Synthese (2023) 201:43
https://doi.org/10.1007/s11229-023-04044-2

ORIGINAL RESEARCH

How to do things with deepfakes

Tom Roberts[1]

Received: 29 June 2022 / Accepted: 5 January 2023

Abstract
In this paper, I draw a distinction between two types of deepfake, and unpack the deceptive strategies that are made possible by the second. The first category, which has been the focus of existing literature on the topic, consists of those deepfakes that act as a fabricated record of events, talk, and action, where any utterances included in the footage are not addressed to the audience of the deepfake. For instance, a fake video of two politicians conversing with one another. The second category consists of those deepfakes that direct an illocutionary speech act—such as a request, injunction, invitation, or promise—*to* an addressee who is located outside of the recording. For instance, fake footage of a company director instructing their employee to make a payment, or of a military official urging the populace to flee for safety. Whereas the former category may deceive an audience by giving rise to false beliefs, the latter can more directly manipulate an agent's actions: the speech act's addressee may be moved to accept an invitation or a summons, follow a command, or heed a warning, and in doing so further a deceiver's unethical ends.

Keywords Deepfakes · Deception · Speech acts

1 Introduction

A range of familiar technologies allow us to communicate at a distance, transmitting a spoken or written message to a person at another location. I can telephone or email you from afar, for example, or leave a message on your voicemail for you to pick up later. These technologies, moreover, allow us to *do* things with our speech at a distance—say, to issue a command or an invitation, ask a question, grant permission, or make a promise. That is, they allow a speaker to perform an illocutionary act (Austin, 1962) without sharing the same spatiotemporal location as the listener. A recorded message can be used, for instance, to accept or make a proposal, to express one's

✉ Tom Roberts
tom.roberts@exeter.ac.uk

[1] Philosophy, University of Exeter, Amory Building Rennes Drive, Exeter EX44RJ, UK

Published online: 24 January 2023

condemnation or approval, or to deliver an instruction. When the utterance receives suitable uptake, its addressee can be moved to act accordingly, and thus to achieve the speech act's perlocutionary intent.

In this paper, I argue that deepfake technologies are capable of manufacturing audio and visual materials that share this power. Deepfakes can make it appear as though a speaker is performing an illocutionary act intended for one or more recipients, and thus can move an audience to action. This entails that some deepfakes are best understood not only as misleading 'evidence' of a person's utterances and behaviour, but as more or less direct manipulators of action—rather like a fake court summons; a fake wedding invitation; or a fake stop sign. Attending to the action-guiding possibilities of deepfakes enables us to more clearly delineate the variety of deceptive projects in which they might be exercised, and the moral import of these projects.

Although deepfake technology is in its relative infancy, I will assume that it is likely to improve and proliferate in coming years, with deepfakes becoming both more convincing and cheaper to produce.[1] If this comes to pass, then we can expect to encounter more fabricated footage of celebrities, politicians, and others in the public eye; and to see the emergence of 'local' deepfakes in which one friend or family-member, say, generates a lifelike depiction of another. Deepfake illocutionary deceptions may thus arise at the level of a wide population (counterfeit commands or declarations made by a government official to the citizenry, for example) or at a smaller scale (instructions from the boss to her employees, or a request from a child to a parent, for instance). In theory, any context in which an illocutionary act is delivered at a distance—over telephone, radio, or Internet—is vulnerable to the deepfake threat.

In Sect. 2, I introduce how speech acts operate when we use ordinary recording equipment that accurately captures what is said and done by a speaker. I argue that even these faithful recordings can be put to use in distinctive forms of deception. In Sect. 3, I extend the discussion to fabricated materials: deepfakes that make it appear as though a speech act such as a command, request, invitation, or plea has been addressed to a certain recipient, in order to mobilise their actions. Section 4 reflects upon the moral harms that flow from such deceptions, and Sect. 5 sums up conclusions.

2 Faithful recordings

Faithful recordings allow an observer to witness speech and action that occurred at a separate time and place. Tape recorders, voicemail services, digital camcorders, dictaphones, security cameras, and so forth can capture an utterance and make it available to those who didn't hear it first hand.[2] It will be useful to introduce a distinction between what I will call "open" and "closed" recordings in this section, as this will help us to make sense of the various possible (mis)uses of deepfakes to be examined in the remainder of the paper.

[1] This assumption is the default in the literature, see e.g. Citron and Chesney (2019), Rini (2020), and Harris (2021).

[2] Martin (2012) argues that audio recordings allow us to literally hear the voice of the speaker in question, because they enable individual sounds to be reproduced at a later date. My use of 'available' here is intended to be neutral on this matter.

2.1 Open and closed recordings

Consider, as an introductory example, a video recording of a wedding ceremony that took place several years ago. From our retrospective point of view, we can hear two people offer their marriage vows to each other, and hear the officiant making a declaration to the effect that the couple is married. In neither case does the speaker in the video say anything to *us*—we are not the intended recipient of their utterances. This is an example, then, of a "closed" recording: the speech acts delivered within the recorded scene are not directed to listeners out here in the wider world. It is as though we have overheard a verbal transaction in which we did not ourselves participate; where this overhearing takes place at a temporal distance.[3] Closed recordings are very common in modern culture—they document how speakers behaved and what they said to one another (including what they asked and asserted, what they promised and declared, and so on), in the way that a hidden camera, a documentary crew, or a tape recorder might do.

