Philosophy & Technology (2024) 37:120
https://doi.org/10.1007/s13347-024-00812-1

RESEARCH ARTICLE

"Deepfakes and Dishonesty"

Tobias Flattery[1] · Christian B. Miller[1]

Received: 16 April 2024 / Accepted: 26 September 2024

Abstract

Deepfakes raise various concerns: risks of political destabilization, depictions of persons without consent and causing them harms, erosion of trust in video and audio as reliable sources of evidence, and more. These concerns have been the focus of recent work in the philosophical literature on deepfakes. However, there has been almost no sustained philosophical analysis of deepfakes from the perspective of concerns about honesty and dishonesty. That deepfakes are potentially deceptive is unsurprising and has been noted. But under what conditions does the use of deepfakes fail to be honest? And which human agents, involved in one way or another in a deepfake, fail to be honest, and in what ways? If we are to understand better the morality of deepfakes, these questions need answering. Our first goal in this paper, therefore, is to offer an analysis of paradigmatic cases of deepfakes in light of the philosophy of honesty. While it is clear that many deepfakes are morally problematic, there has been a rising counter-chorus claiming that deepfakes are not *essentially* morally bad, since there might be uses of deepfakes that are not morally wrong, or even that are morally salutary, for instance, in education, entertainment, activism, and other areas. However, while there are reasons to think that deepfakes can supply or support moral goods, it is nevertheless possible that even these uses of deepfakes are dishonest. Our second goal in this paper, therefore, is to apply our analysis of deepfakes and honesty to the sorts of deepfakes hoped to be morally good or at least neutral. We conclude that, perhaps surprisingly, in many of these cases the use of deepfakes will be dishonest in some respects. Of course, there will be cases of deepfakes for which verdicts about honesty and moral permissibility do not line up. While we will sometimes suggest reasons why moral permissibility verdicts might diverge from honesty verdicts, we will not aim to settle matters of moral permissibility.

Keywords Deepfakes · Artificial Intelligence · AI · Generative AI · Machine learning · Technology ethics · Ethics · Honesty · Dishonesty · Deception

Extended author information available on the last page of the article

Published online: 04 October 2024

1 Introduction

Less than a month after Russia's invasion of Ukraine in early 2022, a video began spreading on social media in which Ukrainian President Volodymyr Zelensky appeared to be standing at a podium, urging Ukrainians to surrender. Although the video looked real enough, at least to a casual observer, it was a fake—a *deepfake*. A typical deepfake video is a digital video in which, using deep neural network-powered artificial intelligence techniques, one person's likeness has been at least partially superimposed over the likeness of another person originally appearing in a video.[1] Deepfake techniques also can be used to produce realistic audio fakes of a person's speech. In the Zelensky deepfake, Russian disinformation agents (or their proxies) presumably recorded an actor, of roughly the same bodily proportions as Zelensky, standing at a podium giving a speech. Then, leveraging extant (authentic) video and audio recordings of Zelensky, they trained a deep learning system to model Zelensky's facial movements, and mapped a simulation of Zelensky's face overtop the actor's face in the video, adding in similarly simulated and synchronized voice audio.[2]

Obviously, one immediate goal of the producers and initial distributors of the Zelensky deepfake was dishonest: to deceive Ukrainian citizens and soldiers into believing that their president directed them to cease their resistance to the Russian invasion. Deepfake technology is even more well known, and notorious, for its use in producing pornographic videos in which famous actresses or actors appear to feature (Cole, 2017). In at least some—perhaps many—of these cases, too, the deepfakers' intentions were surely dishonest, at least insofar as they intended their audiences to believe that their deepfakes were authentic.

In both the political and pornographic cases of deepfakes just mentioned, there are numerous reasons for concern: risks of political destabilization and even the fall of a sovereign nation (e.g., Ukraine), explicit depiction of an actor or actress without consent, erosion of trust in video and audio as reliable sources of evidence, and more. These concerns have been the focus of recent work in the philosophical literature on deepfakes.[3] However, there has been almost no sustained philosophical analysis of deepfakes from the perspective of concerns about honesty and dishonesty. That deepfakes are potentially deceptive is not surprising and has been noted.[4] But under what conditions does the use of deepfakes fail to be honest? And which human agents,

[1] This rough and ready characterization of deepfakes is sufficient for our purposes. But for a more careful discussion of deepfake technology and its close cousins, see Paris & Donovan, 2019. Millière, 2022 argues that deepfakes and similar techniques constitute "a genuine paradigm shift in media synthesis" (2022: 24). For a general philosophical discussion of deep neural networks, see Buckner, 2019.

[2] There have been more recent deepfakes of political figures, but so far none has been as globally high profile or had as much potential for political upheaval.

[3] For discussion of political and moral concerns about deepfakes, see Chesney & Citron, 2019, Cole, 2017, Floridi, 2018, Öhman, 2019, de Ruiter, 2021, Harris, 2021, Hosanagar, 2021, Kerner & Risse, 2021, Young, 2021, and Rini & Cohen 2022. For discussion of epistemic concerns about (or that can be extended to) deepfakes, see Hopkins, 2012, Cavedon-Taylor, 2013, Fallis, 2020, Rini, 2020, Harris, 2021, and Kerner & Risse, 2021.

[4] See, e.g., Rini, 2020, de Ruiter, 2021, and Hosanagar, 2021.

involved in one way or another in a deepfake, fail to be honest, and in what ways? If we are to better understand the morality of deepfakes, these questions need answering. Our first goal in this paper, therefore, is to offer an analysis of paradigmatic cases of deepfakes in light of the philosophy of honesty.

Even though it is clear that many uses of deepfakes are morally problematic, there has been a rising counter-chorus claiming that deepfakes are not *essentially* morally bad, since there might well be uses of deepfakes that are not morally wrong, or even that are morally salutary, for instance, in education, entertainment, activism, and other areas.[5] However, while there are surely reasons to think that deepfakes can supply or support moral goods, it is nevertheless possible that these uses of deepfakes are dishonest. Our second goal in this paper, therefore, is to apply our analysis of deepfakes and honesty to the sorts of deepfakes hoped to be morally good or at least neutral. We conclude that, perhaps surprisingly, in many of these cases the use of deepfakes will be dishonest in some respects.

Of course, there might well be cases of deepfakes for which verdicts about honesty and moral permissibility do not line up. It might be that the use of one deepfake is honest but nevertheless morally wrong. And it might be that the use of another *is* dishonest but nevertheless morally permissible, all things considered. While we will sometimes suggest reasons why moral permissibility verdicts might diverge from honesty verdicts, we will not aim to settle the question of overall moral permissibility, which would require consideration of potentially many factors unrelated to honesty.

But if the dishonesty of dishonest deepfaking is neither necessary nor sufficient for moral wrongness, why does it matter if we have an analysis of its dishonesty? There are at least two reasons why having this analysis is philosophically significant. First, it is one of the standard goals of moral philosophy to provide philosophical explanations of, and grounds for, commonly shared moral intuitions. As noted above, one of our primary goals is to do precisely this, for the intuition that various deepfakes are (or are not) dishonest. Second, even if dishonesty might not be necessary or sufficient for moral wrongness, nevertheless it is an important factor for determining overall verdicts concerning moral permissibility or wrongness.[6]

How important a factor? We are inclined to think there is at least a *pro tanto* moral reason against dishonest actions—perhaps even a *prima facie* moral obligation to be honest—in which case a good, overriding reason would be required for thinking that dishonesty is morally permissible. Moreover, how honesty ought to be weighed surely will depend on readers' background moral theoretic commitments, which we cannot hope to address here. But we believe that the following analysis and arguments will be useful in helping to arrive at more philosophically clear and comprehensive all-things-considered moral decisions concerning the use of deepfakes.

[5] See, e.g., Dhillon 2019, Fallis, 2020: 626–627, Kerner & Risse, 2021: 97f, and Rini & Cohen 2022.

[6] It is not unusual in applied ethics, including technology ethics, to advance moral reasons that are weighty despite being neither necessary nor sufficient for overall moral verdicts. A recent example from a different subfield: some philosophers have argued that one reason in favor of giving some rights to robots is that, if we don't, we might end up treating them harshly, which might increase the odds that we'll do the same to humans. But this reason is likely overridable if, e.g., giving rights to robots would cause other problems. See Flattery 2023.

