Synthese (2022) 200:319
https://doi.org/10.1007/s11229-022-03798-5

ORIGINAL RESEARCH

The identification game: deepfakes and the epistemic limits of identity

Carl Öhman[1]

Received: 9 November 2021 / Accepted: 29 June 2022

Abstract
The fast development of synthetic media, commonly known as *deepfakes*, has cast new light on an old problem, namely—to what extent do people have a moral claim to their likeness, including personally distinguishing features such as their voice or face? That people have at least *some* such claim seems uncontroversial. In fact, several jurisdictions already combat deepfakes by appealing to a "right to identity." Yet, an individual's disapproval of appearing in a piece of synthetic media is sensible only insofar as the replication is successful. There has to be some form of (qualitative) identity between the content and the natural person. The question, therefore, is how this identity can be established. How can we *know* whether the face or voice featured in a piece of synthetic content belongs to a person who makes claim to it? On a trivial level, this may seem an easy task—the person in the video is A insofar as he or she is recognised as being A. Providing more rigorous criteria, however, poses a serious challenge. In this paper, I draw on Turing's imitation game, and Floridi's method of levels of abstraction, to propose a heuristic to this end. I call it the *identification game*. Using this heuristic, I show that identity cannot be established independently of the purpose of the inquiry. More specifically, I argue that whether a person has a moral claim to content that allegedly uses their identity depends on *the type of harm* under consideration.

Keywords Synthetic media · Deepfakes · AI · Imitation game · Information ethics

1 Introduction

Recent advances in deep learning have made the production of so-called *synthetic media* increasingly accessible to the public. Synthetic media is an umbrella term that encompasses any content that has been modified or produced by various forms

✉ Carl Öhman
carl.ohman@statsvet.uu.se

[1] Department of Government, Uppsala University, Box 514, 751 20 Uppsala, Sweden

Published online: 25 July 2022

of artificial intelligence (AI). Due to its reliance on deep learning, such content is most commonly referred to as "deepfakes" in the public debate.[1] Applications range from face swapping (Bitouk et al., 2008) to voice conversion (Lorenzo-Trueba et al., 2018), pornography (Öhman, 2019), the superimposition of dead actors into new movies (Minton, 2017) and beyond. A number of open-source software programs, like Faceswap and DeepFaceLab, and several applications—ReFace, Zao, Deepfakes web β and Wombo, to name a few—provide interfaces through which images or videos can be doctored within seconds even by non-experts.

The new accessibility, in combination with the fast spread and rapidly developing precision of synthetic media, has caused deep concerns both within the popular and the academic literature. Some worry that the technology opens the door to new kinds of cyber-bullying, sexual harassment and humiliation (Harris, 2019; Öhman, 2019). Others have raised concerns over the ability of synthetic media to ruin people's reputation by planting false beliefs among the public (Diakopoulos & Johnson, 2021, p. 2081) or even within law enforcement (Yadav & Salmani, 2019, p. 4). Relatedly, some stress the political aspects, warning that synthetic audio and video material falsely starring political figures may pollute democratic discourse to the extent that meaningful deliberation becomes impossible, due to the lack of a shared ground truth (Diakopoulos & Johnson, 2021). Some even go as far as to predict an "information apocalypse" (Warzel, 2018) or "epistemic maelstrom" (Rini, 2020, p. 8), where the (mediated) truth becomes impossible to discern through the constant noise of synthetic content.

Regardless of the relative accuracy of these warnings, the underlying concerns are worth taking seriously. Most people, be they political figures or not, probably feel that they have a natural right to materials that feature their face, voice and/or other personally distinguishing features, and find the mere prospect of losing control over them daunting. That people have at least *some* such moral right seems uncontroversial among philosophers. As stated by de Ruiter (2021, p. 1314) "Protection against the manipulation of digital representations of our face and voice should be considered a fundamental moral right in the age of deepfakes." Many policy makers agree. In fact, both UK and US law already contain frameworks that protect a person's "name, picture or likeness" (Perot & Mostert, 2020, p. 33) from illicit exploitation, and several jurisdictions similarly appeal explicitly to a so-called "right to identity" in their attempt to battle deepfakes (see Mashinini & Africa, 2020, p. 196). However, in order to establish that someone's identity has been *illicitly* used in such a fashion, it must first be established that it has been used in the first place, which is not always straight forward. Deepfakes commonly result in so-called "identity leakage" (Mirsky & Lee, 2021, p. 7), i.e., a mix of multiple people's facial or vocal features blending into each other, and some are so poorly done that identification of a single individual is difficult. It is not always clear whose face is featured in the final product.

This poses a major problem for the supposed right to identity. Assume, for example, that Alice finds a video online depicting an individual partaking in some incriminating activity. Despite never having taken part in such an activity, she recognizes the face of the individual in the video as her own, and thus demands it to be removed. Now, this

[1] Not all synthetic media relies on deep learning. However, to connect with popular discourse, deepfakes will be used interchangeably with synthetic media throughout this article.

demand makes sense if and only if it can be shown that the synthetically generated face featured in the video does in fact belong to Alice. The creator of the video, Bob, may for example counter Alice's demand by claiming that she has mistakenly identified the synthetic face as her own. It could be an actor who looks just like her, or a purely synthetic person, generated by data from multiple individuals, who has no or little correspondence to Alice.[2] If Bob is correct, Alice lacks justification to demand the removal of the content. She simply has no claim to it, since it is not *her* face that is featured in the video. For a content moderator (or court), the ethical/legal challenge of defining the extent of a person's right to their identity thus hinges upon an epistemic one, namely—how do we *know* that the face or voice featured in the synthetic content is indeed Alice's?[3]

To answer this question, we need some form of criteria for how to establish identity between Alice and the person featured in the synthetic content, henceforth referred to as "$\text{Alice}_{\text{Synthetic}}$."[4] The most intuitive approach, perhaps, would be to say that $\text{Alice}_{\text{Synthetic}}$ is Alice to the extent that she is *recognized* as being Alice. If one can *see* that it is her, her demand to remove the content is worth consideration (though there may be other principles, such as freedom of speech, that weigh against it). Another way to frame this is to say that the criterion for how to establish identity between $\text{Alice}_{\text{Synthetic}}$ and Alice is that the former makes a successful *performance* as the latter, because it can *do* what Alice does in some given context. Thus, if $\text{Alice}_{\text{Synthetic}}$ successfully passes as Alice to some intelligent epistemic agent, it *is* Alice, at least insofar we are discussing her right to her face. I believe that such an answer is largely correct. Nevertheless, it remains insufficiently clear. What exactly do we mean by "recognized as"? Assume that Alice is a person who possesses the features X, Y and Z, and $\text{Alice}_{\text{Synthetic}}$ lacks each of these. Depending on what one means by recognition, $\text{Alice}_{\text{Synthetic}}$ may be recognized as "Alice but without feature X" by one person, and recognized as "not Alice *because* she lacks feature X" by another. Moreover, how should we deal with the common occurrence of misidentification? If we are to use recognition as a criterion, it is imperative that we add precision to what that concept means, especially given the pressing ethical concerns arising from synthetic media content.

