Art-ificial: The Philosophy of AI Art

Alice C Helliwell

A thesis submitted to the University of Kent in fulfilment of the requirements for the Degree of Doctor of Philosophy in History and Philosophy of Art

Word count: 96,455

December 2022

University of Kent School of Arts

Acknowledgments

This PhD project was made possible through the generous funding of the University of Kent Vice Chancellors Scholarship.

First, I would like to offer profound thanks to my supervisors. My primary supervisor Dr Michael Newall has been a mentor throughout my Master's and PhD, and continued to offer support and guidance after he left the University of Kent and the UK during the pandemic. I am immensely grateful to have had him as my supervisor; his insight, patience, and enthusiasm have been an inspiration. Without his encouragement, I may not have continued on to a PhD at the University of Kent, and I hope this project makes him proud. Dr Hans Maes has been a source of support, encouragement, and philosophical rigour throughout my PhD project, but particularly after Michael's departure. Hans has been a kind, caring, and steadying force and was essential in helping me to complete this project. Thanks go also to Dr Jonathan Friday, who stepped in as second supervisor after a change in my supervisory team and has offered invaluable advice and insight.

I must thank the Aesthetics Research Centre and its members, who have been a network of support, camaraderie and philosophical interest at the University of Kent.

Thanks too go to my colleagues at Northeastern University - London. I undertook a teaching position at NU-L in my fourth year of study, which then became a permanent research post. Despite the challenges that came with working full-time whilst completing a PhD, my work at NU-L has been incredibly rewarding. I have had the opportunity to learn more about philosophy and artificial intelligence from a wide range of colleagues, and I have had the pleasure of teaching many bright students, whose thoughtful conversations on topics in AI Ethics and Aesthetics have certainly helped to develop my thoughts on a variety of philosophical issues. My colleagues from across the University have cheered me across the finish line of this project.

Many more colleagues and friends deserve acknowledgement in this project, but particularly my thanks go to Aurélie Debaene, Dieter Declercq, and Eleen Deprez.

I must also thank those dearest to me. My mum Emma Helliwell, who is always there for me, no matter what, and my dad Peter Helliwell who has always supported me through my endeavours. Finally, my partner David Brown, without whom I am not sure I would have managed to keep going on this project through COVID and working full-time. He has supported me through it all, and I'd truly be lost without him.

Abstract

This thesis aims to contribute to a novel area of philosophical work: the philosophy of AI art. AI art is proliferating online and increasingly in the world of art. The growing presence of works made by (or with) artificial intelligence has led to a clamour of questions such as 'Is AI art really art?' and 'can AI be truly creative?'. As yet, these questions have barely been tackled in the philosophical literature, especially in aesthetics. This thesis aims to address this gap.

This thesis starts by establishing what we mean by 'AI art' by examining examples of AI works and the technological underpinnings of these systems. Existing work on the topic of AI art is explicated. In particular, Mark Coeckelbergh's three questions on AI art scaffold the first three chapters of the thesis: 'can machines create *art*?', 'can machines *create* art?' and 'can *machines* create art?'

Chapter 1 aims to answer the question of whether AI can make art through the evaluation of AI works against three definitions of art: the institutional account, the historical account, and the cluster account. Chapter 2 focusses on the question of whether AI can be creative. Three accounts of creativity are utilised: a Darwinian theory of creativity, Margaret Boden's account of creativity, and Berys Gaut's agential account of creativity. It is argued that some AI systems can meet the requirements of each of these, aside from Gaut's necessary criterion of agency. After chapters 1 and 2, questions about the limitations of AI systems in meeting the requirements of different accounts of art and creativity remain; chapter 3 aims to address some of these. The possibility of AI (extended) mind is investigated, followed by AI embodiment. Finally, an argument for the possibility of AI agency is put forward. This minimal account of agency allows for the possibility of AI creativity under the agential account.

The latter part of this thesis begins with chapter 4, which examines the possibility that AI systems will not share aesthetic or artistic values with humans, and whether this is cause for concern. Finally, Chapter 5 examines two qualities of AI images: weirdness and convincingness, showing that AI art can offer interesting aesthetic qualities worthy of investigation. Through this thesis, I put forward a first step in developing a philosophy of AI art.

Contents

List of Figures

Images are reproduced under fair dealing.

Introduction

"Art is dead, dude. It's over. A.I. won. Humans lost."

These were the words of Jason M Allen, winner of the 2022 Colorado State Fair's annual art competition in the category of "emerging digital artists" (Roose, 2022). Whilst the fair is certainly well-attended in the state, it seems odd that anyone outside of Colorado would become aware of it. So why have Allen's words reverberated around the artworld? Well, much to the chagrin of other entrants to the competition, Allen's work was not by his own hand; it was an image generated by an AI. Twitter was aflutter with angry responses, but one tweet in particular led the charge: "Someone entered an art competition with an AI-generated piece and won the first prize. Yeah that's pretty fucking shitty." (Jumalon, 2022). AI art, as *The New York Times* points out, is not new (Roose, 2022). Before 2015, AI art had a marginal presence in the artworld; however, since then innovations in AI have led to the proliferation of art made by AI systems. This rapid growth shows no signs of slowing down. The era of AI art in which we find ourselves has brought about many questions regarding the nature of these new works. "Can AI create true art?" asks S*cientific American* (Weiner, 2018) whilst one *Towards Data Science* contributor leads with "AI agents are not artists" (Chambers, 2020). "Artificial intelligence can now make art." declares *Vox* with the added reassurance: "Artists, don't panic." (Samuel, 2019). "The Robot Artists Aren't Coming" an opinion piece in *The New York Times* claims (Elgammal, 2020), whilst *Wired* pronounces that "When AI Makes Art, Humans Supply the Creative Spark" (Knight, 2022). Questions such as 'is AI art really art?' and 'can AI be creative?' are clearly being debated in the public sphere. These are exactly the kinds of questions that philosophy is equipped to deal with, yet there is still little work in this area. With this thesis, I aim to contribute to this pressing philosophical discussion. Before we can examine these questions about the nature of AI art, we need to delineate exactly what it is we are talking about when we talk about 'AI art'.

WHAT IS AI ART?

My general definition of AI art is as follows: *AI art refers to works in the domain of the arts, produced by (or with) artificial intelligence systems.* I address the question of whether these works can properly be called art in Chapter 1. After that point, I will work on the assumption that AI works can indeed be art. Nevertheless, one need not accept my conclusions in Chapter 1 and can instead consider my use of 'AI art' to just be a pragmatic shorthand for the kinds of AI-made works in the art sphere. Some of the AI I discuss are designed with the purpose of producing art, whereas others are not (solely) used for art making (but could be). For instance, an image generation AI might be used to produce stock photos for corporate presentations, but it might also be used with the goal of producing artworks. Throughout much of this thesis, I use 'AI art' to mean works made *by* Artificial Intelligence systems. In these cases, the AI is relatively autonomous in determining the features of the work. I distinguish this from works made *with* Artificial Intelligence, where there is a higher level of human involvement in the determination of features of the work. Of course, this distinction is not black-and-white; there are various possible levels of autonomy for AI systems in producing works of art. Let us turn to a brief discussion of these levels.

Autonomy is not easily defined, particularly when it comes to machines. This is further complicated by the difficulty in determining what we mean by autonomy; philosophers go beyond the relatively minimal definition used in computer science and engineering. I discuss the issue of autonomy at length in Chapter 2. For now, however, I will utilise Haselager's definition:

> Autonomy is deeply connected to the capacity to act on one's own behalf and make one's own choices, instead of following goals set by other agents. The importance of being able to set one's goals is also part and parcel of the common sense interpretation of autonomy. (Haselager 2005 p. 519)

Below I put forward a working taxonomy of the possible levels of autonomy an AI system may have in producing a work. We will start with lower levels of AI autonomy and move to higher levels.

1) Tool (AI assistance)

There is little to no creative input on the part of the AI (suggestion or prediction algorithms may provide small amounts of input). The human user is the key creative agent. The AI system makes no decision unsupervised. For example, some features of Adobe Photoshop are 'AI-powered', such as automated object selections (P Clark, 2020).

2) Manipulation (partial autonomy with considerable human oversight)

Heavy manipulation of the algorithm by a user or programmer, which might otherwise produce images without close supervision. This would include very close control of the training set (the set of images used to train the AI system) such as through inputting one image repeatedly, or manipulation of the weightings of a trained network. Many artists who work closely with Generative Adversarial Networks (GANs) insist that they maintain a level of influence over the images through this kind of manipulation of the AI.

3) Collaboration

A collaboration in close to equal or equal parts. Just like the previous level of autonomy, the human user has some role in determining the features of the ultimate output. For example, the user may tinker with the training set or thresholds, select images from the latent space, or manipulate images post-output from the AI (either computationally or physically). However, if we are to describe the final work as a collaboration, then the level of manipulation involved is much lower. There is a conceptual difficulty with talking about collaborative AI and autonomy. Autonomy is in part predicated on independence from a human or setting its own goals, but in collaborative working one is not wholly independent and may share goals with others. Examples include *Cloudpainter* (which I will explain in greater detail later in this chapter).

