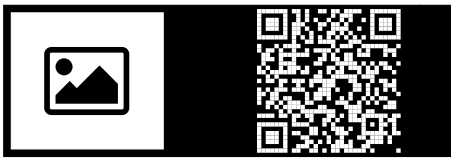

A New Hermeneutics of Suspicion?
The Challenge of deepfakes to Theological Epistemology

Clifford Anderson
Vanderbilt University
clifford.anderson@vanderbilt.edu

Clifford Anderson provides an introduction to "deep fakes" and related machine-learning technologies for theologians, assesses their danger as well as potential uses, and advocates for developing a spirituality of critical empathy in response. He relates deep fakes to a theology of mediation, pushing us to ponder the relation between εικών and είδος (icon and idea).

1. A New Hermeneutics of Suspicion? The Challenge of deepfakes to Theological Epistemology

"Jesus said to him, 'Have you believed because you have seen me? Blessed are those who have not seen and yet have come to believe.'" (John 20:29; NRSV)

What does it mean to see and yet not to believe? Is this inverse of the Johannine pericope of 'doubting Thomas' a virtue or vice in the age of synthetic videos, better known as deepfakes? How will the growing use of deepfake videos affect theological epistemology, that is, our ability to discern the truth about God, our neighbors, and ourselves?

In this paper, I provide an introduction to deepfakes and related machine-learning technologies for theologians, considering their potential use and misuse in theology.

Zeitschrift für explorative Theology
CZeTh_3 (2022) S. 153–176
DOI: https://doi.org/10.17885/heiup.czeth.2022.3.24531
Zuvor veröffentlicht auf PubPub: https://doi.org/10.21428/fb61f6aa.771d30b7

As we explore the topic, we will find that the phenomenon of deepfakes brings us deep into the theology of mediation, pushing us to ponder the relation between εικών and είδος (icon and idea). As Christians, we learn that appearances can be deceiving, misleading, or at least obscure underlying reality. The paradox of the form that is other than its substance is at the heart of Christian faith, from the mystery of the Lord's Supper to the crisis of the Cross.

Paul Ricœur introduced the term "hermeneutics of suspicion" as a counterpart to the "hermeneutics of faith." Whereas a "hermeneutics of faith" seeks to discern and bring the meaning of a text to light, a "hermeneutics of suspicion" questions its meaning, looking beneath the surface for repressed or suppressed significance. "This hermeneutics is not an explication of the object, but a tearing off of masks, an interpretation that reduces disguises."[1] While he apparently changed his mind over the course of his long career about the relation between these hermeneutical modes,[2] when he discussed them in *Freud and Philosophy*, he argued that they are necessary and complimentary. As with Hegelian dialectic, suspicion turns into its opposite, namely, faith, when seeking for meaning behind the mask. Ricœur famously described this so-called "second faith" as "postcritical" or a "second naïveté."[3]

Will the proliferation of deepfakes push us as a society to become more critical about the mediascape about us? Will this critical perspective lead us into a postcritical stance that opens onto deeper vistas of meaning? Or will we become more suspicious, refusing to believe the evidence of our eyes even when all signs suggest we are facing the truth?

2. Assessing deepfakes

When people learn about the technology behind deepfakes, they tend to become fearful and for good reason: the origin story of deepfakes is murky and unseemly, starting with an anonymous member of reddit who called himself 'deepfakes' and applied off-the-shelf machine learning techniques to swap the faces of celebrities into pornographic videos. As Samantha Cole, senior staff writer at Motherboard and Vice, ex-

[1] Ricoeur (1970), 30.
[2] Scott-Baumann (2012), 59–77.
[3] Ricoeur (1970), 28.

plained in twin articles from 2017[4] and 2018,[5] a community has developed around the production of these videos.

Creating deepfake videos to blackmail people is also on the horizon. As Samantha Cole writes, "It isn't difficult to imagine an amateur programmer running their own algorithm to create a sex tape of someone they want to harass."[6] The majority of states now have laws against the circulation of 'revenge porn,' that is, of sexually-explicit images or audiovisual records.[7] While these laws criminalize nonconsensual sharing of sexually-explicit photographs or videos, their application to synthetic images and videos is another question. Internet trolls have used Photoshop to create and spread degrading images of women for more than a decade, emerging as a public issue during the harassment of technologist Kathy Sierra in 2007.[8] The personal and social harm caused by the release of synthetic pornography is no less real than organic pornography. And the more realistic it becomes the worse the impact on its victims. For this reason, producers of deepfake pornography may find it lucrative to threaten people with its release. In fact, this kind of blackmail has already started to take place.[9]

If deepfake pornography threatens to cause victims personal anguish and social harm, the dissemination of fake videos in charged political situations might prove fatal. In October 2019, for instance, four protesters in Bangladesh died in riots sparked by a post on Facebook criticizing the Prophet Mohammed.[10] The Hindu citizen who supposedly published the post complained about the hacking of his account. In fact, it turned out that police corroborated his complaint and arrested the hackers. The quick action did not stop the riots. In regions where there is little trust between communities, people who see incendiary videos may act without waiting for confirmation of their veracity (or falsity). While digital media forensics (see below) might reveal malicious doctoring, such evidence would come too late to prevent violent disturbances on the ground. The production of deepfake videos about political figures has become a popular sub genre of the deepfake community; the channel r/SFWdeepfakes/ on reddit features synthetic videos of Donald Trump, Barak Obama, and Hilary Clinton, among others. The majority of these videos function as parody, inserting Trump into the film, *The Wolf of Wall Street* or having Obama sing and dance to the tune

[4] Cole (2017).
[5] Cole (2018).
[6] Cole (2017).
[7] Electronic Privacy Information Center (2019).
[8] See MacKinnon (2013), 87–88.
[9] Liotta (2019).
[10] Gupta (2019).

of "Spooky Scary Skeletons." These applications of the technology of deepfakes are innocuous, clever, and funny, but more sinister applications could make real political impact. As with foreign interference in the 2016 presidential elections in the United States, no straightforward remedy exists for undoing the immediate social and political aftermath of faked images and videos.