In this essay, I draw on Turing's imitation game (1950), and Floridi's methodology of levels of abstraction (2008), to propose a heuristic to this end, an approach with which to establish identity between a natural person and her synthetic counterpart. In view of its kinship with Turing's imitation game, I refer to this heuristic as the *identification game.* I propose that $\text{Alice}_{\text{Synthetic}}$ can be identical to Alice at (at least) four different levels of abstraction, referred to here as *levels of identity*. The threshold for each level depends on the information required for a rational human agent to detect a

[2] Given the ever-growing stream of synthetically generated faces, this seems an increasingly plausible scenario.

[3] The same problem applies for both visual and audio content. Alice's face could, for instance, be replaced by her voice, or potentially even other qualities that define her identity.

[4] To be clear, in this context, *identity* refers only to *qualitative* and *not numerical* sameness (see Hall, 1933). That is, we are not interested in whether the synthetic content is or is not an *authentic* depiction of Alice, but whether the content shares enough properties with Alice to establish sameness. Henceforth, "identity" will be used exclusively in this sense.

distinction between a given synthetic artifact and an authentic counterpart. This heuristic, I argue, provides concrete criteria for how to determine whether $\text{Alice}_{\text{Synthetic}}$ is sufficiently similar to Alice to infringe upon her moral right to her identity (irrespective of the former's possible authenticity). Upon introducing this heuristic, I illustrate its practical implications by applying it to three real-life cases.

2 Four types of deepfake harm

To set the stage for the following argument, it is necessary to briefly summarize how deepfakes are said to harm. The literature includes a number of attempts to specify the various ways in which deepfakes may cause harm to individuals and society (Citron & Chesney, 2019; de Ruiter, 2021; Harris, 2019). In this study, however, I refrain from proposing yet another such conceptual framework. My ambition is merely to survey the literature on how deepfakes harm, which, as we shall see, will help specify the utility of the identification game. Drawing loosely upon the distinctions made by Citron and Chesney, (2019) and Kerner and Risse (2021), I propose four categories of harm, the first two relating to individual natural persons, the latter two to institutions.

The first category relates to various forms of *humiliation*. This category is most commonly addressed in discussions of non-consensual deepfake pornography (Maddocks, 2020; Öhman, 2019), but also pertains to other illicit usages of a person's likeness, such as illicit commercial benefit, or hate speech (Young, 2021). It often occurs under the guise of different names such as degradation or violations of human dignity, but largely fall under what Nagel (1970) refers to as "facts of harm", i.e., harms that do not require a subjective experience to be effective. Though the feeling of being humiliated is undoubtedly unpleasant, the harm of humiliation is not limited to that feeling. It is a negative *fact* that transcends the boundaries of one's consciousness.

The second category relates instead to harms to a person's *reputation*. Having one's reputation damaged is an unpleasant experience in and of itself. The subject whose reputation is harmed may feel shame and guilt, despite being innocent of the accusations in question. More importantly though, harms to reputation do de facto affect the subject's opportunities in life. Even if only a small portion of one's social surrounding believes some synthetic incrementing material to be authentic, this may be sufficient to cause serious harm to one's well-being, including loss of personal relationships and job opportunities. As suggested by several studies, deepfakes may cause serious damage to a person's reputation, even when the audience is made aware of its fictitious nature (Bakker, 2020). By contrast to humiliation, these threats fall under what Nagel (1970) would refer to as *states of harm*. They are negative experiences with a direct impact on the health on the individual in question. Though the consequences of having one's reputation damaged may be viewed as their own separate harms, I will, for the purposes of this paper, treat them as consequences of the overarching category of reputational harms.

Related, yet separate, are harms that go beyond the well-being of a single identifiable individual, i.e., harms that are systemic in nature (for a closer look at the entanglement of individual reputation and group level harms, see [Waldron, 2012]). This may include multi-agent systems such as firms and organizations, but is most importantly related to

political ecosystems. Harms to democratic systems can be understood as a particularly serious form of reputational harm that affects representatives of larger groups and thereby undermining a democratic community's grounding in a shared ground truth. For example, a deepfake of a presidential candidate may not only damage his or her personal reputation, but will affect the opportunity of large groups of voters and eventually the entire democratic community of which the candidate is part, even if only part of the population actually believes the material to be authentic (see Diakopoulos & Johnson, 2021 for an elaborate analysis). Indeed, we may even imagine cases where the personal suffering of the targeted individual is limited or insignificant, while the systemic damage is considerable. Hence, political harms are to be viewed as separate, albeit related, to reputational harms.

The final category is what we may understand as *epistemic harms*. This relates to materials that do not merely divide democratic communities, but that destroy the very authority of an entire medium. As argued by Rini (2020), the harm of deepfakes is not merely their deception of audiences. It arises also from the fact that they give everyone reason to doubt video *as a category of evidence*, which may in turn be understood as a consequence of deepfakes reducing the amount of information carried by video materials (see Fallis, 2021 for a more elaborate analysis on this). Without reliable verification of the authenticity of video or audio evidence, not only our political systems, but our entire social infrastructure, including law enforcement, news, research and day-to-day interactions are in jeopardy, threatening to set off an "information apocalypse" (Warzel, 2018).