In contrast, consider a recording that I have made of myself announcing my wishes for what you should do in the event of my death. When you view this recording following my suspicious demise, you receive my instructions to you: I deliver a series of posthumous requests, permissions, and directives through my recorded speech. Furthermore, if you take my utterances to heart, you will *act upon* them and fulfil the various plans I had intended to set into motion. This is a case, then, of an "open" recording—one that captures an utterance that breaks the fourth wall and affects the external world in material ways, through executing a sequence of illocutionary acts. Other "open" recordings include a voicemail message left for you by your sibling, requesting that you come and pick them up at 6 o'clock; a multilingual video-host at an airport who directs each visitor to baggage reclaim; or a despot who declares war at time t_2 by releasing a statement that was pre-recorded at time t_1. In each of these cases, the recorded speech is addressed to a certain audience, just as ordinary, in-person episodes of speech typically are. The acts have illocutionary force, such as enjoining others to action; making a promise, threat, or apology; asking for a favour; or granting permission. The acts can have perlocutionary force when they have their intended effect upon the addressee, such as persuading, inducing, scaring, or flattering.[4]

Let us make this distinction more precise by filling in some details. Firstly, notice that a closed recording is not one that has no effect at all upon its audience. Observing a recorded scene can, after all, have epistemic consequences such as altering our beliefs about what went on in the past; as when we watch security-camera footage of a crime that was committed yesterday, for example. The dialogue in a faithfully recorded scene may have a variety of other psychological effects: it may insult, baffle, scandalise, terrify, or enrage its audience, depending upon what it reveals to them about the speaker. Moreover, the recording may move us to wider patterns of action. A candid clip of a politician's gaffe may motivate us to vote them out of office, for instance, while a recording of a friend's betrayal may lead us to end the relationship. So

[3] For discussion of the epistemic differences between overhearing and being addressed, see Hinchman (2005), and McMyler (2013).

[4] Austin (1962) counts both the intended and unintended consequences of an utterance among its perlocutionary effects. For simplicity I will focus on the former (for discussion, see Gustafsson, 2020).

closed recordings are by no means inert. The key distinction between closed and open recordings is that the latter but not the former address their constituent speech acts to the observer—say, asking a question, making a promise, or issuing an instruction.[5] These acts can have corresponding perlocutionary effects upon the listener (Austin, 1962; Cohen, 1973), such as soliciting from them an answer to the question, or persuading them to follow an imperative. It is this active, mobilising role that will be our focus in what follows, because what is significant is that deepfaked speech can exert itself in these ways too.

Next, let us unpack some similarities and differences between ordinary in-person speech acts and those delivered via faithful recordings. Notice that open recordings are like many standard illocutionary acts in that they have an intended recipient in mind; sometimes an individual person, sometimes a collective. My posthumous instructions are directed to *you*, my trusted confidante, and not to any others with whom you might share the tape. Elsewhere, an open recording may have a wide intended audience—perhaps including everyone who consumes it, present and future. An environmentalist may sincerely urge every human citizen to take care of the planet and nurture its resources, for example, in a video that is distributed on a popular public platform such as TikTok. An open recording, being something we can recycle again and again, may have different intended recipients at different times. I could use a single recorded message to fire several employees in sequence, for example, or to invite multiple guests to my birthday party. There may be a substantial time-lag between uses of an open recording, for instance when I repurpose last year's asynchronous revision lecture to advise this year's class on how to prepare for their final exam. An open recording may never find an audience, for example when it lies dormant in forgotten hardware, in which case it is rather like an ordinary vocalisation that goes unheard.

Recorded speech acts, like their in-person counterparts, can be more or less successful. Austin (1962, Lecture II) distinguishes two fundamental ways in which an illocution can fail. In a *misfire*, the speech act is not committed at all; for example when a speaker lacks the relevant authority to perform it. I may declare "you are now husbands together", for example, or "I sentence you to life imprisonment"—but unless I have a particular legal standing I do not thereby alter your status. In an *abuse*, the illocutionary act is performed but suffers from a defect. An insincere or hollow promise, for example, is an abuse because the speaker has no intention of abiding by the conventional procedure of promise-keeping (Austin, 1962; see also Searle, 1969). For an illocutionary act to be fully successful depends upon its receiving *uptake* from the listener—that is, on its bringing about "the understanding of the meaning and of the force of the locution" (Austin, 1962, p. 116).[6] This entails that the hearer must not only grasp the semantic content of the utterance, but also the type of act to which it belongs—e.g. plea, announcement, promise, or warning. An illocutionary act that is delivered via a faithful recording may succeed or fail on similar grounds: it may be

[5] See Searle (1976) for an influential taxonomy of illocutionary acts. My focus here is on those acts that, when they are successful, most obviously move the listener to action, such as directives and declaratives.

[6] There is controversy over whether Austin's own treatment of uptake is consistent (e.g. de Gaynseford, 2011; Longworth, 2019), and there are complexities concerning how uptake is negotiated between speaker and addressee (e.g. McDonald, 2020). What matters for the current discussion is simply that some speech acts can be successful in the robust sense that they move listeners to action, and others can fall short of this.

insincere, for instance, or the speaker may not possess the institutional authority to perform an action of the kind in question. Technology, moreover, introduces further potential obstacles to uptake: a tape that has degraded over time, say, might contain a message that is hard to understand, or that makes a speaker's intentions difficult to discern.

As with live utterances, open recordings may be more or less successful in mobilising an agent to action—that is, in bringing about their intended perlocutionary effects. A listener may grasp that an invitation is being offered, for example, but refuse to accept it. A speaker may deliver a set of instructions, but find that nobody is willing to comply with them. Later, we will see that entirely fabricated speech acts may fail to move their audience because their technology is insufficiently convincing—if they do not capture a speaker's true likeness, for instance, or are uncanny or off-putting. For now, what is important to notice is that recorded speech acts, just like ordinary utterances, can be successful: they can be understood and acknowledged, and they can move listeners to action by persuading, encouraging, exhorting, convincing, deterring, and so forth.

2.2 Error and deception

Next, we will consider certain deceptive strategies that are made possible by faithful recordings. By stipulation, these are accurate records of events—audio and video footage that hasn't been doctored, dubbed over, manipulated, cut or otherwise edited in a misleading way. Faithful recordings enable distinctive forms of deceit to be enacted by a malevolent agent, and it is worth characterising these in order to set the scene for the account of deepfake deception that will follow.