Coverage of deepfake videos tends to dwell on their negative potential. Given their origins, use for harassment, and potential for spreading disinformation, the media's alarm over deepfakes seems justified. As with any new technology, the advent of deepfakes comes with positive and negative potential. As the authors of *Blown to Bits*, a textbook used in high school and college level courses in computing across the country, opine, "the key to managing the ethical and moral consequences of nourishing economic growth is to *regulate the use* of technology without *banning* or *restricting its creation*."[11] In the introductory computer science course I teach at Vanderbilt University, *The Beauty and Joy of Computing*, I cover the moral panics that periodically sweep through the media, school boards, and Congress, ranging from the worries about children's exposure to pornography that led to the passage of the Communications Decency Act of 1996 to the battles over copyright and fair use that prompted the Stop Online Piracy Act (SOPA) and PROTECT IP Act (PIPA) in 2012. As we face the prospect of legislation after the passage of the Deepfake Report Act of 2019,[12] will it be possible for us to overcome anxieties about the genuine threats deepfakes pose in order to consider and safeguard the positive applications?

As theologians, we have particular reasons to take care. We must think beyond the economic, legal, and even ethical dimensions of deepfakes to consider their spiritual implications. We also have to avoid falling into our socially-assigned role of conservators of the status quo even as some hyperbolically speculate that "AI may be the greatest threat to Christian theology since Charles Darwin's *On the Origin of Species*."[13] The best way to assess the spiritual impact of any new technology is to spend time exploring its potential for good and bad, examining its components, and exercising our theological imagination.

The majority of publications about deepfakes address their potential for spreading disinformation. But the technology also has positive aspects. Deepfakes can serve legitimate ends by bridging cultural divides and forging emotional connections. But

[11] Abelson, Ledeen and Lewis (2008), 14.
[12] Portman (2019).
[13] Merritt (2017).

the boundary between such valid uses and virtual creepiness may be difficult to discern. In what follows, I present three brief scenarios, grounded in contemporary technology, for us to consider.

2.1. Editing Sermons

Sharing audio or video recordings of sermons online is common today. If you are like me, you prefer not to listen to the sound of your own voice. At Vanderbilt, I am one of the team members who collectively produce *Leading Lines*, a podcast about educational technologies. I am grateful that our team also includes Rhett McDaniel, an educational technologist who also happens to have a M.S. in Music Technology. Rhett skillfully edits every episode to smooth over verbal stumbles and tics. If you are a pastor, having your worship services broadcast increasingly comes with the territory. But, from my experience, mainline churches do not edit the recordings they put online, making them difficult at times to listen to. If you stumbled while reading a biblical passage, made an impromptu joke that fell flat, or neglected to mention one of the volunteer leaders of vacation Bible school, your gaffe will linger for the longterm in the congregation's digital library.

A company called Descript markets audio editing software that makes it straightforward and easy to edit out mistakes, pauses, and other problems in podcasts and other kinds of recordings. Descript generates a transcript from the audio and, by keeping text and speech in sync, allows you to edit the audio by changing the transcript. If you want to get rid of that bad joke, you strike it from the transcript and it vanishes from the audio too. Of course, while Descript provides an attractive interface, it does not differ qualitatively from other audio editing and transcription tools, which also provide sophisticated software for correcting errors.

What makes Descript distinctive is the integration of a technology called Lyrebird to enable audio overdubbing. The researchers collaborating on Lyrebird highlight similar scenarios for its use. Drawing on an area of study called "text-informed speech inpainting," Lyrebird uses deep learning techniques to allow editors to insert new text into the transcription and to produce new audio in the recording that blends seamlessly with the words that came before and after.[14] In other words, if you forgot the name of that volunteer, you do not have to live with the mistake – by editing the transcript, you insert mention of that person into the audio and, to all the world who lis-

[14] See Brébisson (2019).

tens to the recording, it sounds as natural as it would have had you said it on Sunday morning.

2.2. Preaching in Tongues

What about using deepfake technology to bridge linguistic divisions in congregations? In churches serving immigrant communities, pastors commonly hold services of worship in different languages. There may, for instance, be one service in Spanish and another in English. A Methodist congregation in my neighborhood in Nashville holds simultaneous services of worship in English in the main sanctuary and Karen, English, and Thai in the community center next door. While accommodating the linguistic difference of parishioners is admirable, maintaining separate worship services might lead to divisions within the congregation. The alternative, combining services with the aid of simultaneous translators, is problematic because of its cost and its potential for increasing the length of the service. What if we could draw on deep learning to create versions of the same sermon in English and any other language spoken in the congregation?

Synthesia is a company based in London that specializes in what it terms "video synthesis technology."[15] Synthesia uses "Generative AI" to "reduce the cost, skill and language barriers to creating, distributing and consuming content." On its website, Synthesia also highlights its ethical commitments, promising to "never re-enact someone without explicit consent" and to work with partners of all kinds "to develop best practices" on the use of "video synthesis technology."[16]

The Synthesia website features exemplary stories about the potential of "video synthesis." Consider the story of a cross-cultural marriage proposal using Synthesia's technology: "I Used AI To Propose To My Wife In Her Native Language."[17] In the video, a white man from the United States agrees to ask his Chinese spouse to marry him again, this time proposing in Mandarin. How can he pull off this feat without speaking Mandarin? Technologists from Synthesia film him delivering the proposal in English, creating a computer model of his facial expressions as he speaks. A Mandarin-speaking vocal actor then reads the translation of his proposal in Mandarin. The technology then maps the vocal sounds and facial expressions onto the man's face, allowing him to "speak" to his spouse in her native language.

[15] See http://web.archive.org/web/20190428185007/https://www.synthesia.io/.
[16] See http://web.archive.org/web/20190428185005/https://www.synthesia.io/ethics.
[17] Kanter (2018).

2.3. Museum Informatics

The emerging field of museum informatics seeks to inform and engage visitors about works of art through new media and digital technologies. Developments in augmented reality will make the current audio tours with the bulky headsets and players seem woefully dated. Imagine coming across Lucas Cranach the Elder's portrait of Martin Luther. By holding up your phone in front of the painting, you might see Luther turn to face you and begin to describe his ongoing efforts to reform the church, his intention to translate the Bible into German, and his sorrow at the loss of his daughter, Elizabeth. Through augmented reality, the portrait becomes a window into another time, another place, educating viewers about the people, places, and events they find depicted in oil.

