Deepfakes, Intellectual Cynics, and the Cultivation of Digital Sensibility

 Taylor Matthews *University of Nottingham*

 *Forthcoming in Royal institute of Philosophy Supplement.[[1]](#footnote-1)*

**1. Introduction**

On the 25th December 2020, the British television company, Channel 4, broadcasted an ‘alternative Christmas Address’, in which Queen Elizabeth II spoke of the challenges of the outgoing year (2020). Given the trajectory of 2020, it was unsurprising to hear her urge caution about the Covid-19 pandemic. Perhaps more surprising, though, was the Queen’s admission that she wished to participate on the hit BBC television programme *Strictly Come Dancing*. Indeed, anyone who has listened to the Queen speak, let alone watched a Christmas Address on television, would have realised that this admission was out of character. In fact, the Queen’s admission was not just out of character – it was entirely fake. Channel 4’s ‘alternative Christmas Address’ took the form of a so-called *deepfake*.[[2]](#footnote-2)

A deepfake is a video that depicts events, people, or states of affairs that never happened. Occasionally, the term is used to refer to any video that has been manipulated to create a false pretence, but what people are usually gesturing at is a so-called *shallowfake*. These are videos that are slowed down, sped up, or whose audio is tampered with, all to depict people doing or saying things they never did. This occurred, for example, when the Speaker of US House of Representatives, Nancy Pelosi, was caught ‘slurring’ her words (Reuters, 2020). In reality, somebody had slowed down her speech on a recording, but it nonetheless convinced many that she was drunk. *Deepfakes*, by contrast, are far more sophisticated than this. The name itself hints at just how. Deepfake is a portmanteau of ‘*deep* learning’ and ‘*fake*’: the ‘deep learning’ aspect corresponds to a specific kind of ‘artificial intelligence’ (AI), and the fake element speaks for itself.

Despite their relative infancy, an increasing number of philosophers have begun expressing concerns about deepfakes. For example, Don Fallis (2020) claims that deepfakes will potentially reduce the amount of evidence videos impart to their viewers; Regina Rini (2020) worries that deepfakes will ‘erode the credentials of all videos’, leading us to increasingly distrust what we watch from videos; while the writer and political advisor Nina Schick (2020: 9) warns that deepfakes could usher in an ‘information apocalypse’ – where our digital environments are characterised by mis- and disinformation. These concerns are cleverly woven into Channel 4’s deepfake. In the tradition of setting out a clear message to the Commonwealth, the Queen cautions about a world increasingly characterised by ‘fake news’, echo chambers, and disinformation, only for her actions and facial movements to manifest the very thing she warns against. In a rapidly changing world, whom we decide trust and what sources of information we deem credible has never mattered more. With the advent of deepfakes, never has this task been harder.

My aims in this chapter are twofold: first, to show how sophisticated deepfakes are likely to make us more prone to develop what I call *intellectual cynicism*, a habitual distrust and disengagement towards epistemic practices such as deliberation, inquiry, and information-sharing. Second, to offer a novel way of addressing and offsetting this cynicism in the form of a trained *digital sensibility*, a calibrated sensitivity to the merits (and demerits) of online content. I start in the next section by setting out the technology behind deepfakes and demonstrating how these videos mark a watershed moment in media manipulation. Then, in section 3, I draw on a philosophical framework in contemporary epistemology to begin illuminating how deepfakes might put us in a position to develop intellectual cynicism. Section 4 argues for this conclusion, as well as articulating what such cynicism amounts to. In section 5, I develop the concept of a trained digital sensibility and explain how best to cultivate it. Finally, section 6 concludes by briefly considering the relationship between a trained digital sensibility and the role of ethical and intellectual virtues in navigating an increasingly technological future.

**2. Deepfakes and Deep-learning**

Image manipulation has a rich history, pre-dating the advent of the internet and digital media. Under Stalinist rule, for example, political opponents and adversaries were not just executed but routinely removed from official photographs, deleting any trace of their existence from the public eye. The classic example is that of Nikola Yezhov – at one point, Stalin’s right-hand-man and secret police official - who was notoriously removed from a photograph of Stalin and himself walking along the Moscow canal after he was assassinated. More recently, the advent of computing technology has ushered in programmes like Adobe Photoshop and CGI, which allows users to creatively edit images and video footage, inserting people and objects into places they never originally were. This has led to the creation of slightly more sophisticated versions of shallow-fakes called *cheap-fakes*, which rely on ‘cheap, accessible software’ to manipulate media (Paris and Donovan, 2019: 2). Given the extensive use (and misuse) of Photoshop and CGI, one might question whether deepfakes do pose a distinctive threat. That is, haven’t people been able to subvert the truth with deepfakes and cheap-fakes alike (Paris and Donovan, 2019: 6)? Keith Raymond Harris (2021) has recently questioned whether the pessimism surrounding deepfakes is ‘overblown’. To an extent, Harris is right. Currently, many deepfakes are blurry, costly to produce, and easily identifiable. However, this should not fool us into thinking that this will continue to be the norm. As I mentioned in the introduction, part of the name ‘deepfake’ derives from *deep learning*, and it is this technology that arguably sets deepfakes apart from its predecessors.