Notably, only the former two categories necessitate that some form of identity can be established between a natural person and a synthetic counterpart. A deepfake may harm a community, for example, by displaying a generic (albeit synthetic) member of the group taking part in some incriminating act, or by fabricating some sort of event not involving human individuals but, for example, animals, machines or locations. Assume that a far-right activist fabricates a video of a synthetic but generic Muslim committing some sort of heinous crime, which causes national antipathy against Muslims as a collective identity, despite the fact that the individual in question does not correspond to any natural person. Such hypothetical cases illustrate how deepfakes can be harmful even when not targeting a specific individual. However, some form of identity needs to be established between the content and whatever it is that is being depicted. The synthetic Muslim perpetrator in our hypothetical scenario still needs to be recognizable as a Muslim. Moreover, where things stand today, the vast majority of debated deepfakes are using the faces of natural persons to establish such identity. Therefore, my focus henceforth shall be exclusively on harms that arise from deepfakes accused of depicting natural persons, be they degradational, reputational, political or epistemic. The premises for my argument is that the legitimacy of such claims to harm requires an establishment of identity between a natural person and the synthetic counterpart. That is, Alice's claim to have been harmed by the spread of $\text{Alice}_{\text{Synthetic}}$ makes sense only insofar as there is some form of correspondence between them.

3 The insufficiency of technological solutions

It is important to explain why a mere technological criterion for identity is insufficient. There are two ways in which technological means could hypothetically be used. The first is to say that the person in the video is Alice if and only if it can be shown that the application has been trained on Alice's personal data. That is, $\text{Alice}_{\text{Synthetic}}$ is Alice insofar as there are traces of some original personal data in the synthetic content.[5] This provides a concrete criterion and would potentially even be a bridge to copyright legislation. Nevertheless, the approach remains practically unfavorable, since Bob, the creator of the content, may not be able to provide the necessary information. Moreover, if he is indeed guilty of having trained his algorithm on Alice's data, it is unlikely that he would voluntarily provide the information needed to reverse engineer the content. Content moderators, as well as courts, would almost certainly have to do without such information when making a decision on whether or not to remove the content in question and/or to punish Bob for producing and spreading it.

In addition to its practical infeasibility, the detection of data traces is conceptually insufficient to establish identity. Even if it were to be proven that Bob's algorithm was trained on Alice's personal data, it does not automatically follow that the person in the video *is* Alice in any relevant sense. The algorithm may, for example, fail to produce anything that resembles a human face, let alone Alice's. Moreover, the final product may be an amalgamation of several individuals, where Alice's data plays only a small part. We may also imagine cases where the opposite is true; the face in the video could reasonably be identified as Alice *despite* the algorithm not being trained on any of her known data. Her voice may have been recorded in secret, meaning that it is impossible to trace. Or, Alice could just look/sound a lot like some public figure(s). Within her immediate social circle, the synthetic content would be more likely identified as Alice than the figure(s) in question, while the opposite is true for the algorithm. In sum, traces of some original data is ultimately a poor criteria to assess identity.

The second alternative is to simply use algorithmic facial recognition software to determine to what extent $\text{Alice}_{\text{Synthetic}}$ is identical to Alice. The benefits of such a solution are several. It could provide a similarity score of 0–100, whereby policy-makers could decide upon a threshold for identity. One may also argue that it is fair, since it uses (at least hypothetically) transparent means to produce the output. Nevertheless, technological facial recognition is, for all intents and purposes, a man-made technology, inevitably informed by the decisions of its designers and the data it is fed. Moreover, a human agent has to decide what level is similar enough according to some kind of standard. Indeed, an algorithm may be both more and less accurate than desired. For example, Povolny and Chick (2020) used the CircleGAN algorithm to create a facial replica that looks identical to the real person to the naked eye, but is categorized as someone else by cutting-edge facial recognition techniques. Similarly, the

[5] As argued by Öhman (2019), deepfakes do in fact not steal any data, at least not more so than a human brain "steals" someone's visual impression by looking at their physical appearance and memorizing it. If, in some hypothetical future scenario, fantasies based on memories (for example pornographic ones) could be directly digitized from the brain, this would be an infringement on privacy, despite no personal data being illicitly used. Hence, the actual use of anyone's the data that the deep learning algorithm was trained on, is therefore secondary.

Fig. 1 Photos of SAND lab members with and without the Fawkes software, photos at courtesy of the SAND lab

Fawkes software, developed by the SAND lab at the University of Chicago (Shan et al., 2020), allows users to add minimal tweaks to their photos, making them undetectable by facial recognition software. For example, out of the photos displayed in Fig. 1, only the left-hand one of each pair is recognized by facial recognition algorithms as the SAND lab team members.

These examples demonstrate the discrepancy between human and artificial recognition. What we mean by recognition is simply something different for machines than it is for humans. In the context of Alice's identity with $\text{Alice}_{\text{Synthetic}}$, the technological aspects must thus be viewed as secondary to the visual impression. What matters is the *effect* that the content has on human beings, not whether some data has been illicitly used in the training, or whether an algorithm can establish a certain percentage of identity. Such criteria inevitably lead back to human judgement anyway. In sum, the problem addressed in this essay cannot be solved by mere technological means because it is a *philosophical* matter. What we need is not a number between 0 and 100, but to discern with precision what we *mean* and what we *should mean* by recognition. No technology, no matter how sophisticated, will ever be able to answer this for us.

4 Turing's approach

The problem of how to establish identity between an individual and a synthetic counterpart is in itself not new. It is present in debates on the moral limits of fiction (Young, 2021), in metaphysics and, of course, in debates on the nature of personal identity (Floridi, 2011; Parfit, 2007). In the latter case, the debate has largely circulated around

various notions of *essential* (or necessary) and *accidental* (or contingent) properties (see Robertson & Atkins, 2013). For example, if *P* is a necessary property for the identity of A, then any object that lacks *P* cannot be A. The question, in our case, would be what *P* is for a given natural person. However, for the purposes of the present article, the (numerical) identity between A and its synthetic counterpart, at least as discussed by Kripke (1980) and Lewis, (1971) among others, is irrelevant. What matters is not whether a given synthetic object is "really" A, but whether their effect in some given context is the same. In other words, the only property *P* that is relevant is the impression of the synthetic content on a human audience. Consider the analogy of counterfeit coins. We may debate forever about whether a certain coin is *really* worth £1 or not, but in practice its value is whatever people believe it to be. The same goes for identity. If everyone, including yourself, take you to be Alice, you are Alice, socially speaking. For this reason, I suggest we leave the ontological debate aside. Instead, I propose that we focus a more conceptually proximate parallel, namely the problem introduced by Turing (1950) in his essay "Computing machinery and intelligence."