In the following section, we sketch the most prominent recent account of honesty in the philosophical literature. In § 3, we give an analysis of the honesty of paradigmatic uses of deepfakes, distinguishing between the main phases of a deepfake's lifecycle (production, distribution, and viewing) and the main agents involved in each stage. In § 4, we extend our analysis to several proposals for what might be taken to be morally salutary uses of deepfakes.

2 Honesty

Our goal is to examine whether and how the use of deepfakes might fail to be honest. But what *is* honesty? In this paper we will take on board Christian Miller's (2021) recent account of honesty, since it is the most prominent recent account, and since we are inclined toward it. Fortunately, most of our discussion could be adapted for other recent accounts of honesty, if readers find specific aspects of Miller's account problematic.[7]

Miller's focus is on honesty as a moral virtue, understood along traditional Aristotelian lines as involving cognition, motivation, and outward behavior.[8] Hence an honest person is disposed to think honest thoughts, have honest motives, and behave honestly, both across a variety of situations relevant to honesty and stably over time. In this paper, however, we are focused on evaluating a person's *actions* as honest or as dishonest, but not their character.[9] With respect to actions specifically, Miller's core approach can be captured as follows:[10]

(i) An agent acts honestly when she does not intentionally distort the facts as she sees them, in contexts she understands to be factual.
(ii) An agent acts dishonestly when she does intentionally distort the facts as she sees them, in contexts she understands to be factual.

We shall try briefly to unpack a few features of this approach. First, "intentionally", for Miller, is to be understood as the opposite of "accidentally." Thus, behavior that arises from subconscious mental states can still count as dishonest. Second, Miller does not have an account of what "distorting the facts" amounts to, but he does offer as a close synonym "misrepresenting." If Sam tells his teacher that his dog ate his homework, he would be misrepresenting or "distorting the facts" in communicating to the teacher what happened to his homework. Third, Miller's account ties honest and dishonest behavior to subjective construals of the facts. So, on his approach, peo-

[7] See, e.g., Smith, 2003, Guenin, 2005, Adams, 2006, Carr, 2014, Baehr, 2017, and Wilson, 2018.

[8] For more on character more generally, see Miller, 2013 and Miller, 2014.

[9] Of course, honest or dishonest actions might well be *evidence* for underlying honest or dishonest character. But an isolated action is not conclusive evidence for underlying character. Dishonest people might act honestly at times, and even a fairly honest person might, on occasion, act dishonestly.

[10] What follows is derived from Miller, 2021: 71, 134. Over the course of his book, Miller provides a variety of revisions to the proposals above to handle various complexities, but those revisions do not bear on our discussion in this paper.

ple five thousand years ago were not acting dishonestly when they reported that the Earth was flat, even though they were seriously mistaken about the objective facts.

Finally, by "factual contexts", we mean contexts in which the agent believes that representing the facts is relevant and likely to be expected by many people. This condition is important, since in some contexts distortions of facts do not plausibly amount to dishonesty. For instance, an author writing historical fiction is representing historical times and places to her readers, but representing non-factual events as occurring in those times and places. The author reasonably believes that readers will not take the context—i.e., reading historical fiction—to be a purely factual context.[11] Similarly, when filmmakers use CGI to represent fictional superheroes flying through space or tossing train cars with ease, they reasonably believe that viewers will not take the context—i.e., watching superhero movies—to be a purely factual context.[12]

What forms do *dishonest* actions take? We note first that we are concerned here with communicative actions that result in the formation of propositional attitudes (e.g., beliefs, doubts, etc.). Often these kinds of actions take linguistic forms (e.g., speech, writing), but sometimes not. To use one of Miller's examples (2021: 13), if a shady salesman applies a fresh coat of paint atop a car's rusty roof, the action communicates propositional content about the car's condition to prospective buyers. Or if a corrupt police officer plants drugs in a person's pockets during a raid, the action communicates propositional content to other law enforcement officials. For convenience, however, our examples will typically involve linguistic actions.[13]

While there are a number of ways to be dishonest, perhaps the ways most likely to be relevant to discussions of deepfakes are lying, misleading, and bullshitting.[14] On the traditional account of *lying*, a person lies in communicating *that P*, in typical cases, when she intends her audience to believe that *P* is true, while herself believing that *P* is false.[15] For instance, if we told you we loved free climbing sheer rock faces, and intended for you to believe us, we would be lying.

A person *misleads* in communicating *that P*, in typical cases, when he believes that *P* is true, but intentionally communicates it in such a way that, he hopes, his audience will form a false belief about *Q*.[16] For instance, in the 1990 film, *Quigley Down Under*, the protagonist, Quigley, is asked by the antagonist whether he has any skill with a pistol. Quigley replies, "I never had much use for one," which was true, despite his being quite skilled indeed with a pistol.

A person *bullshits* in communicating *that P*, in typical cases, when in doing so he is mainly trying to get his audience to take some course of action, but is not concerned about whether *P* is true. "He does not care," as Harry Frankfurt put it, "whether the

[11] Of course, most readers of historical fiction expect *some* historical facts to be represented accurately, even if the characters and main plot are fictional. So, even with historical fiction, the context is at least partly factual.

[12] Pierini (2023: 18) raises a similar point about CGI.

[13] We thank an anonymous referee for suggesting the evidence fabrication case.

[14] Miller (2021: 7–22) also discusses, as actions incompatible with honesty, stealing, cheating, and promise-breaking, and notes more briefly others as well.

[15] See Miller, 2021: 8. For an overview of philosophical accounts of lying, see Fallis, 2010.

[16] See Miller, 2021: 10–14 for further discussion of misleading.

things he says describe reality correctly. He just picks them out, or makes them up, to suit his purpose" (Frankfurt, 2005: 56). For instance, a politician might tell his audience, "My opponent doesn't care a whit about inflation!" If the politician's concern is gaining his audience's votes, and he does not care one way or the other about the truth of his claim, he is bullshitting.[17]

We will take a broad view of dishonesty as encompassing all those sorts of actions, though we do not claim that producing and distributing deepfakes could not fail to be honest in *other* ways as well.

Finally, it is important to note again that, in our view, from an action's being dishonest it does not immediately follow that the action is morally wrong. For instance, most people would judge that families who harbored Jews during the Nazi occupation, but lied to the Gestapo about doing so, did nothing morally wrong, despite the fact that their doing so was dishonest. Similarly, from an action's being honest it does not immediately follow that the action is morally permissible, as telling the truth to the Gestapo in this case illustrates.

3 Paradigmatic Cases

In this section, we will use Miller's account of honest action to introduce a basic model for evaluating deepfakes vis-a-vis honesty, and then we will apply that model to a paradigmatic kind of deepfake. Not all deepfakes are paradigmatic ones, of course, but for simplicity and clarity's sake, we will begin by considering such a case. Our conclusion about the paradigmatic case will be unsurprising: those who produce and distribute these sorts of deepfakes engage in dishonesty. But, in examining the paradigmatic case, we give a more precise explanation of *why it is* that the use of paradigmatic deepfakes is dishonest, even if there is little doubt that dishonesty was afoot. Moreover, in so doing we introduce our general method of examining the use of deepfakes with an eye to honesty and dishonesty, which lays the groundwork for examining a number of other deepfakes in the following section, some of which are not so obviously dishonest and perhaps even seem morally above board in all respects.

3.1 A Paradigmatic Deepfake

Although nothing critical turns on what qualifies as a "paradigmatic deepfake", and although not all deepfakes will satisfy fully the following description, we will understand a paradigmatic deepfake to be a digital video; based on an original video recording; intended to deceive some viewers; produced without the consent of the people featuring in the original recording, or the people falsely represented as featuring in the deepfake, or the people who produced the original video; and for which the act of production and distribution is morally dubious (or worse). For the sake of convenience, we will typically talk about deepfakes using visual language, and thus use video deepfakes as the paradigmatic medium of deepfakes. But deepfakes need

[17] See Miller, 2021: 19, 53–54 for further discussion of bullshitting and honesty.

not be videos; they can be, for instance, digital audio files. Our analysis of deepfakes is intended to apply as well to non-video deepfakes.