The ability to produce this kind of animation is not novel. Using game development platforms like Unity or Unreal Engine, skilled animation artists create and animate sprites from static images. But deep learning promises to automate the process and make it scalable. In "Few-Shot Adversarial Learning of Realistic Neural Talking Head Models," a team of scientists from the Samsung AI Center in Moscow and the Skolkovo Institute of Science and Technology have created algorithms for generating animated representations from photographs. What is innovative about their technique is the ability to produce these animations from a single image: "Our system can generate a reasonable result based on a single photograph (one-shot learning), while adding a few more photographs increases the fidelity of personalization."[18] The team used their deep learning algorithms to generate animated models from images of the *Mona Lisa*, Fyodor Dostoevsky, Albert Einstein, and Marilyn Monroe. The title of the article in *Art News* covering the achievement encapsulates the response from curators and art historians: "Russian Researchers Used AI to Bring the Mona Lisa to Life and It Freaked Everyone Out."[19]

As it happens, artists are already using deepfake technology in their works. In a 2020 exhibition at the International Center of Photography in New York, James Coupe created a series of installations that permit visitors to insert themselves digitally into the 1979 film, *The Warriors*.[20] But, as Jason Farago contends in *The New York Times*, the artistic potential of deepfakes remain essentially unexplored. Beyond the "janky tech," Farago labels Coupe's installations as "tech for tech's sake," and remarks that the In-

[18] Zakharov u. a. (2019), 2.
[19] Dafoe (2019).
[20] See https://www.icp.org/exhibitions/james-coupe-warriors.

ternational Center of Photography "should expect artists to examine life as shaped by new photographic technologies, rather than simply announce new technologies exist."[21] Whatever we think of Farago's judgment of *Warriors*, he is undoubtedly right that we find ourselves at the awkward beginning of creative inquiry into the artistic potential of deepfakes.

3. The Technology of deepfakes

How do deepfakes work in practice? The technology of deepfakes belongs to a subfield of machine learning called "deep learning." As Gary Marcus succinctly defines it, "Deep learning...is essentially a statistical technique for classifying patterns, based on sample data, using neural networks with multiple layers."[22] In less abstruse terms, the goal is to take a set of inputs and map its contents to a labeled set of outputs.[23] At the beginning of the process, we start with a bunch of unlabeled data we want to label, and the algorithm's job is to draw lines between the data and the correct labels. As Marcus indicates, a typical application of deep learning is taking a digitized set of manuscripts and mapping the handwritten letters to some canonical alphabetic representation. The thing with deep learning is that the lines are not drawn directly from the input set to the labeled data. Rather, the lines from the initial data pass through interim layers until they converge on the labels. Forward and backward propagation algorithms allow for input and output layers to communicate through sets of interim layers, making adjustments between the "neurons" (or provisional mappings) until the fit between inputs and outputs becomes satisfactory. "It's like a giant game of telephone" explains Andrew Trask in *Grokking Deep Learning*, "at the end of the game, every layer knows which of its neurons need to be higher and lower..."[24] Unlike the game of telephone where communication frequently goes hilariously wrong, these web of connections often wind up producing uncannily accurate outcomes.

The development of a technique termed "Generative Adversarial Networks" (GANs) reduced the computational expense of producing deepfake videos.[25] The leading idea is to pit two deep learning models against one another. The first model (the "generative model") presents its output data to the second model (the "discriminative"

[21] Farago (2020).
[22] Marcus (2018).
[23] Ibid., 4.
[24] Trask (2019).
[25] Goodfellow u. a. (2014).

model), which seeks to classify that data as a product of the generative model or a sample of the data-to-be-modeled (i.e. the training data). As the authors of the 2015 publication that introduced the concept explain, "The generative model can be thought of as analogous to a team of counterfeiters, trying to produce fake currency and use it without detection, while the discriminative model is analogous to the police, trying to detect the counterfeit currency. Competition in this game drives both teams to improve their methods until the counterfeits are indistinguishable from the genuine articles."[26] The innovative aspect of this technique is that both generator and discriminator are learning as the game proceeds. The generator continues to create data distributions that approximate the training data and the discriminator learns to distinguish between the generator and the training data more accurately. The competition between the models concludes when, as the analogy suggests, the generator produces models that the discriminator can no longer reliably distinguish from the training data. GANs are not guaranteed to bring generator and discriminator into equilibrium; they sometimes oscillate between suboptimal solutions. Researchers have put forth pragmatic techniques to prevent the models from collapsing before converging.[27]

3.1. The Democratization of Manipulation

The ability to produce image-to-image translations is not new.[28] Major movie production studios already have technologies to produce realistic body doubles. As Patrick Shanley and Katie Kilkenny wrote in *The Hollywood Reporter*, "Hollywood has long swapped faces – just using different tech."[29] For example, studios have used these methods to create continuities in fictional universes like *Star Wars*, bringing back characters like Princess Leia and Grand Moff Tarkin after the deaths of Carrie Fisher and Peter Cushing.[30] If special effect studios in Hollywood possessed the technology for creating synthetic videos, then intelligence agencies in the United States and abroad must have too. After all, intelligence agencies around the world have produced propaganda, manipulated media, and planted 'false flags' for decades. Among the materials from the National Security Agency that Edward Snowden released in 2014 is a document listing the British Government Communications Headquarters' (GCHQ) digital manipulation tools.[31] While the ability to alter digital video may not be new,

[26] Ibid., 1.
[27] Goodfellow (2016), 34.
[28] Shen u. a. (2018), 1.
[29] Shanley and Kilkenny (2018).
[30] Kemp (2019); Shanley and Kilkenny (2018).
[31] Ball (2014).

organizations and agencies lacked the wherewithal to pull off these transformations. What is new about 'deepfakes' is the democratization of video manipulation.

