# Beyond deep fakes: Conceptual framework and research agenda for neural rendering of realistic digital faces

Mike Seymour
University of Sydney
mike.seymour@sydney.edu.au

Kai Riemer
University of Sydney
kai.riemer@sydney.edu.au

Ligyao Yuan
Iowa State University
lyuan@iastate.edu

Alan R. Dennis
Indiana University
ardennis@indiana.edu

## Abstract

*Neural rendering (NR) has emerged as a novel technology for the generation and animation of realistic digital human faces. NR is based on machine learning techniques such as generative adversarial networks and is used to infer human face features and their animation from large amounts of (video) training data. NR shot to prominence with the deep fake phenomenon, the malicious and unwanted use of someone's face for deception or satire. In this paper we demonstrate that the potential uses of NR far outstrip its use for deep fakes. We contrast NR approaches with traditional computer graphics approaches, discuss typical types of NR applications in digital face generation, and derive a conceptual framework for both guiding the design of digital characters, and for classifying existing NR use cases. We conclude with research ideas for studying the potential applications and implications of NR-based digital characters.*

## 1. Introduction

There has been a steady progression in recent years towards creating photo realistic digital characters, avatars and agents [1]. These digital human entities have been widely adopted in many industries, such as entertainment, gaming, education and communication.

The field of digital humans extends from digital representations of people in videos, to fully synthetic digital agents and virtual avatar representatives. These agents are appearing as digital influencers and organizational representatives, as well as complex avatars that stand in for their owners in virtual meetings and events.

Driven most recently by the COVID-19 pandemic, many digital production companies are looking for cost-efficient ways to generate fully digital characters, with only minimal involvements from real actors. This has led to significant growth and the emergence of a range of start-up businesses, many with significant sums of venture capital investment.

Virtual influencers, like Lil Miquela, have allowed Brud to raise over $125M from investors. With applications such as virtual police officers, mental health professionals and assistant professionals, embodied agent company Soul Machines successfully raised over $40M investment. Virtual mentors like Digital Deepak Chopra and the AI Foundation, raised over $10M. The popularity of virtual performers seen in events such as the Travis Scott's Fortnite concert or the John Legend digital concert, allowed The Wave to raise over $30M, and virtual real-world celebrities company Genies to raise over $35M [2, 3, 4, 5, 6].

The capital market success of these entities is accelerating technological developments in the field of neural rendering (NR), a technology for inferring realistic human faces from training data using deep learning techniques, such as generative adversarial networks (GANs).

The most prominent, but also quite limited, example of NR is its use in the now infamous deep fakes of famous people. Deep fakes present a form of face 'hijacking', whereby video material of a public person (the source material) is used to train a machine learning (ML) program to generate and overlay the source person's face onto existing video footage (the target material), such that the target person's facial expression drives features of the source face. This has been used to place faces of celebrities into compromising, often pornographic, video material.

Importantly, NR as a technique, while popularised by deep fakes, has since outgrown this space, with many serious and innovative applications emerging. This technology is said to revolutionise the field of natural face technology [1], presenting a much more speedy and cost-effective rendering technique than traditional computer graphics (CG). Whereas in CG, a face needs to be painstakingly built up from scratch, NR affords inferring faces from existing video footage as training data.

In this paper, we provide a conceptual overview of this emerging field as the basis for research in Information Systems (IS) into the emerging applications of NR and their implications. We begin by contrasting NR with traditional CG face rendering.

We then present a number of typical NR examples that represent the state of the art in NR-based face rendering. Our main contribution is a framework that conceptualises and orders the field of NR face generation with a view to support both research into emerging uses of novel NR applications, as well the practice of NR application design.

Consequently, the paper addresses two questions:

*RQ1: How can neural rendering approaches be categorised into a useful taxonomy?*

*RQ2: How can the socio-technical aspects of neural rendering be researched in IS?*

## 2. Background

NR is a new and rich area for research that combines fast-moving technological advances with a lack of social science research on their implications for individuals, groups, organizations and society as a whole. Our focus is on the sociotechnical research implications of the different types of NR, rather than the underlying technical research of NR.

In this section we first briefly introduce traditional approaches for generating and animating realistic digital humans based on CG techniques that have emerged from the film, entertainment and gaming industries. We then introduce NR, from the field of artificial intelligence (AI), as an emerging powerful alternative that rivals CG for quality but with much lower costs and certain operational advantages. We finish this section with an overview of techniques using NR in the production of digital characters with high visual realism.

