, as an environment to collect accurate information and develop authentic interactions, then Deepfake will certainly constitute a challenge. Just like our biological senses, which allow us to inhabit a world by collecting information about our environment to make decisions that will define our interactions with its different entities, we are using digital technologies more and more as a digital sense to inhabit cyberspace. However, our biological senses can deceive us as famously argued by René Descartes (1641) [14], and we should be aware of their imperfectability as sources of knowledge when facing an optical illusion or a mirage in the desert. Likewise, not only can our digital sense deceive us, but there also are 'evil demons' actively seeking to fool us in cyberspace. It is thus of paramount importance for us, not only to be aware of this, but also to process it and make an informed use of this sense, just like someone can learn to live with the contradictory messages of a phantom limb. To this end, I believe Deepfake can be of great help, by training us to not just passively believe the signals received by our sensitive entries, but rather raising our critical mind and actively searching for a trustworthy source of information.

A century ago, a photograph would have been considered as irrefutable proof, while it does not prove anything today, since photo editing software is available to anyone. From dynamic pricing, persuasive design and nudging strategies, manipulation techniques are already highly sophisticated, raising most digital technologies to the status of captological interfaces. GPT-3 [9] can also produce convincing human-style articles, which could be used for the purposes of deception. Tomorrow, the internet of things will multiply the number of connexion points with cyberspace present in individuals' ecosystems and the possibilities of virtual reality and invasive technologies such as brain–computer interfaces will doubtlessly make Deepfake seem like prehistoric techniques. For this reason, it is vital to train our brain to overcome the passive credulity we have in videos, in the same way that we familiarised it with photos, and to treat entry points to cyberspace as bargaining spaces, considering that even access to information has become an adversarial game. Only in such a way will it be possible for us to free ourselves from the drifts of Bernays' 'invisible government' ([7], 31), and establish a true bilateral dialogue between the people and their decision-makers: 'public opinion becomes aware of the methods used to model its opinions and behaviours. Better informed about its own functioning, it will exhibit all the more receptivity to reasonable announcements aligned with its own interests […] If it formulates its commercial demands more intelligibly, companies will satisfy its new exigencies' ([7], 141).[12]

Instead of harming trust, Deepfake could on the contrary promote it and help us prepare ourselves for the challenges of the digital age. This calls for a communication effort to inform the public opinion about the performance of such deceiving techniques. It also requires acknowledging that Deepfake cannot only be used to put someone's worlds in someone else's mouth, but also offers an alibi for people to deny the veracity of embarrassing words from an accurate recording, or to fake someone's identity in a meeting. Still, ceasing to believe in everything does not result in distrusting everyone, and this is why social relativism on truth does not necessarily lead to nihilism on trust. Reducing our passive systematic benevolence towards all information coming from cyberspace should also lead us to search more actively for trustworthy sources and redesign the map of our trust relationships around a network of key people. With the condition of securing authentic identification together with information traceability—for instance through blockchain solutions to rapidly identify the original source of a piece of information—we should observe the emergence of a new kind of authority, personified by actors sharing well-verified information on a regular basis. Both journalists and influencers, these new actors will not only be considered as reliable based on the history of accurate information they have shared in the past—and always at risk of losing this reliance capital by the sharing of one single piece of false news [3]—but also trusted because of the personal engagement underlying their articles, exposing their individual reputation to public shaming in case of failure perceived as betrayal.

The reputation cost was already discussed by Hobbes [23] (1651, 239), who makes it an argument against the posture of the 'fool', a pure *homo oeconomicus* with no consideration for past conventions and making decisions based solely on its immediate interests. This emphasises the distinction

[12] My translation of 'le grand public prend conscience des méthodes utilisées pour modeler ses opinions et ses comportements. Mieux informé de son propre fonctionnement, il se montrera d'autant plus réceptif à des annonces raisonnables allant dans le sens de ses intérêts […] S'il formule plus intelligemment ses demandes commerciales, les entreprises satisferont ses nouvelles exigences'.

between the *conditions of social trust* which are enabled by the State as a reliable third party, allowing people to interact safely, and *trust* itself, which relates to individuals' reputation. Although such a posture is irrational for Hobbes—who considers the risk too high and synonymous with social suicide, as nobody would be disposed to contract with the fool anymore—it was, however, still possible for someone with a bad reputation to flee their city or country and start a new life with the money made on the betrayal in Hobbes' times. The opportunities for such efficient-breach-of-contract-like postures are less and less possible in the age of the 'global village', when a stranger's reputation can quickly be verified on Google from virtually any country in the world.