Faithful closed recordings might be deployed in deceptive ways despite their contents being truthful, when they are taken out of their original context. For example, I might misrepresent your current political views by showing others a video of a speech you made some years ago; or I might record you saying something in jest or with irony, and then present it as though you were speaking sincerely. In cases like these, I am the deceiver even though you are the recorded speaker. It is me who makes the attempt to mislead others through my deployment of a recording of you. If I succeed in my aim, the primary consequence of my deceit is epistemic, in that my audience comes to hold a set of false beliefs.[7]

This epistemic dimension can be seen, too, in simple forms of deceit that are made possible by faithful open recordings—that is, those that address a speech act to an external audience. I can leave a series of statements that I believe to be false on your answering machine, for example, and so lie to you from a distance. In this case, it is the assertoric content of my utterance, conveyed in the speech act of telling, that is designed to mislead you. Once again, the purpose of my deceit is epistemic—it is for you to come to believe something I take to be untrue. Of greater interest for current purposes, however, are faithful open recordings that are deceptive in virtue of their *non-assertoric* content; that is, in virtue of some other illocutionary dimension such as

[7] For discussion of the ways in which static images might be used to lie and mislead, see Bátori (2018), Cooke (2019), Dixon (2022), and Viebahn (2019).

being a command or an invitation. Here, the deception is not driven by the falsity of a recorded assertion, because this element of the utterance ("go upstairs!", "please come to the ball") isn't assessable for truth or falsity in the first place. Instead, the agent is made to act upon an order, instruction, or request that is delivered in a deceptive way.

As a preliminary step, notice how consumers of a faithful open recording may be susceptible to two kinds of error even when there is no deliberate subterfuge at work. First, one can be mistaken about the addressee of a particular speech act—to think, for example, that a question, instruction, or invitation that was directed to another person was directed to you. This can happen with in-person speech, such as when you hear a cry through the window and think that you are being addressed, when in fact the call is for somebody out of sight. And it can happen through recorded materials, such as when you share an answer-phone with your housemate and you think that an invitation left for them was meant for you. Second, one can mistake a recording that one has already responded to, or which is otherwise out of date, for a fresh recording whose illocutionary dimension has not yet been acted upon. Suppose I have missed a voice message you left for me yesterday, inviting me for lunch at the cafeteria at noon. Then, today, I listen to it and mistakenly proceed to our rendezvous 24 hours late.

For these errors to occur, circumstances surrounding the recording and its wider context have to co-operate. An open recording that clearly addresses a person by name, for example, is unlikely to confuse anyone with a different name. Commands, requests, or invitations that specify a particular time and place, similarly, will tend to be too idiosyncratic in their details to be open to misapprehension. Who the speaker is often matters, too: the message left by your housemate's mother is unlikely to be for you, even if she does not address her child by name. Facts about the vehicle of a recording, such as whose device a video message is left on, are often a reliable guide to its intended target—it's a fairly safe bet, for instance, that messages heard on my voicemail are for me. And there may be visible or audible markers of the time or date upon which a recording was captured, that counteract the possibility of errors of the second kind. A loss of fidelity, for example, may signal that a recording is some years out of date, or the speaker may simply appear younger on screen than they are today.

Now we are in a position to see how errors of these kinds might be exploited in the service of deceptive ends. In short, it is possible to take an open recording out of its original context and to use it to manipulate the behaviour of a person who takes its illocutionary significance at face value.

In the first scenario, you work at a fast food restaurant. Your manager, seeing that she will be short-staffed tomorrow, sends you a video message instructing you to come in for the early shift. Your deception consists in forwarding this footage to a co-worker and letting them follow the manager's command in your place.

In the second scenario, you apply for a prestigious academic job, and the chair of the search committee leaves a message on your voicemail service, soliciting further documents in support of your candidacy. You choose to play a malicious prank on a professional rival, by forwarding the chair's request to *their* answering machine. Your rival hears the message, sends their own materials, and commits an embarrassing faux pas.

In the third scenario, we are all back in high school competing for peer approval. A popular student posts a recording on your social media page, inviting you to a

fashionable party this weekend. The ruse in this case is to recycle the recording by transferring it to a less popular friend's page, and to watch as they turn up to a social event at which they are not welcome.

For deceptions of these kinds to be successful, circumstances must again align rather neatly. There has to exist a suitable recording in the first place; it needs to avoid mentioning the original recipient by name; and it must be possible to disguise the recording's provenance, such that the victim cannot easily detect that it has come from you, second hand. So the perpetrator needs some skill, or luck, to pull the gambit off.

The three deceptions are distinctive in that their victims are not simply misled about matters of fact. The deceit is not exhausted by its epistemic effects, but has a targeted behavioural outcome. Each recording brings about some true beliefs ("we're short on staff", "there's a party on Saturday") and some false beliefs ("I've been ordered into work", "I'm invited to the party") in the victim. When the deceit is successful, the latter give rise to a corresponding behavioural response, which is the deceiver's ultimate aim. Notice that the false belief in each case concerns being the addressee of an illocutionary act that is sincerely delivered by someone with suitable authority. And the victim comes to that belief because the recording presents a realistic appearance of such an act. That is, the victim is not simply *told* that they have been requested, ordered, or invited to do something; they are put in a position where it seems to them that they are being directly addressed by a speaker. The deceiver exploits the non-assertoric illocutionary character of the original recorded utterance: repurposing the speech act in order to manipulate another agent's conduct.[8] As we will see, these types of artifice are made possible by deepfakes, too, with the advantage of this technology being that there is no need to recycle existing footage, and any voice, any name, and any choice of words may be put to work in the service of the deception.

3 Deepfakes

A deepfake is a "recording" in which "utterances" are made by one or more "speakers". In order to dispense with the scare quotes, let us clarify the structure of deepfake speech acts. Deepfakes are audio and visual representations of real persons, but they depict behaviour and speech that has been selected and composed by somebody else.[9] Deepfake utterances can thus be usefully compared to other instances of "nonserious" speech (Searle, 1975), such as fictional dialogue.

When a stage actor delivers their lines, these speech acts are attributed by the audience to the fictional character being portrayed (Alward, 2009). It is Hamlet who makes certain declarations, requests, and commands, for example, rather than Sir Laurence Olivier, even though the words come out of the latter's mouth (Eaton, 1973, p. 45). In cinema, the dramatic utterances are captured on film, and the same thing happens at one remove: we attribute particular acts of speech to Dorothy, even though

[8] Compare this to a case in which I dress in a police uniform and stand in the road directing traffic. I haven't lied to the drivers I deceive, because I haven't asserted anything to them. Instead, I exploit their impression of my authority to make them do my bidding.