### 2.1. Traditional CG approaches

With traditional CG approaches to digital face generation, the target face is built up step by step using labor-intensive modelling and computer animation techniques. It typically starts with a 3D model, either built digitally or obtained through scanning a real face. The model is then textured, lit and rendered [7].

A central goal of CG face generation has always been realism. Thus algorithm advances and simulations have aimed to better emulate reality with ever more accurate algorithmic models. For example, higher model fidelity and skin response was made possible by more advanced facial scanning approaches, such as Light Stage [8]. More complex lighting and rendering approximations to the way light interacts with materials became more common place as light transport systems moved from earlier forms of rasterization to advanced sampling techniques in bi-directional path tracing, ray tracing solutions. Coupled with these progressive improvements to solving the lighting equation, computer hardware has advanced, and ever more complex equations.

For example, in the MEETMIKE project [9] the fully digital MIKE character was rendered using rasterization in the Unreal Engine (UE4). The MIKE character looked realistic and was interactive, meaning that it was controlled as a digital puppet and rendered in real-time. This project used advances in games engines, specifically the Unreal Game engine to achieve a high frame rate and realistic representation of the source actor.

As survey of traditional CG face generation is out of the scope of this paper; prior works have been published elsewhere [10]. However, we note that such CG pipelines can contain ML subsystems. For example, while digital MIKE itself is not generated with NR, it uses ML techniques to aid in both the modelling and the real-time facial decoding of the expressions of the source subject. This captured performance is then reconstructed in 3D with traditional CG approaches, running in real-time to allow interactivity with the digital character. The CG MIKE face on the left in Figure 1 is controlled by a Cubic Motion pair of infrared computer vision cameras, the output of which can be seen on in the two images on the right. As the actor moved his face and spoke, the digital version of MIKE mimicked and emulated both the actor's expressions and responses to light. The set of images on the right allows a three-dimensional reconstruction of the facial performance and an interactive human character. While MIKE uses computer vision and ML, the actual face was produced with traditional computer graphics approaches.

Figure 1. MEETMIKE (left). A UE4 traditionally rendered character controlled using AI.

A traditional CG pipeline has a linear, if not exponential, effort requirement to achieve near perfect realism, requiring both highly skilled artists and long render times [7]. Coupled with vast data size compared to the final output, traditional methods have proven particularly challenging for real-time or interactive applications. While real-time digital humans are a rapidly developing area, especially within games, full realism is still an unsolved problem.

## 2.2. Neural Rendering

Unlike the iterative improvement to modelling, texturing, lighting and rendering of a traditional CG approach, ML approaches infer digital faces using statistical techniques from large amounts of face data.

Most often a source face is sampled or used as training data to apply a new face or expression to a target face in the final video or interactive experience. Additionally, other training data is used as input, and at times additional control inputs are also applied to moderate or enhance the final re-imagined target character. While the source data is typically video, in the simplest form, a face can be rendered from a single image as shown in figure 2. Where possible, we have aimed to use the same actor throughout to illustrate the range of NR examples. This aims to provide a valid visual comparison to the various characteristics of each example. Video samples are also provided via links, as facial movement is a key component of NR. Using wildly different source actors makes any such direct comparison much more difficult.

Figure 2. Real (left) and an animated face produced from the single Jpeg (right). Using the Pinscreen App, updating in real-time on a mobile. Video: https://tinyurl.com/HICSSNR1

Using deep learning approaches, it is possible for the computer to learn both the source and target human faces and then to infer what a human face would look like in a particular position with particular lighting and a specific human expression. Related techniques can go even further and infer or invent a completely realistic new face that has never existed [7]. This broader class of approaches is now referred to as NR.

While NR techniques do not yet allow for a full range of complex animations and movements, and have restrictions in certain visual situations, the level of realism in NR has proven to be often equal to, or exceeding, that of the more traditional approaches. At the same time, due to its technical approach of using training data, NR is often characterized by reasonable comparative preparation training times, but with much faster final rendering durations [7].

Much of the published research on NR has been into the computer algorithms and big data analysis techniques for generating digital faces. The CVPR State of Neural Rendering [7] does an excellent job in summarising many of the current technical innovations. It defines NR techniques as "deep image or video generation approaches that enable explicit or implicit control of scene properties such as illumination, camera parameters, pose, geometry, appearance, and semantic structure." Key to this definition is that it goes beyond simple generative ML approaches, in that NR encapsulates controllable image generation.