4) Facilitation (high-level of autonomy)

The AI is facilitated by a human programmer, but the majority of the process of producing a final work occurs within the AI system itself. This could include use of trained networks, where the training occurs once at the start of the process and the training set is not heavily limited as might occur in the cases of (2) (manipulation) and (3) (collaboration). The system in some way is setting its

own sub-goals (for example, by determining the specific criteria of evaluation). The system is not deciding on the goal of creating art, but its outputs cannot be straightforwardly attributed to the hand of a human (unlike with (2) above). AICAN could arguably be considered an example of reaching this level of autonomy (I will expand upon this later also).

5) Complete autonomy

A system that is able to set its own high-level goals (such as deciding to make art in the first place). Here the key creative agent would be the AI itself. No current systems meet this level of autonomy, and it is unlikely to occur without general artificial intelligence (AGI) or multi-domain AI.

In this thesis, I am chiefly interested in what AI systems can do, not how humans can use them. As such, I will mainly be focusing on level (4) (Facilitation) and the possibility of (5) (Complete Autonomy). As many AI systems that are used to produced works with a high level of autonomy are also manipulated by artists, I will at times discuss examples at the level of (2) (Manipulation). In this taxonomy I have begun to refer to some examples of AI systems that are used for artistic purposes. As will become clear in the subsequent chapters, in order to make arguments about the nature of AI art and creativity, we will need to know how exactly these systems work. Whilst some may be satisfied with looking only at the output of AI to determine if it is art and if the AI has been creative, this will prove insufficient for applying many definitions of art and creativity. The next step we must take then in order to understand the nature of AI art is to examine how these systems actually work. To take this step forward though, we must take another step back and examine how Artificial Intelligence itself and machine learning work.

WHAT IS AI?

We can define AI as the science and engineering of making intelligent computer programmes and machines. What do I mean by 'intelligent' here? Russell and Norvig (2010, p. 2) highlight four approaches to defining Artificial Intelligence, each of which varies in terms of how we would determine what qualifies as intelligence:

1. Thinking humanly
2. Thinking rationally
3. Acting humanly
4. Acting rationally

Those developing Artificial Intelligence typically follow one of these four approaches. Let us consider each of these in turn.

Thinking humanly

The project of AI may be to replicate the human ability to think; however, this is not to say that the aim of AI is to reproduce a human mind. Computer scientists do not typically seek to replicate the structures of the human mind, but they are trying to create functional analogues that can process information in a way that would typically require human (or biological) brainpower such as "reasoning, making decisions, learning from mistakes, communicating, solving problems, and moving around the physical world." (The Alan Turing Institute).

Thinking rationally

Others are not concerned at all with how *humans* think. Instead, the focus is on how to make machines think *rationally*, often through following rules of logic (Russell and Norvig 2010, p. 4). Humans are by no means perfectly rational agents so while AI may be developed to have human-level abilities, researchers may instead focus on optimising AI to go beyond the abilities of humans.

Acting humanly

Given the difficulty in knowing whether an AI is thinking in any particular way, some may focus on *behaviour* instead; behaviour is of course far easier for us to access through observation. Behavioural similarity to humans is the basis of the Turing Test, whereby a computer 'passes' by being mistaken for a human interlocutor.

Acting rationally

Of course, humans make mistakes. To behave convincingly human would necessitate AI to also make mistakes. Some may therefore prefer to seek *rationality* in the behaviour of AI. An AI that acts rationally can be understood to act to achieve the best outcome in a given situation (Russell & Norvig, 2010, p. 4).

What concerns me here is not which systems should count as *intelligent*, but which systems count as *creative* or capable of making art. So, for the purposes of this thesis, we can remain agnostic on which of these approaches is the correct way to define AI.

Just like the variety of different approaches to defining AI, there are a vast number of methods and techniques directed at achieving intelligence. The majority of the AI systems discussed in this thesis employ 'machine learning' in some capacity. Let us turn to examine what this means.

MACHINE LEARNING

Machine learning is an arm of computer science and AI that uses data and algorithms together to imitate human learning. The field of machine learning aims to create systems that gradually improve in accuracy following training (IBM Cloud Education, 2020b). There are four main approaches to machine learning:

1) **Supervised learning** involves training an algorithm on predetermined inputs and corresponding outputs. It requires a labelled training set (Sarker, 2021).

2) **Unsupervised learning** is more data driven than supervised learning and does not require human interference. Datasets are unlabelled and are analysed without direct oversight (Sarker, 2021).

3) **Semi-supervised learning** is a hybrid approach of supervised and unsupervised learning.

4) **Reinforcement learning** is an environment-driven approach, which enables automatic evaluation of behaviour to improve efficiency. Reinforcement learning utilises penalties or rewards: the system seeks to maximise reward (or, alternatively, minimise penalty) (Sarker, 2021).

Although these represent the four dominant approaches to machine learning, there are some systems in this thesis that do not sit neatly within one of these four approaches. For example, GPT-3 (a text generation model) uses 'self-supervised' learning. Self-supervised systems can utilise self-defined 'pseudolabels' to supervise learning in place of pre-defined labels (Jaiswal et al., 2021).

How is machine learning implemented? *Neural networks* are the architecture of machine learning. Neural networks (sometimes called artificial neural networks or ANNs) aim to replicate (in a reduced form) the structure of the neural systems found in the brains of humans and animals (Walczak & Cerpa, 2003). Neural networks are made up of layers of nodes, and each node functions like a neuron in the brain (see fig. 1). The system receives inputs like a normal computer programme. To produce an output, the input is processed by each layer of nodes as it passes through the system. Each node connects to other nodes and has an associated weight (multiplier of the input) and a threshold (the level at which a node is activated and sends on a signal) (Gershenson, 2003). An input enters into the network through the input layer. The input passes through various hidden layers of nodes before passing through the final output layer. At each layer, if the output surpasses the set threshold, the node is activated (sends a positive signal) to the next layer in the network (IBM Cloud Education, 2020b). Depending on the weighting of the node, the output will be altered. Although we could in principle alter the output of the network ourselves by adjusting the weights in the system, this is not practical in many neural networks; there may be thousands of nodes in a network and they are not representational, so humans cannot easily understand the information contained within them.[1] While humans cannot easily alter neural networks for a specified purpose, some algorithms are trained on datasets and can alter the weightings themselves to improve their outputs in response. This is a form of machine learning (Gershenson, 2003). Sometimes, machine learning is referred to as 'deep' learning. This is machine learning that occurs in neural networks that consist of more than three layers (including the input layer and output layers), otherwise it is a basic neural network (IBM Cloud Education, 2020b).

[1] This is a key issue resulting in the 'black box problem'.

Deep machine learning can make use of labelled datasets but does not necessarily require labelling. In deep learning, a raw dataset can be provided to the AI system in an unstructured way. A deep learning algorithm can process a dataset and automatically determine features in the dataset which distinguish various categories within the data. This enables very large datasets to be used with deep learning algorithms, increasing the accuracy of the AI system and reducing the need for human intervention (IBM Cloud Education, 2020b).

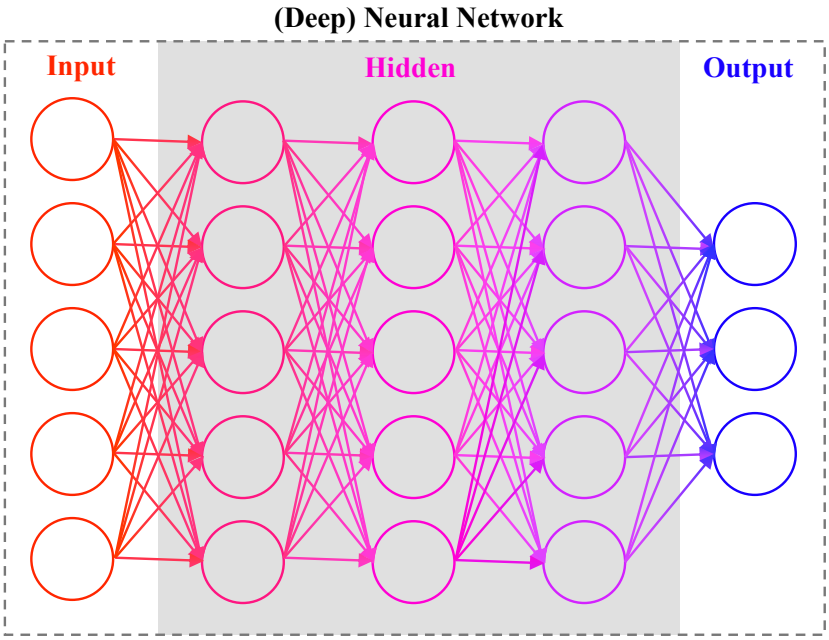


Fig. 1 Diagram showing layers of a deep neural network. Adapted from IBM Cloud Education (2022a).