This is how Deepfake, by challenging the passive reception of the signals received from our senses and confusing our appreciation of the possible and the probable, may increase the need for trust and thus prepare us to navigate in the digital age while avoiding manipulative enterprises. Left incapable of reasonably relying on received information, such a constructive scepticism may then force us to build a network of trustworthy relationships with personified nodes, engaging their reputation to ensure the integrity of the structure. In such a configuration, we would then ground our judgment less and less on the processing of our increasingly imperfect perception of the world by our also imperfect instrumental rationality, and more and more on the social rationality of trust, enabled by the feedback loop of public shaming, and substituting social faith to increasingly impossible rational expectations.

5 Conclusion

Questioning the widely accepted assumption that holds Deepfake to be a threat to democracy, due to its potential to back sophisticated disinformation campaigns and bury the conditions of possibility for online trust, I exposed here the reasons which prompt me to reject it in the absence of solid ground, and to consider, on the contrary, that Deepfake could help promote trust and prepare us for the digital age. I started by recalling the social justification of trust in traditional societies, as a necessity enabling the positive reciprocity of any social life, followed by the introduction of a sacred third party to solve the social dilemma of the gift. After the progressive institutionalisation of this third party leading to the modern conception of the State, I introduced the contemporary theories of trust based on rational choice theory to make decisions in situations of involuntary vulnerability associated with a lack of information. I then rejected these on the ground that they are based on a confusion between expectations and trust, are invalidated by the reality of social interactions and fail to understand trust not only as a means to an end, but also as an end in itself. This led me to formulate a distinction between instrumental rationality, based on perceived reliable information to formulate rational expectations, and social rationality of trust which goes beyond an action theory to associate an information processing with a social end for self-realisation. I finally countered the claim that Deepfake poses a unique threat to democracies, arguing that it is only a manipulatory instrument among others and likely not even the most efficient one, that the real issues it raises relate to privacy, not misinformation, and that it ultimately does not challenge trust but reliance in information perceived digitally. At a time when digital perception simultaneously grows in importance and uncertainty, as we are increasingly experiencing our world through the mediation of cyberspace—which also is a competition space between powerful actors with sophisticated manipulative instruments for the shaping of reality—Deepfake can help us enhance our collective critical mind to reduce our gullibility towards false news and promote source verification. In the long run, it can also help us conduct the deliberate choice of shifting from instrumental rationality towards the social rationality of trust, considering faith in people as a more viable way to ensure one's self-realisation within a coherent network of trusted individuals, than reliable information-based expectations.

References

1. Ajder, H., Patrini, G., Cavalli, F.: Automating image abuse: deepfake bots on telegram. Sensity (2020)
2. Ajder, H., Patrini, G., Cavalli, F., Cullen L.: The state of deepfakes: landscape, threats, and impact, deeptrace (2019)
3. Altay, S., Hacquin, A.-S., Mercier, H.: Why do so few people share fake news? it hurts their reputation. (2019). https://doi.org/10.1177/1461444820969893
4. Anspach, M.R.: A charge de revanche: figures élémentaires de la réciprocité. Seuil, Paris (2002)
5. Augustine [ca. 397]: Confessions, Volume II: Books 9–13. Ed. & trans. by Hammond C.J.-B., Harvard University Press, Cambridge (2016)
6. Baier, A.: Trust and anti-trust. Ethics **96**(2), 231–260 (1986)
7. Bernays, E.: [1928] Propaganda. Zones, Paris (2007)
8. Blackburn, S.: 'Trust, cooperation and human psychology. In: Braithwaite, V., Levi, M. (eds.) Trust and Governance. Russel Sage, New York (1998)
9. Brown, T.B., Mann, B., Ryder N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G, Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners'. Preprint at arXiv:2005.14165
10. Boneh, D., Grotto, A.J., McDaniel, P., Papernot, N.: Preparing for the age of deepfakes and disinformation, Stanford HAI (2020)
11. Bourdieu, P.: L'opinion publique n'existe pas. Les temps modernes **318**, 1292–1309 (1973)
12. Le Bouter, F.: 'Formes et fonctions de la confiance dans la société moderne. Implications philosophiques. http://www.implicatio

ns-philosophiques.org/actualite/une/formes-et-fonctions-de-la-confiance-dans-la-societe-moderne/ (2014). Accessed 21 Dec 2020