A number of different sorts of deepfakes could qualify as paradigmatic, but most of the scholarly and popular attention has concerned pornographic deepfakes. This is not because deepfakes have any intrinsic affinity with pornographic media. Rather, this is largely because of the sad truth that, so far, the vast majority of deepfakes are pornographic (Simonite, 2019), and also because these sorts of deepfakes raise glaring moral concerns about the treatment of and attitudes toward women.[18] We will use a case of a pornographic deepfake as a paradigm case. However, since we suspect that many of those falsely represented as featuring in real pornographic deepfakes would prefer the existence of those deepfakes not to be advertised, we will not cite such cases and instead use a fictional case, including only enough detail needed for our analysis. Consider the following fictional case, which certainly has real analogs:

> Adult Film: Adult Productions produces a pornographic film featuring Amy and Bob engaging in an explicit sexual encounter. [Call this recording *Adult Film*$_A$, where the subscripted "A" stands for "authentic".] Both Amy and Bob are adult film actors, and both consent to being recorded for the film. Some time later, Chuck, a proficient user of deepfake software, sets out to produce a deepfake that will use *Adult Film*$_A$ as its primary source, and that will falsely represent Dana, a public figure having no association with pornographic films, as the female lead. Chuck's purposes for producing the deepfake are to see what it would look like for Dana to do what Amy does in *Adult Film*$_A$, and also to distribute the deepfake to others, with at least some desire that others will believe it to be an authentic recording. Chuck gathers a number of authentic video and audio recordings of Dana, trains the deepfake software to model her facial likeness and voice dynamically, and replaces Amy's face with Dana's, resulting in a deepfake video that appears to feature Dana and Bob. [Call this deepfake *Adult Film*$_D$, where the subscripted "D" stands for "deepfake".] Chuck stores *Adult Film*$_D$ on his personal computer, views it, and then uploads it to Shady Forum, a popular and unregulated online discussion forum. Ed, a Shady Forum user, views *Adult Film*$_D$ and then posts a link to it on Social Site, a popular social media platform, exposing the deepfake to a broader audience.

3.2 Phase-agent Analysis

We can think of a typical deepfake's lifecycle as having three very general phases: the *production phase*, during which the deepfaker gathers source materials and produces the deepfake; the *distribution phase*, during which the deepfaker (or others) distribute the deepfake, making it available to people for viewing; and the *viewing phase*, during which people view the deepfake. Of course, in some cases phases might overlap (e.g., a deepfaker might view the deepfake while producing it), recur (e.g., a deepfake

[18] For discussion about pornographic deepfakes' effects on women, see Cole, 2017, Öhman, 2019, Kerner & Risse, 2021, Harris, 2021, de Ruiter, 2021, Young, 2021, Viola & Voto, 2023, and Rini & Cohen 2022.

might be distributed multiple times), or be missing (e.g., a deepfake might never be distributed or viewed).

We can also think of any deepfake as involving agents who fill (in some cases unwittingly) some or all of the following roles, which we will outline using the Adult Film case above. Chuck is the *deepfaker*, that is, the person who produced the deepfake. Dana is the *faked-in* person, that is, the person whom the deepfaker artificially inserted into the deepfake (*Adult Film$_D$*), and thus who was misrepresented as featuring in the deepfake, but who was never actually recorded for the primary source recording (*Adult Film$_A$*) or the deepfake (*Adult Film$_D$*). Amy is the *faked-out* person, that is, the person who was recorded for and featured in the primary source recording (*Adult Film$_A$*), but whose likeness was removed or overwritten in the deepfake (*Adult Film$_D$*). Bob is a *faked interactant*, that is, the person who was recorded for and featured in the primary source recording (*Adult Film$_A$*), and during which he interacted with the faked-out person. Adult Productions is the *faked-from* party, that is, the party having a legitimate authorial right to the authentic content in the primary source recording (*Adult Film$_A$*). Chuck and Ed are both *distributors*, that is, both made the deepfake (*Adult Film$_D$*) available to additional audiences. Shady Forum and Social Site might also be considered distributors, even if only unwittingly. Chuck and Ed are both *consumers* of the deepfake (*Adult Film$_D$*), that is, they both view the deepfake.

Note, however, that not for every deepfake will there be agents who fill all these roles. For instance, in the Zelensky deepfake noted above, only the faked-in Zelensky appears, so there is no faked interactant. Also, in the case of audio deepfakes in which there was no original audio recording serving as the base media onto which faked content is overlayed, there is a faked-in party, but no faked-out party, since no other person's auditory likeness is removed.[19]

While surely not every case of a pornographic deepfake involves dishonesty, we take it to be obvious that dishonesty was afoot in the Adult Film case. But using our phase-agent analysis, along with Miller's account of honest actions, we can, in a more precise way, try to answer the following questions about the Adult Film case: were there any actions that failed to be honest? If so, whose actions were these, and why were they not honest? When discussing other deepfakes in subsequent sections of this paper, we will not step through each and every phase and agent involved. But it is important to do so for this first, paradigmatic case, to make our general method clear.

3.2.1 Distribution Phase

Consider the distribution phase first, since it is the phase during which deepfakes are shared with a broader audience. In the Adult Film case, clearly none of the faked-in (Dana), faked-over (Amy), faked interactant (Bob), or faked-from (Adult Productions) parties failed to be honest with respect to actions in this phase, since none of these parties was aware of the distribution of the deepfake, or even that the deepfake existed or would exist. So, none of these parties intentionally distorted any of the

[19] For an example of an audio-only deepfake, see § 4.1 below.

facts relevant to what appears in *Adult Film*$_D$, though of course all these parties were misrepresented in some way by other parties in the case.

Clearly, the deepfaker, Chuck, failed to be honest. After creating the deepfake, Chuck distributed it by posting it in an online forum. This was a communicative act, since he knew that posting the deepfake would make it viewable to other users. In the case as described, Chuck intended for other users to believe that the deepfake was an authentic recording. Since Chuck knew that the deepfake was inauthentic, he thus intended to distort the facts as he understood them. Thus, Chuck's act of distributing the deepfake was dishonest.

But what facts, precisely, did Chuck distort for his intended audience? Several. Most obviously, Chuck misrepresented Dana as engaging in sexual acts with Bob, acts which Dana never engaged in. But Chuck misrepresents more than this. To the extent that Dana's bodily appearance differs from Amy's, Chuck misrepresents Dana's bodily appearance, since it will appear that Amy's body is Dana's.[20] Chuck also misrepresents Dana as being recorded engaging in these acts with Bob. Further, since *Adult Film*$_A$ was recorded with Amy's and Bob's knowledge and consent—let us suppose it is evident that the film was not recorded using hidden cameras—Chuck also misrepresents Dana as both knowing that she was being recorded and consenting to being recorded.[21] These are all distinct facts. Having a particular bodily appearance is a fact about one's physical features. Doing something is a fact about one's actions. Being recorded doing something is a fact about one's actions being captured by a recording device. Knowing that one is being recorded is a fact about one's epistemic mental states. Consenting to being recorded is a fact about one's volitional mental states. Chuck intentionally conveyed all these distorted facts about Dana to viewers of *Adult Film*$_D$, and thus was dishonest about all these matters.

Chuck also intentionally distorted facts about Bob, Amy's sexual partner in *Adult Film*$_A$ and the faked interactant in *Adult Film*$_D$. Since Chuck's deepfakery did not target Bob, Bob's authentic likeness still appears in *Adult Film*$_D$; and so Chuck did not distort the facts about Bob's likeness itself, nor about his engaging in sexual acts. But Chuck did misrepresent Bob as engaging in sexual acts *with Dana*, which Bob did not do, and so Chuck's misrepresentation was dishonest. Chuck did not distort the facts about Bob's being recorded, knowing he was being recorded, or consenting to being recorded. But Chuck did misrepresent Bob as being recorded *with Dana*, as knowing about being recorded *with Dana*, and as consenting to this.

Chuck also may have intentionally distorted facts about Adult Productions, the company that produced the original film, *Adult Film*$_A$. If Chuck's deepfake, *Adult Film*$_D$, includes the display of "Adult Productions" in the opening or closing credits of *Adult Film*$_A$, then Chuck will have misrepresented Adult Productions as having recorded Dana and Bob (rather than Amy and Bob), knowing they recorded Dana and Bob, and perhaps other distortions as well. Or, if Chuck removed all mentions

[20] This will often be the case, too, in non-pornographic deepfakes, though of course less of the faked-in person's bodily appearance is misrepresented. For instance, in the Zelensky deepfake mentioned in the introduction, since the actor in the primary source recording is not Zelensky himself, his body will have slight differences, even if they are not obvious to casual viewers.