[9] In typical cases at least; although we might imagine a person making a deepfake of themselves in order to, say, deliver a polished performance of a political address.

they were spoken by Judy Garland. In a cartoon series or a ventriloquist act, we attribute particular speech acts to a fictional character who may look entirely unlike the voice artist responsible for what they say.

Deepfake speech, I suggest, has a parallel structure to these cases but involves an additional level of artifice: the utterance in question is not delivered by one person posing as another, but by an entirely fabricated version of a speaker.[10] Just as the deepfake contains moving images that represent a person and their bodily actions, so it contains audio that represents their acts of speech. When we view the manufactured footage, we attribute actions of both kinds to the person depicted therein. It is, for example, a fictionalised Obama who delivers a promise and a thumbs-up in a deepfake of the former president. When the deepfake is especially convincing, moreover, an audience may fail to recognise that what they are seeing or hearing is a fiction, and instead take themselves to be witnessing real footage of a real person. Although the deepfake speech acts are fictional, then, they are apt to deceive. In what follows, I will qualify this speech as "fabricated", "counterfeit", "simulated" and so on, in order to flag that it is not delivered with intent and sincerity by a real human speaker.[11]

It is useful to distinguish these speech acts—the ones 'in' the deepfake, as it were—from any speech act that is performed by the creator or distributor of that deepfake. There are many speech acts that I might execute by making and sharing doctored footage of a powerful person, for example. I might express my moral disapproval, or make a joke at their expense, or defy their authoritarianism, or commit an act of protest.[12] And in the core cases to be advanced below, there are open deepfakes whose authors undertake a deceptive act—an act of fooling or misdirection. But these are not generally to be identified with the speech acts that appear to be delivered by the speaker in the film. If I construct a satirical deepfake of Donald Trump, for example, then I may use it to perform a public act of mockery, but I personally do not say the things that he appears to say in the footage. My focus here will be on the manufactured speech acts that are presented in the content of a deepfake; and only where necessary will I comment on the speech acts of its author.

Now we can revisit the distinction between "closed" and "open" recordings, as it applies to the category of deepfakes. In a closed deepfake, the depicted speakers talk only to one another and not to an external audience. For example, when a fake video makes it appear as though you and I are conspiring to commit a crime together. Any illocutionary acts that form part of the content of the recording—tellings, promises, apologies, etc.—are not directed to viewers in the outside world.

An open deepfake, meanwhile, is one in which an agent *does* appear to address the audience through the fourth wall, and to direct to them an illocutionary act such as an instruction, invitation, request, or declaration. The intended recipient of the act may be a general audience, for instance with hoax footage of a political candidate making a promise to the electorate. Or it may have a more narrow target, such as when it appears

[10] They are thus not the same as speech acts in acting or impersonation, which involve pretence (Searle, 1975; Alward, 2009).

[11] Consider, by comparison, a deception that operates through the use of a hoax document, such as a fake court summons or wedding invitation. Do these documents deliver a 'real' request or command? No: they are fictional versions of those speech acts, attributed to the court or to the engaged couple.

[12] For discussion of how artworks can be used to do things like this, see Levinson (1995).

that one family member makes a request of another, or a teacher seems to deliver an instruction to her class. In either type of case, the simulated speech act can be taken up by its recipient, spurring them into action. A convincing plea or command, placed in the mouth of an authoritative speaker, can mobilise its audience's behaviour. It follows that novel forms of deception are made possible by deepfake technology—the ability to sway a listener's actions, in order to further a malevolent end.

3.1 Closed deepfakes

Existing literature on the emerging ethical, political, and epistemic dangers of deepfakes has tended to take closed deepfakes as its subject matter. Here, I briefly survey the major concerns that have so far been raised in response to the imminent threat of powerful deepfake technologies.

As with other forms of misinformation, such as doctored photography or fabricated eyewitness testimony, a foundational worry is epistemic. It is the concern that audiences will be misled by deepfakes and come to hold false beliefs about the events they depict. As technology improves and deepfakes become more prevalent, the growing stock of untrustworthy footage will lead to widespread doxastic error (e.g., Rini, 2020; Diakopoulos & Johnson, 2021; de Ruiter, 2021). Given the ease with which digital media are shared across the Internet, a convincing deepfake may quickly take hold in the public consciousness, with deleterious consequences for democracy and civic debate (Citron & Chesney, 2019). These epistemic and political concerns arise from thinking of deepfakes largely on the model of misleading *evidence*—a convincing but unfaithful record of past events, that is apt to deceive the observer into holding inaccurate beliefs about a subject.

A more subtle threat to doxastic practice is identified by Rini (2020), who argues that deepfakes have the potential to jeopardise the positive role that faithful recordings play in regulating the testimony of those in the public eye. The ever-present possibility that what you say is being recorded by a smartphone or a camera crew, she argues, provides "good reason to be as sincere and competent as possible" (p. 3), lest the recording be later used against you. In the event that lifelike deepfakes become endemic, however, it will be easier for a speaker to deny that they said and did what's on the tape, because it will be easier to discredit it as a fake. Indeed, the very act of labelling material as a deepfake is likely to generate enough public controversy to undermine its credentials whether it is real or not, and so "[w]e will all confront a suddenly plausible skepticism about the knowledge-bearing potential of video and audio" (p. 8).