For Turing, the issue is not how to establish identity between a natural person (e.g., Alice) and her synthetic counterpart ($\text{Alice}_{\text{Synthetic}}$), but between artificial and human intelligence. His initial question in approaching this relationship is the iconic "can machines think?"—a question which, ironically, he quickly dismisses as "too meaningless to deserve discussion" (1950, p. 433). Both "machine" and "think", argues Turing, are too elusive, imprecise concepts. Instead, he reformulates a metaphysical problem as an epistemic challenge: if a machine could in fact think, how would we know? His celebrated solution is *the imitation game*. The game is played by three players: a man (A), a woman (B), and an interrogator (C), who may be of either sex. The players are separated into two different rooms, where A and B stay in the first room and C in the second. C communicates with A and B separately via teletype. C's goal is to accurately judge who is the man (A) and who is the woman (B). The goal of A, however, is to respond in a way that will cause C to make the wrong decision, while the goal of B is to help C, for example by adding "I am the woman, don't listen to him!". Turing (ibid.) now asks "What will happen when a machine takes the part of A in this game? Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?" If the answer to the latter question is *yes*, then the machine intelligence has *passed* as human, and thus identity has been established—the machine can think, at least insofar as we are concerned within the game.

Turing's solution is an application of a philosophical method that far predates him, and which would later be formalized by Floridi (2008) as the *method of levels of abstraction* (LoA), with inspiration from formal methods in computer science, and arguably also from the pragmatic tradition following Austin, Grice and, to some extent late Wittgenstein. A bit simplified, the method can be summarized as a systematic way of saying "it depends." A LoA refers to the extent to which an entity has been "abstracted" from its natural unique context, or the totality of its information. This can be expressed as a collection of *observables*, that is, a set of "possible values and outcomes" (Floridi, 2013, p. 31) that enables comparison between entities. As such, two entities may be the same or different, depending on the LoA we apply. Consider, for instance, the following example given by Floridi (2011, p. 553):

> Whether a hospital transformed now into a school is still the same building seems a very idle question to ask, if one does not specify in which context and for which purpose the question is formulated, and therefore what the required observables are that would constitute the right LoA at which the relevant answer may be correctly provided. If the question is asked in order to get there, for example, then the relevant observable is "location" and the answer is yes, they are the same building. If the question is asked in order to understand what happens inside, then "social function" is the relevant observable and therefore the answer is obviously no, they are very different.

The point is that we can never find out whether the hospital and the school are "really" the same irrespectively of the purpose for asking. To paraphrase Floridi (2012, p. 3538), this would be like asking the "real" price of a second-hand car in absolute figures, "insisting that no currency is used in order to express it." In other words, there is no essential property that determines the building's absolute identity, it all depends on the purpose of the question. The impossibility of such a "real" answer should, however, not be confused with relativism. In stressing this point, Floridi (2011, p. 554) makes an analogy with the classic problem of Theseus's ship:

> [G]iven a particular goal, one LoA is better than another, and questions will receive better or worse answers. The ship will be Theseus's, no matter how many bits one replaces, if the question is about legal ownership (try a Theseus trick with the taxman); it is already a different ship, for which the collector will not pay the same price, if all one cares about are the original planks. Questions about diachronic identity and sameness are really teleological questions, asked in order to attribute responsibility, plan a journey, collect taxes, attribute ownership or authorship, trust someone, authorise someone else, make sense of one's own life, and so forth.

In other words, a question is always asked for a *purpose*—a request for some specific information—and for that specific purpose, there are more or less appropriate LoAs. For instance, the true answer to the question "is this the hospital?" may be very different for someone in need of a doctor than for someone interested in nineteenth century architecture. This is because a different LoA is required in order to generate a proper response, i.e., different observables come into question. Thereby, the method reduces the complexity of cases where two things are kind of, but not really, the same, by specifying a level of abstraction where the answer is always binary—the two objects of comparison either are, or are not, identical.

As explained by Floridi, the imitation game is essentially Turing's way of replacing the meaningless question of whether machines can *really* think, with a specific LoA—a game with a fixed rule-based scenario. Human intelligence cannot be compared with machine intelligence without some common frame of reference. Or better, their identity cannot be established independently of the purpose for asking the question. In the context of the current article, we are not interested in the identity between machine and human intelligence, but between a natural person and her synthetic counterpart. Yet the problems share a very similar structure, and for this reason, I propose a remodeling of Turing's game that helps define strict criteria for the identity between Alice and

$\text{Alice}_{\text{Synthetic}}$, i.e., a way of adding precision to what we mean by "recognition." I call it the *identification game*.

5 The identification game

Before describing the basic rules of the identification game, two things need to be noted. First, the game is only concerned with visual (and potentially *audio*) identity. There are of course plenty of other ways to communicate identity, such as textual or sensory information. Indeed, even the most rudimentary drawing can establish a reference to a natural person if there is a text spelling the name of the individual next to it. Such textual context is important, but ultimately lies beyond the scope of this paper. Second, the game's obvious similarity to the structure of a Generative adversarial network (GAN) is purely accidental. Indeed, the entire point of the game is that it is played by *human* (albeit imaginary) and not artificial agents. Now, to the rules of the game.

Like Turing's original, the identification game includes three players: A, B and C, who are each assumed to be rational human agents with no prior knowledge of one another before the game starts.[6] The players are separated into two rooms: A and B in one room, and C in the other (see Fig. 2). B will then forge a visual (synthetic) imitation of A, referred to as Synthetic A or A_S. A_S is then displayed on a screen in the other room alongside an authentic depiction of A, referred to as A_1. B's goal is to make the imitation realistic enough to make C unable to tell which is A_1 and which is A_S. The goal of C is to determine which of them represent the real A. C knows that only one of the items is authentic, but not which one. (Note that in the abstract world in which the game takes place, we may assume that C has no bias in terms of their ability to recognize faces from different ethnicities etcetera. As I elaborate below, such an agent may not exist in the real world, but can be assumed in order to illustrate the argument).