At its most basic, deep learning works by computing a process, then attempting to carry it out. If the programme fails, the technology revises the process and tries again until it is successful. In this way, deep learning operates using a similar ‘trial and error’ method by which the human brain learns. It is for this reason that the deep learning techniques behind many of the emerging deepfakes today – including that of Queen Elizabeth II - are called deep *neural networks*. Roughly, these can be understood as algorithmic models that predicts and generate an output – in our case a video – based on an input of images or video content. As the cyber-security researchers Yisroel Mirsky and Wenke Lee (2020) explain, the neural networks often used to make deepfakes are so-called Generative Adversarial Networks (GANs). GANs consist of two algorithms – one called the ‘generator’ and the other called the ‘discriminator’ – which *compete* against each other to manufacture highly authentic videos. After being fed the same input of images or recordings, the generator creates new samples based on the input that are good enough to trick the discriminator, while the discriminator works to identify which samples are fake. Eventually, through trial and error, the GAN produces a highly sophisticated video that is capable of manipulating the speech inflections and facial movements of a target person doing and saying things they never did.

In this process, GANs are able to manipulate the attributes on a target’s face – adding or removing features, lightening or darkening the skin colour, or swapping certain expressions on one person and superimposing them onto another. By doing this, deepfake purveyors can quite literally swap the identity of one person with another, and in some cases, even generate entirely new, non-existent faces from scratch (Tolosana et al., 2020).[[3]](#footnote-3) This sort of deep learning manipulation is popular in the video game and film industries, but it has also been exploited to create faces to match fake online profiles on social media platforms from Facebook to Twitter. In the early days of GAN technology, the photographic and video content required to create a compelling deepfake was often in the hundreds or thousands, but recently Samsung has developed a technology that allows purveyors to do this with only a handful of photos of videos (Zakharov et al., 2019). As the authors of the study report, a reasonably sophisticated deepfake could be created using only a single image. Alongside this, researchers at Princeton University have collaborated with Adobe to develop ‘VoCo’ technology, which allows video creators to modify the audio of a recording by typing the desired words into a transcript. The authors of the research observed how the algorithm was able to quickly analyse voice samples from a target, then synthesise what that person’s voice would sound like, were they to say the words typed into the transcript (Jin et al., 2017). As the lead author, Zeyu Jin, (2017: 96:10) reports, when the synthesised words were inserted in the context of a spoken sentence, the modified sentence was ‘often perceived as indistinguishable from other sentences spoken in the same voice’.

Together, the use of VoCo technology and GANs marks a watershed moment not only in deepfake technology, but in media manipulation more broadly. For one, use of GANs has meant that purveyors can efficiently produce deepfakes that are increasingly indiscernible from authentic video footage. As the technology learns to improve, and the amount of video input required to produce a sophisticated fake decreases, deepfakes will inevitably become more widespread. This is observable, for example, in the rapidly growing number of deepfake apps and programmes available including Zao, Reface, Wombo, and DeepFace Lab just to name a few.[[4]](#footnote-4) As a result, deepfake production no longer lies necessarily in the hands of specialists as with CGI or Photoshop, but in the hands of a broader, non-specialist audience. Moreover, when VoCo technology is factored into this assessment, deepfakes become more credible than older forms of media manipulation because purveyors can design deepfakes with an accompanying, indistinguishable voice. In turn, we are more likely to accept the videos’ testimony unless there are significant reasons not to. Considering these developments, its unsurprising that the pioneer and computer scientist Professor Hao Li (Stankiewicz, 2019) has warns that we’re going to get to a point where there is ‘no way that we can actually detect them [deepfakes] anymore’.

**3. Vice Epistemology and Epistemic Corruption**

If there is any truth to the above statement, then it’s understandable why philosophers worry that these videos could jeopardise our trust in videos (Rini, 2020) or the evidential status we gain from such a ubiquitous source (Fallis, 2020). My aim is not to press these issues any further but to address a currently unexplored problem that deepfakes risk generating. This problem is exposed by offering a *corruptionist* critique of deepfakes. Ordinarily, the verbal meaning of ‘corrupt’ refers to way in which the character of something or someone is made worse or degraded - like when we say that somebody has been ‘corrupted by power’. When we talk of people becoming corrupted, we often invoke the language of virtue and vice, specifically that a person has lost certain virtues or gained particular vices in their place. This is the function of a corruptionist criticism: to demonstrate how an environment causes one to acquire vices. To get a better sense of the kind of vices that I’m interested in here, we need to say something about *vice epistemology.*