4. The Cognitive Science of Deepfakes

Over Labor Day weekend in 2019, I attended Dragon Con, an annual gathering of more than 85,000 fans of science fiction, fantasy, gaming, and other forms of contemporary geek culture. Alongside all sessions devoted to exploring Dr. Who, Harry Potter, Star Trek, and the latest anime, there is a Dragon Con Skeptic Track devoted to "critical thought, extraordinary claims, and promotion of good science." This year, the track sponsored a session titled, "How Deep Is Your Fake?" on the challenge of identifying and debunking deepfake videos. The presenter, Teddi Fish, who cosplayed as Teddy the Flying Spaghetti Monster while giving her talk, provided an overview of the state of the problem from a technical as well as a social perspective, concluding with a slide advising "Question before you share. Question that with which you agree. Stay skeptical." The advice sounds laudable and, certainly, nobody wants to fall prey to fraud.

According to Karen Hao, our anxiety about being mislead by deepfakes may be creating the negative effects we are seeking to avoid. In "The Biggest Threat of Deepfakes Isn't the Deepfakes Themselves," she notes that overly skeptical viewers have already come to regard authentic videos as potential fakes, leading to serious political consequences.[32] In other words, we are becoming so concerned about the potential of fraudulent video that political agents are using that anxiety against us, discrediting videos as misinformation and 'fake news.' As Hao quotes Aviv Ovadya, an expert in misinformation: "What [disinformation actors] really want is not for you to question more, but for you to question everything."[33]

Skepticism runs counter to core principles of human psychology and information economics. As Fish herself remarked during her presentation at Dragon Con, "human beings are wired so that what we see sticks in our brain as something that is, in fact, reality."[34] If we doubt everything, our ability to act degrades. A major reason we have trademarks and service marks is, in fact, to save us the trouble of evaluating sources.

[32] Hao (2019).
[33] Ibid.
[34] http://video.skeptrack.org.

As the American pragmatists taught us more than 150 years ago, absolute skepticism is a practical impossibility. We cannot suspend belief in all our convictions simultaneously. In "Some Consequences of Four Incapacities" (1868), Charles Sanders Peirce argued that Cartesian skepticism foundered on this practical inability. Peirce noted that Cartesianism "teaches that philosophy must begin with universal doubt," but countered that such a standpoint is self-deceptive.

> We cannot begin with complete doubt. We must begin with all the prejudices which we actually have when we enter upon the study of philosophy. These prejudices are not to be dispelled by a maxim, for they are things which it does not occur to us *can* be questioned. Hence this initial skepticism will be a mere self-deception, and not real doubt; and no one who follows the Cartesian method will ever be satisfied until he has formally recovered all those beliefs which in form he has given up.[35]

A problem with advocating sweeping doubt about the veracity of every digital image or audiovisual recording we encounter is that, if we followed that advice, we would lose our ability to act. We cannot be skeptical about everything we see. At best, we can train ourselves about when to become skeptical. To become skeptical about something we thought we knew, as Peirce indicated, we need to have genuine grounds for doubting its veracity; cultivating artificial doubt will not lead us to the truth about what we are seeing.

If casting doubt on everything we see until it is proven true does not constitute a workable strategy, what can we do to prevent ourselves from falling for misinformation? From the standpoint of cognitive science, the task may actually be more difficult than it appears. In "Believing that Humans Swallow Spiders in Their Sleep: False Beliefs as Side Effects of the Processes that Support Accurate Knowledge," Elizabeth J. Marsh, Allison D. Cantor, and Nadia M. Brashier of Duke University examine how errors become integrated into our "knowledge base" through what they term "adaptive processes." These processes "normally support accurate knowledge, but sometimes backfire and allow the introduction of errors into the knowledge base."[36] In their article, they review five such adaptive processes. Of these, I'd like to highlight three processes that connect directly with the question of deepfakes.

[35] Peirce (1868).
[36] Marsh, Cantor and Brashier (2016), 107.

First, the authors note that disbelieving something we learn takes more cognitive effort than believing it.[37] Our cognitive wiring is such that we tend to accept novel information as true; it takes mental effort to flag it as false. As they point out, this strategy makes sense given that human beings evolved in an environment where perceptions are generally grounded in the truth. Of course, we do have cognitive systems for rejecting perceptions as untrue. But psychologists have demonstrated that short-circuiting these higher-level evaluative systems is not difficult.[38] As we distractedly scrolled through social media feeds in 2017 during Hurricane Harvey, who among us paused to reflect on the likelihood of a shark swimming along one of the flooded aqueducts, as depicted in a heavily-shared image? Who of those who saw the image on Twitter later read Linda Qiu's admonition in the *New York Times*, "Don't believe it. This fake image is an old hoax that circulates routinely after major hurricanes."[39]

Another "adaptive process" that inhibits our ability to screen out errant beliefs is what the authors term the "fluency-based heuristic for judging truth."[40] The effect stems from confusion between our ability to process information and the truth value of that information. If we can recall something readily to mind, we are more prone to judge it as true. As the authors note, advertisers exploit this effect by exposing people in certain markets again and again to certain claims, making it easier for them to remember those assertions and, hence, to assume their truth. On a related note, pairing an image with factual assertions amplifies people's tendency to accept those assertions, even if the image is factually unrelated.[41]

A final "adaptive process" worth noting is that we often accept "partial matches" when making connections between facts. The authors note that speech communication is fraught with parapraxis and other forms of verbal disfluencies. When someone is struggling with communicating an idea, we generally try to make sense of what that person is saying, filling in the gaps while reassuring him or her that we "know what you mean." But, as it happens, employing this strategy also means that we tend to gloss over factual errors. The authors point to an effect that Thomas D. Erickson and Mark E. Mattson described as the "Moses Illusion" to illustrate this tendency. As Erickson and Mattson demonstrated, when asked " 'How many animals of each kind did Moses take on the Ark?' most people answer 'two"', overlooking that Noah built

[37] Ibid., 108.
[38] Ibid.
[39] Qui (2017).
[40] Marsh, Cantor and Brashier (2016), 108.
[41] Ibid., 110.

the Ark, not Moses.[42] The etiology of this effect is not certain, but Marsh, Cantor, and Brashier follow Erickson and Mattson by assuming that "monitoring [for errors] takes effort, and accepting 'good-enough' representations is a shortcut that normally works."[43]

The upshot of this research is to show that our cognitive processes balance efficiency and accuracy when assimilating new information. To my knowledge, researchers have not yet studied how these adaptive processes will affect our ability to judge the veracity of deepfake videos, but we might readily imagine that their producers will draw on this research to make them slip past our cognitive defenses. While adopting a skeptical attitude toward what we see may help us to screen out errors, doing so will also slow down our assimilation of new information.