[21] If attributing knowledge to Amy and Bob is too strong, we can say instead that Amy and Bob believed they were being recorded.

of Adult Productions, he will have distorted the facts about who produced the video, perhaps making it seem as if he was responsible for all the content of the video (not just the faked parts).[22]

We said above that Chuck *intentionally* conveyed a number of distorted facts to viewers, and thus was dishonest about a number of facts. But is Chuck really dishonest about *all* the facts he distorts? While he surely was aware that he was misrepresenting Dana as *doing* certain things—namely, engaging in sexual interactions with Bob—it is unlikely that he thought expressly about each of the other facts he distorted. If dishonest actions are intentional distortions of the facts as one understands them (in contexts believed to be factual), how could Chuck be dishonest in representing Dana as having a certain bodily appearance, knowing she was being recorded, and so on, if he had no specific intention to convey those particular distorted facts?

Miller's account of honesty alone does not settle this question.[23] But since we think it is worth seeing how this question might be settled, we will sketch a couple of approaches for further specifying the account. On either approach, Chuck fails to be honest to some extent. On the first approach, one can intentionally distort a fact only if one expressly intends to distort that particular fact.[24] The distortion of that particular fact must be part of the content of one's intention. On this way of specifying intentional distortion, Chuck intentionally distorted the fact about Dana's interaction with Bob—since Dana did *not* interact with Bob—and so Chuck is dishonest for doing so. Similarly, he intentionally distorted facts about Dana's not being recorded interacting with Bob, about Amy's interactions with Bob, and so on. But since Chuck did not *intentionally* distort some other facts, he was not dishonest in distorting those particular facts.

On the second approach, one acts intentionally with respect to outcome O_1, if either (a) O_1 is part of the content of one's intention (broadly construed)[25] in acting, or (b) O_1's happening follows from the outcomes O_2-O_n that *are* part of the content of one's intention, and one reasonably ought to have known so (e.g., if one were to have paused for a moment to think, one would have seen the entailment), and if one had been aware of the entailment, one still would have done the action. Consider an example unrelated to distorting facts. Suppose Gary is with a group of people, all of whom are his friends except for Harry, whom Gary does not care for. Gary wants to make fun of Harry, to make Harry look like a fool. Gary succeeds. Harry looks like a fool. Gary's friends laugh, and as a result, Harry is embarrassed. While Gary did not explicitly consider that making fun of Harry also would embarrass him, had Gary thought for a moment longer, he *would have* seen that making fun of Harry also would embarrass him. Moreover, had Gary taken that extra moment to consider, thus having both outcomes in view—viz., that Harry would look like a fool *and* that he

[22] Roberts (2023: 43) expresses similar ideas, but about how viewers are deceived by the content of deepfakes.

[23] See Miller's discussion of the term "intentionally" (2021: 30–31). Following Miller, our usage of this term is broad, including not only acting from an intention, but also other kinds of mental states like desires, hopes, and feelings.

[24] This intention would be understood *de re* as opposed to *de dicto*.

[25] See n. 23 above.

would be embarrassed—Gary would have acted no differently. Both outcomes would have been part of the content of Gary's intention.

On this second approach to specifying "intentionally", Gary intentionally *both* made Harry look like a fool *and* embarrassed him, which at least seems plausible to us. Similarly, it seems plausible that, had Chuck taken a moment to consider what facts about Dana he would be distorting with his deepfake, he *would have* seen that the deepfake misrepresents not just Dana's behavior, but also some or perhaps all of the following: her bodily appearance, her being recorded, her knowledge of being recorded, and her consenting to it. And it seems plausible that Chuck nevertheless would have produced the deepfake and distributed it. So, on this second approach, Chuck intentionally distorted some or perhaps all these further facts about Dana, and thus was dishonest on each count. We prefer this second approach, but perhaps others will prefer the first approach. On either approach, Chuck dishonestly distorts at least some facts about Dana.

Finally, what about Ed, the other distributor of *Adult Film*$_D$ in this case? Did he act dishonestly? If Ed believed *Adult Film*$_D$ to be a deepfake, but posted a link to it on Social Site at least in part so that others would believe the deepfake to be authentic, then what he does parallels lying. Or if, while believing *Adult Film*$_D$ to be a deepfake, Ed posted a link to it without concern for its distortion of the facts—e.g., if he simply wanted people to enjoy a video he enjoys, and the video's authenticity did not matter to him—then what he does parallels bullshitting. Either way, by promoting the deepfake, he, too, intentionally distorted the facts (mis)represented in the video to his audience, and thus acted dishonestly. If instead Ed believed *Adult Film*$_D$ to be authentic, then, while his distributing it might well be a moral failure on other grounds, it would not be a failure of honesty.

3.2.2 Production Phase

Again, it seems clear that none of the faked-in (Dana), faked-out (Amy), faked interactant (Bob), or faked-from (Adult Productions) parties failed to be honest with respect to any actions relevant to the production phase. None was aware of the existence of the deepfake or, we can suppose, of any intent to produce a deepfake based on *Adult Film*$_A$.

While it is clear, we think, that Chuck acted dishonestly in distributing *Adult Film*$_D$ to others, did his act of *producing* this deepfake fail to be honest? After all, producing the deepfake media file is distinct from posting it online, just as writing a (physical) letter is distinct from mailing it to someone. But because, in this case, distributing the deepfake with at least some hope of viewers believing it to be authentic was one of Chuck's intentions *for* producing it, Chuck's act of producing it is dishonest. This is because, it seems to us, the correct account of Chuck's action of producing the deepfake involves his intentions for doing so. And since his intention to get others to believe the deepfake to be authentic is, in part, an intention to distort the facts as he understands them, Chuck's act of producing the deepfake is dishonest.[26,27]

But suppose we alter the case slightly. Suppose that, while Chuck did intend to distribute *Adult Film$_D$* (e.g., by posting it on Shady Forum), he had no specific intention to deceive anyone. Rather, he was simply unconcerned about whether anyone would believe that *Adult Film$_D$* was authentic. Perhaps Chuck relished the thought that others would view his digital creation, or that others would enjoy it as he does. Would Chuck, in this version of the case, be dishonest? While the case in the previous paragraph parallels lying, this version of the case parallels bullshitting: the fact that his deepfake distorts the facts is of no concern to him with respect to distributing the deepfake to others. So it is bullshit. But bullshit is not honest.[28] If Chuck were to distribute his deepfake just as he did in the original version of the case—i.e., without any indication that the video is a fake—he would be intentionally distorting the facts as he understands them, even if he had no specific intention to deceive.[29]

Consider another variation of the Adult Film case, in which Chuck produced *Adult Film$_D$*, but without any intention to distribute it, at least not at the time he produced it. Would Chuck's act of producing the deepfake, in this variation of the case, be dishonest?[30] Although we are uncertain about this, we lean toward answering "no". On the one hand, some people might have the intuition that simply *producing* a deepfake like the one Chuck produced amounts to some sort of failure of honesty. On the other hand, Chuck did not, in this variation of the case, distort any facts *to* anyone. So it is unclear to us how merely producing a deepfake—despite its having fictitious or distorted representations of real events—can be, strictly speaking, dishonest, *absent* any intent to distribute it *to* anyone. But if one still seems to have the lingering intuition—as at least one of us might have—that Chuck is dishonest in producing the deepfake even if he did not intend to distribute it, perhaps this merely reflects the conflation of a more general intuition of *moral wrongness* with the recognition that deepfakes involve distortions of facts. But again, since we are uncertain about this sort of case, we leave it to readers for further consideration. And, of course, we are not at all suggesting that Chuck's act of producing the deepfake was not a moral failure for *other* reasons.[31]

[26] Alternatively, one might understand Chuck's producing the deepfake to be, not a distinct action in its own right, but rather a *part* of the more drawn out action of trying to deceive some viewers into thinking that Dana did what she was falsely depicted as doing. On that sort of view, it wouldn't be apt to evaluate Chuck's producing the deepfake as being honest or not, since producing the deepfake was not a distinct action. We are, of course, taking for granted positions on issues debated in the literature on action theory. For a useful overview of action theory, see Wilson & Shpall 2016.