At the technical implementation level NR uses deep neural networks built on the seminal work by Ian Goodfellow on GAN [11], combined with Variational Autoencoders (VAEs) from researchers such as Pushmeet Kohli [12]. Importantly, the original Deep Fake image manipulation was primarily based on VAEs and is no longer the most advanced of the NR technical approaches [13], but it remains the catch-all term incorrectly applied to most NR applications.

By being able to direct the rendering with explicit and implicit controls, the range of possible techniques is wide and encompasses novel view synthesis, semantic photo manipulation, facial and body re-enactment, relighting, and free-viewpoint video [7]. This in turn allows for a wide range of use cases from the widely discussed image manipulation (deep fakes) to the creation of photo-realistic avatars for virtual and augmented reality, virtual telepresence, digital assistants and others (see below).

## 2.3. Example techniques using NR

There are three main techniques for face generation and animation that utilize NR: 1) face swapping using existing video clips, 2) face synthesis generating fully artificial faces, and 3) hybrid solutions combining traditional CG modelling and face swapping.

**2.3.1. Face Swapping and "deep fakes".** The approach that has most commonly been used for high quality face swapping is image-to-image translation using conditional GANs [14], where training data is provided for the target face and the source face. To place a source face onto a target face the system ideally requires adequate training data of both faces in similar poses and lighting. More recent research advances have allowed for innovative approaches to real-time face swapping without extensive prior target face training data, however quality can be improved with

more training data [15].

This type of face hijacking was made possible and extensively disseminated due to the open-source publishing of several early implementations. The original 'DeepFakes/faceswap' program (github.com/deepfakes/faceswap), and the subsequent DeepFaceLab are both readily accessible, (github.com/iperov/DeepFaceLab). The original Deepfake program uses 3D morphable models. Deepfakelabs approach uses a dual Y-shaped autoencoder implementation architecture [14].

While effective, early versions often had difficulty generating high-resolution imagery due to memory limitations or failure to believably blend lighting and skin texture (see Figure 3). Other versions exhibited temporal instability, looking acceptable on still frames but flickering on moving clips [15].

Over time improvements have included better temporal contrast matching and skin blending, but insufficient training data at the appropriate resolution can still lead to images that appear gaussian or don't match the correct head orientation satisfactorily. The training data provides the training space within which the best solution can be provided. While some extrapolation is possible, the best results require training data at the same or higher resolution than the target and with appropriately similar facial orientation and lighting. For example, a side view is not easily or plausibly inferred from front looking training data. Strong daylight from the side does not allow for the best night-time facial image synthesis.

Figure 3. A face swap using DeepFaceLab (Permissions granted for this example, but the videos were not shot for this purpose).
Video: https://tinyurl.com/HICSSNR2

The DeepFaceLab example seen in Figure 3 used the Stylized Autoencoder (SAE) option to face swap a source face from an exterior 'grabbed' clip onto an interior grabbed clip. Note that the top lighting of the source face with its' heavier shadows is not reflected well in the shadowing under the chin or on the ears of the target. The Poisson blending in DeepFaceLab cannot address fully this training data mismatch. The result here is visually upsetting while not looking like traditional computer graphics.

Newer solutions improve on many aspects of earlier software programs but are still dependent on relevant training data. For example, footage shot with the correct training data in terms of lighting, resolution and camera angles can still produce inadequate results if the training data includes facial occlusion or motion blur. The training data is considered as a discontinuous set of frames and not a continuous clip, as such, training data often requires cleaning before it can be data warehoused for later use. Frames are scanned for frames where the face blurs due to rapid movement and are excluded. Algorithms are sufficiently complex to not require the target footage to be clear of any motion blur, but automatic occlusion in the target video remains an unsolved problem and often requires hand keyframe adjustment or segmentation-based image correction and compositing.

Importantly, it is also possible to produce a plausible face from as little as a single frame image but in several cases this approach supports a solution that explicitly solves the problem with the use of an underlying 3D mesh, vs, a full inferred image synthesis.

**2.3.2. Face synthesis**. Deep generative models excel at generating random realistic images using statistical big data approaches that result in a face resembling a member of the training data set [16]. For example, in such a system the output face would look like a fashion model of the same gender, age and classically defined beauty if built from a data set of classically beautiful fashion models and actresses. The faces in Figure 4 were generated from the simplest data noise, via primarily statistical data and GANs, into plausible human looking faces.