5. Potential Countermeasures

If deepfake videos threaten to undermine the cognitive processes we use to process information, how can we as a society address the threat? The two primary lines of counterattack at present are technological and legal. As we will discover, these two means of counteracting the threat of deepfakes are promising, but inherently limited.

On the one hand, technologists recognized the threat posed by the widespread availability of tool sets for creating synthetic videos and began to develop forensic software to detect such videos. The techniques range from looking for simple physiological tells, like unnatural patterns of eye blinking,[44] to sophisticated "ensemble" models.[45] In September 2019, Facebook announced a "Deepfake Detection Challenge" to incentivize the study of detection methods;[46] Amazon and Microsoft, as well as several academic institutions, have since joined on.[47] As a part of this initiative, Facebook released a dataset of 100,000 videos, some of which are the products of audovisual manipulation, for researchers to use a proving ground for detection algorithms.[48] A competition on Kaggle for the most effective detection algorithms promises awards of up to half a million dollars.[49] The social networks have intrinsic interest to expose

[42] Erickson and Mattson (1981).
[43] Marsh, Cantor and Brashier (2016), 116.
[44] Li and Lyu (O A 2018).
[45] Yu, Chang and Ti (2019).
[46] Cole (2019).
[47] See https://deepfakedetectionchallenge.ai.
[48] Dolhansky u. a. (2019).
[49] See https://www.kaggle.com/c/deepfake-detection-challenge/overview.

fake news, including fake videos, to stave off increasing political scrutiny and additional regulation. But the effort will benefit noncommercial entities as competitors must release their code under open source licenses to qualify for the prizes. An intrinsic problem is that developers of deepfake toolkits can also use these improvements in detection methods to refine and enhance their algorithms. "Battling deepfake algorithms with detection algorithms using CNNs [Convolutional Neural Networks], RNNs [Recurrent Neural Networks], and other methods ultimately leads to a perpetual machine-learning cat-and-mouse game."[50]

On the other hand, politicians and legal scholars are investigating ways to inhibit the spread of deepfake videos through regulation and legislation. In a recent law review comment, Elizabeth Caldera surveys potential legal and regulatory approaches. She argues that "while it is likely too late to control the actual technology behind deepfakes, it is not too late to regulate the videos actually produced."[51] But this goal may prove elusive. As noted above, these proposed remedies should address the likely harms of deepfake videos without prohibiting their potential benefits. Caldera's quick survey of applicable areas of law, ranging from right of publicity, copyright law, and laws against "revenge pornography,"[52] shows the difficulty of lining up our ethical intuitions with existing legal frameworks. Caldera is more sanguine about the possibility of administrative regulations, either from the Federal Trade Commission (FTC), the Federal Communications Commission (FCC), or perhaps a to-be-established Agency of Artificial Intelligence,[53] despite the current administration's general disinclination to create new federal regulation. Noting that whatever regulation comes to pass will need to pass muster with the free speech protections of the First Amendment, Caldera suggests the federal government might require deepfake videos to label themselves as modified. While this proposal sounds modest, would it also require users of social media and dating sites to admit to fixing up their selfies when distributing them online? Given the pervasive use of photographic filters, such a regulation might well require us to label nearly all photographs on social media as enhanced, allowing deepfakes once again to circulate unnoticed.

While technologists and legislators seek appropriate measures to counteract misleading and harmful deepfake videos, I suggest that we need to add a third approach based in spirituality to complement technology and the law.

[50] Greengard (2019).
[51] Caldera (2019), 203.
[52] Ibid., 192–3.
[53] Ibid., 193–97.

6. A Spirituality of Media Iconoclasm

The hermeneutics of suspicion is a kind of latter-day iconoclasm. As we have seen, Paul Ricœur described the hermeneutics of suspicion as a "tearing off of masks." Like any iconoclasm, the goal is not solely to destroy, but also to see. By tearing away the mask, we hope to behold the face behind it: the truth behind the appearance. But the aggressive act of tearing a mask away clashes with a more subtle form of revelation we find in the biblical narrative. In the Song of Songs, for instance, the lover perceives his beloved through a veil: "How beautiful you are, my love, how very beautiful! Your eyes are doves behind your veil" (Song of Solomon 4:1; NRSV). Here, the veiling reveals as well as conceals. As Paul J. Griffith notes in his commentary on the passage, the beloved's eyes "are veiled because their beauty would otherwise be too radiant: the world, and the gaze of the lover, must be protected from them."[54] The veil serves a purpose, obscuring in order to reveal. While a hermeneutics of suspicion would rid us of masks and veils, we risk becoming blinded as a consequence. Not all truths should be looked on directly.

Philosophically, the notion of the body as veil takes central place in the phenomenological philosophy of Edmund Husserl. In the fifth chapter of his *Cartesian Meditations* (1931), Edmund Husserl explores the phenomenology of intersubjectivity.[55] Husserl tackles the question of our perception of the other. How do we experience another consciousness in the word of objects? The experience of an other differs from the experience of an object, but we never encounter the ego of the other directly. If we did, Husserl wrote, the other would become ourselves. To maintain the distinction between ourselves and the other, we encounter the other through some mediating form, whether a physical body, a voice, a moving image. Husserl describes the intuition that an ego exists behind the form as a "mediate intentionality." As he explains in §50,

> *A certain mediacy of intentionality* must be present here, going out from the substratum, "primordial world" [...] and making present to consciousness a "there too", which nevertheless is not there itself and can never become an "itself-there."[56]

Husserl develops the concept of apperception to articulate this form of mediated intentionality. In perceiving the other, we perceive first the body of the other and then,

[54] Griffiths (2011), 90.
[55] Husserl (1960).
[56] Ibid., 109.

by way of analogy, the "I" of the other. The apperception of the other does not function as a temporal two step whereby we first see a body, and then analogize to the presence of an ego. The body and the ego become paired in apperception, but nevertheless remain conceptually distinct and never fused or collapsed. The veil of mediating form cannot be stripped away but through its fabric we perceive the other "I" who stands before us.