[27] Thanks to Raphael Mary Salzillo for discussion of these options.

[28] See Miller, 2021: 19, 53–54.

[29] What if Chuck had included a label or disclosure that the deepfake was a fake? We discuss cases with labels/disclosures in § 4.3.

[30] We are content to assume, without evidence, that at least some pornographic deepfakes are produced for private viewing, even if the deepfaker later decided to distribute the deepfake.

[31] For related discussion, but from a different angle, see Öhman's discussion of the "pervert's dilemma" (2019: 134–135). A version of the dilemma (or trilemma) adapted to the question of dishonesty might look like this: (i) privately fantasizing about a person isn't dishonest; (ii) producing a private deepfake of a person isn't morally different from privately fantasizing about that person; (iii) producing a private

3.2.3 Viewing Phase

In the Adult Film case, Chuck and Ed each view *Adult Film$_D$*. But could the mere *viewing* of a deepfake be a failure of honesty? We think this is very unlikely. Chuck, the deepfaker, is dishonest both in distributing the deepfake and in producing it—so long as his reasons for producing it involved distributing it—but not in merely viewing it. Similarly for Ed: even if he exhibits other moral failings in his activity on Shady Forum and Social Site, merely viewing the deepfake does not itself seem dishonest. We are not suggesting, however, that viewing a deepfake is never morally wrong, just that it is does not seem *in itself* to be a failure of honesty.

3.3 Intentions and Non-Factual Contexts

According to our stipulated characterization of "paradigmatic deepfakes" above, the deepfaker intends to distribute the deepfake, and intends for at least some viewers to believe the deepfaked content to be veridical, which implies that the deepfaker believes that he is distributing the deepfake into a factual context (i.e., a context in which representing the facts is both relevant and expected). But if this is *not* the case, *then* would producing or distributing a deepfake be dishonest?[32] This will come up again when we consider special cases of deepfakes in the following section. But it will be helpful to address the question in a general way here.

Adapting a term from Viola and Voto (2023), let us use "overt deepfake" to refer to a deepfake whose content most reasonable viewers would judge as obviously both not representing reality and not so intended by the deepfaker. For instance, in a viral video in 2019, Steve Buscemi's face was deepfaked onto Jennifer Lawrence's body (Kelleher, 2019). It is obvious that the video's content was not authentic, not so much because of the quality of the deepfake, but rather because of the clear mismatch of Buscemi's face with Lawrence's body; and it is thereby also obvious that it was not intended to be taken as genuine. It was intended to generate laughs, and succeeded. On our analysis, the deepfaker most likely was *not* dishonest in distributing the Buscemi/Lawrence deepfake. This is because the overtly non-genuine content of the deepfake converts the context of distribution into a non-factual context, i.e., a context in which people would not reasonably expect the content to represent the facts. And, as we noted in § 2, an agent acts dishonestly when, among other things, she believes the context to be a factual one.

It is also possible to distribute a deepfake whose content is *not* overtly fake, and yet do so without dishonesty. In yet another variation of the Adult Film case, suppose instead that Chuck posted his deepfake—the content of which is not overt—on a rather unsavory online forum called "Deepfake Celebrity Videos", which every-

deepfake about a person is dishonest. If (i) and (ii) are true, then it seems (iii) should be false, in which case merely producing deepfakes can't be dishonest. See also Kerner & Risse (2021: 134–135), where some of what they say might be read as supporting the claim that a deepfake using a person's likeness without their consent is, so long as it is kept private and not distributed, no more morally wrong than a private fantasy about that same person. We think concerns could be raised about both positions.

[32] We thank two anonymous reviewers for pressing us further on this issue.

one knows is a place for posting deepfakes rather than authentic videos.[33] If Chuck indeed knew that those who frequent this forum expect the videos disseminated there to be deepfakes, then Chuck believed the context to be non-factual, in which case Chuck did not intentionally distort the facts as he understood them in a context he believed to be factual. And so Chuck's posting of the deepfake, while perhaps morally problematic for other reasons, most likely was not dishonest. It is still *possible* that Chuck's action was dishonest, however, depending on other factors concerning the content of his intention. For instance, if he hoped that his deepfake would be further distributed beyond the unsavory forum and into factual contexts where people would likely believe the deepfake's content to be genuine, then his posting of the deepfake was dishonest.

Some authors—e.g., Viola and Voto (2023), Cavedon-Taylor (2024), and Fallis (2020)—have argued (or at least suggested) that, if deepfakes become widespread and people become widely aware of this development, it might well "[shift] our default attitude toward…photographs and videos, possibly including genuine ones, toward a skeptical and (possibly) detached attitude" (Viola & Voto, 2023: 30). And, if this happens, they reasonably suggest that perhaps most people will not be deceived by deepfakes. Adapting this line of thought to our position in this paper, if that happens, then someone who produces and distributes a deepfake in such an environment will not be dishonest in doing so. Why? Because, so long as the deepfaker is aware that most people have a general skepticism toward digital recordings, she will not satisfy the factual context condition for dishonesty. Her intent would be similar to that of filmmakers who use CGI: they reasonably expect that people will not believe the CGI to be representing reality. We agree that, *if* this sort of general skepticism about digital recordings takes root, dishonesty in deepfaking would be far less common. However, we think it is clear that, *at present*, most people do not have this sort of general skepticism. Moreover, we think it is unlikely that most people will adopt this form of skepticism any time soon, if ever. But time will tell.

The above sort of analysis can be deployed easily to examine the honesty of uses of other paradigmatic kinds of deepfakes as well, for instance, the Zelensky political deepfake noted in the introduction. But now we want to extend our analysis to some interesting kinds of *non*-paradigmatic deepfakes.

4 Special Cases

What makes the following kinds of deepfakes interesting, from our point of view in this paper, is that each has been (or could be) claimed to be, on the whole, morally salutary, or at least not morally objectionable. Rather than taxing the reader with, for each example, the more extended sort of analysis undertaken in the previous section, we will focus on the special features of these cases of deepfakes. We hope it will be clear how our verdicts in these cases might extend to a wider range of other deepfakes with similar features. Since our focus is honesty, we do not aim to give an all things considered verdict concerning the bigger-picture question of the moral permissibil-

[33] Thanks to an anonymous reviewer for this example.

ity of using these deepfakes. But since considerations of honesty ought to be at least part of—probably a significant part of—all-things-considered-verdicts about moral permissibility, verdicts about the honesty of using deepfakes will be important for answering that bigger-picture question.

4.1 Deepfakes Communicating Important Truths

What features of a deepfake might render its production and distribution honest, or at least not dishonest? Perhaps a deepfake, in virtue of its content, can communicate a message that is important and true.[34] For such a deepfake, communicating this message might well be the primary *goal* of the deepfaker and distributor(s). For instance, a number of deepfakes have been used to communicate important humanitarian or political messages. Malaria No More UK released a deepfake aimed at raising awareness about the death toll caused by malaria each year, and challenging world leaders to act.[35] David Beckham delivers the message in the video—presumably because his fame on the pitch would attract international attention—and in doing so he appears to speak nine different languages, one after the other. But Beckham himself spoke only in English; the rest was faked.[36] Taking a different approach, Solidarité Sida, a French charity organization, released a deepfake in which it appears that Donald Trump gives a short speech dismissing concerns about AIDS. Their stated goal was to move world leaders to act, "by broadcasting a piece of fake news. The first piece of fake news that might eventually become true." (Skinner, 2019)

We are happy to grant that these deepfakes' primary intended messages are true and communicate important facts. Insofar as the deepfakers' and distributors' intentions were to represent these facts, they were, to that extent, honest. However, in producing and distributing these deepfakes, they communicated more than just their primary messages. They also distorted other facts as they understood them, and in factual contexts. Malaria No More UK was aware, of course, that Beckham did not—and, we assume, cannot—speak all nine of the languages he appeared to speak. In misrepresenting him as if he did speak those languages, they intentionally distorted the facts about what Beckham did speak and about what languages he was able to speak.[37] Thus, Malaria No More UK was partially dishonest in producing and distributing their deepfake. Solidarité Sida, too, acted dishonestly, at least if they intentionally distorted the facts about Trump, even if there was no failure in honesty in their underlying message about AIDS.

