Artificial Intelligence Is No Match for Human Stupidity: Ethical Reflections on Avatars and Agents.

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

Abstract

What should our ethical concerns be in a future with 'Artificially Intelligent' agents?

The zeitgeist of AI agents often envisions a future encompassing a hyper intelligent singularity. In this worldview, AI "monsters" appear very separate from us as, abstracted, ethically ungrounded omnipotent overlords. A world of superintelligences that have moved beyond our comprehension, with no ethical restraint.

In this polemic, I explore a different future. I discount the 'Robopocalypse' initially depicted in Science Fiction. Instead, I examine how realistic digital humans do pose a very real and different ethical dilemma, as we assume intelligence based on their appearance, leading to an abdication of responsibility.

I phenomenologically explore the future of realistic digital agents and avatars, and ask: what does this human-like form say about *us*? How will we judge ourselves when the computer, *looks like* us? I argue that the singularity is unlikely and thus the primary ethical concern is not some superhuman AI intelligence, but in how we, ourselves, treat these digital humans.

Keywords : Embodied cognition, Virtual humans, Avatars and Digital Agents, Existentialism.

1 Introduction

Who are the AI actors we see in a digital future? The spirit of our age in popular culture would suggest that many people are horrified by the notion of realistic digital people, be they robots or computer simulations. The media landscape is littered with intelligent robots gone wrong, computers taking over and the notion of an eventual robopocalypse. In this narrative, our digital overloads rise up and replace our emotionally imperfect humanity with a cold, rigid, and uncaring machine logic. Is this view of the world probable, or are we more likely to see a different digital actor, one that is more a reflection of ourselves? What, or who, will we *actually* be concerned with in the future? Are we troubled by a future vision of realistic digital humans? Are they uncaring, unethical monsters or should we be more concerned with how we will treat and react to these digital humans?

I will show that while computers may appear as human, they will not have super intelligence in the foreseeable future. The threat they pose is not from ruling us unmercifully. The risk is in how we treat them; what that says about us and how easily we may be willing to empower them beyond their actual capabilities. A human lack of understanding or 'stupidity' may be of far more concern than unethical robo-rule and annihilation.

There exists in the zeitgeist the notion that scientists are driven, single minded idealists. In popular culture this is articulated in the film *Jurassic Park*, when the character Ian Malcolm states, "your scientists were so preoccupied with whether or not they could, they didn't stop to think if they should". This builds on the popular notion that researchers work without thinking in advance about moral or other consequences. Or if scientists do ask ethical questions, they continue anyway, resigned to the inevitability that science cannot be stopped, and someone will always take the next step. This 'precautionary principle' of lack of informed prudence, dates back to Mary Shelley's *Frankenstein* and is seen today in TV series such as *Westworld*. There is some evidence that this Hollywood image may not be without some basis in truth[1]. Under this view, the world will be inhabited by technologically enabled monsters, these monsters threaten our liberty, control our destiny and our very existence. This narrative would have these AI monsters as the natural and inevitable product of continued computer advances towards computer based consciousness. I do not dispute the advances of technology, just the nature of virtual human agents and actors that are likely to be in our world.

Against this backdrop I explore the phenomenology of the next generation of synthetic, digital actors. Phenomenology comes from the Greek word "to appear", and based on the work of Husseri, Hegel, Heidegger and the French Existentialists such as Merleau-Ponty.

Ethics are at the core of our belief system and govern behaviour. Ethics therefore sit at the nexus of the concepts I am exploring, digital humans, being human/having consciousness and virtuous behaviour. The work of the 18th century Immanuel Kant is relevant today in this modern setting. While Kant's view, equating humanity with consciousness is at odds with later philosophers, his work on ethics is relevant. Kantian Deontological ethics such as the Categorial Imperative, are a key part of commonly held views on ethics today. While Kantian views of philosophy are moderated or questioned by later philosophers (Riemer & Johnston, 2017), his ethical perspectives remain a useful normative or prescriptive tool. It is with this lens that we explore a future inhabited with digital humans.

I question ethical and societal implications of realistic virtual humans, whereby digital agents and avatars will stand in for us, represent us and reflect us.

I start by dispelling the myth that a super intelligent singularity is likely. To do this, I rely on the work of philosophers such as Merleau-Ponty. They disputed the Cartesian view that the mind and the brain/body can be considered anything other than one non-decomposable whole.

I then point to the natural and likely development of realistic cognitive agents. I argue that these agents will succeed not because they are logical, emotionless automatons. They will gain acceptance for exactly the opposite reason. Their affective emotional base (or rather, simulation thereof) will be what drives their acceptance and use. Rather than emotions being viewed as a flaw, the emotional engagement of such agents, combined with realistic human-like appearances, will drive wide scale adoption.