Husserl described the apperception of the other, that is, the perceiving of a spiritual alter "ego" through the veil of physical presence, as "*transcendental theory of experiencing someone else*" or "a transcendental theory of so-called 'empathy' [*Einfühlung*]."[57] The role empathy plays in constituting our perception of the other has been the subject of philosophical discussion.[58] For our purposes, what is crucial is the distinction between intentional experience of the physical presence of the other and empathetic perception of the spiritual "I" of the other. For this distinction allows us to imagine exercising empathy to perceive a spiritual other with a completely different surface form than our own. Or, conversely, confronting a form that, though familiar in its external features, proves impenetrable in fact – a form that does not lead to a spiritual reality, no matter how empathetically we seek the other behind the veil.

Strangely, Husserl's meditations on intersubjectivity from 1931 bring us close to Alan Turing's reflections on artificial intelligence from 1950. In the 'imitation game' that Turing described in *Computing Machinery and Intelligence*, the goal is to discern whether the messages you receive across a physical barrier come from a spiritual "I" (presumably, a intelligent being) or a vacuous mechanical device.[59] The goal of the machine is to convince you that it is not a machine but a person. The machine employs subtle deceptions to achieve this effect, making blunders in chess, taking longer than expected to calculate numbers, responding poetically with allusions to Shakespeare. The question behind the test is whether the human interlocutor can see through these guises, correctly identifying and unmasking the machine. Effectively, Turing is identifying intelligence with empathy. That is, he asks us to empathize with the sender of the messages, seeking to find a spiritual other behind the veil. As we find the surfaces of perception becoming increasingly diverse and deceptive, we may find that empathy, as conceived by the philosophers in the phenomenological tradition, becomes key to exposing or exploring the spiritual dimensions of deepfakes.

[57] Ibid., 92.
[58] Zahavi (2015), Chapter 10 on the phenomenological analysis of empathy according to Max Scheler, Edmund Husserl, Edith Stein, and Alfred Schutz.
[59] Turing (1950).

The growing alarm over the impact of deepfake videos correlates with the media saturation of contemporary culture. A partial solution to the threat of deepfake videos would be to remove ourselves from the theatre of contemporary media, stepping away from Times Square into quieter backstreets. Jaron Lanier has delivered modern day jeremiads against social media, arguing that social media has deleterious effects not only on our ability to discern the truth but to cultivate our souls.[60] Certainly, limiting our exposure to social media reduces our personal vectors of attack. When we imagine participating in a Turing test, we picture ourselves in the controlled setting of a research laboratory, attentively scrutinizing the messages arriving at intervals for our inspection. In the online world, though, we have to balance multiple Turing tests at once. While deepfakes remain rare, the number of so-called 'shallow fakes,' that is, images and videos subtly and not-so-subtly manipulated to achieve certain effects, have become ubiquitous.[61] Scrolling absent-mindedly through social media feeds, we devote scant attention to whether a bot produced some controversial tweet or whether a shocking image might have been photoshopped. In these circumstances, most fail the Turing tests, as shown by the number of politicians, journalists, and others who unwittingly interacted with bots on Twitter during the 2016 election. But, as Darren Linvill and Patrick Warren argue, these twitter bots are engineered to play to our biases and slip through our cognitive defenses.[62] The more confident we feel of our ability to suss out shallow fakes, deepfakes, and other forms of disinformation online, the likelier we will unwittingly fall prey to them as none of us can process and evaluate so much (dis)information at once.

A spirituality of iconoclasm imposes distance from the cascading series of images that surround us online to cultivate empathy. The purpose of fostering this remove from visual culture is not to reject images wholesale as false representations, but to consider them with greater intentionality, thoughtfulness, and perspicacity. By fostering a reserve, whether ironic, intellectual, or spiritual, toward visual media, we gain facility in reading and interpreting their cultural logic. This philosophical reserve toward visual culture has roots in Platonism, as Edith Wyschogrod noted.

> In the new age of images there are only images. Could it not be argued that the promiscuity of the image was already present in Plato's philosophy? From the

[60] Lanier (2018).
[61] Johnson (2019).
[62] Linvill und Warren (2019).

> Platonic standpoint, art objects, shadows, and the reflections of things are the wanton and wild images that escape regimentation by the logos.[63]

As Wyschogrod also anticipated,[64] far from being unregimented, the logos saturates deepfakes. The synthesis of disparate objects, the swapping of body parts, switching voices, and juxtaposition of dissonant elements reflect the mind of a creator, carried out through data, algorithms, and processing power. The overabundance of logos in deepfake videos is perhaps the best 'tell,' as the design is so perfect that it becomes uncanny. But where does this saturating logos lead? To the void or to a genuine spiritual "I" communicating through its computational veil? Only empathetic intuition may tell. But we cannot exercise empathy "at scale." Cultivating empathy online requires us to tarry and dwell, not to rush and react.

What would an epistemology of iconoclastic empathy look like in practice? A little science fiction might assist our imaginations by way of conclusion. In his story "Liking What You See: A Documentary" (2002), Ted Chiang imagines a medical condition called 'calliagnosia' that disrupts the recognition of beauty.[65] Those afflicted with this condition still recognize others but they cannot discriminate between ugliness and beauty. Chiang builds the narrative from the documentary reports of various agents, ranging from college students to neuroscientists, exploring the advantages and limitations of taking a drug to induce calliagnosia. A major question of the story is why physical beauty should shape our perception of the spiritual "I." As a student in the story avers, "Calli doesn't blind you to anything; beauty is what blinds you. Calli lets you see."[66] Chiang asks his readers to examine the degree to which their social interactions transpire on the surface. We all know the truism "beauty is skin deep" and, when pressed, will readily agree that beauty should not blind us to character. Yet the pursuit of beauty remains central to our lives off- and online, as witnessed by the dominance of Instagram and dating apps like Tinder. Would iconoclastic empathy have Christians placing personals at the back of literary magazines instead of circulating photoshopped images on OkCupid?[67]