Epistemology is the branch of philosophy concerned with knowledge, beliefs, and thinking while vices are often considered to be defects, faults, or failures. Together, it is the job of the *vice epistemologist* to study what makes one’s beliefs and thinking bad or defective. Generally, when vice epistemologists approach this task, they tend to focus on defects of *intellectual character* – that is, the traits, attitudes, and habits of attention that hinder good thinking or knowledge, and which reflect badly on us. Some classic examples include close-mindedness, arrogance, and snobbery. Intellectual character defects such as these are what Quassim Cassam (2016, 2019) calls the ‘vices of the mind’, or simply *intellectual vices*. These are contrasted with what Linda Zagzebski (1996) calls the ‘virtues of the mind’, such as open-mindedness, intellectual humility, and creativity. Intellectual vices are bad for several reasons. First, they often reflect poor or deficient motivations towards and other knowledge-related practices. Call this the *motivational account* (Tanesini, 2018, 2021). Second, they frequently, although invariably, obstruct responsible inquiry and the transmission of knowledge. Call this the *obstructivist account* (Cassam, 2016, 2019), Third, they reflect entrenched patterns of bad judgement. Call this the *judgement account* (Baehr, 2020; Crerar, 2018).

Regardless of how one’s vices play out in the actual world, they’re usually the product of what we might call our *intellectual environments*. Think of these as physical or virtual spaces where *epistemic practices* take place. Amongst other things, these include practices such as inquiry, deliberation, educating, and evidence-sharing. They are ‘epistemic’ in the sense that they usually aim at discovering the truth of some matter. While some of us are lucky enough to engage with relatively healthy intellectual environments - where these practices are encouraged and facilitated – it’s more likely that many of us will inhabit less than healthy intellectual environments; one’s where inquiry is facilitated up to an extent then shut down after a certain point, where questioning authority and information is a risky business, and where the evidence available to us can come from dubious or untrustworthy sources. In a world of disinformation campaigns, ‘fake news’ and ‘alternative facts’, it’s hard not to find oneself in steadily deteriorating intellectual environments.

Ian James Kidd (2019b, 2020) introduces the concept of *epistemic corruption* to draw attention to environments that encourage the development and exercise of intellectual vices or lead to the loss of intellectual virtues. Those that fail to nurture intellectual virtues are what Kidd calls ‘passively corrupting’, while environments that encourage the exercise of intellectual vices are ‘actively corrupting’. Importantly, this distinction is not clear-cut and the two will typically reinforce each other in a mutual fashion. For example, suppose you attend an elite debating club that encourages and rewards ‘kill or be killed’ argumentative styles. Within that environment, arrogant or dismissive tendencies are likely to become entrenched in one’s intellectual character. Simultaneously, the longer one is exposed to such conditions and styles, the harder it will become to cultivate humble or attentive behaviour and traits towards others. Of course, not all environments are equally corrupting.[[5]](#footnote-5) Some might contain structures that are conducive to the development of many vices or lead to the loss of numerous virtues. At other times, the environment might only feed a handful of vices in a small number of domains. My participation in a football fan club might cause me to develop a close-mindedness towards other football clubs but leave my political or music tastes wholly unscathed.

How might these environments contribute to the corruption of intellectual character? Kidd says that epistemic corruption can occur in multiple ways. A particular environment might contain conditions that cause one to simply *acquire* certain traits or habits that one didn’t previously possess. Only after joining and engaging in the activities of the football fan club might I realise that I have acquired an intellectual vice - assuming I’m even in a position to recognise it in the first place! (Cassam, 2019). Other environments might draw out and *ignite* one’s previously managed intellectual vice(s). Suppose that you recognise your arrogant tendencies but nonetheless keep it contained. By participating in the debating club, the norms, rules, and structures present could unlock your arrogance and allow it to flourish. Furthermore, the same environments could well *amplify, stabilise, or intensify* particular intellectual vices, thereby corrupting one’s intellectual character from many different angles (Kidd, 2020: 72). In short, epistemic corruption is a complex phenomenon with real implications for our character and conduct.

**4. Deepfakes and Intellectual Cynics**

In this section, I will use the concept of epistemic corruption to show how deepfakes encourage us to develop what I call *intellectual cynicism*. First, though, I want step back and consider Regina Rini and Don Fallis’ concerns, as this will illuminate why and how deepfakes might lead to cynicism in the first place. According to Fallis (2020) a video carries information if there is a low probability of watching a false positive: a video that looks and sounds just like a genuine video but turns out to be fake. Accordingly, a video will carry *more* information if the probability of watching false positives is low. However, if deepfakes become sufficiently sophisticated and wide-spread enough, they will inevitably increase the probability of our watching false positives. As such, the evidence a genuine video carries will have to increasingly compete with the ‘counter-evidence’ depicted in a deepfake. Consequently, we will either be forced to suspend judgment about what we watch or accumulate false beliefs.