In taking this view, rather than evoking futures of bipedal gun-laden cyborg machines lumbering emotionlessly towards us, I offer an alternative. A better analogy is a 'future mirror' that reflects us, embracing emotional engagement with extensions of our humanity that encompasses artificial agents and digital avatars. These "digital humans" are simulations that not only accurately reflect our

[1] Cognitive scientist, and pioneer of AI, Marvin Minsky at MIT stated to his students, "ethicists are people who give reasons to *not* do a thing"

appearance, but also indirectly our humanity. They are not capable of actual emotion, but they more than adequately simulate emotion effectively.

I will show that our future will not have independent AI thinking monsters. While there is a popular move to extrapolate from significant gains in Machine Learning (ML)and especially computer vision to actual intelligence, current ML does not indicate a path to general intelligence. Collins argues that the public perception has exceeded reality and that the non-(social)embedded nature of computers means they lack social learning. He outlines how computers are far from being able to actually understand (Collins, 2018).

We will not have general super human intelligence in a computer agent in the foreseeable future. The science fiction narrative would have computers learning from computers in a closed loop. The most effective actual approach of this concept of AI learning from another AI engine has Generative Adversarial Networks (GAN) (Goodfellow et al., 2014). While leading researchers such as Deep Learning pioneer and Facebook AI director Yann LeCun have called innovations GANs "This, and the variations that are now being proposed is the most interesting idea in the last 10 years in ML" (Mayo, 2016), it is not a path to general intelligence. LeCun is one of the co-creators of convolutional neural networks (CNN) and yet LeCun believes society is far from any form of general intelligence. "We're very far from building truly intelligent machines. All you're seeing now — all these feats of AI like self-driving cars, interpreting medical images, beating the world champion at Go and so on — these are very narrow intelligences, and they're really trained for a particular purpose" (Vincent, 2017).

There can be no general human intelligence without some notion of consciousness. Since early last century, philosophers such as Merleau-Ponty (Merleau-Ponty & Landes, 2013), have argued that one can not have human cognition without embodiment. No humanness can exist separately, or isolated. Against this notion, we will have computer agents that are rendered to look real and that we will co-create such that we treat them as if they are their own embodied cognitive entities. While they will not have intelligence in any true sense, it might not matter. How we interact with them will reflect who we are. They will reflect us and what we bring to the experience. I suggest that many people will not stop short of imbuing them with far more intelligence, far more depth than they have. This ability to believe from the strong illusionary clues that the agent is 'smart' will only be emphasised by a general sense in modern society of technology being constantly innovative and wonderous. A sense that 'they can do anything' will feed an irrational belief in some that the realistic agents are, in fact, exhibiting genuine general intelligence. For others the cloak of intelligent human simulation will allow a greater trust in the agent's judgement than might be rationally attributed.

The central ethical question is not how we instil ethics in machines, but what our interactions say ethically about their use. In this context I argue that realistic agents reflect each of us ethically. This re-orientation is critical to how we present agents, legislate and discuss safeguards. Our current orientation is to waste our efforts on protecting ourselves from a future AI monster that is improbable, while neglecting other real ethical issues.

2 BACKGROUND

I start by defining what I mean by digital human agents, or avatars. An *agent* is a fully computer-based entity that exhibits, at least to some degree, autonomous behaviour (Seymour, Riemer, & Kay, 2018). Agents can carry out specified, recurring tasks in a self-directed way, or interact with human actors in various ways, including natural language understanding and/or dialogue. For example, *Conversational agents* are based on machine-learning and natural-language processing, generation technology, and they interact with human users "in natural language while being sensitive to their cognitive and emotional states" (Graesser, Li, & Forsyth, 2014). Examples of current conversational agents are personal digital assistants, such as Amazon's Alexa, Apple's Siri or for example Google's Duplex Assistant (Leviathan & Matias, 2018). Moreover, conversational agents that are visualized and "combine speech with non-verbal modalities for intelligible multimodal utterances" (Kopp & Wachsmuth, 2004) are commonly referred to as *embodied conversational agents* (ECA) (Cassell, 2000).

Avatars refer to visual representations of human actors. An avatar can be thought of as a digital puppet, a character that is instructed by and acts on behalf of a human actor for whom the avatar acts as a digital 'stand-in'. For the purposes of this paper I will assume it is an interactive avatar and not a pre-rendered depiction of a human, for example a digital actress or actor in a motion picture.

2.1 Advances are not an inevitable path to computer intelligence

If our fear is AI monsters with super-intelligence running amok, how possible or likely is that future?

I assume the advances in computer graphics will be able to solve the appearance of digital human agents. Already digital characters are closely resembling the visuals of their real human counterparts. In facial visual simulation, the progress of graphics technology shows no signs of slowing. In this respect, the power of computer graphics to render a human far outstrips our ability to simulate the responses and intelligence of a human.