When a deepfake acts as a deceitful record of what a person has done, this agent may be the victim of one or more morally significant harms. If a deepfake makes it appear as though one has said something contentious, slanderous, bigoted, or sectarian, for example, one may become a target of criticism, abuse, or social stigma. One may suffer reputational damage, the loss of employment, election defeat, and public disgrace. One

may be regarded as hypocritical, malicious, or unwise, and suffer the interpersonal consequences of these negative attitudes.[13]

An additional class of harms arises in an especially pernicious form from the creation and distribution of pornographic deepfakes (e.g., Franks & Waldman, 2019; Harris, 2021; Rini & Cohen, 2022). Firstly, these materials can objectify women. Women who are depicted in pornographic deepfakes may be subject to "virtual domination" (Rini & Cohen, 2022, p. 143) by men who manipulate their likeness for sexual satisfaction and feelings of power, having gained "the ability to treat women's images as playthings" (op. cit. p. 146). Secondly, a person who is forced to explicitly deny that they spoke or acted in the ways that a deepfake depicted has thereby been coerced into issuing unwanted testimony, and so are wronged in their capacity as a speaker. Thirdly, Rini and Cohen hypothesise that a person who is depicted in a deepfake may come to lose faith in their own autobiographical memory (op. cit. p. 153). Footage that repeatedly appears to show me acting in a way that contradicts my personal memory is a form of gaslighting that may confuse and disorient me, and eventually cause me to lose my grip on reality.

All of these epistemic and moral harms are made possible by closed deepfakes. In each case, the effect lies in the consumer of the deepfake treating it as a faithful document of something that has happened in the past.[14] In no case is there a speech act addressed to the viewer or listener from within the recording. The deception lies only in the content of the deepfake, which is a misleading representation of what a person or group of people has said and done.[15] In the next section, we will see that open deepfakes make novel forms of deception possible—in short, those that more directly mobilise an agent's actions in support of some duplicitous end.

3.2 Open deepfakes

Open deepfakes are not only misleading fabrications of a person's behaviour; they also involve counterfeit illocutionary acts that are directed to the outside world. Open deepfakes can make it appear to you as though a person, such as a figure of authority or a trusted acquaintance, has delivered to you a particular proposal, request, offer, instruction, or invitation. It follows that open deepfakes can be used to wield power over their audience: to cause them to follow an apparent command, for example, or to treat some course of action as permissible or forbidden. When the deception is successful, that is, its primary perlocutionary effect upon the victim is not simply to

[13] Harris (2021) argues that, although worries about a deepfake-induced 'epistemic catastrophe' are overblown, deepfakes may nonetheless exert implicit effects upon the viewer, leading them to have negative attitudes towards the person depicted in the video even though they know it is not real.

[14] In the pornographic cases, the mode of engagement may be imagination or fantasy, rather than belief.

[15] Rini and Cohen (2022, p. 154) consider "a fake video showing you making a disadvantageous promise or bet", but this is not yet an open deepfake: it is a fake record of an illocutionary act having been made in the past, to someone other than the current audience (the promise or bet is thus like the declaration of marriage in the old wedding video). Diakopoulos and Johnson's "scenario 4" (2021, p. 2095) is an instance of an open deepfake: a public figure appears to encourage citizens to go out and vote, but provides false information about polling station locations.

propagate false beliefs about past events and utterances, it is to more directly influence their conduct.

With improving technology, we can expect deepfakes to become *bespoke*, in the sense of being tailored to a specific recipient. We can thus expect them to avoid some of the barriers to success faced by deceptive recordings in our examples above (the restaurant worker, job applicant, and high school party cases). They can address an individual by name, for example, and their vocabulary, tone, and delivery can be artfully manufactured. It is not difficult to construct examples—some hypothetical, some closer to real events—that illustrate the shape of what would be involved in a deepfake command, permission, plea, or invitation, nor to see what would make them persuasive. In this section, I introduce a set of cases that will allow us to draw out salient dimensions of open deepfakes and their consequences.

Consider the ballistic missile alert that took place in January 2018, during which a warning of incoming nuclear attack was sent, erroneously, by text-message to phone users across the state of Hawaii.[16] Before it was revealed to be false alarm some 38 min later, the alert received substantial uptake—with many recipients seeking emergency shelter. It is not hard to imagine that future warnings might be delivered to mobile phones in the form of a short video address by the US President or Defence Secretary. An authoritative speaker stressing the urgency of the threat and issuing clear and immediate instructions would provide a compelling call to action. This possibility, in turn, makes room for a deepfake subterfuge: a malevolent agent could manufacture a clip of the presidential command, distribute it across the cellular network, with the perlocutionary effect of causing recipients to flee for safety. The efficacy of the deepfake could be improved by refining its details. It could, for example, name the time and place at which the missile strike is due, and give directions to local points of shelter. This would amount to a more sophisticated deceit than, say, simply sounding a fake air-raid siren: it would give the appearance that a specific command had been delivered.

A second context in which orders are issued is within a military hierarchy. In March 2022, following the Russian invasion of Ukraine, a deepfake was released in which Ukrainian president Volodymyr Zelenskiy appeared to instruct his troops to lay down their arms and return to their families.[17] In this instance, the deepfake was crude and easily debunked: Zelenskiy's vocal delivery is stilted; his face is oddly expressionless; and the proportions of his head and body are off. It is likely that few viewers were taken in by the deception. However, there is a not-too-distant possible world in which the deepfake is far more persuasive, capturing Zelenskiy's likeness and mannerisms, the cadence of his speech and tone of voice, and so on. If carefully deployed, a convincing deepfake could have targeted an audience of Ukrainian forces, and addressed to them the appearance of a request or instruction to surrender. It could have sown confusion and hesitation; wasted time; and convinced soldiers to abandon their positions.

The suggestion is not, of course, that every combatant would follow fake-Zelenskiy's order, however authentic the footage might appear. After all, military

[16] See, e.g., Wong and Barney (2018). Citron and Chesney (2019, pp. 1781–1782) also note the relevance of this case for discussions of deepfakes.

[17] See, e.g., The Washington Post (2022).

directives are usually delivered through a formal chain of command, not via online video messages. But it is not hard to imagine an authentic-looking deepfake of this kind having some degree of efficacy in the turbulent conditions of warfare, irrespective of its provenance. And there are details we could add to the scenario to magnify the deepfake's chances of success—such as making it appear as though the president is in captivity, or by making him address certain staff by name and rank. Possibilities therefore exist for one side of a military campaign to deploy open deepfakes to exert a disruptive effect upon the other through falsifying a sequence of commands.