Deepfakes also have been proposed and already used for broadly educational purposes. For instance, for an exhibit called "Dalí Lives", the Dalí Museum in St. Petersburg, Florida developed impressive deepfake videos of the already-deceased

[34] For arguments or suggestions to this effect, see Dhillon 2019; Fallis, 2020: 626–627; Kerner & Risse, 2021: 97–98, 102; and Rini & Cohen 2022.

[35] Davies, 2019. For discussion, see Dhillon 2019, de Ruiter, 2021: 1316, and Kerner & Risse, 2021: 98.

[36] In a similar kind of case, New York mayor Eric Adams recently used robocalls featuring his own deepfaked voice speaking the native languages of various communities. See Fitzsimmons & Mays 2023.

[37] They also distorted other, less obvious facts, e.g., that he was recorded speaking in nine languages, and so on. But for ease of discussion, in this section we leave aside those less obvious distorted facts.

Salvador Dalí appearing to introduce himself, welcome visitors, and interact with them, all with his distinctive flair. The purpose of the deepfake is "to have visitors empathize with Dalí as a human being" (Lee, 2019), and to help introduce them to Dalí's art.[38] Deepfakes have been proposed to realistically represent events that did not actually happen, but could very well have happened, and that tell us something worth knowing about history. For instance, CereProc, a technology company specializing in text-to-speech software, produced and distributed an audio deepfake in which it sounds as if John F. Kennedy is delivering the speech that his assassination prevented.[39] Even more creatively, it might be possible to create deepfake videos in which historical authors appear to deliver lectures, using only their texts, photographs, or perhaps other authentic videos of them (Fallis, 2020: 627).

Again, we are happy to grant that these kinds of deepfakes would communicate worthwhile messages, and insofar as communicating those messages is the intent behind producing and distributing these deepfakes, to that extent doing so would not be dishonest. However, producing and distributing deepfakes of this kind communicates more than just the intended historical facts. Dalí Museum decision-makers knew, of course, that Dalí himself never said or did what he appears to be saying or doing in their deepfakes, and so their conveying these distorted facts was intentional. Thus they were partially dishonest in producing and distributing their deepfake. Similarly, even though JFK wrote the speech he did not end up giving, the deepfakers knew he did not give it, and so the deepfaked audio is a distortion of the facts, and thus producing and distributing it was partially dishonest. A similar verdict would result for any deepfake representing someone as giving a lecture they did not give.

One might reply that, for some or all the deepfakes noted above, using a deepfake was the *only way*—or at least the most effective way—to communicate the relevant truths about those important facts.[40] And so, surely communicating them via deepfakes was not dishonest. First, for many deepfakes, we doubt that this is true. For instance, Malaria No More UK might have communicated their message about malaria with an authentic video of Beckham speaking in English, with subtitles in the other languages. Or they might have taken the time to coach Beckham to speak a few sentences in those other languages. Or they might have brought on additional celebrities who are native speakers of the other languages to deliver those lines. In the case of Solidarité Sida's deepfake, we doubt there are many people who would honestly claim that a deepfaked Trump dismissing the seriousness of AIDS is the most effective method of spurring world leaders into action. And while the Dalí Museum's deepfakes of their namesake are undeniably impressive, their attraction seems more in their AI-assisted artistic achievements than in their ability to convey the facts about Dalí's art and personality. After all, it is not as if there are no extant genuine videos of Dalí one can watch.

Second, however, even if it is true that a given deepfake is the most effective way to convey the intended facts, it is also nevertheless true that any other facts intentionally distorted by the deepfake are intentionally distorted. For instance, it might well

[38] Lee, 2019. For discussion, see de Ruiter, 2021: 1316 and Kerner & Risse 2021: 97.

[39] BBC 2018, Floridi, 2018: 319, de Ruiter, 2021: 1316, Kerner & Risse 2021: 97–98.

[40] Kerner & Risse (2021: 98) suggest this sort of reply.

be true that Beckham speaking those nine languages indeed would be the most effective way—which is not to say the *only* effective way—to convey to an international audience particular facts about malaria. Nevertheless, this deepfake misrepresents Beckham as having spoken those languages, and clearly Malaria No More knew this when producing it and distributing it. Perhaps there were outweighing moral reasons in favor of using this deepfake, but even if so, that does not mean the conditions for dishonesty were not met.

Similarly, watching a deepfake video of a past occurrence might well be a superior way to learn certain facts about that occurrence. For instance, to our knowledge there is no video recording of the origin of American baseball's "seventh inning stretch" tradition, when, in 1910, then-President William Taft stood up to stretch midway through the seventh inning.[41] But, despite video cameras being relatively rare at the time, there are extant video recordings of Taft. So it seems possible to produce a reenactment of the first seventh inning stretch using deepfake technology to fake Taft's likeness into the video. But even if doing so would provide a unique insight into that historical event, thus in some sense conveying the facts about that event, doing so would also involve intentionally distorting other facts. Taft would be misrepresented as being video recorded; as having a particular precise body shape—it is unlikely a perfect body double can be found, after all—as standing in a particular place, with a particular pose; as standing next to a person with the likeness of the actors used in the reenactment; and so on. And, depending on the particulars of production and distribution, the deepfake might foreseeably wind up appearing in contexts in which many believe they are viewing a factual representation of the past. Thus, to the extent that considerations of honesty are important when engaging in moral deliberation about producing and distributing any of these kinds of deepfakes, those considerations might well count against moral permissibility.

4.2 Deepfakes for Entertainment

While the deepfakes considered above were intended primarily to communicate important facts to audiences, not all deepfakes are meant for this sort of communication. Some deepfakes are meant merely for entertainment. For instance, Sway, a small technology startup, recently developed a smartphone app enabling users to star in their own deepfaked dance videos.[42] The app instructs users to record themselves from different angles, using the smartphone's built-in camera. The app then fakes the user's facial likeness into a number of pre-recorded videos in which an actor performs a dance routine. With a tap and a swipe, users can then share their dance deepfakes directly to a number of social media platforms. Some have also proposed that deepfakes might equip a wider range of media producers (e.g., small companies without the big budgets of Hollywood studios, or even individuals) with inexpensive

[41] There is, we have heard, some doubt about whether this is indeed how the seventh inning stretch originated. If this story is but a myth, that would be a shame, since it is a charming story. But other examples of unrecorded historical events would serve just as well.

[42] See Dhillon 2019. Sway's website is https://getsway.app/.

but powerful tools for creative visual storytelling (Kerner & Risse 2021: 103–104, Dhillon 2019).

If one produces and distributes a deepfake for these or other sorts of entertainment purposes, but not to convey important information, could doing so involve any failure of honesty?[43] We think so. We agree that using creative media to entertain ourselves and others is often laudable. But even in producing and distributing media intended for entertainment and not information sharing, we can still convey facts to our audiences, and thus we can also distort facts. For instance, even when one uses Sway's app to post on social media a self-deepfaked dance video for the purpose of having some fun, one nevertheless intentionally distorts the facts—the video represents one as performing those dance moves, when one in fact did not—often in contexts in which viewers would, by default, expect videos to be authentic. This would be a form of bullshitting: one intentionally distorts the facts about one's dance performance for one's audience, but without particular concern for the facts themselves. Rather, one's goal—not a bad goal in itself—is to entertain. But in the process one does indeed convey distorted facts. Similarly, if a movie studio, large or small, deepfakes actors' likenesses into scenes in which these actors themselves did not act, the studio (mis) represents the actors as having performed in those scenes, and thus the studio intentionally distorts the facts to the audience.[44] We do not, of course, claim that these sorts of uses of deepfakes always would stem from bad intentions. But failures of honesty sometimes result from harmless or even worthy goals.