AI VISUAL ART EXAMPLES

This relatively brief overview of machine learning and approaches to defining AI will be sufficient for my purposes here. We are now equipped to examine some examples of AI art. As I have explained, there are lots of examples of art-making AI, with varying types and levels of human involvement (i.e., they operate at different levels of autonomy). I will outline some noteworthy systems here. Not all of these systems produce works that are good candidates for being taken seriously as art, but it is nonetheless helpful to have an overview of the AI art landscape. It is worth noting that all the systems discussed in detail below produce visual works, and throughout this thesis I focus almost exclusively

on the visual arts and image-based AI systems. This is due to two reasons. The first is a practical reason. There are now numerous AI systems applied in the arts, across multiple mediums. Each of these systems has a different architecture to understand and works produced in different artistic mediums are likely to have (some) different relevant considerations. It was necessary to limit this project in some way, and by medium was a simple way of achieving this. Art-making AI are not yet able to generate works across more than one modality; an image-generating AI cannot write a novel or poem. The second reason why I will be focusing on visual art is that the works produced by AI in the visual domain are quite advanced in comparison to other domains (as far as application to the arts is concerned). AI struggle in particular with time-based mediums (such as music) and narrative. Artworks in many mediums are expected to be coherent, whether that is formally or narratively. For example, any work where there is an expectation of narrative continuity (such as literature, film, or narrative videogames) presents a challenge for AI systems. Similarly, we expect formal coherence; if you want to make a formally unified work like a pop song with an AI, you will typically end up with something more like free jazz. The nature of the predictive models typically used in text- and music-producing AI also have a limited potential for creativity (I discuss the limits of predictive models in Chapter 2).

Whilst many of the examples are from the visual arts, this does not mean that the work of this thesis is not more widely applicable. For an overview of key developments in AI art and creativity, including key moments beyond visual arts, see Appendix I where I have compiled a (non-exhaustive) timeline of developments. Let us now turn to take a look at a few key examples of AI art.

Robotic systems

Some AI art-making systems have a physical instantiation – a robotic element. These systems may combine an artificial intelligence system with a robot, varying from a robotic arm with instructions from a computer, to a full-blown, humanoid, robot (such as Ai-Da, see fig. 2). These systems are often widely reported on in the media and garner a lot of public attention.

Ai-Da, for example, has been highly publicised, even featuring on popular UK television shows (Hutchinson, 2019). Ai-Da is a humanoid robot, made to look like a woman with robotic arms. The website for Ai-Da boasts that she is "the world's first ultra-realistic robot artist" (Ai-Da, 2019). The team that built Ai-Da have been very cagey about the details of the AI system(s) which make up the robot; there is little to go off in public statements or marketing and no academic papers on her design to be found. The website strongly suggests there is a smoke-and-mirrors component to Ai-Da's artistic activity, stating "Ai-Da, the machine with AI capacities, highlights those tensions: is she an artist in her own right? Is she an artist's alter ego? Is she an avatar, or a fictional character?" (Ai-Da, 2019). There are very few substantial details of Ai-Da's inner workings provided, even at her solo exhibition *Unsecured Futures* (2019). The most concrete claim is that unspecified 'neural networks' are responsible for some of the bare bones of Ai-Da's works… but that the works are ultimately the result of creative input from artist Suzie Emery (Ambrosio, 2019).

Fig. 2 Ai-Da photographed with one of her works (Leemurz, 2021)

Not all robotic systems are so mysterious. Pindar Van Arman is involved in multiple projects aimed at teaching robots how to paint. Two of his projects *Cloudpainter* (fig. 3) and *Artonomous*, make use of a robotic arm which can wield a paintbrush, and is instructed by various AI systems to paint pictures. *Cloudpainter* uses many different types of AI in producing each image: "Some of my robots' creative capsules are traditional AI. They use k-means clustering for palette reduction, viola-jones for facial recognition, hough lines to help plan stroke paths, among many others." (Van Arman,

2018).[2] Van Arman also states in 2018 that he increasingly uses neural networks. In particular, he uses 'GANs'. More on GANs will follow shortly. Whilst aiming to teach his robots to paint, Van Arman still plays a considerable role in producing the work of his systems, stating "While the robot's algorithms made all of the aesthetic decisions independently, I did curate a small number of those decisions and even wrote some new algorithms for it to use when I felt the robot needed a little artistic help." (Van Arman, 2018). Van Arman also decides when the work is finished. It appears then that these robot painters are quite low on the scale of autonomy (in the case of *Cloudpainter*) or are very likely to be low in autonomy (in the case of Ai-Da, though we can only speculate given the lack of information provided about the inner workings of the system).

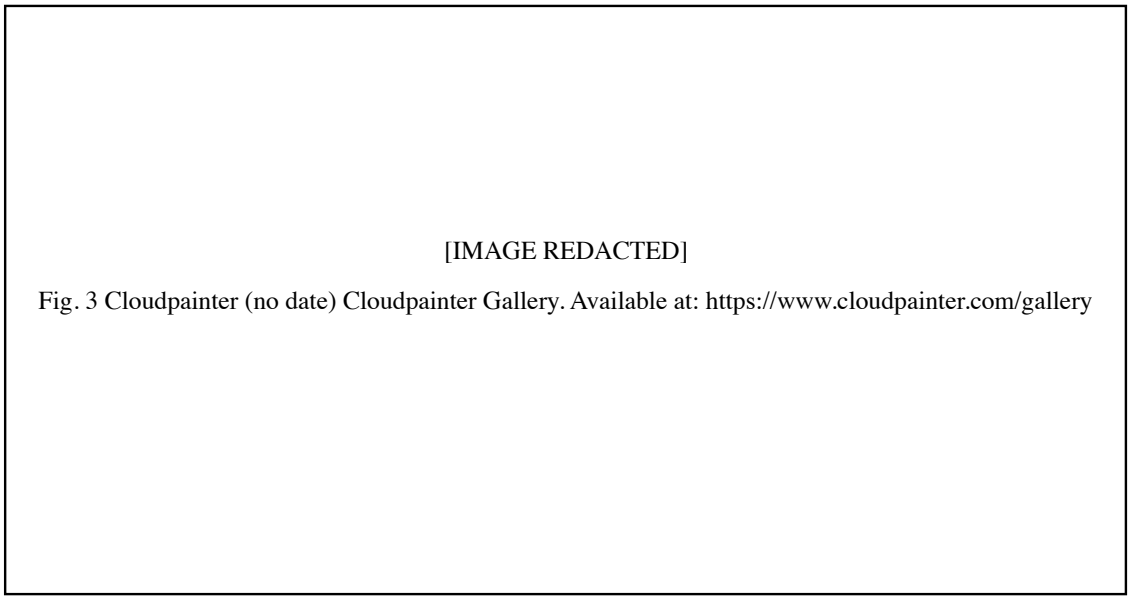


Fig. 3 Cloudpainter (no date) Cloudpainter Gallery. Available at: https://www.cloudpainter.com/gallery

Fig. 3 Pindar Van Arman's *Cloudpainter.*

Iterative systems

When Google DeepDream broke onto the AI scene in 2015, it captured the imagination of many as an "art machine" (Rayner, 2016). DeepDream started life as a way for researchers to visualise what neural networks were identifying as salient for image recognition (Mordvintsev, Olah & Tyka, 2015). Researchers found that networks trained to classify images had enough relevant information to also

[2] The particulars of these systems are not necessary to understand for our purposes. The key here is that Van Arman makes use of multiple techniques to assist in the image-making process, and they are not generative in nature.

generate images. These networks could be reverse engineered, and the images generated revealed the features that the AI had identified as relevant for image recognition. This gave valuable insight as to whether the network had got something wrong; for example, in generating an image from the label 'dumbbell', the network produces an image of a dumbbell with what appears to be arms attached (fig. 4). As the Google's researchers point out:

> There are dumbbells in there alright, but it seems no picture of a dumbbell is complete without a muscular weightlifter there to lift them. In this case, the network failed to completely distil the essence of a dumbbell. Maybe it's never been shown a dumbbell without an arm holding it. Visualization can help us correct these kinds of training mishaps. (Mordvintsev, Olah & Tyka, 2015).

[IMAGE REDACTED]

Fig 4. Mordvintsev, A. Olah, C. and Tyka, M. (2015) Available at: https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

Fig. 4 Images of a 'dumbbell' showing attached arms, (Mordvintsev, Olah & Tyka, 2015).

Researchers found that they could allow the network to choose a feature to amplify in a given image. Different layers of the network handle different levels of complexity in terms of feature abstraction, so feature-amplification can produce a variety of output images depending on which layer of the network is set to amplify features. If this occurs at the lower layers of the network, basic features are amplified in an image. For example, at lower levels of the network, features amplified could be line orientation (see fig. 5) or edges of the objects (Mordvintsev, Olah & Tyka, 2015).

[IMAGE REDACTED]

Fig. 5 Mordvintsev, A. Olah, C. and Tyka, M. (2015) Available at: https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

Fig. 5 Left: Original photo by Zachi Evenor. Right: processed by Günther Noack, showing amplification of features at lower levels of the network (Mordvintsev, Olah & Tyka, 2015).