Away from science fiction, the greatest risk of super-intelligent monsters is from the hype surrounding the discussion, not the reality. Much is written about computers taking all our jobs as they become 'intelligent' (PwC, 2017). While job losses and reallocation of jobs is a real issue, the extent of computers becoming super-intelligent is often vastly over played. Our future will not be defined by human like intelligence in a computer, and thus we should not be focused on the ethical concerns of computers becoming so powerful that they replace most of our jobs and render us irrelevant.

I will now show why it is flawed to assume one can build human intelligence, or download a mind, from the perspective of Technical and Conceptual points of view. Finally, I will look at the special case of avatars. Human Avatars introduce a new ethical problem centred around identity

2.1.1 Technically: We are not a computer

I start by questioning the base assumption that it is possible, with any extrapolation of known technology, to produce a separate digital human intelligence or download a human intelligence into a computer. I challenge the view of many leading AI experts including Ray Kurzweil, Google's Director of Engineering.

Kurzweil stated in 2017, that by 2029, computers will have human-level intelligence. He added that by 2045 we will have a 'Singularity', and that this superintelligence will abruptly trigger runaway technological growth, resulting in unfathomable changes to human civilization.

The silicon utopian vision of a general application deep AI (as opposed to the limited pattern-matching software that the term often refers to currently) is based on the extrapolation of the past exponential growth of computer power. This growth extrapolation is core to the belief that anything is possible. Authors have stated it will lead to "the extermination of the human species by godlike artificial intelligences" (Anonymous, 2018). Kurzweil's view of the same inflection point is that it will provide vast extended computer enhancement with brain implanted humans directly connected to a vast hive mind of knowledge. I will now explain how this is based on a false assumption. The problem is a not numerical one. It is wrong to assume that human intelligence is a computationally bound problem.

Our ability to explain our being, has always been influenced by the technical language of the day. In the industrial revolution, this meant we saw a person as a machine, who 'worked up a head of steam' and who's muscles were like pistons. Today, this metaphoric language is replaced with computer terms. It is not uncommon to hear of someone needing to stop 'networking' and 'download' the events of the day. A mild loss of one's 'train of thought' is now a 'glitch' or an 'overload'. The mind and physical brain can be referred to as the "software and hardware" of humanity. The senses are "inputs" and behaviours are "outputs". Children are taught in schools that neurons are "processing units" and synapses as "circuitry" to life. The very nature of cellular replication plays into this concept, with our own molecular DNA as a double helix biological software script for life. This language implies we are similar in our thinking to computers, are we are not.

The problem with the 'we are like a computer' lens of humanity is that it colours our view of what could be. In this worldview, the mind could be downloaded to a computer, should the computer be powerful enough. It is argued that we just need to match the capacity of a human brain when measured in petaflops of neurons or terabytes of memory. But this ignores the reality of how vastly unlike a computer the human experience is. Our memories are not stored in a file, sequentially in a human cerebral database (Greenfield, 2016). There is no direct equivalency of human thought and memory with CPUs, GPUs or RAM. Consciousness itself is poorly understood and there is no clear computer equivalent, or even a roadmap to attempt to research broad human level consciousness away from science fiction (Greenfield, 2016).

Kurzweil explained that current computational growth trends will soon lead to "computers having human intelligence". Leaving aside the timeline, to agree or disagree, we need to explore the nature of human intelligence (Reed & Galeon, 2017). I have disputed this technology extrapolation and now I challenge how different human intelligence and consciousness is from computers conceptually.

2.1.2 Conceptually: We are more than our minds

To match human intelligence in an agent, we would need to address consciousness. This is not a computational problem that will be solved as Moore's Law continues playing out over time. There is no level of computer power where we have any indication that a computer would magically have consciousness. It is not a processing or memory limit bound problem. It is a different kind of problem, and finding a solution is not a natural consequence of just increased computational power.

Human intelligence linked completely to human consciousness. This is not just a high enough level of intelligence, it is something different. The mind is not the same as consciousness. I can easily change my mind while conscious. I do not 'lose my mind' when asleep, or if I am rendered unconscious or I am under anaesthesia. Consciousness is 'me-ness', it is part of me, not just intelligence stored in a container.

Science has yet to solve the nature of human consciousness. We must do this before we can match it. I acknowledge that this has been recognized in research and philosophy previously.

A great contribution to the alternative non-Cartesian perspective is found in the writings from early last century of authors such as Merleau-Ponty (Merleau-Ponty & Landes, 2013). His work, The Phenomenology of Perception, written many decades ago, can provide an alternative view. Here we find justification for accepting just how improbable and unlikely a future is that suggests building human intelligence in a machine or downloading of consciousness into a computer. At the heart of Merleau-Ponty's *The Phenomenology of Perception* (Diprose & Reynolds, 2011; Merleau-Ponty & Landes, 2013) is the assertion that the body is an inseparable part. As opposed to the views of Descartes and a common contemporary view that the body is just a container for consciousness," Merleau-Ponty's view is that one is one's body, "*I am my body*". A body with "*momentum of existence*", that exceeds any biometrically objectified body. This body is understood to be the locus of "being-in-the-world*", être au monde*, which is a direct adaption of Heidegger's *in-der-Welt-sein*. He believed that the body is our general medium for experiencing the world. This negates the notion that there can be a true and full 'Embodied Cognition' for an agent, as there is no body. At best, we can just emulate it.