Figure 4. Two Neural Renders of synthetic people generated from data alone. Image Credit: Karras et al. 2019 NVIDIA.
Video: https://tinyurl.com/HICSSNR3

User control and interactivity play a key role in

image synthesis and manipulation, but current face synthesis approaches offer only limited user control. Their remarkable realism highlights the power of ML, but without adequate controls their usefulness is currently limited.

**2.3.3. Hybrid NR and CG solution.** A recent and powerful NR approach is the hybrid solution that combines both real-time CG and a VAE-GAN style face overlay. In the Figure 5 from Pinscreen, Digital MICHAEL is animated in real-time with the base video being fully digitally CG and the inferred synthetic face blended in to increase realism. This is distinctive as the result exhibits both photographic qualities and is completely digitally generated. A comparison in Figure 6 allows the difference in image quality to be compared between a real-time traditional character and a hybrid version.

Figure 5. A fully digital MICHAEL produced as a Hybrid. Video: https://tinyurl.com/HICSSNR4

Figure 6. A comparison of a CG only face (left) and the hybrid version using paGAN 2 (right).

Combining the face swap inside the render engine producing the base CGI character has three advantages. Firstly, the software must track the face and orientation of the person being replaced. Figure 5 is produced using the paGAN2 (photoreal avatar GAN) approach developed by Pinscreen [17]. Here the underlying head is fully digital, so the computer already knows the head position and orientation information perfectly, thus avoiding the need to approximate the head, reducing errors and improving the efficiency of the process.

Secondly, the earlier methods had issues of occlusion, which required additional processes such as segmentation to allow the face to blend behind a handheld in front of a face, or under hair that fell across the face. In the hybrid model, the paGAN2 engine is aware of the position of such objects as they are traditional 3D. It can therefore perform a face replacement without the need for special occlusion allowances that would normally be required to deal with objects hiding or obscuring the face.

Thirdly, the VAE only works well if the source and target lighting approximately match, and there is ample training data available. Obtaining relevant training data can be time consuming and difficult. Most public open-source face-swapping approaches such as Deep Fake/ Faceswap and DeepFaceLab tend to rely very heavily on VAE and often do not even use GANs. To make a more generalized, more robust, and temporarily stable solution, a number of additional steps, which include the intelligent uses of GANs at multiple stages in the process is required.

## 3. NR examples

NR is not limited to faces and people but there are a wide variety of ways it can be applied in the emerging field of digital humans. We present four typical examples of how the technology works in practice to illustrate how the inferencing is applied. We will then draw on these examples in the next stage when developing our conceptual framework.

### 3.1. Face swap impersonation (type 1)

This type of application subsumes traditional deep fakes and refers to the replacement of a face with the face of another person, where the replacement face is animated by the motions and expressions of the original (now hidden) face (see Figure 7). This allows for the impersonation of one person by another, either in pre-recorded footage or live animation.

Figure 7. NR impersonation (left), source (right).

Video: https://tinyurl.com/HICSSNR5

This application makes it appear as if someone has said or done something that they never did and is not uncommon in the entertainment industry. For example, a stunt performer could have the source actor's face incorporated into the stunt, masking the identity of the stunt performer completely. Here the inference is used to produce the new face of the actor, whereby the face shows the expressions of the source actor yet placed into the context of the target video.

### 3.2. Video Dialogue Replacement (type 2)

In Video Dialogue Replacement, the face identity remains the same; the face is not replaced with a different one but is now animated by another person whose voice now determines face movements and facial expressions.

Figure 8. Original unaltered left and VDR (right).
Video: https://tinyurl.com/HICSSNR6

This allows someone to appear to say things in another voice and language. An example would be as a replacement for dubbing or sub-titles (see Figure 8). An actor could be seen to accurately speak in another language, and the audience would also hear that other person's voice as if it was the main actors.

It is conceivable that the new voice could also be synthetically generated to simulate sounding as if it was the same actor. This additional audio processing is required for it to appear as if the new clip is actually the actor speaking and not just an audio overlay.

This differs from the first type as the target individual maintains the same face and body identity but is now controlled by the source performance.