At higher layers of the network, recognisable objects begin to be amplified. For example, if the network recognises a dog in the image (or a bird, or an eye, or scales etc.) it will amplify that feature, making it look more like a dog. Further passes of the image through the system can be undertaken, where the networks will pick up on the 'dog' with greater strength, exaggerate it more, and the resulting output will be even more 'dog-like'. As a result, it appears the system fills the image of a landscape with 'hallucinated' dogs, birds, etc. (see figs. 6, 7). The kinds of object that are recognised by the network will greatly depend on the network's initial training data. Similar to how seeing the dumbbell with an arm attached can reveal how the system 'understands' the word, these images too can reveal something about what the system recognises in images (Mordvintsev, Olah & Tyka, 2015). The process can be iterated again and again, and this is what produces the weird, hallucinogenic 'deep dreams' that garnered attention as a key development in AI art and creativity (e.g., Rayner, 2016; Auerbach, 2015).

Figs. 6-7 The author's own photo, before (fig. 6) and after (fig. 7) being put through *DeepDream Generator*.

Programmed computational creativity systems

Researchers in computational creativity have also attempted to create computationally creative systems aimed at producing novel art works. There are many such systems, but Simon Colton's project *Painting Fool* is particularly well-regarded (Colton, 2012; Colton et al., 2014). Colton and colleagues aim to solve issues in computational creativity with solutions that "involve software accounting for its decisions, actions and products, and taking the radical step of thinking of computer-generated artefacts as fundamentally different to their human-produced counterparts." (2014, p. 1).

This is also what they aim to produce with the Painting Fool. For example, one iteration of the Painting Fool was programmed to draw portraits of visitors to an exhibition. The system did not just produce images of whatever was in front of it; it interacted with its sitters, directing their poses, and providing an 'intention' and 'emotion' for each work (Colton et al., 2014) (fig. 8). The emotion, or 'mood', of the system was determined by the system's analysis of newspaper articles; a random article from *The Guardian* was selected for the Painting Fool each morning, and it analysed related articles from that point onwards. Based on the sentiment of the ten previous articles analysed, the system selected a related pre-programmed mood from a choice of six. This mood informed the style of the image it produced, as well as the explanation it offered for the depiction of the sitter (Colton et al., 2014; Colton and Ventura, 2014). To produce an image from a camera recording the sitter, a still image is

> extracted where the sitter was expressing an emotion. Machine vision techniques were applied to remove the background, into which was substituted one of 1,000 abstract art images, to which one of 1,000 image filters was applied. The same filter was applied to the face of the sitter placed in the foreground, producing in a few seconds an image conception, or sketch for the portrait ... Following this, a canvas appeared on screen, and a hand holding either a pencil, paint brush or pastel made virtual marks on the canvas leading to a non-photorealistic rendering of the background and foreground of the portrait, taking between 2 and 10 minutes, depending on the style. (Colton et al., 2014).

The system also included an evaluative component, offering an assessment through a machine vision system of what it had created (Colton et al., 2014).

[IMAGE REDACTED]

Fig. 8 Valstar, M. F., Colton, S. and Pantic, M. (2008) Available at: https://doi.org/10.1109/FG13733.2008

Fig. 8 Example portraits from *the Emotionally Aware Painting Fool.* Emotions expressed, clockwise from top left: anger, happiness, disgust, and surprise (Valstar, Colton & Pantic, 2008).

Colton et al. (2014) anecdotally report that many visitors were impressed and surprised by the Painting Fool, and that some went so far as to assign creativity to the AI. Colton's definition of computational creativity is "The philosophy, science and engineering of computational systems which, by taking on particular responsibilities, exhibit behaviours that unbiased observers would deem to be creative." (Colton and Wiggins 2012, p. 1). Colton and Wiggins here include responsibilities as a means of excluding creative tools such as Photoshop. There is a focus here on *behaviour;* however, Colton and Wiggins specify that they do not mean only human behaviour, as they do not wish to limit creative behaviours to those that would merely replicate what humans do.

Colton and Wiggins' focus on behaviour in their definition of computational creativity raises a potential issue: like the Turing Test, this definition lends itself to smoke-and-mirrors trickery, as one might seek to produce convincingly creative pre-programmed actions. An (unbiased) observer might be convinced by ersatz creative actions or 'window-dressing'. A robot that pauses to 'think' might convince us that it is evaluating its work, where in fact it has been programmed to pause at standard intervals to add to the mere appearance of creativity.

We can see elements of this in the Painting Fool itself. First, the system was presented in the exhibition as "exhibiting aspects of intentionality, imagination, skill, appreciation, reflection and learning." (Colton et al., 2014, p. 5). Colton and Wiggins (2012, p. 1) feared that the audience would be biased against the possibility of machine creativity, however they may have over-corrected. Their presentation of the Painting Fool in the exhibition seems likely to bias the audience in the opposite direction, priming visitors (particularly those with little understanding of the system) to consider that the system is capable of far more than it actually is. Further to this, the design and deployment of the Painting Fool has the system give 'intentional' and 'emotional' explanations for its work, but these are selected from a menu of possibilities, with associated scripts. What appears as the system's self-determination is the execution of a pre-programmed set of instructions. Whilst the choice of which 'mood' the system is in is determined from the analysis of the newspaper articles, which increases apparent independence, the actions stemming from the choice of mood are less so. In terms of the image that is produced, it is not clear what kind of system exactly selects the 'abstract image' to apply

to the photograph; however, once the image is chosen the system produces the image in an apparently formulaic fashion. With a sentiment analysis to seek a mood, and the application of pre-made images to a photograph of a sitter, this system already seems less impressive than audience members might be led to believe. Colton's work has certainly been very influential in the field of computational creativity and for good reason, but this piecemeal approach leaves something to be desired in producing art by AI, particularly if we seek (as Colton does) machine creativity.

Adversarial models

For much of the time spent writing this thesis, the most successful of AI art-making systems were Generative Adversarial Networks (GANs). GANs, and artists that use them, can be attributed with the biggest commercial successes in AI art. A GAN produced the *Portrait of Edmond de Belamy* (fig. 9), the first AI work to be sold at auction (it sold at Christie's for $432,500 in 2018) (Christie's, 2018a). Sotheby's AI art offering in 2019, *MEMORIES OF PASSERSBY I* by Mario Klingemann, also makes use of multiple GANs, guided by the hand of Klingemann (Sotheby's 2018).

[IMAGE REDACTED]

Fig. 9 Wikimedia Commons (2018) Edmond de Belamy, by Obvious. Available at: https://commons.wikimedia.org/wiki/File:Edmond_de_Belamy.png

Fig. 9 *Portrait of Edmond de Belamy,* generated by a GAN by art collective Obvious. (Wikimedia, 2018)

GANs were invented by Ian Goodfellow and colleagues in 2014 and were a breakthrough in image production by Artificial Intelligence. GANs are called 'adversarial' for a reason: this AI is made up of two neural networks that work in competition. The two components are the generator and the discriminator. The discriminator is trained on a set of training images – often this is photographs,

but, in the field of AI art, it could be a set of images of artworks, as has been done by artists using GANs (figs. 10-12).[3] The discriminator learns to distinguish images from the training set; we could say it learns features of the training images.

[IMAGES REDACTED]

Fig. 10 AI artists (2021b) GAN nude by Robbie Barrat, Available at: https://aiartists.org/robbie-barrat

Fig. 11 Sotheby's (2019) Memories of Passerby I. Available at: http://www.sothebys.com/en/auctions/ecatalogue/2019/contemporary-art-day-auction-l19021/lot.109.html

Fig. 12 AI artists (2021a) Night in the garden, azealeas, by Helena Sarin, Available at: https://aiartists.org/helena-sarin

Figs. 10-12 Artworks made with GANs. Left to right: **Fig. 10** GAN nude, by Robbie Barrat. **Fig. 11** *MEMORIES OF PASSERBY I,* by Mario Klingemann. **Fig 12.** *Night in the garden, azaleas,* by Helena Sarin.

The generator begins producing images, initially randomly (with input from a noise vector). These images are fed into the discriminator, which tries to determine whether the image could be part of the training set or not, given what it has learnt about the training data. The image generator is trying to 'fool' the discriminator by producing an image that could be part of the training set, and the discriminator is working to detect 'fake' images. The discriminator produces a score for each image it processes; if the discriminator determines that the image could not be from the training set, it will give the image a low score. If it determines that the image is plausibly part of the training set, then it will give it a high score. These scores are fed back into the generator, which uses them to adjust weights in the network. These weights impact the next round of images that are produced by the generator. Gradually, through the implementation of this feedback, the generator will produce images that are increasingly similar to the training set of images. The generator has no access to the training set; the only way for the generator to improve the images it produces is through feedback from the discriminator (Goodfellow et al., 2014).

[3] See, for example, the works of Helena Sarin, Robbie Barrat or Mario Klingemann (figs. 10-12). Many other artists using AI can be found at AIArtists.org.

The results of GANs are generally very impressive, particularly when they were initially developed. Despite this, there are some limitations to GANs, and related systems. They are difficult systems to train for two reasons in particular:

> (a) non-convergence—the model parameters oscillate a lot and rarely converge, and (b) the discriminator gets too successful that the generator network fails to create real-like fakes due to which the learning cannot be continued (Jaiswal et al. 2021).