I believe the phenomenology of virtual human identity can be built on the foundations of Merleau-Ponty, regardless of his writings preceding modern computers (Matthews, 2010). Here we are best thought of as just one non-decomposable whole. For Merleau-Ponty the body is not merely occurring in a space or a time, but "inhabits ('habite') space and time". I do not live based on a representational conceptual map of the world in a traditional, old school AI sense; I am in the world and interacting with it, and in so doing, not just making sense of it, but the very interactions define my conscious reality.

We have no computation model for this existentialist consciousness in the world. It cannot be modelled or simulated. I might go so far as to say this is the existential crisis of the Singularitarians when predicting super-intelligence. Without actual general intelligence we have just the illusion of life, unless the digital human is driven by a real person, much like a puppet.

2.1.3 Avatars: the illusion of life and the issue of identity with digital humans.

It is easier to simulate appearance than reproduce intelligence. With various types of input and tracking, it is possible to drive a simulated human with natural movement and effectively puppet a digital human.

Such puppet style manipulations can lead to deception. As computer graphics produce increasing realism, the ability to detect such forgery is becoming more difficult. The deception can take various forms

1. Giving the illusion of life and artificial intelligence when the 'intelligence' is really human,
2. Showing an inaccurate persona e.g. presenting as a child, when an adult,
3. Presenting as someone other than yourself or impersonating another.

The ethical issues are complex. There are examples in each of these three categories, that move beyond the core ethical deception of mis-representation:

1. Presenting financial advice as if sourced from a mathematical analysis compared to a personal preference could greatly influence a potential investor. The illusion that the computer could provide an objective recommendation, free from personal opinion would influence the framing of that advice.
2. Representing one's self differently to how you really are. If someone had a facial scar from an accident, would it be ethical to use technology to present themselves without the scar, if this helped their self-confidence? Would it be appropriate to present one's self without a major

disability? Does the benefit for the individual out weight the societal issue of hiding 'perceived' human imperfections and thus avoids normalizing disfigurements?

3. The issue of impersonation creates two problems. The harm of disinformation (and its evitable breakdown of trust) and the excuse that any embarrassing filmed action, or transgression could have been faked, and thus even authentic footage is discounted in its probative value.

The discussion thus far has been regarding a digital human agent or avatar, but there is a particular case to be made for examining a physical presence, namely a robot. I will now explore the robotic artificial human.

3 Embodied Cognition: Domo Arigato, Mr. Roboto

We now turn to the issue of agents with a real-world presence as it applies to AI and robotics. This excludes industrial robots that one might find in a car factory, which are intended to lack independence and perform only repetitive prescriptive tasks.

In robotics, success in practical terms has not come from advances in super isolated intelligence. Success has come from a simpler form of emulation. Yet the success of emulation and the appearance of intelligence, has created strong populist concerns of a robopocalypse. This appears to be an unfounded andromorphic extrapolation. The robots now appear more natural and thus they are assumed to be radically more intelligent.

A form of embodied cognition has proven successful in dealing with complex real world technical issues such as mobility in robotics. Honda's ASIMO robot was built using an older traditional cognitive, computational approach, sometimes called Computational Theory of the Mind (CTM), which relies on a representational view of the world (Chemero, 2011). Honda was able to make the robot walk and climb, but even minor issues could disrupt the robot as it tried to interpret data, update its model of the world, and decide on a reactionary course of action. Compare this to the work of Boston Dynamics, which built a series of robots starting with *BigDog* using a very different approach(Thompson, 2012). The *BigDog* robot can handle complex terrain and later versions even managed to recover from violent knocks or walking on slippery ice. Boston Dynamics decided that a computational strategy would be too slow and so opted for a Dynamic System. They built a robot with springy legs and joints that mimic those seen in animal quadrupeds. *BigDog* has a comparatively small computer and its success was not due to it having access to a more powerful computer than ASIMO. "The specific movements he produces at any given time emerge from the interaction between his moving legs, the surface he's on and any other forces acting on him. If you knock *BigDog*, he doesn't need to re-compute his behavior; he simply responds to the new force and the details are left up to his anatomy"(Thompson, 2012). Dynamic Systems Theory (DST) is a broad approach imported from the physical sciences and used in cognitive science as an alternative to the computational and information-processing approach. DST is best described, according to Chemero as "*complex, non-linear, self-organizing and emergent* and whereby cognition develops over real-time as a probable description of many possible alternatives instead of linear-assembly-of-symbolic-processes". It works by assuming real time embodiment and not a dualist CMT model.