### 3.3. De-aging or Digital Beauty work (type 3)

The third application type also maintains facial identity, but aims to make alterations to the appearance of the face, such as de-aging or digital make-up, while also maintaining identity of body and voice. A range of tools already exist for making digital alterations to captured footage. For example, on static images photo manipulation is commonplace using Adobe Photoshop. On moving footage, alterations can also be performed using visual effects tools such as The Foundry's Nuke or Adobe AfterEffects. However, these solutions are not fully automated and require human artistry. With live imagery there are automatic real-time image processing tools, such as Snapchat filters. It is worth noting that while the original 2014 Snapchat Geofilters were simple overlays, these have developed into quite complex augmented reality tools using GANs. The primary intent of these filters is often a humorous one, and they are very focused on mobile use. But filters such as the 'Burgundy Makeup' Snapchat Lens or 'Skin Smoother' Snapchat Filter aim to provide realistic digital makeup augmenting people in real-time with skin, eye and lipstick enhancements.

Advanced non-real-time digital makeup has been shown to reliably de-age actors using a suite of NR tools. In Figure 9, a person is de-aged, and the NR solution compares favourably to the ground truth of the same line of dialogue, recorded ten years earlier by the same actor. As with most similar processes, the benchmark is not identical reproduction, given the different environment lighting and cameras used, but a plausible and realistic synthetic inference based on appropriate training data.

Figure 9. Real (left) and a ten years younger Neural Render.
Video: https://tinyurl.com/HICSSNR7

### 3.4. Digital Assistants (type 4)

Digital Assistants that have a visual facial representation have thus far been achieved with traditional CG approaches. It is also possible to create a fully digital human CG entity that is driven by an AI engine. There are also a variety of virtual companions, intelligent assistants and chatbots which could drive a digital face, but are currently only text or audio, such as Amazon's Alexa and Apple's Siri.

Technologies have recently been demonstrated such as NVIDIA's Audio2Face: to produce audio-

driven AI-based facial characters [18]. This allows for a synthesized voice in real-time to drive a realistic facial animation system or a stylised Animojis system. The digital human framework, named Jarvis, harnesses NVIDIA's specialist GPU hardware for a range of digital agent enterprise solutions.

Companies such as Soul Machines in New Zealand have implemented high level digital humans as customer facing organisational representatives, such as Madera Residential's digital agent Mia, who is an autonomous, AI-driven Digital agent that is presented as a virtual leasing consultant. See a video example at https://youtu.be/6K85bUFtOSo. In the future, digital assistants with expressive faces could reasonably be expected to be synthetically inferred, in addition to the current CG approaches.

## 4. A Framework for NR Use

We have introduced NR as an emerging technology revolutionizing the field of face generation and animation in film, entertainment and gaming, but also with potential for application in business, education and health, as well as malicious uses as demonstrated by the deep fake phenomenon. Having provided example applications, we will now introduce a framework that conceptualises and orders NR applications with a view to support both research into uses of emerging and novel NR applications, as well the practice of NR application design. Our framework distinguishes 1) different source materials for NR faces, 2) how such animated faces are deployed onto different target material, 3) how faces are controlled, and 4) use scenarios with different intentions for use.

### 4.1. How a face is inferred (source material)

Much of the technical research on NR explores the algorithmic nature of face inference. While such technical details are out of scope for our purposes, it is important to know how a particular NR application is derived as it points to pertinent questions about use. For example, if faces available via photographs in the public domain can be used to generate malicious face swaps (e.g., deep fakes). This raises important copyright, ethical and responsibility questions well within the purview of the IS research agenda.

Inputs to the NR process are the source material (containing the face) and more general training data used in the process of deriving the animated face. Three types of source data can be distinguished as inputs to the inference process, (1) a still image, (2) video clip(s) or (3) a statistical model derived from large amounts of training data.

While still images usually do not result in high quality face renderings, the availability of still images makes programs easy to use, with applications in entertainment and social settings. It also raises questions about unwanted use of one's face in face swapping applications.

Most NR rendering applications today use video material as their input. This can either be pre-existing material obtained from public or third-party sources or bespoke material that is deliberately shot and lit for better quality control. Types 1,2 and 3 in section 3 all use video as source material. In type 1, the actor's face was generated from existing material in the public domain; for type 2, bespoke material (i.e., custom made) was used; whereas type 3 again utilised existing footage from the private stock of the subject.

A third way to bring into existence source faces for NR applications is digital face synthesis from statistical models, as discussed in section 2.3.2. While the generation of faces in terms of quality and realism is impressive, such faces are as of yet not suited for NR face swapping because of the lack rig control of facial features. However, as research in this field is fast progressing it is worth including fully synthetic face generation as a potential source in our framework.

