Deepfakes and Dishonesty

Tobias Flattery, Wake Forest University (United States)
Christian Miller, Wake Forest University (United States)
International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL

Keywords: deepfakes technology ethics, honesty, dishonesty, deception, AI

Extended Abstract

Deepfakes raise various reasons for concern: 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 yet to be sustained philosophical analysis of deepfakes from the perspective of the philosophy of honesty. 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. Therefore, our first goal in this paper 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 by some 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.

In §1, we take on board and outline the contours of the most prominent recent account of honesty in the philosophical literature, that of Christian Miller (2021). On Miller's account, an agent acts honestly when she does not intentionally distort the facts as she sees them, and an agent acts dishonestly when she does intentionally distort the facts as she sees them. We then sketch three different general sorts of dishonest action that are also potentially relevant to the production and distribution of deepfakes: lying, misleading, and bullshitting.

In §2, we use Miller's account of honest and dishonest actions to introduce a basic model for evaluating the use of deepfakes vis-a-vis honesty, and then we apply that model to a paradigmatic kind of deepfake, viz., a fictional case of a pornographic deepfake (no graphic details are employed). While not all

deepfakes are pornographic or even paradigmatic ones, beginning with a such a case helps to make the our general approach clear. We employ a phase-agent analysis of a deepfake, separating the lifecycle of a deepfake into its production, distribution, and viewing phases, and identifying a number of roles various agents might play during each phase. Our conclusion about the paradigmatic case is 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 what makes these agents' uses of paradigmatic deepfakes to be dishonest, even in a case where 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.

In the final section, we extend our analysis to a range of interesting non-paradigmatic cases of deepfakes. What makes these 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. We examine a number of these sorts of deepfakes in this section, but rather than rehearsing the more extended sort of analysis undertaken in the previous section, we focus on the special features of these cases of deepfakes. Our verdicts in these cases can be extended to a wider range of other deepfakes with similar features.

In some cases, the content of the deepfake might be thought to supply good moral reasons for—or at least to not supply moral reasons against—producing, distributing, and viewing it. We consider, for instance, deepfakes aimed at communicating important truths, such as humanitarian, political, or educational messages, and deepfakes intended for entertainment purposes. We highlight the conditions under which producing and distributing these sorts of deepfakes, even if well-intended, would be honest or dishonest.

In other cases, the consent of those whose likenesses appear in the deepfake might seem to render honest the use of those deepfakes. We give an account of why consent might be thought to render a deepfake honest, but then argue that in many cases, including many of the sorts of increasingly common cases of deepfakes used in marketing, consent will likely not succeed in doing so.

Finally, some argue that the method of packaging or distributing a deepfake—for instance, by including labels or disclosures—might help to avoid dishonesty. This is indeed a promising approach. But we argue that, while it is certainly possible to avoid dishonesty by using labels or disclosures, it is more difficult than it might seem. Our discussion of the difficulties should help to advance the more practical conversation about honestly packaging and distributing deepfakes.

Since our focus is the philosophy of honesty, we do not aim to give an all things considered verdict concerning the bigger-picture question of the moral permissibility 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. Further, while our account is primarily a theoretical one, establishing the conditions under which producing and distributing deepfakes is dishonest, it also provides implicit practical guidance and has critical practical implications. For instance, an account like

ours ought to be in hand prior to attempting to identify dishonest uses of deepfakes or crafting policies that take dishonesty into account. This will help to reduce the risks of misidentifying persons as being dishonest with deepfakes (false positives), misidentifying persons as being honest with deepfakes (false negatives), and failing to identify persons who were indeed dishonest with deepfake (missed positives).