Beyond acknowledging the superior approach, the reaction to *BigDog* and its successors has been particularly informative. The DST solution produces a very natural looking physical solution to movement, especially if the robot is interfered with. This has resulted in the robots being described as "very biological" (Thompson, 2012). A wide scale reaction to videos of these tests can be characterized by the quote that the company makes "autonomous pack mules like *BigDog* that inspire fear with every step". This narrative directly plays into our monster view of the future. Testing of these robots recovering from humans pushing them, prompted comments of fear of retribution when the computers 'rise up'. Comments such as those made by the BBC, that some version of *BigDog* might be used in the future for "terrorizing humble civilians ... - especially if we go around treating them as badly as this!" The fear appears to be connected to how natural and biological the robot's interactions are in the world. One can only speculate how dramatically people would react to this style of robot if they had realistic human faces.

Our background affects our view in this context. Depending on your point of view, the *BigDog* is close to being a monster, or it is an extension of my affordance with regards mobility. Both are reflections of who we are. What *BigDog* is, reflects me. I co-create what it is. My worldview for it as either a monster or an invaluable step towards Army troop mobility. *BigDog* has the illusion of 'life', which one can start to anthropomorphize as it is moves more realistically in the world. It is not profoundly more intelligent, but many people may believe it is, due to its natural simulation of movement.

This Robotic example highlights a bigger issue. The isolated success of a particular AI or Machine Learning application is often seen as indictive of a greater intelligence. But as Oxford Professor Floridi wrote, “the truth is that climbing on top of a tree is not a small step towards the Moon; it is the end of the journey” (Floridi, 2016). He goes on to point out that isolated advances will continue to be made and in their isolated areas, we *are* going to see increasingly smart machines able to perform more tasks that we currently perform ourselves. Such isolated success such as face recognition systems seems to the uninitiated, to represent great intelligence, but it is a narrow domain specific success and cannot be used to extrapolate other more general problem solving. The advances especially in deep learning, as a specialist subset of Machine Learning have provided great advances. It has especially done this in the area of digital humans. Facial reconstruction, tracking, emulation and rendering have all seen Machine Learning advances, but just as with *BigDog*, a more believable agent or avatar does not denote a vastly more naturally intelligent actor.

3.1 The HCI development of realistic cognitive agents

Until now I have not made the case that digital agents will come to naturally encompass biologically realistic appearances. I will now state that this is not only very probable, but desirable. I will explore how an agent with an expressive photorealistic face may change the nature of Human-Computer Interaction (HCI).

A central part of research into digital humans has been to produce a photo-real digital double, or duplicate image. Given how important faces are to communication, the goal is a direct reflection of one's facial appearance that would be imperceptible from the human original.

Far from being ‘user interface tools’ that recede into the background, these new human-looking tools will focus attention. Such agents will aim to communicate verbally and non-verbally, with emotionally laden facial expressions. They will seek the user's attention, engagement and trust. They will face us and look at us, literately.

Such a future is reasonable to imagine as the four largest American corporations by market capitalization are all investing heavily in such cognitive assistant technology (Statista, 2015). Apple's Siri, Google's Assistant, Microsoft’s Cortana and Amazon's Alexa are all currently faceless, but it would seem reasonable that a joyous countenance on these AI bots is a logical next step. Not to be outdone, the sixth largest company, Facebook, has also embarked on an ambitious program of avatars that mimic us directly (Seymour, 2016).

While some areas of the popular Press may voice anxiety and luddite resistance (Kletzer Lori, n.d.), research has also shown an overwhelming majority of white-collar workers are genuinely excited and optimistic about what technology can do for them. In an industry study, the majority of the 4,000-plus surveyed office workers believed AI technology will make them more productive and help them (Abramovich, 2017). This suggests there is momentum for strong potential adoption in organizations.

Nicolas Negroponte postulated several decades ago, "*in a foreign land, one uses every means possible to transmit intentions and read all the signals to derive even minimal levels of understanding. Think of a computer as being in such a foreign land... ours*" (Negroponte, 1995). In this foreign land, the notion is that a computer should be made more aware of human means of communication, and the face is one of the most expressive mediums for nonverbal communication. "*Interface is not just about the look and feel of a computer. It is about the creation of personality, the design of intelligence, and building machines that can recognize human expression"*. The challenge, he stated, was to make *"computers that know you, learn about your needs, and understand verbal and nonverbal languages...."* Negroponte felt that the humans in the HCI equation were bearing the "burden of interaction". Computer power has grown steadily, and exponentially under Moore's Law. In the reasonable future, we could have both the innovation and available computer power to move the burden from "*the shoulders of the human party"* to the computer.