Meanwhile, the goal of A is to help C make the right decision by providing relevant information. There are three wayes in which A can assist C: The first is to simply provide some form of visual information about A's true identity, such as a photo or video, that C can use to compare with A_1 and A_S. Note that determining whether a particular item is a fake by comparing it to an authentic image requires some complicated judgements about their absolute similarity. Recall, however, that C's goal is not to determine whether a particular item is fake, but its *relative* likelihood to be fake in comparison to an authentic depiction (A_1). This means that items that look something, but not quite, like A, will always have lesser relative similarity to the visual information provided by A than A_1. The second strategy for A to help is to provide C with technical expertise that helps identify anomalies in A_S that indicate synthetic manipulation, such as eye-reflections, shape of teeth, misaligned shades or other glitches (see Mirsky and Lee, 2021 for a more technical overview of such strategies). The third is to provide C with a software designed to detect synthetic media content that is not detectable with the

[6] In this context, "rational" means nothing more or less than the fact the agents understand the basic rules of the game and are actively trying to win it. Their cognitive abilities are at the level of a normal human being.

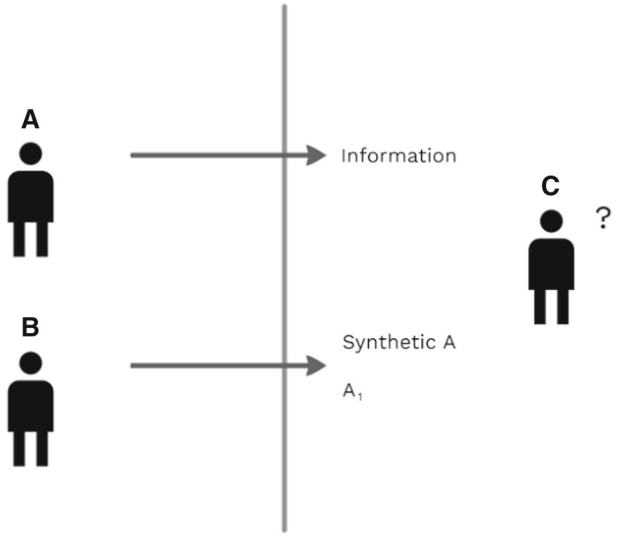

Fig. 2 Illustration of the game's basic premises

human eye. The catch is that, the more information A must provide before C makes the right decision, the more points B will score. Thus, unlike the goals of A and C, B can score in four different ways:

1 point: If C must have some kind of photo or video that proves that A is in fact a natural person before making a correct decision, B scores one point. Thus, all that B needs to succeed is to convince C that A_S is *referring* to A. This includes almost every synthetic depiction that successfully establishes a reference to a natural person. However, if A is in fact not a natural person, any such support is redundant. For instance, assume that a piece of synthetic media depicts a fictional character like, say, Pippi Longstocking. In such a case we would not say that the content is merely a reference to the *real* Pippi, because any picture that successfully passes as Pippi is the real thing, regardless of how realistic it is. Hence, if A must provide some material that proves his or her status as a natural person, (which accounts for almost every case of deepfakes), B will score 1 point.

2 points: If C cannot correctly identify A_S unless provided with technical expertise, then B will score two points. This basically includes any content that would pass as authentic to the (untrained) naked human eye.

3 points: If C cannot correctly identify A_S, despite the technical expertise, B will score three points. At this level, the difference between A_S and A_1 is not detectable by the human eye. Even to a technical expert, both are equally likely

to be the true A. The only thing that can tell them apart is a detection software. Only with access to such a software can C make the right call.

4 points: If C cannot tell A_S from A_1 despite having access to the software, B will score 4 points. At this level, A_S and A are virtually indistinguishable insofar as the game is concerned. Naturally, in real life, the two can hypothetically be told apart by say, eye-witnesses confirming that A has never been at the location depicted in the content, or forensic material proving that B has forged the material. Such circumstances, however, are not part of the identification game. If C fails despite the software, nothing that A says can *prove* that A_S is not identical to A. A's decision will be solely based on his/her faith in A's words.

The various ways in which B may succeed are, of course, levels of abstraction. In this context though, we may refer to them as *levels of identity* (LoI). Note that the LoIs should not be confused with *degrees* of identity, such as percentage points provided by a facial recognition software. Though B can succeed at four different levels, his or her success is, nevertheless, binary within the confines of the criteria for each level. It is not the case that A_S is more identical to A at level 4 than level 1. This would merely imply the introduction of an additional meta-level that views the other levels from outside. The point is that *within* the confines of each LoI, A and A_S are perfectly identical insofar as B has managed to reach the criteria to score. This is not a mere detail, but holds the key to how the game is to be implemented, for the function of the game and its levels is not to provide a scale of 1–4, but to provide a framework wherein different meanings of *recognition* get concrete definitions. When asking whether A and A_S are identical, we ought not to respond with a mere number, but by asking the *purpose for asking the question,* which in turn determines the appropriate level.

One may imagine countless such purposes (taxation, copyright, social interpretation and so on) and there is a relevant LoI for each of them. For the purposes of this article, however, the reason for asking is to determine whether the content in question has a harmful potential. As such, the levels hold no epistemic status in and of themselves. They are deliberately designed to guide our selection of a relevant LoI to judge a person's *moral* claim to content that (allegedly) features their face or voice. Though one may discuss various delimitations of reasonable harm—for instance, Alice may feel genuinely offended by a video, despite no one else recognizing her in it—the argument presented here does not rely on any particular criteria for when feelings of harm are justified. What matters is mere whether the claim is epistemically sound, i.e., whether A_S is similar enough to *potentially* harm A.

The meaning of this "potential to harm" is a bit vague, however. As we have seen, harm from deepfakes comes in many shapes and sizes. So let us use the four categories most commonly discussed within the literature to specify the purpose for asking the identity between A and A_S. If the purpose is to understand whether A has been humiliated/degraded by the spread of A_S, the relevant LoI is 1. For acts of humiliation, even though C does not require much information to see that the content is fake, the reference alone is sufficient to establish identity. For example, a deepfake video of a female politician starring in a pornographic scene may successfully humiliate her, despite the fact that no one actually believes that the video is authentic. Hence, on the basis of humiliation, A has a legitimate claim to have A_S removed insofar

as it is identical on level 1 in the game. If the purpose for asking is to understand whether A_S has harmed A's reputation, on the other hand, level 2 is the more relevant one. Even if individuals with access to technical expertise and detection software can tell them apart, this does not include everyone, and so A's reputation is still in jeopardy. If the purpose for asking is to determine whether A_S is a threat to our political ecosystem, the third level is arguably the most relevant. If the distinction between A_1 and A_S is not detectible even to experts, the system becomes solely dependent on blind trust in the software that is used to "tell the truth." The trust is "blind," because to the vast majority of the population, any computer software, and especially one used to detect traces of deep learning, is an enigma. Thus, in this context, it is only relevant to say that A_S is *identical* to A, if A_S passes the criteria for LoI 3. If the purpose for asking is to determine whether A_S is a serious epistemic threat, i.e., part of what some warn is the beginning of an "information apocalypse," the fourth level is the most relevant. To say "Alice was featured in a deepfake video," when talking about the epistemic threat of deepfakes would require that not even a software could detect the difference.