Finally, we want to point out that for our purposes it is less important how the actual face inference and animation is performed. Hence, we will not examine the internal structure of VAE and GAN use, but focus on the source data they require to infer a result.

### 4.2. How the face is deployed (target material)

The face data generated via NR is then deployed onto target material, whereby the inferred face is placed onto an existing body/head performing a face swap. This can be done with different kinds of target material. We distinguish three different types, which offer increasing degrees of control.

First, a swap face can be performed on existing video footage available in the public domain, which was done in the original deep fakes, where faces of celebrities where placed onto adult video materials.

A second way to derive target footage is to create bespoke video material onto which the inferred face is placed. This allows for better control of the scene and a wider range of expressions.

A third way is to perform real-time rendering and animation of the target material, allowing for live puppeteering of the inferred face. The hybrid NR and CG example shown in figure 6 makes use of this type of target material, as the target body and face are CG generated live using a face rig worn by a subject controlling the avatar character.

### 4.3. How the face is controlled (animation)

The next consideration is what drives the NR face. Again, there are three possible ways to achieve this. The first is to have the face be driven by the facial expressions in the target video footage, regardless of whether the footage is pre-existing or created as bespoke material. Deep fakes make use of this form of animation, but so do types 1 and 3 in section 3.

A second way is to insert a third party whose facial expressions drive the face, either in a recording or in a live animation. Type 2 in section 3 makes use of this type of face control, as the facial expressions of the third party drive the NR face animation to match their voice during video dialogue replacement. A real-time puppeteering of a digital avatar using a face rig falls equally in this category.

The third category refers to the control of the face using a synthetically generated voice or other AI. In other words, it is not controlled by a real video, but by an external device. This is the most experimental form of face control as purely voice-based solutions are only just emerging from research labs [18]. This form of face control will be necessary for the creation of virtual assistants as shown in type 4 in section 3.

### 4.4. How the character is used (usage intent)

There are three main ways that characters inferred using NR will be used: (1) to impersonate another character and to make them appear to be doing or saying something they had previously not done, (2) to animate a version of oneself, or present oneself differently to one's real appearance, or (3) to generate an entirely fictitious character.

In the first case of impersonating another person, the intent could be to bring back to life a deceased actor in film, or it could be malicious, deceptively impersonating a person without their knowledge or permission. Such deception can be for comedic effect, fraud, or harm. Type 1 in section 3, which was created for educational demonstration purposes falls into this category, but so would typical deep fakes.

The second case is to animate one's own face either to create a simple digital avatar for real-time use in online contexts, such as virtual reality [9], or to present one's self differently than one actually is. This might be done deceptively but includes a range of other use contexts, such as de-aging (type 3), or replacement of audio to "speak" a different language as with video dialogue replacement in type 2.

The third category is the creation and use of an entirely new synthetic character that is not based on a real person. This could be to hide an identity when using an avatar or to put a face on a virtual assistant.

### 4.5. Summary

Our framework is intended to provide conceptual clarity and structure to the emerging field of face NR and animation. We distinguish four separate areas of design concerns that need to be taken into account when creating digital characters using NR (Table 1).

| **1. Face inference (source material)** | | |
|---|---|---|
| a. still image | b. video | c. statistical model |
| **2. Face deployment (target material)** | | |
| a. existing video | b. bespoke video | c. real-time puppeteering |
| **3. Face control (animation)** | | |
| a. target video | b. third party video | c. external device |
| **4. Character use (usage intent)** | | |
| a. character impersonation | b. self-alteration | c. synthetic character |

Table 1. NR classification framework.

When designing a digital character, the four areas can be interpreted akin to a process model, as a sequence of steps representing different design choices. At the same time, the model is also useful to analyse and classify existing examples, such as the ones presented in section 3.

Figure 10 shows the classification of the four types of NR application presented in section 3, and additionally the example of a simple deep fake. We note that deep fakes present only one 'pathway' through our framework, which signifies again the difference between the technology of NR for digital face inference and animation, and deep fakes as one particular example application.

## 5. An emerging research agenda

NR has been closely associated with the deep fake phenomenon, so that in common parlance, the term 'deep fake' is being used to denote this emerging field more generally. We advocate to view deep fakes as one form of application of NR, as NR promises to open up a range of new fields of application that are not associated with the often malicious and harmful uses of deep fakes in the public space.