The question we explore is not what is required to create a digital human, but what opportunities does that bring in relationships and what ethical questions about our humanity would such a development reveal? Affective computing explores two-way emotional communication as a new medium for HCI. But to paraphrase Negroponte, the medium is not the message. These new user interfaces are not about interfaces - it is about living with the new possibilities.

3.1.1 Appearances matter

George Zarkadakis claims that a computer agent or avatar that behaves emotionally intelligent will be "considered intelligent even if it is a philosophical zombie (Zarkadakis, 2016)". People have anthropomorphized their 'dumb' devices without any engaging human appearance (Reeves & Nass, 1998). With an expressive human face, the natural tendency will be to react to the computer agent with emotional sincerity.

BabyX is an unscripted simulation of a child pioneered by Mark Sagar. BabyX is both a simulation of neurochemistry and the appearance of a real baby (Sagar, Seymour, & Henderson, 2016). The simulation is highly advanced and uses various forms of audio and visual input to emulate a child on the screen. It is one of the most advanced such attempts. 'She' is a very good simulation of a child's interactions, neuro-chemical responses, and emotional emulation based on input from someone standing in front of her screen's input devices.

Observing real interactions of people with a live BabyX demonstration, audience members anticipate and seek subtle and emotional responses from BabyX as the engagement happens. People imbue the agent with human traits based on their personal pre-learnt real-world interactions with infants.

During these technical demonstrations ("DISRUPT.SYDNEY 2014 Short Talk 4 by Mark Sagar," n.d.),, Sagar and his team have modelled the BabyX receiving a pain response from the tap of a key. In multiple sessions audience members have urged the demonstrator to stop so as to not 'hurt the baby'.

This, after at least fifteen minutes of explanation by Sagar of the underlying simulation including visualizations of the faceless brain model of BabyX. There is no sense that the audience rationally thinks the baby is real, but they immediately emotionally react as if she is, as Sagar moves to press the P key for Pain. The session was part of an academic conference presentation, the pain response a valid part of a comprehensive brain model, yet the audience's response was akin to Sagar actually hurting a real child.

This emotional response is very real, and is often followed by laughter, interpreted as a stress release and an acknowledgement of the surprisingly immediate visceral reaction. Interestingly, Sagar, then naturally laughs, as laughter is infectious and then BabyX laughs as "Dad" (Sagar) is laughing.

The implication of the observed behavior is that it demonstrates the agent, is part of the holistic environment and the co-defined dynamic interaction. What the simulation 'is', is co-defined by Sagar's research, along with the audience. The spontaneous emotional responses indicate that the audience's worldview very much co-defines the agent. This raises valid ethical questions about our perception of realistic agents, and the way simulated humanity will invoke deeply embedded responses.

Human interaction remains mostly about building and sustaining social bonds. I find that our casual language implies something that is not accurate. Picard, in her work defining Affective Computing, points out that people are known to comment that, if we could be 'less emotional, more logical' we would function better in society, but the opposite is true (Picard, 2003). Affective computing and computers emulating emotion is an important aspect to be explored for future digital human agents.

In everyday life, rather than being a limiting factor, emotions moderate and facilitate our behavior. Someone 'being very emotional' can be stated as an assumed negative, but without emotional responses people would not be able to function in society. I can find no scientific proof that removing all our emotions would make us better people or even more effective in society, furthermore there is an inherent assumption that human biases and emotion are faults, but I would argue they are a rich part of our humanity. Future digital agents will be an emotional phenomenon. With virtual humans, this will appear as an emotionally aware extension of us and by extension, a reflection of us. They will not be cold and logical. The agents will appear to be emotionally engaging. Another real-world example illustrates this point.

Encoding Agents with the appearance of emotion will not make the agent human, but it will increase emotional authenticity. It will do this in the same way Google's Duplex audio-only agents incorporated human speech 'disfluencies' such as "hmms and uhs" (Leviathan & Matias, 2018) to seem natural. In Google's tests, an agent called various restaurants to book a dinner. The agent does not understand what makes a meal enjoyable or anything about the nature of food, it just appears to be human and engages the staff of the restaurant to talk to with it in the same way they would a normal human.

What is critical to our discussion is that I judge the rudeness or helpfulness of the restaurant phone staff independently of the fact they are talking to an agent. If the staff had been rude and dismissive of the Google agent's enquiry, I would judge them poorly, even knowing myself they were not offending anyone, (any person). I would also personally judge them poorly if I knew they were talking to a

conversational agent. It would inform my opinion of them if they swore and insulted the agent. This may not be another person's reaction, but then as I have shown, what that agent is co-defined by is my worldview *and* the agent's interaction.

It is important to go one step further and acknowledge that the illusion of humanity of the agent will affect my response to it. I may only feel this way as the voice appeared so human, at least in part due to its 'hmms and uh's'.