For Rini, recorded videos form a sort of ‘backstop’ to our testimonial practices: they allow us to acutely correct what others say, and the threat of being recorded often passively regulates what we say in public. But here’s the catch: if anyone can get their hands on the technology canvassed in section 2 and create a sophisticated deepfake, then the pressure to testify sincerely and competently will gradually dwindle. After all, what’s stopping somebody ‘crying deepfake’ on a recording if these videos become indistinguishable from genuine videos? This ‘get out of jail free’ card – what Rini calls a ‘backstop crisis’ (2020: 7) - is likely to erode any impetus to testify sincerely and competently if there are little, if any ramifications, for our behaviour. So, rather than cause people to form false beliefs, Rini thinks that deepfakes will lead increasingly savvy information consumers to ‘reflexively distrust *all* recordings’ (2020: 8, my italics).

We do not need to endorse the full extent of Fallis and Rini’s arguments to recognise that deepfakes could lead us to distrust videos more frequently or cause us to suspend judgement about what a given video depicts. Another way to think about those who suspend judgement about deepfakes is to think of them as disengaging from the risk associated with forming false beliefs from videos. According to Samantha Vice (2011), distrust and disengagement are central components of *moral cynicism*. What makes cynicism specifically ‘moral’ is that it primarily concerns our attitudes towards other human beings and ourselves. For Vice (2011: 173), cynicism can be thought of as a habitual and cultivated attitude of ‘scepticism or distrust regarding people’s professed values and motivations’ or their ‘strength and nobility of character’. Underpinning this distrust is the cynic’s core belief that humans are ultimately self-interested or of little worth. In this way, cynicism ‘structures our perception, interpretation, evaluation, and expectations of others’, while also affecting our own actions. In its more developed forms, cynicism can even lead one to disengage from humanity and its institutions due to a habitual wariness of what drives people (2011, 172). Those who adopt this stance are what Vice calls ‘pure cynics’, someone whose ‘pervading and fundamental character is cynical’ (2011: 174).

According to Vice, pure cynicism is ‘perhaps impossible to find in reality’ (2011: 174). Whether or not this is true of the moral domain, I want to suggest that there *is* a very common pure cynicism when it comes to dealing with intellectual or epistemic matters. Call it *intellectual cynicism*. In its milder form, intellectual cynicism might appear analogous to its moral counterpart, that is, be understood as a ‘stance’ or attitude one takes towards a particular object. The object in question will not necessarily be epistemic goods – things like truth, knowledge, or understanding. Instead, intellectual cynicism is a negative attitude of distrust towards epistemic practices such as inquiry, deliberation, and the like. Yet, if cynicism ‘structures our perception, interpretation, evaluation, and expectations of others’ (2011: 172), then this suggests that pure cynicism is not simply an attitude but a *sensibility*, a characteristic pattern of seeing the world. Importantly, our sensibilities are often informed by the character traits we cultivate. Consequently, I contend that the more one adopts a cynical stance towards epistemic practices, the likelier one is to habituate the trait of intellectual cynicism with accompanying sensibilities. Accordingly, this trait will be partly grounded in a belief that epistemic practices never serve intrinsically good purposes or indeed serve dubious agendas. Rather than attend to the possibility that an inquiry or deliberation aims at the truth, the intellectual cynic’s sensibilities are too highly attuned to the possibility that epistemic practices are instrumental to the success of a particular agenda. In an increasingly digital world, videos often serve as a direct means of engaging in these practices. As such, the intellectual cynic’s initial reaction is to distrust and disengage from videos.

However, the risk with adopting this stance or attitude is the real possibility that it becomes an entrenched feature of one’s intellectual character. That is, those who continually adopt this stance habituate the *pure* trait of intellectual cynicism. The worry is that the growing sophistication of deepfakes and their application across our digital environments will prompt people to develop and exercise the kind of intellectual cynicism under consideration here. Imagine that you find out that you’ve spoken to a deepfake, been repeatedly duped by one, or worse been the target of a deepfake, then a natural reaction will be to adopt the reflexive distrust towards videos that Rini speaks of. Of course, videos will still play their role within our epistemic practices but if anything we will increasingly shift our attention to the foregone conclusion that all videos could be deepfakes out to deceive us. Over time, as deepfakes become more and more sophisticated and widespread, the cynical stance will plausibly harden into the trait outlined above, which will subsequently inform an increasingly cynical sensibility towards videos as a whole. At first, intellectual cynicism may prevent one from trusting online media, but just as the cynic’s distrust and disengagement is fuelled by a belief in the sub-optimality of humanity, the ‘pure’ intellectual cynic ends up with a ‘‘corrupt sense of value’ (Vice, 2011: 175). In other words, the stronger the intellectual cynicism, the easier it bleeds into other aspects of one’s intellectual character, *corrupting* one’s characteristic way of seeing not just videos but practices that generate and transmit information, evidence, and knowledge. In essence, ‘pure’ intellectual cynicism causes one to be drawn away from considering the intellectual worth or *merit* of epistemic practices *en masse*.