Next, consider how a simulated invitation might be delivered by a deepfake speaker. Suppose that you and I operate rival retail businesses, and I construct a deepfake of you inviting (or urging, or tempting) customers to come and take advantage of a generous discount at your store today. I circulate the footage widely on local social media pages and mailing lists, and many people take up the promotion, before becoming disappointed and hostile when you refuse to honour it. The deception is a success insofar as the fake invitation has the perlocutionary consequence of enticing buyers to your premises, only to suffer a poor consumer experience that damages your reputation. Again, we can stipulate certain details that would add to this deception's efficacy. For instance, that you are a trusted figure who regularly uses video as a marketing tool; that the discount is time-limited, and so forth. Notice, too, that the stakes are lower for the recipients of the deepfake in this more prosaic scenario, compared to those on the battlefield. The penalty for taking the promise of a discount at face value is small—it is just the effort expended in making a fruitless trip to the shops. So the addressee of the speech act may not be on their guard for the possibility of deception, or may be willing to ignore the minor risk that the deepfake is not what it appears.

The next example involves a fabricated plea or request. Online crowdfunding platforms enable their users to solicit targeted financial support by inviting others to donate towards the cost of a business venture, a sports team, a school project, a medical or legal bill, and so forth. Often, the funding request is delivered in the form of a video that aims to persuade the viewer that the cause is a good one—for instance that the business will be a success, or that there is an urgent need for life-saving care. The opportunities for crowdfunding fraud are not difficult to identify: an agent could appeal for funds for a scheme that doesn't exist, invent a charity campaign, or exaggerate the sums required for a legitimate project. Deepfake technology opens a further avenue for deception, by letting us fabricate the speaker who delivers the plea. Suppose that we create footage of a popular YouTube host, actor, or sportsperson, for example, and have them appear to issue the request for donations. The video would include details of how to make a payment to the cause, and if successful the deepfake would channel funds into an account of the deceiver's choosing. Were the deceiver to bypass official crowdfunding sites and allow the deepfake to circulate as widely as possible, its (dubious) provenance might be largely hidden to the average Internet user. Online media content can gain a life of its own and so reach a large volume of consumers in a short space of time. Even a relatively small uptake of the deepfake's appeal, therefore, might generate a non-trivial income.

Lastly, consider how a deepfake might be used to make it appear as though some course of action has been permitted or forbidden. I might play a prank on a colleague, for example, by generating a deepfake of our Head of Department granting them an

extra week off work; or fool my sibling by making a deepfake of our parents setting them a curfew. More seriously, I might make it appear to an employee that they have been authorised by their CEO to make a substantial transfer of funds to me.[18]

These one-to-one deceptions might be engineered by purely auditory means, through cloning a particular speaker's voice, and so they don't require the more elaborate multimedia resources of our previous examples. The means by which these fakes might be distributed, too, could be comparatively straightforward: it might be enough just to leave the message on the recipient's voicemail. Their efficacy is likely to depend upon considerations of timing, word choice, the terms by which the recipient is addressed, and whether the speech act fits into the wider narrative of the relationships involved (e.g., whether our parents are in the habit of setting boundaries like this, or whether business is routinely conducted via voice message). Permissions and prohibitions, when sanctioned by a suitably authoritative speaker, can have wide-ranging perlocutionary effects. All but a narrow suite of behaviours might be made to appear out of bounds, for instance, and so an addressee may be induced to take a certain predictable course of action. Conversely, a whole array of options might be freed up after a period of restricted choice, with the recipient being spoilt for choice. By creating the appearance of an open or closed space of behavioural possibilities, the deepfake can nudge an audience this way or that.

Notice, in closing this section, that a deepfake that delivers a command or prohibition might exhibit a *self-reinforcing* character. A message that demands its own immediate deletion, for example, or that forbids the recipient from answering the phone or accessing the internet, might thereby disguise its fraudulent purposes. Similarly, a request that the message remain private, classified, or otherwise undisclosed might allow it to escape the sceptical scrutiny of others. In reverse, a deepfake might command its own widespread distribution: in the crowdfunding case, for example, the speaker might include a plea that the message be forwarded to others, in order to maximise exposure. A carefully designed deepfake might thus sow the seeds of its own success.

4 Threats and harms

We have seen that in the scenarios above, the deceptive role of each open deepfake is not a purely epistemic matter. That is, they are not only attempts to infix false beliefs in the audience to which they are addressed. They are attempts to manipulate that audience's behaviour through the non-assertoric content of represented speech acts: to give viewers the convincing impression that an instruction, invitation, request, permission, or prohibition has been delivered by a suitably authoritative speaker. The moral badness of the deepfakes, therefore, does not reside simply in the extent to which they compromise their victims' epistemic condition by leading them away from the truth. It lies, principally, in the perlocutionary consequences of the counterfeit speech acts they contain: most obviously, in the complex human behaviours they bring about in their audience.

[18] See, e.g., Stupp (2019) for a real life case.

In this section, I articulate the kinds of moral crime that would be committed *if* the deepfake is successful, before summarising some of the considerations that improve that chance of success. The guiding thought is that if open deepfakes are unlikely to fool anybody, or if there are clear strategies to combat their efficacy, then we can be optimistic about the future threat they pose. I will argue, however, that there are at least some grounds for pessimism.

4.1 Deepfake crimes

Deepfakes have several potential victims: the person whose likeness has been exploited; those who are misled by the material; and wider stakeholders, institutions, or the community at large. In Sect. 3.1, we surveyed the moral wrongs that might be committed when closed deepfakes are made public—where these are understood largely in terms of various harms that might be suffered by the person who is depicted therein, such as damage to their reputation or their political prospects, gaslighting, and objectification. Many of the same harms might be endured by those who feature in open deepfakes. For instance, it might be made to appear as though a speaker has delivered a cowardly, vindictive, or reckless command; asked an ignorant or embarrassing question; or set a rule or permission that is prejudiced or unfair. The harm in each case would reside in any false impression that consumers of the fake reach regarding the speaker's character and motivations, and the consequences of this false impression.