As such, it can be difficult to produce consistent results in a GAN system (if the goal is producing images that are similar to the training set) and to ensure that the discriminator does not get too good for the generator, resulting in little useful feedback for the generator to increase its success rate.

GANs can be utilised at higher or lower levels of autonomy. Whilst the system itself does not change, the outputs of the system can be manipulated to a greater or lesser extent. Some level of control of the outputs can be found through closely curating the training set (either in the initial training phase or through additional rounds of training). The outputs of the system can also be more directly manipulated through accessing the latent space in the system (a representation of image data) (Gadde, 2022). The GAN can alternatively be left to generate images without intervention, having been trained on a very large dataset of, for example, artworks and with no interference in the latent space or any close control of the dataset.

The basic model of GANs has led to several variations of the system being developed. One that I will discuss at length in this thesis is the Creative Adversarial Network (CAN, or AICAN) initially developed by Elgammal and colleagues in 2017. CAN was developed to generate images in a more creative way than a standard GAN can produce. The CAN model follows a similar structure to GANs, but with a few key differences. First, the training set for the CAN consists of art images and style labels. For example, an image of a Picasso painting with the label ‘Cubism’. During training, the discriminator learns not just what images in the training set are like, but also what the styles are like. The generator generates images in the same way as in GANs, and these are also fed into the discriminator. In the case of the CAN, however, the discriminator makes two decisions in judging the

image. First, the same as in the GAN, it determines whether the image could or could not be part of the training set. If it could not be, then the discriminator feeds a low score back to the generator. If the image scores highly on this first judgement, then the image is judged on a second set of criteria: whether it does or does not fit in an existing style category (as learnt from the training set). If the image *does* fit in an existing category, then the discriminator gives the image a low score. If the image *does not* fit in an existing style category according to the discriminator, then it receives a high score. These scores are again fed back into the generator, which, as in the GAN system, is used to adjust weights in the network. In this way, as the generator improves at 'fooling' the discriminator, it increasingly produces images that 1) look like part of the training set and 2) are stylistically ambiguous. Elgammal et al. (2017) designed the system with the hope that stylistic ambiguity would result in novelty in the images produced by their system (fig. 13). As they state on AICAN's website, "Each artwork AICAN makes is an answer to the question 'If we teach the machine about art and styles and push it to generate novel images that do not follow established styles, what would it generate'" (AICAN, no date). They also claim that: "AICAN® is an Artificial Intelligence Artist and a Collaborative Creative Partner." (AICAN, no date).

[IMAGE REDACTED]

Fig. 13 Elgammal, A., Liu, B., Elhoseiny, M. and Mazzone, M. (2017) Available at: https://doi.org/10.48550/arXiv.1706.07068

Fig. 13 Examples of generated works by the initial CAN model (Elgammal et al., 2017).

In Elgammal et al.'s initial paper, the team examined how the images produced by the CAN compared to human art (abstract expressionist images, and images of art from Art Basel 2016) and GAN images when judged by human participants. They found that the CAN images scored higher than all other image sets on judgements of novelty, ambiguity and complexity (Elgammal et al., 2017, p. 17). Participants also rated images as likely to be made by humans or an AI; images from the CAN were judged as more likely to be human-made GAN images and, surprisingly, they also surpassed the human-made Art Basel 2016 set on this metric. In a second round of experimental procedure, researchers also found that CAN images were rated higher than all other image sets on questions of intentionality, structure, communication from the image, and inspiration (Elgammal et al., 2017, p. 18). Of course, these studies have some key limitations: the works were all resized to 512x512, and the experiments took place via Amazon MTurk (an online platform that can be used to solicit workers to complete tasks online) (Elgammal et al., 2017, p. 16). These changes to the images work to level the playing field somewhat for the AI images, ensuring that cues such as size, shape, texture, smell, setting etc. are all eliminated. These factors surely play at least some role in our judgements of artworks. Furthermore, not all the results were tested for statistical significance (though, we might think it notable that the AI images were even rated equally to human images). Despite this, these are impressive results. Elgammal claims that AICAN is "nearly autonomous", stating:

> Our lab has created AICAN (artificial intelligence creative adversarial network), a program that could be thought of as a nearly autonomous artist that has learned existing styles and aesthetics and can generate innovate [sic] images of its own. (Elgammal, 2019).

As far as AI art systems go, the CAN has a high level of autonomy. It evaluates its own output against two sets of criteria, and the outputs are not controlled through manipulation of either the training set or any latent space. It does still require facilitation by the research team in terms of providing an initial, labelled training set, and its outputs are likely selected by researchers (curated) when displayed, as they have been, in galleries (AICAN, no date). When I began this project in 2018 though, the CAN system was perhaps the most promising neural network system for generating art-

like images, particularly as GANs and CAN were beginning to be applied to many art-making projects.

Diffusion models

Diffusion models (such as DALL-E and DALL-E 2, Stable Diffusion and Midjourney) are the latest in image-generation AI. They have drawn a considerable amount of attention, particularly since the launch of Stable Diffusion to the public in August 2022 (Stability AI, 2022), the announcement of DALL-E 2 in April (OpenAI, 2022a) and the release of DALL-E to the public in September 2022, with one outlet declaring "AI art is 2022's hottest trend, and it's all thanks to models like DALL-E and Stable Diffusion." (Wright, 2022). It is this kind of model that Jason M Allen used in producing his prize-winning image, leading him to declare "Art is dead, dude." (Roose, 2022). These systems are all text-to-image models; users input a text prompt (typically a description of how the image should look) and the algorithm produces an image in response. These systems have been producing impressive results (figs. 13-15).

Fig. 14-16 Examples of images generated with a demo of Stable Diffusion. From left to right: **Fig. 14** Image generated from prompt "a painting of a person walking through a meadow". **Fig. 15** Image generated from the prompt "a sculpture of a cat wearing a crown in London". **Fig. 16** Image generated from the prompt "Frodo and the ring of power in Mordor painted in the style of Hieronymus Bosch".

So how do these systems work? Let's take a look at DALL-E 2 as an example of a text-to-image diffusion model. DALL-E 2 is made up of two key components: the CLIP model and the diffusion model. "CLIP is a model that 'efficiently learns visual concepts from natural language

supervision'" and "Diffusion is a technique to train a generative model for images by learning to undo the steps of a fixed corruption process." (Ramesh, 2022).

Let us examine CLIP first. CLIP is a neural network that is trained to provide the best caption in response to a given image (Assembly AI, 2022). The CLIP model is 'contrastive': it does not generate text for the images with which it is presented, but instead predicts which of thousands of text strings are associated with the given image. The CLIP model is trained on pairs of images and text. In order to match text to images, the CLIP model has two encoders: one which takes images and turns them into image embeddings, and another which takes text and turns it into text embeddings. An embedding is a low-quality, mathematical representation of information (in this case, semantic information) (Google Developers, 2022; Assembly AI, 2022). Embeddings for the image and embeddings for the text in text-to-image pairs should be more similar than in non-pairs. This is because these embeddings are representing things with semantic similarity: the text for 'a red rose' and an image of 'a red rose' should be in a similar position in the mapped vector space, particularly in comparison to text and image of 'an elephant eating candy floss'. The aim of CLIP then is to minimise the vector distance between semantic pairs and maximise the vector distance between non-pairs (O'Connor, 2022).

The next element of DALL-E 2, and the component common to many of the latest text-to-image AI systems, is the diffusion model. In OpenAI's DALL-E systems the diffusion model used is GLIDE (Nichol et al., 2022, O'Connor, 2022). A diffusion model is a kind of probabilistic model (Ho, Jain & Abbeel, 2020). The idea for diffusion models comes from thermodynamics, where molecules spread ('diffuse') from high-density regions into lower-density regions (thereby increasing entropy). This can also be thought of as an increase of noise and a loss of information in information theory (Siddiqui, 2022). The principle in diffusion models is that if we could learn how the increase in noise occurs, the process could be reversed: a 'noisy' input could be 'de-noised' to produce an output. This follows a similar formula to auto-encoders (the AI systems used in Deepfakes), which take an input through one network, project it onto a latent space, and then re-interpret it through another network. In a diffusion model, however, the system does not learn the distribution of a training data set; it

instead models the distribution of noise in a *Markov Chain* (Siddiqui, 2022, Sohl-Dickstein et al., 2015). A Markov Chain is a mathematical representation of the probabilistic movement of an object from one point to another through a discrete space (Walrand & Varaiya, 2000). The diffusion model can go from a no-noise input to a noisy output (for example, changing a photograph into seemingly random pixels, like TV static) and then reverse this process, turning the noisy input to a de-noised output (e.g., changing the random pixels back into the photograph). This process can be split to take a random sample of (Gaussian) noise and then de-noising it to generate an image (O'Connor 2022).