In a manner similar to the self-imposed limitations of Calli, an epistemology of iconoclastic empathy would help us to discern truth from falsity by training us to look

[63] Wyschogrod (1998), 73.
[64] Wyschogrod already pointed to the dislocating possibility of "synthetic human actors" in 1998; see ibid., 83.
[65] Chiang (2010), 237–74.
[66] Ibid., 248.
[67] Rose (2010).

beyond surface appearances when interacting online. Training in such practices takes time, patience, and community commitment. In practical terms, congregations might commit themselves to limiting their social media exposure and to interacting with more profundity with fewer people online. Alternatively, they might eschew sites that rely primarily on videos and images as media of communication, returning to text-based communications. The pragmatics of exercising empathetic communication online remain to be worked out. For old timers, this will feel like a throwback to earlier times, when people dialed up to early bulletin board systems like The Well and Echo for the novel experience of chatting with others across the country, knowing that not everyone was who they purported to be.[68] These social communities continue to exist on sites like Wikipedia, where you gain reputation through the hard work of writing, editing, and improving the encyclopedia.

Iconcolastic empathy might also provide rubrics for developing new forms of interaction online. As Lanier remarks, "I still believe that it's possible for tech to serve the cause of empathy. If a better future society involves better tech at all, empathy will be involved."[69] The artistic activities of Stephanie Dinkins, associate professor of art at Stony Brook University, demonstrates both creative capacity and inherent limitations of empathic engagement. In a series of videos titled *Conversations with Bina48*, Dinkins documents her interactions with an African-American android[70] (or, more precisely, a robotic visage who appearance is modeled after an African American woman).[71] The conversations are elliptical and border at times on nonsensical. When I discuss these videos in class, students debate whether developing an emotional bond with Bina48 is a sensible goal. But they generally appreciate Dinkins' persistent attempts to forge affective bonds with Bina48, despite the awkward and wayward conversations. Empathy is a powerful force, simultaneously capable of unmasking digital fakes and also coaxing digital simulacra to life.

The challenge of deepfakes will require collective effort from multiple parties. Technologists and legal scholars have essential contributions to make. We need more sophisticated algorithms and tools to detect synthetic videos as well as rules and regulations to curb their deleterious social and political effects. The argument of this paper is that, while such efforts are necessary, they are ultimately not sufficient. As Lanier suggests, we have grown accustomed to online environments that produce high vol-

[68] Evans (2018), chap. 9.
[69] Lanier (2018), 76.
[70] See https://www.stephaniedinkins.com/conversations-with-bina48.html.
[71] See Harmon (2010).

umes of disinformation.[72] He argues that we need to distance ourselves from these systems and engage ourselves in the effort to build more empathetic forms of digital interaction. From this perspective, deepfakes present us an opportunity to reexamine our broader engagement with humans (and computers) online. The problem is not synthetic videos *per se*. The ability to create them may, in fact, have positive uses for church and society. The profounder issue is our participation in channels of communication that reduce empathy and occlude truth. Addressing the proliferation of deepfakes cannot just mean becoming more critical and suspicious about everything we see online. As Ricœur understood, the hermeneutics of suspicion should not be an end-in-itself, but a means toward achieving a second naïveté. After any new iconoclasm breaks apart the fake, sterile, and empty images confronting us online, our next task as Christians is to develop digital systems that promote truth, empathy, and genuine depth.

Bibliography

Abelson, Hal, Ken Ledeen and Harry Lewis. 2008. *Blown to Bits: Your Life, Liberty, and Happiness After the Digital Explosion*. Upper Saddle River: Addison-Wesley.

Ball, James. 2014. GCHQ Has Tools to Manipulate Online Information, Leaked Documents Show. *The Guardian* July 2014. https://www.theguardian.com/uk-news/2014/jul/14/gchq-tools-manipulate-online-information-leak (accessed: September 6, 2021).

Brébisson, Alexandre de. 2019. How Imputations Work: The Research Behind Overdub. September. https://www.descript.com/post/how-imputations-work-the-research-behind-overdub (accessed: March 13, 2020).

Caldera, Elizabeth. 2019. Reject the Evidence of Your Eyes and Ears: Deepfakes and the Law of Virtual Replicants. *Seton Hall Law Review* 50: 177.

Electronic Privacy Information Center. 2019. State Revenge Porn Policy. https://epic.org/state-policy/revenge-porn/ (accessed: September 6, 2021).

Chiang, Ted. 2010. *Stories of Your Life and Others*. New York: Knopf.

Cole, Samantha. 2017. AI-Assisted Fake Porn Is Here and We're All Fucked. *Vice* December. https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn (accessed: September 6, 2021).

[72] Lanier (2018), Argument Four.

—. 2018. We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now. *Vice* January. https://www.vice.com/en/article/bjye8a/reddit-fake-porn-app-daisy-ridley (accessed: September 6, 2021).

—. 2019. Facebook Just Announced $10 Million 'Deepfakes Detection Challenge.' *Vice* September. https://www.vice.com/en/article/8xwqp3/facebook-deepfake-detection-challenge-dataset (accessed: September 6, 2021).

Dafoe, Taylor. 2019. Russian Researchers Used AI to Bring the Mona Lisa to Life and It Freaked Everyone Out. See the Video Here. *Artnet News* May. https://news.artnet.com/art-world/mona-lisa-deepfake-video-1561600 (accessed: March 13, 2020).

Dolhansky, Brian, Russ Howes, Ben Pflaum, Nicole Baram und Cristian Canton Ferrer. 2019. The Deepfake Detection Challenge (DFDC) Preview Dataset.

Erickson, Thomas D. and Mark E. Mattson. 1981. From Words to Meaning: A Semantic Illusion. *Journal of Verbal Learning and Verbal Behavior* 20, Nr. 5: 540–51. doi:10.1016/S0022-5371(81)90165-1.

Evans, Claire L. 2018. *Broad Band: The Untold Story of the Women Who Made the Internet*. New York: Portfolio.

Farago, Jason. 2020. International Center of Photography Refocuses in a New Home. *The New York Times* January 30, 2020. https://www.nytimes.com/2020/01/30/arts/design/international-center-of-photography-new-home.html (accessed: September 6, 2021).