If this is the case, then deepfake appear to be actively corrupting: they facilitate and encourage the development of a vicious form of intellectual cynicism. This viciousness manifests in the three ways canvassed above. First, the intellectual cynic might be said to possess *bad motivations* towards wanting knowledge if they gradually come to distrust and associate epistemic practices with suspicion. Second, in cultivating this trait, the intellectual cynic *obstructs* knowledge and information-sharing. Third, by developing the cynical sensibilities associated with intellectual cynicism, the cynic is likely to make poorly judged decisions about their utility because they perceive them in a skewed light. These three manifestations will, of course, overlap and be borne out more strongly in some than others. Regardless, though, by habitually distrusting and disengaging from epistemic practices, the intellectual cynic never engages with the very sources and practices that could potentially shake their cynicism. If this is the case, deepfakes help encourage an ultimately damaging intellectual character trait that deprives us of the knowledge and information needed for a flourishing intellectual life. It’s beyond the scope of this chapter to give a full account of the nature of intellectual cynicism, but I hope it’s at least clear what the trait consists of and how deepfakes could encourage it in becoming an intellectual vice.

**5. Cultivating Digital Sensibility**

In the previous section, I argued that deepfakes are epistemically corrupting because they plausibly encourage us to develop intellectual vices, and specifically the vice of intellectual cynicism. However, as Kidd rightly emphasises, identifying corrupting environments - or in our case technologies - puts us in a position to work out what sorts of corrective measures to put in place (2019: 227). The corrective approach I want to advance is somewhat analogous to our everyday treatment of the testimony we hear from informants. The motivation for this comes from Regina Rini’s worry above. Part of the reason why videos play such an important role in her idea of a ‘backstop’ is because of their justificatory or evidential status. It is widely acknowledged, for instance, that videos and photographs provide us with *perceptual* evidence of states of affairs (Cavedon-Taylor, 2013, Hopkins, 2012), which is more authoritative than, say, testimonial evidence. Rini worries that wide-spread deepfakes will relegate the justificatory status we assign to video recordings from perceptual evidence to that of testimonial evidence. In other words, we will take videos to be less authoritative of the states of affairs they depict.

If Rini’s concerns are correct, then why don’t we treat videos like testimony? So, just as we deal with liars and bullshitters (Frankfurt, 2005) by exercising particular caution, hesitance, and critical evaluation in our credibility assessments of what they say, I suggest that we deal with the threat of deepfakes by cultivating a *digital sensibility* towards videos that counteracts the development of intellectual cynicism. The kind of sensibility that I have in mind comes from Miranda Fricker (2007), who offers the idea of a *testimonial sensibility*. This a carefully cultivated way of experiencing and evaluating other people’s testimonies, informed by an awareness that our ability to properly listen to others is corrupted by all sorts of biases and prejudices. It involves training oneself and gaining more understanding of social prejudices, which is no mean feat. By cultivating this sensibility, one comes to gradually acknowledge and, to an extent, nullify the prejudices interfere with one’s credibility assessments (2007: 72-81). For Fricker, there are two components of this sensibility that stand out for our discussion: its perceptual nature and its *affective* dimension.[[6]](#footnote-6) I will look briefly at each as a way of modelling the digital sensibility required for mitigating the intellectual cynicism generated by deepfakes.

Fricker’s starting point is an idea dating back to Aristotle, for whom the good or *virtuous* person is one whose ‘moral perception’ is drawn towards ethically salient aspects of a situation. This perception is carefully cultivated and trained to be reflexive and spontaneous, so that upon being confronted with a morally charged scenario, the good person comes to ‘just see’ the situation in a certain light and is able to pick out the best course of action to help. Fricker suggests that we do something similar with what she calls an ‘epistemic perception’. Similarly, this is a calibrated sensitivity to salient aspects of a speaker’s situation and is informed by a number of background assumptions about the speaker’s trustworthiness, including cues relating to their sincerity and competence on the matter at hand (2007: 72). Importantly, the work of calibration should also be an interpersonal project, rather than an individual one, though obvious care should be taken in deciding whom to trust. With the help of a conducive intellectual environment – one that encourages and emphasises the virtues of the mind – people can develop an ability to see the world in a certain light, enabling them to assess their interlocutors in a fair, open-minded, and non-inferential manner.