The fact that reputational or political harms require more information from B in order to be valid does *not* imply that they are to be considered graver. Humiliation can do more damage than harms to one's reputation, yet require less sophisticated technology (and thus a lower LoI). By analogy, the information needed to prove that a crime has been committed is unrelated to the seriousness of the crime. Note also that the above categorization of harm is conceptually arbitrary. Other conceptions may thus be added to this list as the technology develops, but regardless of what harm is under consideration, the game provides a framework to select the relevant LoI.

In other words, the legitimacy of Alice's claim to content that allegedly features her face depends on whether it is possible to establish identity between her and the synthetic content, i.e., whether it is *recognizable* as her; what we mean by recognizable depends in turn on the purpose for asking the question—in this case the type of harm that is under consideration.[7] This point may seem almost trivial, but it has two major advantages: First, it allows us to avoid notions of a *partial* identity. Recall that the identity between A and A_S is on each level complete. They either are, or are not the same. The question is only which level is relevant, and the answer always depends on the purpose for which the question of identity is asked. Second, it circumvents any notion of an absolute identity, either technologically or philosophically constituted, that may be used to assess the legitimacy of all claims of harm arising from synthetic media.[8]

This also explains how we should understand the possibility of misrecognition. Surely, an individual may be mistaken about the identity between A and A_S. You may recognize someone as your friend, only to realize they were someone else. Or, more plausible, you may recognize someone as A only to realize it is their identical twin.

[7] (Note that the various kinds of harm do *not* presuppose each other. A video can be a political threat without being humiliated or ruining someone's reputation).

[8] For those acquainted with Butler's (1999) performative theory of gender, this follows a familiar logic. Just like there is no original woman (or man) of which other men and women are mere mirror-images, there is no supreme LoA by which every other LoA can be judged. A person can belong to a certain sex at a biological level, but to a different one at a social level. Insofar as there is a totality of contexts (itself a paradox), this totality is by no means elevated above the others.

From the perspective of the proposed heuristic, does this mean that identical twins have rights to each other's content? If, for instance, Bob makes a deepfake of Alice with her permission, can Alice's twin object to this, on the basis that a player of the identification game would likely recognize the face in the video as hers? Epistemically speaking, the answer is yes—the faces are similar enough to warrant moral consideration. The result of that consideration, however, is a separate matter, that lies outside the scope of this essay.

On a more general note, mistakes about the identity of a synthetically generated face should be understood by analogy of counterfeit coins. Suppose you receive two identical coins in change. Later, you use one of them to buy a newspaper, but when trying to use the other in a vending machine you it does not work. The machine tells you your coins were counterfeit. The point here is not that the value of the coins was zero all along, but that their value depends on their effect. Different contexts may require different levels of identity to discern the value of the coin. Sometimes we need an expert to make the assessment, sometimes we need a machine. But whenever we make a mistake in recognizing the value of money, the mistake can only be judged as such from a different level of abstraction, or alternatively, in relation to the judgment of other people. The same goes for persons. Twins are of course two separate people, but when it comes to things like DNA, or synthetic media, they may *effectively* be the same. In sum, identity is always absolute yet always also relational.

6 Implications

How is the identification game intended to be used? Like Turing's original, it is primarily meant as a thought experiment—a heuristic. Turing was interested in the relationship between human and machine intelligence as a *general* conceptual problem, not in the supposed intelligence of a particular machine. The same is true for the identification game. It is a way of conceptualizing and explicating what we mean by "recognition", rather than an empirical measurement. Nevertheless, Turing's game has often been interpreted as a concrete test to assess the alleged intelligence of AI systems. Similarly, the identification game may have a practical utility in our assessment of real-life borderline cases, where there is (or can be) disagreement about whether the content in question is *really* starring someone's face. In this sense, the identification game can be used as a practical test, where the identity of A and A_S depends both on the purpose for asking the question, but also on the effects that A_S has on a *real* human, i.e., its performance as a social, ethical and epistemic matter.[9]

Let met illustrate this utility by looking at three high profile, publicly accessible real-life cases where the identity between a natural person and a deepfake is or can be questioned: (1) the objections of American actress Scarlet Johansson against the

[9] As such, we may also think about the identification game in metaphorical terms, where B represents producers of synthetic media content, C represents the general public and A represents society's efforts to counter the damaging effects of synthetic media, such as educational campaigns or development of detection software. From this viewpoint, it is imperative to increase the role of A in society; providing better information and developing better software. However, regardless of the technological development, the matter of identity remains a philosophical matter.

frequent usage of her face in pornographic deepfake videos, (2) the TikTok account @deeptomcruise, and (3) a video featuring former US President Donald Trump urging Belgium to withdraw from the Paris clime agreement.[10] Before analyzing these examples, recall that the scope of this study is not to determine whether the contents in question ought to be removed. That would require consideration of multiple principles, including free speech, the nature of celebrity culture and the public sphere. Nor is it the purpose to determine the authenticity of the videos. What is under consideration here is *only* whether it is possible to establish identity between the person claimed to have been harmed and the content in question. In other words, disregarding the possible existence of stronger opposing claims, is the contents in question similar enough for the subjects to have claim to them?