4 Ethics of our Agents and Avatars

4.1 Ethical choices

While a computer agent may appear independent, how an agent acts reflects on us. The fact that the agent is not human does not change this. For example, I am under no illusion that my pet dog is not human, yet my behaviour towards my beloved Labrador reflects greatly upon my humanity. While legally and conceptually he is very different from a person, any cruelty to my animal will draw a visceral rejection from most people in society.

The perceived interaction of an agent will be affected by choices in appearance. The choice of an agent's age, sex, and race can greatly affect perceptions. Would a sexist, demeaning attitude to an agent, that looks like a younger female of a minority group be considered vastly less of an indictment because the agent can be described as not 'a real person'? Would such behaviour not greatly influence the views of fellow workers or employers? If I chose my mother as my assistant, would this not send a message about who I am? It would inform someone very differently than if I selected instead an archetypal English butler, or a provocatively dressed anime school girl? What we choose to represent us, who we choose to 'serve' us, reflects a lot about us as 'users'.

As mentioned, with the advent of advanced digital makeup, the question arises about the desire or otherwise of hiding facial flaws when choosing an online realistic avatar. Does this advance the individual or reinforce stereotypes of perfection? Would this eventually resign the less than perfect amongst us to primarily remote communication. Could this make people who are self-consciously different further isolated?

This digital misrepresentation extends to full impersonation. *Deep Fakes* and other technology already allow believable face replacement. It cannot be long before this can be achieved in real-time to a level that is not identifiable by untrained observers.

Similar facial recognition AI algorithms that can aid in understanding can be used to monitor. Ironically, in China, agents designed to read human traits from people's faces are being used to monitor school children's attentiveness in class (Connor, 2018). The attentiveness monitoring of pupils by three cameras connected to a facial recognition engine achieves the opposite effect of de-humanising the students. The performativity of the monitoring means the children must increasingly come to act in mechanistic ways to comply with the way in which the machine reads and renders humans. This monitoring chips away at the children's very individuality.

The agency one gives to an agent as we resign decisions and cognitive functions to the agent also reflects on our character. If we defer even modest decisions to the wisdom of the AI agent, does this hybrid agency not characterize us and open us to the biases of the deep learning data sets of the AI Agent. In relinquishing our agency, we inherit the biases of the computer. Deep learning approaches solved inside their training space, the abdication of cognitive effort to an agent, is an abdication of one's own personal ethical sense of fairness and what constitutes unbiased reasoning. We are trading not just agency, but privacy with the corporate provider of the agent. The transaction is often predicated on selling some part of our privacy for the luxury of reduced cognitive load. Even one's willingness to trade personal liberty for ease of use, makes a strong statement. It reflects our values and our self-worth.

5 Conclusion

The ethical questions of AI and digital humans are not about a remote future Singularity that would grow to independently command our fate. Rather it is a question of the personal responsibility and level of control that one could have as these new digital opportunities provide a mirror to our personal values.

I have argued that the important and relevant ethical issue facing our future with technology is not to attempt to build into sentient robots some Isaac Asimov style "Three Laws of Robotics" or some other kind of "safe word" or "safety switch". I have shown in this polemic that the issue is rather our ethical

values reflected in our own interactions with and co-constituency of the digital agent. The robopocalypse is science fiction and not worth our serious consideration.

It is important to turn the debate from futuristic scenarios that are highly improbable to a discussion of imminent implications of living with digital agents which are highly believable, expressive emotional parts of our lives, while also not attributing too much significance to their apparent intelligence.

While the timeline is very much open to speculation, the resources and benefits of complex realistic digital agents is tangible. What these Agents will become is something I have shown to be co-defined by our interactions with them in the world. For us to examine the ethics of digital humans and simulations of cognition, we need to understand what it is that we are thinking about and how we think about the ethics of that situation.

6 References

Abramovich, G. (2017). Study: Office Workers Want AI At Work. Retrieved May 27, 2018, from https://www.cmo.com/features/articles/2017/6/28/study-office-workers-want-ai-at-work-tlp.html#gs.y7AFV5g

Anonymous. (2018). What is the Singularity? - The Economist explains. Retrieved May 29, 2018, from https://www.economist.com/the-economist-explains/2018/05/14/what-is-the-singularity

Cassell, J. (2000). Embodied Conversational Interface agent. *Communications of the ACM*, *43*(4), 70–78.

Chemero, A. (2011). *Radical Embodied Cognitive Science*. MIT Press.

Collins, H. (2018). *Artifictional Intelligence* (first edit). Polity Press.