This perceptual sensibility is also supported by an affective dimension. In order to properly judge the competency and sincerity of a speaker, one needs to be able to sufficiently empathise with them. Only by invoking a degree of emotional engagement with a speaker can one fully appreciate the socially situated factors that might be in play when it comes to ‘reading’ another’s body language or the competence one displays in their testimony (2007: 79-80). In order to better determine the level of trust one ought to place in another’s word, Fricker observes that good advice is ‘to listen to one’s emotions’ insofar as the virtuous person’s emotional response to different speakers across different contexts is only improved by training and experience (2007: 80). By exposing ourselves to different speakers in different contexts, we can build up a sophisticated emotional radar for detecting trustworthiness in speakers. In doing so, we not only cultivate a trained testimonial sensibility, but we also manifest the virtue of ‘testimonial justice’ – a disposition to reliably neutralise prejudice and bias from our credibility assessment of speakers (2007: 92).

Taking these points into consideration, we can develop a model for digital sensibility. For starters, this will be a reflective and spontaneous sensitivity towards online content, so that when we’re confronted with certain videos, images, or website we learn to ‘just see’ them as trustworthy. Again, this will turn on features like the content’s sincerity and competence. For instance, it is common nowadays to stumble across a video online or see adverts on websites and immediately withhold judgement about the claims made, or outright flag them as deceptive. One reason for this hesitance is that the content seems insincere: it makes claims that do not accord with the background assumptions against which we ordinarily evaluate videos or other media. Another reason is that the content appears incompetent: its presentation and configuration fail to conform to established norms of publishing and communication. A further factor in our assessments of trustworthiness might also include our worries about clicking on videos and developing computer viruses, being scammed, or losing personal information.

This last factor draws out the affective dimension of a digital sensibility. While Fricker takes empathy to be an important emotion in judging the sincerity of an interlocutor, I think that past, present and future experience with our digital environments will require a dose of healthy scepticism. When our emotions or ‘gut feelings’ tell us that a video is unscrupulous or dubious, we can exercise this healthy scepticism towards the content and take steps to verify its authenticity. This means that an effective digital sensibility will require a level of perseverance to fact-check such content and a degree of humility in recognising that our credibility assessments will not always be watertight. But insofar as we as we draw on past experiences, exposing ourselves to a wide variety of claims, depictions, and sources of digital content, we can practice and improve such assessments. Over time, the scepticism that informs our emotional radar helps us to judge certain content as more trustworthy – established news media and websites that adhere to communicative norms – than other content which fail to adhere to such standards or are at odds with the catalogue of past experiences we have built up. In this way, we continually fine tune the ‘radar’ that comes with cultivating a digital sensibility. Eventually, a trained digital sensibility, like its testimonial counterpart, will develop to be an active, automatic, and critical alertness towards online content.

Crucially, though, the development of a digital sensibility is as much about individual caution and critical awareness as it is about developing a *collective* safeguard against deceptive online media. As Fricker makes clear, the formation of our testimonial sensibilities is partly individual and partly collective (2007: 82). Accordingly, the cultivation, exercise, and regulation of a trained digital sensibility relies on cooperating with our peers across digital environments. Those who are more technologically savvy will have duties to flag potentially untrustworthy content, allowing those less well-versed in digital matters to observe how to decide how much trust to place in websites and online media. Gradually, those with less experience can learn from the ‘digital exemplars’ in their community and slowly fine-tune their perceptual and emotional radar towards trustworthy online content. This collective dimension not only illustrates the educational role that training one’s digital sensibility entails, but it also emphasises the socially-scaffolded nature on which the effectiveness of this sensitivity is premised. This is particularly salient when it comes to regulating the affective dimension of one’s digital sensibility. It’s not hard to imagine somebody who exercises a deficiency or excess of scepticism towards digital content, perhaps sympathising with the claims made by a video or website or being overly sceptical of authentic content. To ensure that one’s healthy scepticism is kept in check, one’s peers will need to hold them account, warning them when their emotional radar fails to track the merits or demerits of digital content.

However, even a collective effort to cultivate a trained digital sensibility might not be sufficient to fend off the cynicism that deepfakes and emerging technologies might encourage. In addition to individual and collective practice and cultivation, a well-trained digital sensibility requires *institutional* and *structural* scaffolding (Anderson, 2012, Medina, 2013). This will require social media companies, news corporations, and fact-checking sites to flag deepfakes, correct misinformation, and raise awareness to other untrustworthy sources of information. For instance, Facebook still permits ‘parody’ deepfakes on its website, and this could be exploited by malicious actors to indirectly spread misinformation and sow distrust (Shead, 2020). If efforts to cultivate a trained digital sensibility are to succeed, then institutions with regulatory powers like Facebook need to work in tandem with their user-base. A good place to start is by cultivating the virtue of *collective responsibility*, a stable disposition of a collective to be morally responsible for its actions (Astola, this volume). Only then, will a trained digital sensibility be sufficiently socially scaffolded and effective in the long-run.