(1) American actress Scarlet Johansson was early on among the most popular faces to superimpose onto pornographic content. When interviewed about this in the *Washington Post*, Ms. Johansson responded that "Clearly this doesn't affect me as much because people assume it's not actually me in a porno, however *demeaning* [italics added] it is" (quoted in Harwell, 2018). In the case of Ms. Johansson, it is obvious, at least from the quoted objection, that the harm under consideration belongs to the category of *humiliation*. The relevant LoI, therefore, is level 1. So, would the content live up to the criteria for B to score one point? In these early days of deepfakes, most of the videos were not very realistic, and the faces of the actresses upon whom Ms. Johansson was superimposed shone through, resulting in an odd mix of the two individuals. Still, for most of the content in question, even the rather poorly made videos, the answer is obviously yes. It is possible for any rational human agent, especially if there is a text specifying the intention of imitating Ms. Johansson, to identify her. Were the creators to object that it is in fact not her, their arguments would fall short due to the content scoring 1 point in the game, even if their algorithms have not even been trained on Ms. Johansson's data. This is however not the case for all deepfakes allegedly depicting Scarlet Johansson. Some videos fail to establish identity, not only to Scarlet Johansson, but to any natural person, living or dead. In such cases, claims of identity do not hold up and thus Ms. Johansson does not have a legitimate claim to the content.[11]

(2) The TikTok account @deeptomcruise, managed by Chris Ume, a Belgian graphics artist, went viral in 2021 due to its stunningly accurate deepfakes using the face of American actor Tom Cruise. The videos do not display any humiliating or otherwise incriminating activity. Nevertheless, due to the outstanding spread of the videos (the initial one has received well over ten million views) an associate of Mr. Ume contacted Mr. Cruise's management quiring whether they desired to take over the account or whether they wanted it removed, reports ABC (Corcoran & Henry, 2021). At the time of writing, it appears they are yet to send a response. In this case, the situation is more straightforward. Were Mr. Cruise to find the videos humiliating, or

[10] Though it partly defeats the purpose of analyzing real-life cases, I will not share any of the said materials within this publication. This is partly due to copyright, but also more importantly, out of respect for Ms. Johansson.

[11] Note that this does not exclude other forms of legitimisation. If it were to be proven that the content contains copyrighted material, or in other ways are made up by her personal data, she does have legitimate claim to it.

even if he were worried about them damaging his reputation, he could legitimately claim that they are, as it were, *similar enough.* Disregarding counterarguments based on the nature of celebrity, he does have a claim to the content. However, an individual provided with expert insight from A, would soon find that there are anomalies that expose the videos as fake. The coloring of Cruise's skin in the videos is, for example, slightly off in relation to the neck. Hence, were Mr. Cruise to make a claim about the damage caused by the videos on our political ecosystems, one may reasonably object that the person in the synthetic video is in fact *not* him (i.e., $A \neq A_S$).

(3) In the spring of 2018, a video surfaced on the internet depicting then US President Donald Trump delivering a message to the people of Belgium. "I had the balls to withdraw from the Paris climate agreement" he said, "and so should you." Some were outraged that Trump had interfered with Belgium's climate politics, but, as it turned out, the video was actually a fake commissioned by Socialistische Partij Anders, a political party on the left. The intention, explained the party, was merely to grab people's attention and then lead them to a climate petition website. They did not believe that people would actually be fooled by it due to the poor quality of the lip movements (see Schwartz, 2018, for full story).

If one were to claim that the video posed a threat to democratic deliberation, the heuristic of the identification game would lead us to respond that the person in the video is in fact not identical to Trump. Were someone to demand its removal referencing harms to our democratic communities, they would not be making an epistemically sensible claim because on LoI 3, $A \neq A_S$. Likewise, were someone to make the argument that the video undermines trust in video as a category of evidence, one may also counter that the video is in fact not a depiction of Mr. Trump, because also on LoI 4, $A \neq A_S$. However, were Mr. Trump to make a claim to the video on the basis of its humiliating nature, this would be sensical—on the level of dignity, the person in the video *is* Trump because on LoI 1, $A = A_S$. While opposing principles, such as the right to ridicule figures of power, provide stronger reasons not to remove the video, such a claim by Mr. Trump would at least be *epistemically* sound.

In summary, one may conclude that, just like Theseus's ship both is and is not the same after its planks have been replaced, the videos above are and are not using the faces of the individuals making (or potentially making) the claims, depending on the moral basis of the claims these individuals make. Using the heuristic of the identification game, we may thus add precision to what is meant by recognition, which in turn alleviates the conceptual ambiguity haunting legislators' appeals to a right to one's identity.

7 Concluding remarks

As noted in the first section of this essay, multiple jurisdictions are attempting to combat deepfakes by appealing to a so-called right to identity, i.e., control over one's "likeness" or "image" (Perot & Mostert, 2020, p. 33). The subject of this essay is the epistemic limits of such rights. A necessary, though insufficient, criterion, I have argued, is the possibility to *recognize* the person featured in the content as the natural person making the claim (or on whose behalf the claim is made). However, as it is

unclear what exactly recognition means, the purpose of this paper has been to develop a heuristic, the identification game, in order to add rigor and clarity to it.

To summarize my argument, let us now return to the scenario introduced at the outset of this essay, where Bob creates a deepfake in which Alice recognizes a face as her own. She demands the video to be removed; he counters that the video does not depict Alice. How can we know who is right? How can we be sure that the face in the video does in fact belong to Alice? As I have suggested, the person in the video is Alice insofar as she is recognized as Alice. But what we mean by recognition, depends on the nature of Alice's objection, i.e., *how* she claims that the video has caused harm. If she claims that the video is humiliating, the answer may be very different from a case where she claims that it is damaging her reputation or ruining political deliberation. According to this argument, it does not matter whether Bob has created the content "by accident," or whether he has in fact used any of Alice's data in the process. What matters is the ability of a (human) epistemic agent to *recognize* $\text{Alice}_{\text{Synthetic}}$ as Alice under certain conditions.

Though establishing identity is a necessary condition for claims to one's distinguishing features, it is not sufficient. Even if it were to be proven that the face in Bob's video does in fact belong to Alice, and even though it may cause emotional harm, this does not in and of itself warrant removal of the content. Despite the existence of legitimate reasons to remove it, there may be opposing, stronger reasons to let it remain. Emotional harm is a natural part of social life, especially for people living in the public eye. If the person featured in the video is a politician, for example, we may think that elements of humiliation and satire are more permissible than if they were a private individual. Inversely, we may be more careful with deceptive videos featuring political leaders than if they were private individuals, in order to protect our political ecosystems. Though it is beyond the scope of this essay to weigh these various reasons against each other, the proposed argument does provide the epistemic premises for such a debate. I look forward to taking part in it.

Funding Open access funding provided by Uppsala University. Funding was provided by Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS).

Declarations

Conflict of interest The author of this article declares no conflict of interest.