Of greater interest for current purposes are those forms of moral wrongdoing that arise from the illocutionary and perlocutionary dimensions of open deepfake utterances, some of which are straightforward and others more subtle. Moral transgressions can be seen, firstly, in the deceptive acquisition of the various material rewards accrued through the manipulative use of an open deepfake, including the financial goods won in cases like the crowdfunding or commercial fraud examples, or the tactical advantage derived in the warfare scenarios. In more local, low-stakes cases, the payoff might consist in triumphing in a domestic dispute, playing a successful practical joke, or improving one's position in a social hierarchy. The victim in these cases is the recipient of the faked illocutionary act in question, or the wider institution they represent; and the deceiver is guilty of profiting from them by illicit means. Secondly, immoral actions may be set into motion by an open deepfake, whose victims may be the population at large or a targeted group. For instance, it might be made to appear as though an influential speaker has called upon their supporters to commit violence, or to discriminate against others on racist or homophobic grounds. There is little subtlety in this kind of case: the wrong committed by those who create and distribute the deepfake is determined by the bodily, financial, and social harm it causes to a third party.[19]

There is, however, a further and more insidious layer to the moral wrongdoing hereabouts, and it hinges upon the relationship of trust that can exist between the speaker and recipient of a successful illocutionary act. In brief, interpersonal acts of

[19] Given that illocutionary acts often depend upon institutional authority, and given that such authority is unequally distributed in society, it is likely that marginalised groups will be unequally vulnerable to these sorts of harm.

speech—tellings, promisings, invitations, apologies, petitions, vows, and so forth—are transactions that take place within ongoing human relationships. Sometimes, these relationships are superficial, temporary, or purely instrumental, as when we ask a stranger for directions, or order a meal at a restaurant. Here, 'trust' may amount simply to an unreflective expectation that one's interlocutor will tell the truth to the best of their ability, or that there is a general norm against deliberate deception in the relevant context. But elsewhere, the success of a speech act rests upon more emotionally rich bonds of trust established over time between persons whose relationship is substantial and enduring (Jones, 1996). A child may follow her parent's instruction because she has implicit faith in their goodwill, for example, and she wouldn't follow a command issued by just anyone. I will accept my close friend's request for a favour because we enjoy a shared history of mutual support and kindness. You will go and pick up your sister from the station, after dark and in the rain, because of an attachment that stretches back to childhood. Similar considerations apply to long-term romantic couples; teachers and their students; a therapist and a patient; and so forth.

There is a special kind of cruelty, I suggest, in invading or exploiting relationships like these by fabricating an act of speech and delivering it from one member of a partnership or group to another. Notice that when a deepfake speech act moves an addressee to action, it does so not by force or threat—it is not a straightforward form of coercion—nor simply by generating a false belief in the listener. It operates by taking advantage of the trust that exists between the listener and the person who appears to be speaking. When I manipulate you into acting upon a plea from your wife or husband by using a deepfake clone of their voice, for example, my deception exploits a cherished marital relationship. It taps into your unreflective willingness to respect your beloved's request or to come to their aid whenever necessary. The underlying moral intuition here is that it is *not my place* to treat this relationship as a resource for my ends; and that it is a violation of privacy and an invasion of intimacy for me to do so. Of two deepfakes that achieve an identical material advantage for the deceiver, one of which does so by impersonating a faceless and unidentified bureaucrat, and the other by fabricating a cry for help from a child to a parent or a dying wish from a lifetime companion, it is surely the latter that constitutes the more grievous moral transgression. It follows that the moral crime committed when an open deepfake is deployed is not always restricted to the illicit acquisition of material gains, or to the mobilisation of unethical behaviour in others—it can also include a form of trespass or encroachment upon valued human relationships.

4.2 The threat of open deepfakes

To finish, let us sum up how the examples in Sect. 3 illustrate the extent of the threat that open deepfakes might come to pose as future technology improves. To evaluate this threat, we need to consider those factors that bear upon the uptake of deepfake illocutionary acts by their recipients. On an optimistic view, barriers are already in place to prevent even the most realistic deepfakes from deceiving their audience. Most obviously, as we saw in the Zelenskiy case, there is the substantial challenge of making

it appear as though the communication is arriving from a credible source.[20] If it will forever remain difficult to convince a rational agent that they are really being addressed by the speaker who is fabricated in a deepfake, then the chances for a malefactor to mobilise an audience to action by this method will be slim.

In some situations, there will indeed be obstacles to the effective deployment of open deepfakes. There are institutional environments in which channels of communication are formalised and subject to continual scrutiny. We cannot expect to issue persuasive instructions to the Prime Minister's staff just by fabricating her voice and playing it to them over the phone, for example. There are contexts in which priority is given to in-person speech or to written communication, leaving little room for deepfake recordings to find purchase. Where video or voicemail messages are not standard commerce—such as in many workplaces, families, and social circles—any such media might arouse their addressees' suspicion. Deepfake illocutionary acts are thus not a magic wand with which we can manipulate a recipient's behaviour at will, and so we should not overstate their risks.

However, there are competing factors that are likely to improve an open deepfake's chances of success. First, there are cases—like the crowdfunding example—where the sheer weight of numbers is in the deceiver's favour. With sufficient exposure online, a deepfake might need only a low rate of uptake in order to achieve a substantial impact. For example, a call for civil disobedience that appears to have been issued by a popular protest group; or a plea for the electorate to spoil their ballots that seems to come from an influential spokesperson. Second, there are cases—like the missile alarm example —where the relevant action is time-critical. If one receives an urgent request or instruction that apparently derives from a trusted or authoritative figure, then one may feel pressured not to delay matters by investigating its provenance too closely. Third, there are cases—like the shopping discount example—where the stakes are low enough to discourage a deepfake's recipient from attending to its authenticity, or where the situation seems too trivial to merit an elaborate deception. A deepfake that is designed to get someone out of the room for a few minutes, for instance, may escape detection because the costs of acting on it seem minimal. Fourth, there are scenarios where the causal origin of a faked illocutionary act *is* easy to disguise, such as when the message is broadcast over citizens-band radio or left on an anonymous voicemail service. The source can be hard to discern, too, when the deepfake is copied and distributed across multiple venues, such as different social media platforms. Fifth, a deepfake might be deployed in conjunction with other, corroborating forms of deception. A letter or email telling the recipient to expect an incoming request or invitation, say, might prime them to act upon the deepfake when it arrives; especially if the first missive can itself be made to look authentic. Sixth, self-reinforcing deepfakes include mechanisms that aim to conceal their deceptive character; for instance by demanding that they never be revealed to a wider audience.