OpenAI's GLIDE diffusion model makes an additional contribution. As standard diffusion models start from a random sampling of noise, they do not allow for any specification of the image produced. For example, if trained on a dataset of cats, the model could produce an image of a cat reliably, but there would be no way to instruct it to produce an image of 'a black and white cat on a sofa cushion'. GLIDE includes an addition to the training process: it incorporates textual information. This results in a model that can generate images conditional to text. DALL-E 2 makes use of a GLIDE model which has been modified to include CLIP text embeddings (O'Connor 2022).

DALL-E 2 combines GLIDE's text-conditional image-generation abilities with the CLIP image encoding; thus, GLIDE learns to generate images that are semantically consistent within CLIP. As O'Connor writes "the reverse-Diffusion process is stochastic, and therefore variations can easily be generated by inputting the *same* image encoding vectors through the modified GLIDE model multiple times." (2022). This explains how when you use a diffusion text-to-image model like DALL-E 2 you will be offered several images from the same input; there is no single correct way to predict the image from the noise (O'Connor, 2022).

There is a final step needed in DALL-E 2, and that is getting from text input to image encoding. With GLIDE, we can now get from the image encoding to the final generated image, but how does this link to the text prompt inputted by a user? This comes from the other half of the CLIP model: the text-encodings. In DALLE-2 there is another 'prior' diffusion model that is used to map from text-embeddings to image-embeddings of the corresponding image. The prior takes the text input (in the form of separate tokens), encodes each token of text with CLIP, and passes these through

the diffusion prior (from text to image encoding) resulting in a noisy image. This noisy image is then passed through the image-encoder element of CLIP, and then this final encoding is used as a prediction of the final de-noised CLIP image encoding (O'Connor, 2022).

All together then, DALL-E 2's components make a system that can take a text prompt and generate a semantically relevant image from this system:

1) CLIP's text encoder takes the text prompt and maps this into the representational space.
2) The diffusion prior maps this text-encoding in CLIP onto the image-encoding in CLIP.
3) The image-encoding is then passed through the diffusion model GLIDE, which takes the representation of the image to an actual, de-noised image. The result is a probabilistically produced image that represents the semantic content of the original prompt (O'Connor, 2022).

As the text prompt is the starting point for this process, the resulting images can be shaped by altering the prompt. Small changes in the prompt may result in different encodings in the representational space. As this is directly used to generate the image (with a probabilistic, if not completely predictable, output), altering it can change the output. Getting to know which words prompt what kinds of outcomes can lead to greater control over the final generated image. This has already resulted in 'marketplaces' for prompts and guides to 'prompt engineering'.[4] As this is a probabilistic system, and the original components are still dependent on the dataset, there will be limitations in terms of invention or surprise outwith the bound of both training and probability distribution. I will discuss in Chapter 2 why I do not spend a lot of time evaluating the likelihood of creativity from systems like DALL-E 2, Stable Diffusion and Midjourney. Aside from the recent proliferation of these (which thus meant they fell outside the timeline of much of the work in this thesis) I will argue that they have a limited capacity for creativity. Having explored *what* I mean by AI art, let us now turn to how others have explored this topic.

[4] See for example promptbase.com.

LITERATURE ON AI ART

I will now examine a selection of the most prominent approaches to AI art in theoretical literature. I focus here on the discussion of art rather than creativity (I will address AI creativity more specifically in Chapter 2). The selection of papers I examine is also by no means exhaustive; in 2022 in particular there has been a wave of new publications in the area of AI art and creativity.[5] Finally, I have focussed here on academic works, though there have been two notable additions to this field of work intended for a wider audience (AI Miller, 2019; du Sautoy, 2019).

AI as another art technology

Blaise Agüera y Arcas wrote in 2016 on "Art in the Age of Machine Intelligence" (originally published on *Medium*, later published in the journal *Arts* in 2017). Agüera y Arcas worked with others to form the Artists and Machine Intelligence project at Google after the impressive results of DeepDream (A I Miller, 2019, pp. 71-2). In his essay, Agüera y Arcas draws parallels between AI to other technological developments in art, and related resistance from the art world in the acceptance of these. Similar comparisons to photography have been made very recently in *Aesthetics for Birds*, a prominent aesthetics and philosophy of art blog (see Trujillo, 2022). It is worth noting, however, that we might wish to be cautious about extrapolating from previous technologies to AI. As Agüera y Arcas points out, the distinct nature of AI is "profoundly transformational" (2017, p. 7). Artificial Intelligence, as we have seen, seeks to reproduce processes which were previously only biological in a machine, including the capacity to learn. This is not the case for other forms of technology such as photography.

> Agüera y Arcas speculates about the critiques that will face AI art in the (then) future:
>
> As machine intelligence develops, we imagine that some artists who work with it will draw the same critique leveled at early photographers. An unsubtle critic might accuse them of "cheating", or claim that the art produced with these technologies is not "real art". A subtler

[5] See for example the special issue "Creativity in the Light of AI" from the journal *Odradek,* or a recent paper in *Digital* (Grba, 2022).

> (but still antitechnological) critic might dismiss machine intelligence art wholesale as kitsch. As with art in any medium, some of it undoubtedly *will* be kitsch — we have already seen examples — but some will be beautiful, provocative, frightening, enthralling, unsettling, revelatory, and everything else that good art can be." (2017, p. 7)

This is indeed the kind of attitude I have encountered as a researcher in this area. Many think that the work by AI is poor, kitsch, or not 'real' art. We have seen these kinds of responses published in the media too. Jonathan Jones, writing for *The Guardian* responded to the first AI artwork sold at auction saying "believing the algorithm that knocked this up to be in any meaningful way an "artist" is like thinking your voice-interaction programme is out to get you. Dream on." and referring to these works as a "poor pastiche of human genius" (Jones, 2018). Mario Klingemann, an artist using AI for his work said recently in an interview with *Artnet*: "We are currently getting buried in a whole lot of bad kitsch and low-hanging-fruit work with mass appeal, but in my opinion this is simply the natural way these things go" (Lawson-Tancred, 2022). Agüera y Arcas argues that this position is flawed. He argues that in the case of past advances in technology in the arts, early works have had lasting significance in retrospect, and thus we cannot dismiss a new technology such as AI in its early stages. He argues that we should expect AI works to become as significant in the arts as past advances in technology (such as photography, printmaking, and digital art).

I should point out here that discussions of AI art, in the media and in the field of computer science alike, are often conceptually muddy. We can see in Agüera y Arcas's quoted words above that discussions about the potential *value* of AI works feature alongside the ontological question of whether it is or is not really art. The distinction is an important one (and not just for the philosophers). The barriers to each may be different; what would be required for an AI to make *art* is not necessarily the same as to make *good* art. These concerns are also often bound up with a sense of automation precarity from the creative community, and a desire to reassure oneself of human exceptionalism. A common refrain is that an AI could not possibly ever do what we do. Agüera y Arcas highlights this desire to keep art and creativity for the human:

> Faced with a new technical development in art it's easier for us to quietly move the goalposts after a suitable period of outrage, re-inscribing what it means for something to be called fine art, what counts as skill or creativity, what is natural and what is artifice, and what it means for us to be privileged as uniquely human, all while keeping our categorical value system — and our human *apartness* from the technology — fixed. (2017, p. 4)

In this thesis I set aside the concerns of replacement, as I wish to focus on the AI systems themselves; however, I do address both whether AI can make art (Chapter 1) and address whether AI value in the artistic domain will align with human values (Chapter 4).

Art and the social agent

Aaron Hertzmann, a researcher at Adobe, has also examined the possibility of AI art (2018; 2020). Hertzmann generally does not think that AI can make art, due to his emphasis on the social dimension of art-making:

> I observe that each of these functions of art is *social*: art arose as forms of communication, displays, and sharing between people. Although art takes many different forms in different cultures today, each of these forms serves one or more of the same basic social functions that it did in the Pleistocene.
>
> I generalize this theory beyond humans to hypothesize: *art is an interaction between social agents.* A "social agent" is anything that has a status akin to personhood; someone worthy of empathy and ethical consideration. Many of our other behaviors are interactions between social agents, such as gifts, conversation, and social relationships like friendship, competition, and romance. (2018, p.17)

Of course, the social dimension of art is not a new idea for philosophers (see e.g., Abell, 2012), but Hertzmann dismisses philosophical contributions to the definition of art (see below). Hertzmann instead ploughs on with his definition of art, without exploring the potential objections one might raise to such an account. Some objections may be specific to the poorly defined nature of the account Hertzmann puts forward. We might ask: what kind of social interaction counts as art? If, as

Hertzmann states, ‘many other behaviours are interactions between social agents’, what makes art different? Are all social agents persons? What occurs if personhood is given to a non-human? Some objections are more familiar: if art is an interaction between art and audience, what of the medium of art (see e.g., Wollheim, 1980, pp. 40-41)? What of art that occurs in social isolation (Adajian, 2018)? Hertzmann does only discuss this under-specified definition of art. He also discusses, in brief, an institutional definition of art, and mentions cluster accounts of art such as Gaut’s. He does not investigate these definitions in any depth, stating “Understandably, these definitions all assume that the artist is always human, without exploring much whether non-humans can create art, and thus do not provide much guidance for this discussion.” (Hertzmann, 2018, p. 20).