13. Chesney, R., Citron, D.: Deepfakes and the new disinformation war. The coming age of post-truth geopolitics. Foreign affairs. https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war (2019). Accessed 21 Dec 2020
14. Descartes, R.: [1641] *Méditations métaphysiques*. Flammarion, Paris (2011)
15. Dupuy, J.-P.: Pour un catastrophisme éclairé. Seuil, Paris (2002)
16 Gambetta, D.: Trust. The Making and Breaking of Cooperative Relations. Blackwell, Oxford (1988)
17. de Garine, I.: Foulina: possédés du pays Mousseye, documentary. CNRS Images, Meudon, cop. 1966 (2005)
18. Giddens, A.: [1990] Les conséquences de la modernité, trans. By Meyer O. L'Harmattan, Paris (1994)
19 Girard, R.: Mensonge romantique et vérité romanesque. Grasset, Paris (1961)
20 Goldhammer, A., Rosanvallon, P.: Counter-Democracy: Politics in an Age of Distrust. Cambridge University Press, Cambridge (2008)
21. Hardin, R.: Communautés et réseaux de confiance. In: Ogien, A., Quéré, L. (eds.) Les Moments de la confiance. Economica, London (2006)
22. High-Level Expert Group on AI (HLEGAI): Ethics Guidelines for Trustworthy Artificial Intelligence. European Commission, Brussels (2019)
23. Hobbes, T.: [1651], Léviathan, ou la Matière, la Forme et la Puissance d'un Etat ecclésiastique et civil, trans. by Mairet G. Gallimard, Paris (2000)
24. Kahneman, D., Slovic, P., Tversky, A.: Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press, New York (1982)
25. Lefort, C.: 'L'échange et la lutte des hommes. Les Temps modernes, 6 (1961)
26. Lévy J.-D., Bartoli, P.-H., Hauser, M.: Les perceptions de la corruption en France. Harris interactive (2019)
27. Lévi-Strauss, C.: [1949] *Les structures élémentaires de la parenté*. De Gruyter Mouton, Paris (2002)
28. Locke, J.: [1690] Traité du gouvernement civil, De sa véritable origine, de son étendue et de sa fin, trans. by Mazel D. Calixte Volland, Paris (1802)
29. Luhmann, N.: La confiance : Un mécanisme de réduction de la complexité sociale, trans. by Bouchard S. Economica, Paris (2006)
30. Mauss, M.: Essai sur le don, forme et raison de l'échange dans les sociétés archaïques. In: L'année sociologique, t. I, Mauss M. (dir.), Paris, Librairie Félix Alcan (1925)
31. Michela, M.: Qu'est-ce que la confiance? Études **412**(1), 53–63 (2010)
32. Pew Research Center (PRC): Why americans don't fully trust many who hold positions of power and responsibility (2019)
33. Posner, R.: Economic Analysis of Law. Wolters Kluwer, Alphen aan den Rijn (1973)
34. Rousseau, J.-J.: Du contrat social ou Principes du droit politique. Marc Michel Rey, Amsterdam (1762)
35 Salmon, P.: Chômage. Le fiasco des politiques. Balland, Paris (2006)
36. Schwartz, O.; You thought fake news was bad? Deep fakes are where truth goes to die. The Guardian. https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth (2018). Accessed 21 Dec 2020
37. Simmel, G.: Sociologie. Etude sur les formes de la socialisation. Presses Universitaires de France, Paris (1999)
38. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. ACM Trans. Graph. **36**(4), 95:1-95:13 (2017)
39. Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2Face: real-time face capture and reenactment of RGB videos. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2387–2395. Las Vegas, NV (2016)
40. Toews, R.: Deepfakes are going to wreak havoc on society. We are not prepared. Forbes. https://www.forbes.com/sites/robtoews/2020/05/25/deepfakes-are-going-to-wreak-havoc-on-society-we-are-not-prepared/#21b35b157494 (2020). Accessed 21 Dec 2020
41. Turton, W., Martin, A.: How deepfakes make disinformation more real than ever. Bloomberg. https://www.bloomberg.com/news/articles/2020-01-06/how-deepfakes-make-disinformation-more-real-than-ever-quicktake (2020). Accessed 21 Dec 2020
42 Vaccari, C., Chadwick, A.: Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social Media + Society **6**(1), 205630512090340 (2020)
43 Whyte, C.: Deepfake news: AI-enabled disinformation as a multilevel public policy challenge. J Cyber Policy **5**(2), 199–217 (2020)