These facts, I suggest, give us grounds to take the potential dangers posed by open deepfakes seriously. In the wrong hands, these are technological artefacts that can manipulate people's behaviour, by exploiting quite ordinary forms of at-a-distance

[20] This is the kind of worry that motivates Harris (2021), who argues that anxiety about the epistemic threat of deepfakes has been overblown.

communication and trading on established interpersonal relationships. A fabricated illocutionary act—a request, command, invitation, warning, plea, or threat—may propel its audience into action, and in doing so help to bring about whatever wider ends the deceiver wishes to pursue. Like all forms of subterfuge, there is no guarantee of success, and some consumers (such as new or inexperienced internet users) may be more vulnerable than others. A savvy recipient will be vigilant to the deepfake's authorship, while outlandish or highly anomalous speech acts are likely to arouse more general suspicion. Even so, open deepfakes represent a distinctive new weapon in the deceiver's arsenal, and it would be complacent to ignore it.

5 Conclusion

This paper has drawn a distinction between two types of recorded speech act, and extended this distinction to the fabricated utterances that are made possible by deepfake technologies. Closed deepfakes are those that act as a quasi-evidential record of a person's speech and behaviour, and their principal measure of success is their effectiveness in generating false beliefs about that conduct—fooling the viewer into thinking that the depicted scenes are real. Open deepfakes are those that deliver a targeted fabrication of an illocutionary act to an external audience, and their success is determined by the act's perlocutionary consequences—in particular, the effectiveness with which they move others to action. A selection of examples has demonstrated the variety of deceptive practices in which open deepfakes might be made to operate; the immoral ends that might be pursued therein; and the measures that might be taken to enhance their success.

Acknowledgements Thanks to Ody Stone, Lucy Osler, and two referees for this journal, all of whose very helpful comments much improved the paper.

Declarations

Conflict of interest No conflict of interest reported.

References

Alward, P. (2009). Onstage illocution. *The Journal of Aesthetics and Art Criticism, 67*(3), 321–331.

Austin, J. L. (1962). In J. O. Urmson & M. Sbisá (Eds.), *How to do things with words.* Harvard University Press.

Bátori, Z. (2018). Photographic manipulation and photographic deception. *Aisthesis, 11*(2), 35–47.
Citron, D., & Chesney, R. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review, 107*, 1753.
Cohen, T. (1973). Illocutions and perlocutions. *Foundations of Language, 9*, 492–503.
Cooke, B. (2019). When art can't lie. *British Journal of Aesthetics, 59*(3), 259–271.
de Gaynseford, M. (2011). Speech, action, and uptake. In M. de Gaynesford (Ed.), *Agents and their actions*. Wiley-Blackwell.
de Ruiter, A. (2021). The distinct wrong of deepfakes. *Philosophy and Technology, 34*, 1311–1332.
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.
Dixon, D. (2022). Lies in art. *Australasian Journal of Philosophy, 100*(1), 25–39.
Eaton, M. M. (1973). Liars, ranters, and dramatic speakers. In B. Tilghman (Ed.), *Language and Aesthetics: Contributions to the philosophy of art.* University Press of Kansas.
Franks, M. A., & Waldman, A. E. (2019). Sex, lies, and videotape: Deepfakes and free speech delusions. *Maryland Law Review, 78*(4), 6.
Gustafsson, M. (2020). On the distinction between uptake and perlocutionary object: The case of issuing and obeying orders. *Inquiry*. https://doi.org/10.1080/0020174X.2020.1724562
Harris, K. R. (2021). Video on demand: What deepfakes do and how they harm. *Synthese, 199*, 13373–13391.
Hinchman, E. (2005). Telling as inviting to trust. *Philosophy and Phenomenological Research, 70*(3), 562–587.
Jones, K. (1996). Trust as an affective attitude. *Ethics, 107*(1), 4–25.
Levinson, J. (1995). Messages in art. *Australasian Journal of Philosophy, 73*(2), 184–198.
Longworth, G. (2019). Illocution and understanding. *Inquiry: An Interdisciplinary Journal of Philosophy*. https://doi.org/10.1080/0020174X.2019.1667869
Martin, M. G. F. (2012). Sounds and images. *British Journal of Aesthetics, 52*(4), 331–351.
McDonald, L. (2020). Your word against mine: The power of uptake. *Synthese, 199*(1–2), 3505–3526.
McMyler, B. (2013). The epistemic significance of address. *Synthese, 190*(6), 1059–1078.
Rini, R. (2020). Deepfakes and the epistemic backstop. *Philosophers Imprint, 20*(24), 1–16.
Rini, R., & Cohen, L. (2022). Deepfakes, deep harms. *Journal of Ethics and Social Philosophy, 22*(2), 143–161.
Searle, J. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press.
Searle, J. (1975). The logical status of fictional discourse. *New Literary History, 6*(2), 319–332.
Searle, J. (1976). A classification of illocutionary acts. *Language in Society, 5*(1), 1–23.
Stupp, C. (2019). Fraudsters used SI to mimic CEO's voice in unusual cybercrime case. *Wall Street Journal*, 30 August. https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402
The Washington Post. (2022). A falsified video of Ukrainian President Zelensky showed how deepfakes can be disarmed. 23rd March. https://www.washingtonpost.com/opinions/2022/03/23/zelensky-geopolitical-deepfake-reaction-disarmed/
Viebahn, E. (2019). Lying with pictures. *British Journal of Aesthetics, 59*(3), 243–257.
Wong, J. C., & Barney, L. (2018). Hawaii ballistic missile false alarm results in panic. *The Guardian*, 14 January. https://www.theguardian.com/us-news/2018/jan/13/hawaii-ballistic-missile-threat-alert-false-alarm