**6. From Digital Sensibility to Techno-Intellectual Virtue**

In a world increasingly characterised by deepfakes, fake-news, and emerging technologies, how can we best flourish? I want to briefly conclude by reflecting on this question. According to the philosopher, Shannon Vallor (2016), the cultivation of what she calls *techno-moral virtues* will help us achieve this end. Techno-moral virtues are an extension of our existing moral capacities or traits, adapted to a rapidly changing techno-socially opaque world, which require a collective moral wisdom on a global scale (2016: 10). Some notable examples include techno-moral courage – ‘a reliable disposition toward intelligent fear and hope with respect to moral and material dangers and opportunities presented by emerging technologies’ (2016: 131), and humility – ‘a renunciation of the blind faith that new technologies inevitably lead to human mastery and control of our environment’ (ibid, 126-7). In the face of emerging technologies, Vallor rightly stresses the need to cultivate these virtues as a means of offsetting any potential harm to our ethical character.

While Vallor is certainly right to draw attention to this prospect, my aim in this chapter has been to demonstrate just how deepfakes can degrade our *intellectual* character. In light of this, we need a way to fortify our thinking against the potential corruption outlined above. I believe that developing a well-trained digital sensibility will be a major step towards doing this. This will not simply lead those wishing to navigate online space to acquire a number of ‘virtues of inquiry’ towards digital content (see Carter, this volume), but specifically cultivate the techno-*intellectual* virtue of digital attentiveness, a corrective disposition to carefully attend to the merits and demerits of digital media. Just as cultivating a trained testimonial sensibility helps one nullify prejudices and biases against speakers, a trained digital sensibility enables one to better fortify their intellectual character against the corrupting effects of deepfakes and intellectual cynicism described above. For those wishing to flourish in a rapidly changing world, a trained digital sensibility and the correlative techno-intellectual virtue of attentiveness will help to make this a reality.

**References:**

Anderson, E (2012) ‘Epistemic Justice as a Virtue of Institutions’, *Social Epistemology*, 26:2, pp. 163-173.

Battaly, H (2018) ‘Can Close-mindedness be an Intellectual Virtue?’, *Royal Institute of Philosophy Supplements*, 84, pp. 23-45.

* (2021) ‘Engaging Close-mindedly with your Polluted Media Feed’ in (eds.) M. Hannon and J. de Ridder *The Routledge Handbook of Political Epistemology*. pp. 312-324.

Cassam, Q (2016) ‘Vice Epistemology’, *The Monist*, 99:2, pp. 159-180.

* (2019) *Vices of the Mind*, Oxford, Oxford University Press.

Cavedon-Taylor Dan (2013) ‘Photographically Based Knowledge’, *Episteme*, 10:3, pp. 283-297.

Crerar, C (2018) *Bad Judgement: An Essay in Vice Epistemology*, Doctoral Thesis, University of Sheffield. https://etheses.whiterose.ac.uk/21243/1/Thesis%20%28library%29.pdf

CBS News (2020) ‘Artificial Intelligence Project Lets Holocaust Survivors Share Their Stories Forever’, *CBS News 60 Minutes Overtime*, April 3rd, 2020, accessible at: https://www.cbsnews.com/news/artificial-intelligence-holocaust-remembrance-60-minutes-2020-04-03/.

Channel 4 (2020) ‘Deepfake Queen: 2020 Alternative Christmas Message’, *YouTube*, 25th December 2020, accessible at: https://www.youtube.com/watch?v=IvY-Abd2FfM

Chesney, R and Citron, D (2019) ‘Deepfakes and the New Information War: the coming age of post-truth geopolitics’, *Foreign Affairs*, 98:1, accessible at: https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war.

Eg, Meenu (2021) ‘Try These 10 Amazingly Real Deepfake Apps and Website’, *Analytics Insight*, accessible at: https://www.analyticsinsight.net/try-these-10-amazingly-real-deepfake-apps-and-websites/.

Lee, D (2019) ‘Deepfake Salvador Dalí takes selfies with museum visitors’, *The Verge*, May 10th, 2019, accessible at: https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum.

Fallis, D (2020) ‘The Epistemic Threat of Deepfakes’, *Philosophy and Technology*.

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

Fricker, M (2007) *Epistemic Injustice: power and the ethics of knowing*, Oxford, Oxford University Press.

Harris, K. R. (2021) ‘Video on demand: what deepfakes do and how they harm’, *Synthese*, https://doi.org/10.1007/s11229-021-03379-y

Hopkins, R (2012) ‘Factive Pictorial Experience: What’s Special About Photographs?’, *­Noûs*, 46: 4, pp. 709-731.