Goodfellow, Ian. 2016. NIPS 2016 Tutorial: Generative Adversarial Networks December.

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio. 2014. Generative Adversarial Networks June. https://arxiv.org/abs/1701.00160 (accessed September 6, 2021).

Greengard, Samuel. 2019. Will Deepfakes Do Deep Damage? *Communications of the ACM* 63, Nr. 1: 17–19. doi:10.1145/3371409.

Griffiths, Paul J. 2011. *Song of Songs*. Brazos Theological Commentary on the Bible. Grand Rapids: Brazos.

Gupta, Swati. 2019. Four Dead in Bangladesh Riot over Facebook Post. *CNN*. https://www.cnn.com/2019/10/21/asia/riot-deaths-facebook-post-intl-hnk/index.html (accessed: March 13, 2020).

Hao, Karen. 2019. The Biggest Threat of Deepfakes Isn't the Deepfakes Themselves. *MIT Technology Review*. https://www.technologyreview.com/2019/10/10/132667/the-biggest-threat-of-deepfakes-isnt-the-deepfakes-themselves/ (accessed: March 13, 2020).

Harmon, Amy. 2010. Making Friends with a Robot Named Bina48. *The New York Times* July.

Husserl, Edmund. 1960. *Cartesian Meditations: An Introduction to Phenomenology*. Leiden; The Hague: Martinus Nijhoff.

Johnson, Bobbie. 2019. Deepfakes Are Solvablebut Don't Forget That 'Shallowfakes' Are Already Pervasive. *MIT Technology Review* March.

Kanter, Steven. 2018. I Used AI to Propose to My Wife in Her Native Language December.

Kemp, Luke. 2019. In the Age of Deepfakes, Could Virtual Actors Put Humans Out of Business? *The Guardian* July.

Lanier, Jaron. 2018. *Ten Arguments for Deleting Your Social Media Accounts Right Now*. New York: Henry Holt.

Li, M. Chang, Y. and S. Lyu. Eds. 2018. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. In: *2018 IEEE International Workshop on Information Forensics and Security*, 1–7. Hong Kong: IEEE Computer Society. doi:10.1109/WIFS.2018.8630787.

Linvill, Darren and Patrick Warren. 2019. That Uplifting Tweet You Just Shared? A Russian Troll Sent It. *Rolling Stone* November.

Liotta, Edoardo. 2019. Student in Mumbai Arrested for Editing Girl's Face onto Porn and Threatening to Share It. *Vice* October. https://www.vice.com/en/article/j5ynkg/student-in-mumbai-arrested-for-editing-girls-face-onto-porn-and-threatening-to-share-it-on-social-media (accessed: September 6, 2021).

MacKinnon, Rebecca. 2013. *Consent of the Networked*. New York: Basic Books.

Marcus, Gary. 2018. Deep Learning: A Critical Appraisal. *arXiv* 1801.00631. http://arxiv.org/abs/1801.00631 (accessed: March 13, 2020).

Marsh, Elizabeth J., Allison D. Cantor and Nadia M. Brashier. 2016. Believing That Humans Swallow Spiders in Their Sleep: False Beliefs as Side Effects of the Processes

That Support Accurate Knowledge. Ed. Biran H. Ross. *Psychology of Learning and Motivation* 64: 93–132. doi:10.1016/bs.plm.2015.09.003.

Merritt, Jonathan. 2017. Is AI a Threat to Christianity? *The Atlantic* February. https://www.theatlantic.com/technology/archive/2017/02/artificial-intelligence-christianity/515463/ (accessed: March 13, 2020).

Peirce, Charles Sanders. 1868. Some Consequences of Four Incapacities. *Journal of Speculative Philosophy* 2: 140–57.

Portman, Rob. 2019. Text - S.2065 - 116th Congress (2019-2020): Deepfake Report Act of 2019. https://www.congress.gov/bill/116th-congress/senate-bill/2065/text (accessed: March 13, 2020).

Qui, Linda. 2017. A Shark in the Street, and Other Hurricane Harvey Misinformation You Shouldn't Believe. *The New York Times* August.

Ricœur, Paul. 1970. *Freud and Philosophy: An Essay on Interpretation*. New Haven: Yale University Press.

Rose, David. Ed. 2010. *hey Call Me Naughty Lola: Personal Ads from the London Review of Books*. New York: Scribner.

Scott-Baumann, Alison. 2012. *Ricoeur and the Hermeneutics of Suspicion*. London: Continuum.

Shanley, Patrick and Katie Kilkenny. 2018. Deepfake Tech Eyed by Hollywood VFX Studios. *The Hollywood Reporter*. https://www.hollywoodreporter.com/news/deepfake-tech-eyed-by-hollywood-vfx-studios-1087075 (accessed: March 13, 2020).

Shen, Tianxiang, Riuxian Liu, Ju Bai and Zheng Li. 2018. 'Deep Fakes' Using Generative Adversarial Networks (GAN). http://noiselab.ucsd.edu/ECE228_2018/Reports/Report16.pdf (accessed September 9, 2021).

Trask, Andrew. 2019. *Grokking Deep Learning*. Shelter Island: Manning.

Turing, Alan M. 1950. Computing Machinery and Intelligence. *Mind* LIX, Nr. 236: 433–60. doi:10.1093/mind/LIX.236.433.

Wyschogrod, Edith. 1998. *An Ethics of Remembering: History, Heterology, and the Nameless Others*. Chicago: University of Chicago Press.

Yu, Chia-Mu, Ching-Tang Chang and Yen-Wu Ti. 2019. Detecting Deepfake-Forged Contents with Separable Convolutional Neural Network and Image Segmentation. https://arxiv.org/abs/1912.12184 (accessed September 9, 2021).

Zahavi, Dan. 2015. *Self and Other: Exploring Subjectivity, Empathy, and Shame*. Oxford: Oxford University Press.

Zakharov, Egor, Aliaksandra Shysheya, Egor Burkov and Victor Lempitsky. 2019. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. *arXiv* 1905.08233. https://arxiv.org/abs/1905.08233 (accessed: March 17, 2020).