In this section we outline a number of potential sociotechnical research avenues that emerge from this broader view of NR as a general enabling technology for generating and animating realistic digital faces. We

omit research on the technical aspects. At the same time however, design considerations for the creation of digital characters employing NR techniques are well within the scope of IS research.

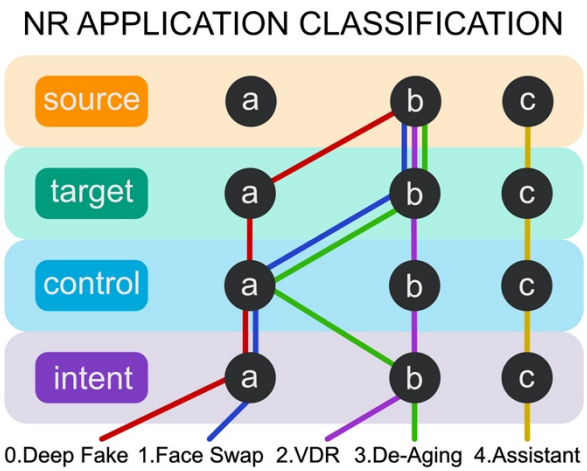


Figure 10. Classification of NR examples using the framework.

### 5.1. Design considerations

One area of research is concerned with considerations that flow directly from our framework. Initial research will be concerned with how different ways of generating and animating digital characters with NR will be perceived in terms of realism and believability. Typical research questions might be:

- What role does interactivity play? Is there a difference in perception of realism between video-animated and real-time puppeted digital characters?
- Does third-party control of digital faces in video dialogue replacement work, in particular when the source character is known to the audience?
- What is the efficacy of controlling digital characters synthetically via an AI engine? How will a discrepancy between high visual realism and low quality 'cognitive' response be perceived?

### 5.2. Areas of application and further research

Notwithstanding the rapid development of NR techniques, and the many design questions that still remain open, a range of applications for NR faces can be envision, each with unique research questions:

- What are areas of application of face-swap impersonation, beyond malicious deep fakes, and their use in entertainment?
- What will the cost impact of NR-generated characters be in entertainment and gaming? Will there be a substitution and salary effect on the use of human actors?
- Can de-aging of subjects be suitably used to support financial decision-making, e.g. when making retirement decision for one's older self?
- Can de-aging be suitably used in health or aged care to help stroke, dementia, ALS or Alzheimer patients by connecting them with younger versions of their relatives or themselves?

### 5.3. User reactions and acceptance

A further stream of research will study how users react to and accept various forms of NR-generated digital characters in different usage contexts. Typical studies might include, but are not limited to:

- Will viewers perceive and behave differently when they are aware that they are interacting with deep fakes versus when they are unaware?
- How (well) will users be able to distinguish between deep fake and real characters in various contexts, in particular advertising and political (mis)information?
- Will individuals trust (or believe) fully digital characters in different contexts, such as education, political speech, advertising or business advisory?
- Will people accept fully digital assistants when they come with a realistic natural face?

### 5.4. Ethical and legal implications

The use of video material to generate digital faces and the use in various contexts raise a myriad of ethical, moral and legal questions, many of which will only emerge in due time; some examples are as follows:

- What is the impact of NR technology in the public sphere on freedom of expression and the integrity of political discourse?
- How can the use of NR with publicly available source and target material be policed and regulated? What policy frameworks are needed?
- What are the ethical implications of digital makeup NR for online self-presentation in social media in propagating unrealistic beauty ideals?

### 5.5. Neural rendering as a research instrument

Besides research into the design, application, use and implications of NR-based digital characters, NR also offers opportunities to advance research in the social and psychological sciences. NR enables the generation of digital faces as research instruments to study human

perception of face features in ways never before possible, such as for the study of bias:

- What are the effects of nuanced variations in skin colour, age, or other facial features in an otherwise identical realistic face on human decision making or perception of character traits?
- What are the effects of different faces when a subject otherwise speaks with the same voice?
- What are the effects of different voices when used with the same face?

## 6. Conclusion

We have introduced neural rendering as a promising new technology for inferring and animating realistic digital human faces from training data. We provided examples and derived a framework that structures the various parts of the NR-based design process. Structuring this emerging field has allowed us to derive a set of initial and tentative research ideas.