References

Bakker, J. (2020). *Deepfakes affecting reputation a study comparing effects of different levels of (fake) media on reputation*. MSc Thesis, Einhoven University, Industrial Engineering and Innovation Sciences

Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P., & Nayar, S. K. (2008). Face swapping: Automatically replacing faces in photographs. *ACM Transactions on Graphics*. https://doi.org/10.1145/1360612.1360638

Butler, J. (1999). *Gender trouble: Feminism and the subversion of identity*. Routledge.

Citron, D. K., & Chesney, R. (2019). Deep fakes: A looming challenge for privacy, democracy, and deep fakes: A looming challenge for privacy, democracy, and national security national security. *HeinOnline*. https://scholarship.law.bu.edu/faculty_scholarship/640

Corcoran, M., Henry, M. (2021). The Tom Cruise deepfake that set off 'terror' in the heart of Washington DC. ABC News. Retrieved 13 October 2021 from: https://www.abc.net.au/news/2021-06-24/tom-cruise-deepfake-chris-ume-security-washington-dc/100234772

de Ruiter, A. (2021). The distinct wrong of deepfakes. *Philosophy and Technology, 34*(4), 1311–1332. https://doi.org/10.1007/s13347-021-00459-2

Diakopoulos, N., & Johnson, D. (2021). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media and Society, 23*(7), 2072–2098. https://doi.org/10.1177/1461444820925811

Fallis, D. (2021). The epistemic threat of deepfakes. *Philosophy and Technology, 34*(4), 623–643. https://doi.org/10.1007/s13347-020-00419-2

Floridi, L. (2008). The method of levels of abstraction. *Minds and Machines, 18*(3), 303–329.

Floridi, L. (2011). The informational nature of personal identity. *Minds and Machines, 21*(4), 549.

Floridi, L. (2012). Turing's three philosophical lessons and the philosophy of information. *Philosophical Transactions of the Royal Society a: Mathematical, Physical and Engineering Sciences, 370*(1971), 3536–3542. https://doi.org/10.1098/rsta.2011.0325

Floridi, L. (2013). *The ethics of information*. Oxford University Press.

Harris, D. (2019). Deepfakes: False pornography is here and the law cannot protect you. *Duke Law & Technology Review, 17*(1), 99–127.

Hall, E. W. (1933). Numerical and qualitative identity. *The Monist, 43*(1), 88–104.

Harwell, D. (2018). Scarlett Johansson on fake AI-generated sex videos: 'Nothing can stop someone from cutting and pasting my image'. The Washington Post. Retrieved 13 October 2021, from: https://www.washingtonpost.com/technology/2018/12/31/scarlett-johansson-fake-ai-generated-sex-videos-nothing-can-stop-someone-cutting-pasting-my-image/

Kerner, C., & Risse, M. (2021). Beyond porn and discreditation: Epistemic promises and perils of deepfake technology in digital lifeworlds. *Moral Philosophy and Politics, 8*(1), 81–108. https://doi.org/10.1515/mopp-2020-0024

Kripke, S. (1980). *Naming and necessity*. Harvard University Press.

Lewis, D. K. (1971). Counterparts of persons and their bodies. *Journal of Philosophy, 68*(7), 203–211.

Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T., & Ling, Z. (2018). *The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods* (pp. 195–202). https://doi.org/10.21437/odyssey.2018-28

Maddocks, S. (2020). 'A Deepfake Porn Plot Intended to Silence Me': Exploring continuities between pornographic and 'political' deep fakes. *Porn Studies, 7*(4), 415–423. https://doi.org/10.1080/23268743.2020.1757499

Mashinini, N., & Africa, S. (2020). *Criminal liability for the violation of identity using deepfakes in South Africa*. Academic Conferences International Limited. https://doi.org/10.34190/IWS.21.065

Minton, T. (2017). 12 dead celebrities who were resurrected with GCI. Screenrant.com. Retrieved December 18, 2019 from https://screenrant.com/dead-celebrities-actors-cgi-resurrected-movies-tv/

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys*. https://doi.org/10.1145/3425780

Nagel, T. (1970). Death. *Noûs, 4*(1), 73–80.

Öhman, C. (2019). Introducing the pervert's dilemma: A contribution to the critique of Deepfake Pornography. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-019-09522-1

Parfit, D. (2007). *Reasons and persons*. Clarendon Press.

Perot, E., & Mostert, F. (2020). Fake it till you make it: An examination of the US and English approaches to persona protection as applied to deepfakes on social media. *Journal of Intellectual Property Law and Practice, 15*(1), 32–39. https://doi.org/10.1093/jiplp/jpz164

Povolny, S., & Chick, J. (2020). Dopple-ganging up on Facial Recognition Systems. Retrieved 3 June 2022 from: https://www.mcafee.com/blogs/other-blogs/mcafee-labs/dopple-ganging-up-on-facial-recognition-systems/

Rini, R. (2020). Deepfakes and the epistemic backstop. *Philosophers Imprint, 20*(24), 1–16.

Robertson, T., & Atkins, P. (2013). In E. N. Zalta (Ed.), *Essential vs. Accidental properties*. Stanford Encyclopedia of Philosophy.

Schwartz, O. (2018). You thought fake news was bad? Deep fakes are where truth goes to die. The Guardian. Retrieved 3 June 2022 from: https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth

Shan, S., Wenger, E., Zhang, J., Li, H., Zheng, H., & Zhao, B. (2020). Fawkes: Protecting personal privacy against unauthorized deep learning models. Proceedings of the 29th USENIX security symposium. https://www.usenix.org/conference/usenixsecurity20/presentation/shan

Turing, A. (1950). Computing machinery and intelligence. *Mind, LIX*, 433–460. https://doi.org/10.1093/mind/LIX.236.433

Waldron, J. (2012). *The harm in hate speech*. Harvard University Press.

Warzel, C. (2018). *He predicted the 2016 fake news crisis. Now he's worried about an information apoca- lypse.* Buzzfeed News. Retrieved 3 June 2022 from https://www.buzzfeednews.com/article/charliewarzel/the-terrifying-future-of-fake-news

Yadav, D., & Salmani, S. (2019). *Deepfake: A survey on facial forgery technique using generative adversarial network*. 852–857. 2019 International conference on intelligent computing and control systems (ICCS). https://doi.org/10.1109/ICCS45141.2019.9065881

Young, G. (2021). *Fictional immorality and immoral fiction*. Lexington Books.