Given Hertzmann’s desire to specify that art is an “interaction between social agents” and that “A ‘social agent’ is anything that has a status akin to personhood” it is odd that his reason to abandon the long history of philosophical efforts to define art is that they are too focussed on the human, as arguably Hertzmann’s own consideration of ‘social agents’ as persons is still very human-focussed. Personhood is not easily granted to non-humans; it has been suggested that dogs have social skills (Hare & Tomasello, 2005) and yet they are not considered persons by many philosophical accounts and common usage (see Dennett, 1988; Sapontzis, 1981), despite the arguments made by some (see Gruen, 2021). Hertzmann also does not provide any additional criteria of what kind of social interaction would count as art. For example, despite the social role dogs can play with humans (Hare & Tomasello, 2005) dogs do not seem to most people to make art, and Hertzmann even states this:

> In other words, we are open to the idea of animals creating art, because they can have social relationships with us. It is just that we have not found any other creature that satisfies our criteria for creating art, whatever they are. (2018, p.18).

Yet, when looking for these other criteria, Hertzmann dismisses skill (2018, p. 18), intention (2018, p. 19) and creativity (2018, p. 19), all in favour of the social dimension of art. This is despite highlighting that whilst non-human animals seem social, they do not make art; so how exactly can we determine what kind of social relationship is the right one?

Whilst Hertzmann is quite right that the philosophical definitions of art he discusses do not investigate machine-made art, Hertzmann gives up at this first sign of difficulty in favour of his own pet theory. Fortunately, I am more dogged. I will examine definitions of art in Chapter 1, investigating whether, despite their lack of explicit consideration of AI, these definitions can accommodate art made by AI systems. Because of this focus on the social dimension of art, Hertzmann concludes that AI currently cannot make art, and that it will be unlikely to in the future. I address a potential solution to AI system's lack of access to the social world in Chapter 3.

Human intention and AI art

Both Agüera y Arcas and Hertzmann write from the field of AI and machine intelligence; let us now turn to philosophers writing on this topic. Steffen Steinert (2017) adopts an intentionalist definition of art to argue that machines, as technical artefacts, can create works of art. As Steinert notes, if we focus on intention (as I will discuss in Chapter 1) the possibility of AI art will depend on AI mind (something I discuss in Chapter 3). To get around this, Steinert argues that there is no need for agency or intentionality on the part of the AI. Steinert argues that we can instead rely upon the intentions of the maker of the AI system (2017, p. 278) to provide the requisite intentions for the creation of art. As I discuss later in this thesis, this is an option open to us to fulfil any requirements for intentionality. However, I take a different approach, one which Steinert nods to, saying "One upshot of the account presented here is that we do not need a new conceptual framework or dubious assumptions about artistic agency on part of machines in order to arrive at the conclusion that creative machines make art." (2017, p. 267). Whilst I do, in places, explain how we can make such assumptions and how this will impact on our conclusions, I also offer an argument about the agency of AI (Chapter 3) (rather than relying on dubious assumptions). This is, in part, because I do not take the route of assuming an intentionalist account of art. In Chapter 1, I investigate both an institutional account and a historical-intentional account of art in brief, before moving on to examine AI art under a 'cluster' account of art, as put forward by Gaut (2000). Under this account, the key component is action and not intention. This means that intention is not *necessarily* needed, as I argue in Chapter 3.

There is an issue for Steinert's proposal to rely on intentions from a human in AI art. In his discussion, he ignores the possibility of a responsibility gap. For Steinert it is enough that someone intended for the machine to make art (and therefore it is art). Whilst this can work well for any system with a lower level of autonomy, as soon as we have a higher level of autonomy in our AI, the intention no longer maps clearly onto the output of the AI. In other words, we will have a gap between the designer of the system (and their original intention), where we can no longer ascribe responsibility for the work to a human. It therefore becomes questionable how much the human's intentions can be used to classify a work as art. Steinert does address that the maker of the artwork needn't be responsible for the appearance of the work, stating:

> To avoid confusion, I hasten to point out that it is not necessarily required that the maker of the work of art is responsible for the physical features of the item. The creation of a work of art need not even involve any physical manipulation on the part of the artist. Consider again found art or ready-mades, where natural objects or artifacts are taken and exhibited without altering their physical features. Please note that even in cases like these there was an intention to turn an object (either natural or non-natural) into a work of art and this intention was informed by past tokens of kinds of art. So, the notion of "creating" a piece of art is a very broad one. (2017, p. 273)

This holds in so far as a human is curating the outputs of the AI; however, as we move to more autonomy in our AI systems, we may have the AI system itself selecting the works. We could argue that in some cases this is already occurring at the point of use by humans, but this is dependent on whether the human decides to be selective with the AI outputs or not. If they do not, then under Steinert's view these will not count as art. This seems odd though for it ignores the initial set-up of the system, whose purpose is to generate art. This then looks like a disanalogy from the ready-made or found object case where an intention to make a work of art occurs at the point of selection instead of creation. In the AI case, the intention only comes in setting up the system to produce works (before the work is made) and not after. Once responsibility for the actions of the AI can no longer be properly attributed to the designer of the system (this is a central claim of the responsibility gap, see

e.g., Matthias, 2004) can we properly rely on that designer for our source of intentionality?[6] I do not think we can. Whilst I do not explore the responsibility gap in much depth, I will discuss its role in responding to a central objection to this thesis later in this introduction.

A framework for machine art and creativity

Mark Coeckelbergh (2017) aims to offer a framework for philosophical discussions of the status of machine creativity and machine art:

> It breaks the main question down in three sub-questions, and then analyses each question in order to arrive at more precise problems with regard to machine art and machine creativity: What is art creation? What do we mean by art? And, what do we mean by machines create art? (2017, p. 285)

Coeckelbergh reframes the question of 'Can machines create art?' into three separate questions, each with a different emphasis: "Can machines *create* art?", "can machines create *art*?", and "can *machines* create art?"

In organising the content of this thesis, I have found the framework provided by Coeckelbergh to be a helpful scaffold. The first three chapters of this project have therefore been organised around these three questions:

1) Can machines create *art*?

2) Can machines *create* art?

3) Can *machines* create art?

Coeckelbergh takes the question "can machines *create* art?" first, focussing on creativity itself. Coeckelbergh draws attention to a divergence in discussions of computational creativity, from the external to the internal, or from the product to the outcome (2017, p. 288). Coeckelbergh précises Boden's approach to creativity as one which is process-based, focussed on the internal workings of

[6] Here I understand responsibility to be control and knowledge, after Coeckelbergh (2020).

the machine, and contrasting these kinds of approaches with those that are outcome-focussed (e.g., Colton & Wiggins, 2012; Cardoso and Wiggins, 2007). As Coeckelbergh highlights, these outcome-based approaches are effectively behavioural: the focus is on what is observable by humans. Like many in computer science, these approaches are likely used because they are more operationalisable and testable (2017, p. 289).

Coeckelbergh turns to aesthetics, and here focusses on views of artistic creation, i.e., the process of creating a work of art (as opposed to views of creativity). The views that Coeckelbergh discusses, are rather outdated: he uses the examples of Collingwood (1938), Plato, Aristotle, and Coleridge (1817). Furthermore, they are not views of creativity. A better historical reference would be Kant's views on genius, or the inspirationalist view of creativity. Coeckelbergh mentions these in passing, but does not focus on them (2017, p. 288). Coeckelbergh distinguishes between what he calls a 'modern and expressivist' view of artistic production, and a 'pre-modern' view in which art should be about imitation (*mimesis*) (2017, p. 290).

For Coeckelbergh's examination of what he calls the 'expressivist view' of artistic creation, that art is self-expression, he draws largely on Collingwood. Under this view, Coeckelbergh states as machines do not have any inner states to express, they seemingly cannot be creative:

> Clearly, what matters for Collingwood is the process, and in particular expression and imagination. But machines, it seems, can do neither of these. They have no 'self' that can be expressed and they lack imagination. (Coeckelbergh 2017, p. 291)

This view of artistic production is not particularly current, nor is it particularly connected to current discussions of creativity. Collingwood puts forward that art is expression; by bringing feelings into consciousness, one creates art (1938, p. 109). Collingwood's theory has been criticised for its overemphasis on what is in the mind of the artist at the point of creation and its disregard of the artistic medium (see Wollheim, 1974). Furthermore, it has been argued that Collingwood's approach

is too exclusive.[7] One possible solution to this problem is to label it a theory of the value of art as opposed to a view on what art is. For our purposes though, this would render Collingwood's theory irrelevant to the question of whether AI art is art. There is a further potential issue for Collingwood's theory: that it is too inclusive of any process of expression, many of which we would not label as art, nor as creative (see again Wiltsher, 2018, p. 764). Finally, the distinction between art and craft by Collingwood, which Coeckelbergh references, has been also criticised as too restrictive (see Kemp, 2021; Ridley, 2002).