Connor, N. (2018). Chinese school installs facial recognition technology to monitor students | afr.com. Retrieved May 22, 2018, from http://www.afr.com/news/world/asia/chinese-school-installs-facial-recognition-technology-to-monitor-students-20180517-h1070p?logout=true

Diprose, R., & Reynolds, J. (2011). *Merleau-Ponty: Key concepts*. *Merleau-Ponty: Key Concepts*. https://doi.org/10.1017/UPO9781844654024

DISRUPT.SYDNEY 2014 Short Talk 4 by Mark Sagar. (n.d.). Retrieved January 9, 2016, from https://www.youtube.com/watch?v=pFjlGiqGrJc

Floridi, L. (2016). Humans have nothing to fear from intelligent machines | Financial Times. Retrieved August 6, 2018, from https://www.ft.com/content/9a6b6536-c372-11e5-808f-8231cd71622e

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative Adversarial Nets Ian. *Mining of Massive Datasets: Second Edition*. https://doi.org/10.1017/CBO9781139924801

Graesser, A. C., Li, H., & Forsyth, C. (2014). Learning by Communicating in Natural Language With Conversational Agents. *Current Directions in Psychological Science*, *23*(5), 374–380. https://doi.org/10.1177/0963721414540680

Greenfield, S. (2016). *A Day in the Life of the Brain*. Penguin Random House.

Kletzer Lori. (n.d.). The Question with AI Isn't Whether We'll Lose Our Jobs — It's How Much We'll Get Paid. Retrieved May 27, 2018, from https://hbr.org/2018/01/the-question-with-ai-isnt-whether-well-lose-our-jobs-its-how-much-well-get-paid

Kopp, S., & Wachsmuth, I. (2004). Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds*. https://doi.org/10.1002/cav.6

Leviathan, Y., & Matias, Y. (2018). Google AI Blog: Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone. Retrieved May 24, 2018, from https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

Matthews, E. (2010). Maurice Merleau-Ponty: Phenomenology of perception. In *Central Works of Philosophy Volume 4: The Twentieth Century: Moore to Popper* (pp. 177–194). https://doi.org/10.1017/UPO9781844653614.011

Mayo, M. (2016). Yann LeCun Quora Session Overview. Retrieved October 21, 2018, from https://www.kdnuggets.com/2016/08/yann-lecun-quora-session.html

Merleau-Ponty, M., & Landes, D. A. (2013). *Phenomenology of perception*. *Phenomenology of Perception*. https://doi.org/10.4324/9780203720714

Negroponte, N. (1995). *Being Digital*. *Media*. https://doi.org/10.1016/0169-7552(94)90144-9

Picard, R. W. (2003). Affective computing: challenges. *International Journal of Human-Computer Studies*, *59*(1–2), 55–64. https://doi.org/10.1016/S1071-5819(03)00052-1

PwC. (2017). How will automation impact jobs: PwC UK. Retrieved June 1, 2018, from https://www.pwc.co.uk/services/economics-policy/insights/the-impact-of-automation-on-jobs.html

Reed, C., & Galeon, D. (2017). Kurzweil Claims That the Singularity Will Happen by 2045. Retrieved May 27, 2018, from https://futurism.com/kurzweil-claims-that-the-singularity-will-happen-by-2045/

Reeves, B., & Nass, C. (1998). *How People Treat Computers, Television, and New Media Like Real People and Places*. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places* (Vol. 34). Cambridge University Press. https://doi.org/10.1109/MSPEC.1997.576013

Riemer, K., & Johnston, R. B. (2017). Clarifying Ontological Inseparability with Heidegger's Analysis of Equipment. *MIS Quarterly*, *41*(4), 1059–1081.

Sagar, M., Seymour, M., & Henderson, A. (2016). Creating connection with autonomous facial animation. *Communications of the ACM*, *59*(12), 82–91. https://doi.org/10.1145/2950041

Seymour, M. (2016). VR on the Lot | fxguide. Retrieved May 21, 2018, from https://www.fxguide.com/featured/vr-on-the-lot/

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

Statista. (2015). Top companies in the world by market value 2014. Retrieved May 14, 2018, from https://www.statista.com/statistics/263264/top-companies-in-the-world-by-market-value/

Thompson, J. (2012). Embodied Cognition: What It Is & Why It's Important. Retrieved May 26, 2018, from https://www.psychologytoday.com/us/blog/beyond-words/201202/embodied-cognition-what-it-is-why-its-important

Vincent, J. (2017). Facebook's AI chief says the public doesn't know how dumb AI really is - The Verge. Retrieved October 21, 2018, from https://www.theverge.com/2017/10/26/16552056/a-intelligence-terminator-facebook-yann-lecun-interview

Zarkadakis, G. (2016). In Our Own Image: Savior or Destroyer? The History and Future of Artificial Intelligence. *Publishers Weekly*, *263*(4), 195. Retrieved from http://search.ebscohost.com.leo.lib.unomaha.edu/login.aspx?direct=true&db=bth&AN=112698238&site=ehost-live&scope=site