Jin, Z, Mysore, G. J, Diverdi, S, Lu, J, Finkelstein, A (2017) ‘VoCo: Text-Based Insertion and Replacement in Audio Narration’, ACM *Transactions on Graphics*, 36: 4, Article 96.

Kidd, I. J. (2019) ‘Epistemic Corruption and Education’, *Episteme*, 16:2, pp. 220-235.

* (2020) ‘Epistemic Corruption and Social Oppression’, in I. J. Kidd, H. Battaly, and Q. Cassam (eds.) *Vice Epistemology*, Abingdon, UK, Taylor and Francis, pp. 69-85.

Lang, F (2019) ‘Adobe Trains AI to Detect Deepfakes and Photoshopped Images’, *Interesting Engineering*, available at: https://interestingengineering.com/adobe-trains-ai-to-detect-deepfakes-and-photoshopped-images.

Medina, J (2013) *The Epistemology of Resistance: gender and racial oppression, epistemic injustice, and resistance imaginations*, New York, Oxford University Press.

Mirksy, Y and W. Lee (2020) ‘The Creation and Detection of Deepfakes: A Survey’, *ACM Computing Surveys*, 54:1, Article 7 (December 2020), pp.1-41.

Reuters (2019) ‘David Beckham’s Malaria Awareness Video’, *Reuters*, April 10th, 2019, accessible at: https://www.reuters.com/video/watch/idOVA9QVUY3.

* (2020) ‘Fact-Check: “Drunk” Nancy Pelosi” video is manipulated’, *Reuters*, August 3rd, 2020, accessible at: https://www.reuters.com/article/uk-factcheck-nancypelosi-manipulated-idUSKCN24Z2BI.

Rini, R (2020) ‘Deepfakes and the Epistemic Backstop’, *Philosopher’s Imprint*, 20: 24, pp. 1-16

Schick, N (2020) *Deepfakes and the Infopocalypse*, London, Monoray.

Shead, S (2020) ‘Facebook to ban ‘deepfakes’’, *BBC News*, January 7th, https://www.bbc.co.uk/news/technology-51018758.

Stankiewicz, K (2019) ‘“Perfectly real”’ deepfakes will arrive in 6 months to a year, technology pioneer Hao Li’, *CNBC*, September 20th, accessible at: https://www.cnbc.com/2019/09/20/hao-li-perfectly-real-deepfakes-will-arrive-in-6-months-to-a-year.html.

Tanesini, A (2018) ‘Epistemic Vice and Motivation’, *Metaphilosophy*, 49:3, pp. 350-367.

* (2021) *The Mismeasure of the Self: A Study in Vice Epistemology*, Oxford, Oxford University Press.

Tolosana, R, Vera-Rodriguez, R, Fierrez, J, Aythami, M, and Ortega-Garcia, J. (2020) ‘DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection’, *Information Fusion*, 64, pp. 131-148. https://arxiv.org/pdf/2001.00179.pdf

Vallor, S (2016) Technology and the virtues: A philosophical guide to a future worth wanting. New York: Oxford University Press.

Vice, S (2011) ‘Cynicism and Morality’, *Ethical Theory and Moral Practice*, 14:2, pp. 169-184.

Zagzebski, L. T. (1996) *Virtues of the Mind: an inquiry into the nature of virtue and the ethical foundations of knowledge*, Cambridge, Cambridge University Press.

Zakharov, E, Shysheya, A, Burkov, E, Lempitsky, V (2019) ‘Few-Shot Adversarial Learning of Realistic Neural Talking Head Models’, accessible at: https://arxiv.org/pdf/1905.08233v1.pdf. For a demonstration of the research, see https://www.youtube.com/watch?v=p1b5aiTrGzY.

1. This is the pre-published version. Please cite the version in *Royal Institute of Philosophy Supplement*, vol. 92, pp. 67-85. [↑](#footnote-ref-1)
2. To watch this video, go to https://www.youtube.com/watch?v=IvY-Abd2FfM. [↑](#footnote-ref-2)
3. Visit thispersondoesnotexist.com to see this in action. [↑](#footnote-ref-3)
4. Other honourable mentions include FaceApp, Deepfakes Web and My Heritage. See https://www.analyticsinsight.net/try-these-10-amazingly-real-deepfake-apps-and-websites/ [↑](#footnote-ref-4)
5. Indeed, it’s plausible that some are not corrupting at all. [↑](#footnote-ref-5)
6. Fricker adds that a trained testimonial sensibility will also be grounded by 1) good judgement that is uncodifiable i.e. it will not be rule-based, 2) be intrinsically motivating, and 3) be reason-giving i.e. generate reasons to act in certain ways (2007: pp. 72-80). [↑](#footnote-ref-6)