4.3 Labels and Disclosures: Packaging Deepfakes to Avoid Deception

Perhaps a deepfake, during its production or distribution phase, can be packaged in such a way that its audiences will not be deceived by it, thus knowingly converting the context into a non-factual one (i.e., one in which most viewers would not expect a representation of the relevant facts). Some writers have suggested, for instance, that requiring that deepfakes be presented with some sort of warning label or disclosure might go some distance toward preventing audiences from being deceived.[45] Indeed, the DEEP FAKES Accountability Act, first proposed in the U.S. House of Representatives in 2019, would require, among other things, that many deepfakes be distributed with watermarks and/or audible disclosures alerting audiences that the video and/or audio contained manipulated content.[46] Would producers and distributers of deepfakes avoid dishonesty, so long as they packaged up their deepfakes with disclosures? It is possible, but not as easy as it might seem.

[43] This case is similar to the second version of the Adult Film case in § 3.2.2.

[44] Even if an actor's recorded performance in a particular scene is heavily altered by, e.g., CGI special effects, the actor was still recorded performing for the scene. Deepfaking an actor into a scene, by contrast, does not involve recording the actor's performance for the scene.

[45] See, e.g., Fallis, 2020: 639, du Ruiter, 2021: 1320, Kerner & Risse 2021: 106, and Harris, 2021: 13,385.

[46] The original bill, H.R.3230, failed to gain traction in the House of Representatives in 2019. See https://www.congress.gov/bill/116th-congress/house-bill/3230. In 2021, another bill, H.R.2395, also called the "DEEP FAKES Accountability Act" and containing precisely the same text, was introduced in the House, where it remains still. See https://www.congress.gov/bill/117th-congress/house-bill/2395.

By distributing a deepfake, or by producing one with the intent to distribute it, one is intentionally distorting the facts for an audience. But if in doing so one *also* intentionally takes steps to *remedy* the distorted facts for the audience—if one intentionally distorts the facts but also informs the audience about the distortions—perhaps one does not fail to be honest. We see three conditions here: if a deepfake producer or distributor packages a deepfake with a disclosure, does so for the purpose of preventing audiences from being deceived by the deepfaked content, *and* believes that the disclosure will prevent audiences (of sound mind) from being deceived, then probably the producer or distributor does not fail to be honest.[47] If any of those three conditions is not met, the act of producing or distributing fails to be honest.

Recall the Adult Film case from the previous section. Suppose Chuck, the deepfake producer and distributor in that case, adds a disclosure to his deepfake and merely believes that the disclosure will prevent audiences from believing that, for instance, Dana in fact starred in the film. But if Chuck did not add the disclosure *in order to* prevent deception, he does not satisfy the second condition noted just above, and thus his action still fails to be honest. Suppose the DEEP FAKES Accountability Act had become law, and Chuck added the disclosure only to stay out of trouble, but all the while still hoping that at least some viewers would be deceived. If so, Chuck would have intentionally distorted the facts for his viewers *without* really intending to remedy the distorted facts. Or, suppose instead that Chuck added the disclosure, and did so for the purpose of preventing deception, but believed it likely that many viewers still would be deceived. In that case, Chuck would be intentionally distorting the facts, but not intentionally remedying the distorted facts to a significant degree by his own lights, since one cannot intentionally φ (e.g., remedy the distorted facts for one's audiences) while believing that one is not φ'ing.

How feasible is it that one could meet the conditions above, and thus produce or distribute deepfakes with disclosures and do so honestly? Although we cannot fully address this question here, we suggest that it is not as feasible as might be assumed. Consider some ways in which a deepfake might be packaged with a disclosure.[48] If a video with deepfaked content is distributed in a physical medium (e.g., a disc), a disclosure label might be added to the video's physical packaging. This would be an ineffective method of disclosure, however, since many viewers would not see the label. Since this is easily foreseeable, it is unlikely the distributor would believe that the disclosure would prevent viewers from being deceived. But then the distributor would fail to be honest in distributing the deepfake, since, in distributing it with a disclosure believed to be ineffective, they would be intentionally distorting the facts but *not* intentionally remedying the distortion.

Alternatively, a disclosure might be added either to the beginning or the end of a deepfake, perhaps in the opening or closing credits in the case of movies or television shows. It takes but a moment's thought, however, to foresee that many, perhaps most, viewers of the video will not see a disclosure added to the closing credits. Similarly,

[47] Why only "probably"? We confess to having a lingering concern that, even in such cases, there might still be a failure of honesty. But we are uncertain about this, and so we will set the lingering concern aside.

[48] We consider only video deepfakes as examples here, but similar examples and concerns can be given for audio-only deepfakes.

many viewers will likely not see a disclosure at the beginning of a video, whether due to inattentiveness, fast-forwarding past the usual piracy warnings and credits, or even viewing edited clips of the deepfaked content on streaming services such as YouTube. If the deepfake distributor is aware of these likelihoods, then they will have, in distributing the deepfake, intentionally distorted facts for their audience without intentionally remedying the distortions.

A more targeted approach might involve displaying a disclosure only during the segment of a video in which deepfaked content appears. For instance, the 2016 film, *Rogue One: A Star Wars Story*, included a scene in which the already-deceased actor Peter Cushing appeared to play a role, but which was accomplished using a stand-in actor and deepfake technology (Lincoln, 2016). No disclosures were given in the film at all, and so probably some viewers assumed that Cushing was alive during the filming of *Rogue One*. If the filmmakers had added a watermark or other visual notice appearing during that scene and disappearing afterward—e.g., a simple "Deepfake!", or a more informative "Dear Audience, be aware that Peter Cushing did not actually appear in this scene. Guy Henry performed for this scene, which was then digitally altered so that Cushing's likeness appears instead of Henry's"—viewers would be less likely to assume they were watching a performance by Cushing, and thus the filmmakers might have avoided a failure of honesty.

But these sorts of disclosures are unlikely to be used much in film or other visual arts. Watermarks or other disclosures covering a significant portion of the screen would be at least somewhat visually jarring, detracting from the aesthetic experience intended by many filmmakers. And for other sorts of visual media less focused on viewers' aesthetic experience, deepfakers with overt intentions to deceive would be motivated to avoid visual disclosures, since it would defeat the communicative purpose of a dishonest deepfake.

What about watermarks covering only a small portion of the video or image, so as not to be significantly visually jarring? As Viola and Voto (2023) note in their discussion of the app, DeepNude—which takes an image of a clothed person, and outputs an AI-generated nude image of that person—less obtrusive watermarks can be cropped away easily. Clearly, doing so before distribution would likely be dishonest. But even if the producer and initial distributor do *not* crop the watermark, it is easy for them to foresee, at least for some types of deepfakes, that someone *else* is likely to redistribute the deepfake with the watermark cropped away. And for deepfakes intended to be distributed broadly to a general audience, or even to a niche audience in widespread or indeterminate locations, those who will see the redistributed version of the deepfake lacking the watermark will indeed still be the intended audience. Such a case would be similar to adding a disclosure to the opening or closing credits in a film, discussed above, where the distributor does *not* actually believe that the disclosure will prevent the intended audience from being deceived.[49] So, at least in a

[49] That the audience's membership is broad and indeterminate is important here. To see why, consider a different sort of case. Suppose I tell my friend Xavier, a known exaggerator, about the fish I caught last weekend. My intended audience is Xavier. Even if I know Xavier is likely to retell a distorted version of my story to others, those others weren't my intended audience. So, even if I was unwise in telling my story to Xavier, I wasn't dishonest. Thanks to an anonymous referee who pressed us for clarification here.

range of cases, visual disclosures either would not be employed or would not clearly prevent the one who added the watermark before distribution from being dishonest.[50]

Finally, another interesting method is to include the disclosure as part of the deepfaked content itself. For instance, in 2018, the Flemish Socialist Party published a deepfake in which Donald Trump appears to call on Belgium to exit the Paris Climate Accords (von der Burchard, 2018; Rini & Cohen 2022). At the end of the video, the deepfaked Trump says, "We all know climate change is fake, just like this video." The Flemish Socialist Party's deepfake was fairly crude, so most viewers would be unlikely to believe it to be a genuine recording of Trump.[51] But had the deepfake been realistic, it seems plausible that at least some viewers would have thought the video authentic, and the disclosure merely a joke. In addition, a disclosure of this sort would not be a practical solution for many deepfakes.

More could be said about how the use of disclosures might prevent dishonesty in producing and distributing deepfakes, but we hope it is clear at least that adding a disclosure to a deepfake does not automatically render honest the production and distribution of the deepfake.