Our research contributes a structured basis for comparing the application of different digital characters that are already emerging research labs and commercial entities, and which have the potential to reshape our interaction with digital characters with high human realism. Such applications come both with great potential in entertainment, business, education and health or aged care, as well as grave concerns for the consumption of information in public discourse, because of its potential for the malicious impersonation of political or otherwise public personalities in the creation of 'fake news'[13].

## 7. References


[1] Seymour, M., Riemer, K. & Kay, J. Actors, Avatars and Agents: Potentials and Implications of Natural Face Technology for the creation of Realistic Visual Presence. *J. Assoc. Inf. Syst.* (2018).

[2] Bradley, S. Even better than the real thing? Meet the virtual influencers taking over your feeds | The Drum. *The Drum* (2020). Available at: https://www.thedrum.com/news/2020/03/20/even-better-the-real-thing-meet-the-virtual-influencers-taking-over-your-feeds. (Accessed: 15th July 2020)

[3] Takahashi, D. Wave raises $30 million for superstars to stage virtual concerts | *VentureBeat* (2020). https://venturebeat.com/2020/06/10/wave-raises-30-million-for-superstars-to-stage-virtual-concerts/. (Accessed: 15th July 2020)

[4] Mitchell, V. Salesforce Ventures part of US$40 million investment into Soul Machines - CMO Australia. *CMO* Available at: https://www.cmo.com.au/article/670172/salesforce-ventures-part-us-40-million-investment-into-soul-machines/. (Accessed: 15th July 2020)

[5] Stone, L. Partnership on AI, Kodiak Robotics, Faraday Future, more take out PPP loans - AI Business. *AI Business* (2020). Available at: https://aibusiness.com/document.asp?doc_id=762264&site=aibusiness. (Accessed: 15th July 2020)

[6] Shieber, J. More investors are betting on virtual influencers like Lil Miquela |*TechCrunch* (2019). Available at: https://techcrunch.com/2019/01/14/more-investors-are-betting-on-virtual-influencers-like-lil-miquela/. (Accessed: 15th July 2020)

[7] Tewari, A. *et al.* State of the Art on Neural Rendering. (2020).

[8] Debevec, P. *et al.* Acquiring the Reflectance Field of a Human Face. *Proc. ACM Siggraph 2000* 145–156 (2000).

[9] Seymour, M., Yuan, L., Dennis, A. & Riemer, K. Facing the Artificial: Understanding Affinity, Trustworthiness, and Preference for More Realistic Digital Humans. *Proc. 53rd Hawaii Int. Conf. Syst. Sci.* **3**, 4673–4683 (2020).

[10] Leo, M. J. & Manimegalai, D. 3D modeling of human faces- A survey. *TISC 2011 - Proc. 3rd Int. Conf. Trendz Inf. Sci. Comput.* 40–45 (2011). doi:10.1109/TISC.2011.6169081

[11] Goodfellow, I. *et al.* Generative Adversarial Nets Ian. *Min. Massive Datasets Second Ed.* (2014). doi:10.1017/CBO9781139924801

[12] Kulkarni, T. D., Whitney, W. F., Kohli, P. & Tenenbaum, J. B. Deep convolutional inverse graphics network. *Adv. Neural Inf. Process. Syst.* 2015-Janua, 2539–2547 (2015).

[13] Agarwal, S. *et al.* Protecting world leaders against deep fakes. *Cvprw* 38–45 (2019).

[14] Naruniec, J., Helminger, L., Schroers, C. & R.M., W. High-Resolution Neural Face Swapping for Visual Effects. *Eurographics Symp. Render. 2020* **30**, 1–15 (2020).

[15] Real-Time "Deepfakes" – fxguide. *fxguide* (2020). Available at: https://www.fxguide.com/fxfeatured/real-time-deepfakes/. (Accessed: 15th July 2020)

[16] Karras, T., Laine, S. & Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. *CVPR 2019* (2019).

[17] Seymour, M. Pinscreen's Advanced Face AI & Neural Rendering. *fxguide* (2020). Available at: https://www.fxguide.com/fxfeatured/pinscreens-advanced-face-ai-neural-rendering/. (Accessed: 15th July 2020)

[18] Tian, G., Yuan, Y. & Liu, Y. Audio2Face: Generating speech/face animation from single audio with attention-based bidirectional LSTM networks. *Proc. - 2019 IEEE Int. Conf. Multimed. Expo Work. ICMEW 2019* 366–371 (2019). doi:10.1109/ICMEW.2019.00069