5 Conclusion

In this paper we have offered an analysis of a range of deepfakes in light of the philosophy of honesty. While most people would readily agree *that* producing and distributing deepfakes of the paradigmatic sort—e.g., pornographic deepfakes—is dishonest, we have tried to explain in some depth *why* this is so. And while we accept that one can have morally good or at least neutral reasons for using other sorts of deepfakes, nevertheless, if what we have argued is correct, it is more difficult than one might have expected to produce and distribute even these sorts of deepfakes *honestly*.

Acknowledgements For helpful discussion and feedback, we thank Tania Flattery, Don Howard, Raphael Mary Salzillo, two anonymous reviewers, the Honesty Project and Wake Forest Philosophy Work In Progress groups, and audiences at the North Carolina Philosophical Society and the CEPE International Conference on Computer Ethics.

Author Contributions The authors are the sole authors of this work.

Funding Open access funding provided by the Carolinas Consortium. This publication was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

Data Availability Not applicable.

[50] For audio deepfakes, there might be audible analogs to watermarks. But analogous concerns could be raised for these, too.

[51] Evidently not all, however. Despite the crudity of the deepfake, von der Burchard (2018) reports that some viewers still were deceived.

Declarations

Ethical Approval Not applicable.

Consent for Participate Not applicable.

Consent for Publication Not applicable. There were no empirical studies conducted for this work, and thus no human subjects studied.

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

References

Adams, R. M. (2006). *A theory of Virtue: Excellence in being for the good*. Clarendon.

Baehr, J. (2017). Honesty's Threshold. In *Moral Psychology*, Vol. 5: *Virtue and Character*, ed. Walter Sinnott-Armstrong and Christian B. Miller, 275–286. MIT Press.

BBC. (2018). John F Kennedy's lost speech brought to life. *BBC*. Accessed 22 July, 2022. https://www.bbc.com/news/uk-scotland-edinburgh-east-fife-43429554

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass* 14 (10).

Carr, D. (2014). The human and epistemic significance of Honesty as an Epistemic and Moral Virtue. *Educational Theory*, *64*(1), 1–14.

Cavedon-Taylor, D. (2013). Photographically based knowledge. *Episteme*, *10*(3), 283–297.

Cavedon-Taylor, D. (2024). Deepfakes: a survey and introduction to the topical collection. *Synthese* 204, 14 (2024). https://doi.org/10.1007/s11229-024-04634-8

Chesney, R., & Citron, D. (2019). Deepfakes and the New Disinformation War: The coming age of Post-truth Geopolitics. *Foreign Affairs*, *98*(1), 147–155.

Cole, S. (2017). AI-Assisted Fake Porn Is Here and We're All Fucked. *VICE*. Accessed 22 July, 2022. https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn

Davies, G. (2019). David Beckham 'speaks' 9 languages for new campaign to end malaria. *ABC News*. Accessed 22 July, 2022. https://abcnews.go.com/International/david-beckham-speaks-languages-campaign-end-malaria/story?id=62270227

de Ruiter, A. (2021). The distinct wrong of Deepfakes. *Philosophy and Technology*, *34*(4), 1311–1332.

Dhillon, S. (2019). An optimistic view of deepfakes. *TechCrunch*. Accessed October 2, 2024. https://techcrunch.com/2019/07/04/an-optimistic-view-of-deepfakes/

Fallis, D. (2010). Lying and deception. *Philosophers' Imprint*, *10*(11), 1–22.

Fallis, D. (2020). The epistemic threat of deepfakes. *Philosophy & Technology*, *34*, 623–643.

Fitzsimmons, E. G., & Mays, J. C. (2023). Since When Does Eric Adams Speak Spanish, Yiddish and Mandarin? *New York Times*. Accessed 13 April, 2024. https://www.nytimes.com/2023/10/20/nyregion/ai-robocalls-eric-adams.html

Flattery, T. (2023). The Kant-inspired indirect argument for non-sentient robot rights. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00304-6

Floridi, L. (2018). Artificial Intelligence, Deepfakes and the future of Ectypes. *Philosophy and Technology*, *31*, 317–321.

Frankfurt, H. (2005). *On Bullshit*. Princeton University Press.

Guenin, L. M. (2005). Intellectual honesty. *Synthese*, *145*(2), 177–232.

Harris, K. R. (2021). Video on demand: What deepfakes do and how they harm. *Synthese*, *199*(5–6), 13373–13391.

Hopkins, R. (2012). Factive pictorial experience: What's special about photographs? *Noûs*, *46*(4), 709–731.

Hosanagar, K. (2021). Deepfake Technology is now a threat to everyone. What do we do? *Wall Street Journal*. Accessed October 12, 2023. https://www.wsj.com/articles/deepfake-technology-is-now-a-threat-to-everyone-what-do-we-do-11638887121

Kelleher, K. (2019). What Is a Deepfake? Let This Unsettling Video of Jennifer Lawrence With Steve Buscemi's Face Show You. *Fortune*. Accessed 29 July, 2024. https://fortune.com/2019/01/31/what-is-deep-fake-video/

Kerner, C., & Risse, M. (2021). Beyond Porn and Discreditation: Epistemic promises and perils of Deepfake Technology in Digital Lifeworlds. *Moral Philosophy and Politics*, *8*(10), 81–108.

Lee, D. (2019). Deepfake Salvador Dalí takes selfies with museum visitors. *The Verge*. Accessed 22 July, 2022. https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum

Lincoln, K. (2016). How Did Rogue One Legally Re-create the Late Peter Cushing? *Vulture*. Accessed 30 July, 2022. https://www.vulture.com/2016/12/rogue-one-peter-cushing-digital-likeness.html

Miller, C. B. (2013). *Moral Character: An empirical theory*. Oxford University Press.

Miller, C. B. (2014). *Character and Moral psychology*. Oxford University Press.

Miller, C. B. (2021). *Honesty: The philosophy and psychology of a neglected Virtue*. Oxford University Press.

Millière, R. (2022). Deep learning and synthetic media. *Synthese*, *200*, 231.

Öhman, C. (2019). Introducing the pervert's dilemma: A contribution to the critique of Deepfake Pornography. *Ethics and Information Technology*, *22*, 133–140.

Paris, B., & Donovan, J. (2019). *Deepfakes and cheap fakes: The manipulation of audio and visual evidence*. Data & Society. Accessed October 12, 2023. https://datasociety.net/library/deepfakes-and-cheap-fakes/

Pierini, F. (2023). Deepfakes and depiction: From evidence to communication. *Synthese*, *201*, 97.

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), 43–160.

Roberts, T. (2023). How to do things with deepfakes. *Synthese*, *201*, 43.

Simonite, T. (2019). Most Deepfakes Are Porn, and They're Multiplying Fast. *Wired*. Accessed 22 July, 2022. https://www.wired.com/story/most-deepfakes-porn-multiplying-fast/

Skinner, H. (2019). French charity publishes deepfake of Trump saying 'AIDS is over'. *EuroNews*. Accessed 22 July, 2022. https://www.euronews.com/my-europe/2019/10/09/french-charity-publishes-deepfake-of-trump-saying-aids-is-over

Smith, T. (2003). The metaphysical case for honesty. *Journal of Value Inquiry*, *37*, 517–531.

Viola, M., & Voto, C. (2023). Designed to abuse? Deepfakes and the non-consensual diffusion of intimate images. *Synthese*, *201*, 30.

von der Burchard, H. (2018). Belgian Socialist Party Circulates Deep Fake Donald Trump Video. *Politico*. Accessed 30 July, 2022. https://www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-video/

Wilson, G., and Samuel, S. (2016). Action. *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (Ed.). Accessed 4 August, 2022. https://plato.stanford.edu/archives/win2016/entries/action/

Wilson, A. T. (2018). Honesty as a Virtue. *Metaphilosophy*, *49*(3), 262–280.

Young, G. (2021). *Fictional immorality and immoral fiction*. Rowman and Littlefield.

Authors and Affiliations

Tobias Flattery[1] · **Christian B. Miller**[1]

✉ Tobias Flattery
flattet@wfu.edu

Christian B. Miller
millerc@wfu.edu

[1] Department of Philosophy, Wake Forest University, Tribble Hall B301, P.O. Box 7332,
Winston-Salem, NC 27109, USA