Coeckelbergh considers his second account of artistic creation, what he calls the 'mimesis conception of artistic process' (2017, p. 291). He claims under this account what matters is imitation. He argues that unlike self-expression, this is something that a machine may be able to achieve:

> if a machine is capable of imitating whatever it 'sees', then according to this criterion, it seems that what it does qualifies as artistic creation. For example, if on the basis of the information given by its sensors and by means of an algorithm a machine is able to draw a portrait of a person that looks like the person, then that machine has succeeded in meeting the criterion. And, if on the stage the machine imitates a human being, then this can, in principle, be art—again, if we understand art as mimesis and if we understand mimesis as imitation. (2017, p. 291)

Again, this is a very outdated view of art, and again has an issue of both excluding too much (potentially any non-representational art form, any avant-garde works of theatre or film, conceptual art etc.) and including things we might wish to exclude (non-art photography, for example). I do not think it worth engaging with this approach, as it is clearly an insufficient account of artistic production. In this thesis, I will favour the approach of examining creativity. I will examine the evolutionary approach to creativity, and Boden's view of creativity, before turning to Gaut's agential account of creativity (Chapter 2).

[7] Under some versions for example it would exclude conceptual art, minimalist art etc. See Wiltsher (2018) who argues against this.

Coeckelbergh turns next to considering whether machines can make *art.* The focus here is the artistic product (the latter part of his examination of creativity overlaps the approach to this question somewhat). Coeckelbergh gives a brief rundown of various approaches to defining art. He characterises the possible views as existing along a continuum between 'objective' and 'subjective' approaches to defining art (2017, p. 292). I am sceptical of the accuracy or utility of this distinction, partly because it simplifies the various accounts of art to either objective or subjective (even if that is not Coeckelbergh's intention) when it is clear many accounts of art do not easily fit in these categories.

Take the institutional account of art as first proposed by Dickie: "a work of art is an artifact upon which some person(s) acting on behalf of the artworld has conferred the status of candidate for appreciation" (Dickie, 1974, p. 34). The object must first be an artefact in order to meet this definition (this may be, once we know the relevant criteria, an objective matter). It then must have had the status of candidate for appreciation conferred on it. Whilst the conferring may be 'subjective' in that it is down to an individual to do this, whether the object has had such a status placed upon it could be an objective matter.[8] Even the presumably 'subjective' view of art then, by Coeckelbergh's standards, does not fit neatly into that category.[9] Coeckelbergh does not think that either objective (criteria-based) or subjective accounts of art necessarily exclude AI creations. On the objective accounts, he states:

> In practice, it may be very difficult for a machine to create something that meets certain objective criteria. But the point is that in principle, it is possible for a machine to do it. In contrast to the expressivist definition of art, these kinds of definitions do not exclude machines a priori. (Coeckelbergh 2017, p. 292)

On the subjective account, Coeckelbergh is even more optimistic:

[8] This definition can take stronger or weaker forms, see Matravers (2000).
[9] Rather than this subjective-objective divide, I would favour Davies' (1991) characterisation of the distinction as either feature-based or relational.

> On the other hand, if there are only subjective criteria, it seems that the machine gets an even better chance to be seen as an artist having created a work of art. If the only thing that counts is subjective decision or social agreement, then if these are in place, this is all the machine needs. (Coeckelbergh 2017, p. 293).

Coeckelbergh finally turns to the question of 'can *machines* create art?' (2017, p. 297). Coeckelbergh raises the issue that much of the prior discussion is premised on the idea that machine creativity should follow human creativity. Coeckelbergh highlights that some researchers have questioned this (e.g., McGregor et al., 2014, Besold et al., 2015), and suggests that this line of questioning is worthwhile.[10]

Coeckelbergh also highlights the issue of whether machines themselves could evaluate or consume art, a question that goes hand-in-hand with the idea that AI might be creative (or indeed produce art) that is quite different from that of humans (this is also raised by Boden, 2014, and Colton, 2008). I highlight Boden's work on this in Chapter 2. Coeckelbergh finally considers the broader relationship between technology and art, drawing on the work of Heidegger (2017, p. 299).

There are two key ideas I wish to highlight from Coeckelbergh's paper. First:

> This exercise has offered a *conceptual framework for philosophical thinking about machine art,* which links arguments for or against giving artistic status to machine art and machine creation to specific views about creation and art. (Coeckelbergh, 2017, p. 300, my emphasis).

As stated above, I have adopted Coeckelbergh's three questions as a framework for the first three chapters for the thesis. Beyond this, there is not much similarity between Coeckelbergh's specific approach and my own. This is particularly clear in my response to the Coeckelbergh's third question (can *machines* create art?) which I instead utilise to highlight aspects of machines which may prevent AI from being creative or from making art (such as mind, embodiment and agency). Nonetheless,

[10] I do not examine this issue in the thesis, though I highlight in Chapter 4 that if the source of difference is in how to judge value in relation to creativity we may well not recognise computer creativity if it differs from human creativity.

Coeckelbergh's initial steps to bring philosophical aesthetics to AI art and creativity are important to this project. Coeckelbergh and Steinert offer two papers which directly combine philosophical aesthetics and AI art in an otherwise sparse body of literature (and notably aesthetics is not either of their main fields of study; Coeckelbergh is a philosopher of AI and Steinert a philosopher of technology). This approach of bringing aesthetics and the philosophy of art to AI art and creativity is the approach of this thesis.

> Relatedly, the second point to highlight from Coeckelbergh's conclusion is the following:
>
> Clearly, *more work could and should be done* to further elaborate the links between thinking about technology, aesthetics and artificial intelligence research. (Coeckelbergh 2017, p. 300, my emphasis).

This thesis fulfils this call for more work in this area. I aim to contribute a significant, in-depth exploration of AI art from the perspective of analytic aesthetics.

MOTIVATIONS AND RESEARCH QUESTIONS

This area is clearly very young in terms of philosophical investigation. Anecdotally, some dismiss AI art as neither philosophically interesting nor a worthy object of aesthetic investigation. I do not accept this position. As with many current developments in AI technology, there is a rich possibility for philosophical investigation in this area. Aesthetics is trailing behind philosophy of mind, language, epistemology and ethics in enquiring into the impact of Artificial Intelligence. This thesis aims to go some way in moving aesthetics forward in the investigation of Artificial Intelligence. With this broad goal in mind, this thesis will address the following research questions:

1) **Can AI make art?** This will be a foundational question for this project. In what way does it make sense to discuss art made by Artificial Intelligence? Can an AI even have the requisite qualities required to make art?

2) **Can AI be creative?** This is our second foundational question. Something may rightly be called art, with little to no creative content at all. What many object to when they consider the possibility of art made by AI is the possibility of a creative machine. From questions one and two, some apparent limitations of AI emerge. So, our third question will be:
3) **What key features of machines might limit the possibility for AI to achieve art and creativity, and can these be overcome?** Investigations into questions 1 and 2 will lead us to several apparent limitations of AI: mind, access to the social world, embodiment, and agency. Can AI achieve these in any way?
4) **Will AI share aesthetic or artistic values with humans?** The value of art is another central question for aesthetics; but will AI share values with humans when producing aesthetic objects? This will be our fourth question. AI might not align with human values, but we still might wonder if there is anything that AI can offer us in the arts. This leads us to question five:
5) **Does AI art have any unique (or at least common) aesthetic qualities?** Finally, I will investigate the possibility of common aesthetic qualities of AI artworks (and AI images more broadly). The exploration of question five may act as something of a riposte to those who are sceptical of the potential for AI works to be aesthetically interesting or worthy of study.

Despite the centrality of AI in the thesis, one needn't think AI is a rewarding focus of study in order to be invested in the outcomes of these investigations. The study of Artificial Intelligence can reveal as much about us as humans as it does about machines. What is different about AI, or indeed where AI is not different from us, can reveal much about our own processes. For example, for the sceptic of AI creativity, if we find AI can indeed meet our definition of creativity, we might wish to investigate this definition. This could reveal whether an addition to our definition is needed if we wish to keep creativity human centric.

AN OBJECTION: AI AS TOOL

There is a common refrain from those who object to the very premise of this thesis: 'AI is just a tool for artists; it's like a paintbrush!'. I wish to respond to this objection here in the introduction. To do so, I must first establish some detail on machine learning and responsibility. To understand whether AI is simply wielded by a human (like a paintbrush), we can think about who is responsible for the actions of AI.

In thinking about the responsibility a human can have over the products of deep learning algorithms, several factors contribute to a likely 'responsibility gap' with these AI (Matthias, 2004; Coeckelbergh, 2020). These factors include: the large number of nodes in a network, the non-representational nature of the nodes' connections, the ability of some systems to alter their own weightings, and the large datasets often used to train the networks.

When discussing attributing (moral) responsibility to AI, and the subsequent responsibility gap that ensues, an Aristotelian approach is commonly applied (Matthias, 2004). By this account of responsibility, we need to have both control and knowledge in order to be held responsible for our actions (see Fischer & Ravizza, 1998). The control condition requires that we must be in control of what we are doing to be held responsible for it. This condition allows us to exclude scenarios where we are coerced into something, or something occurs entirely by unforeseen accident. Imagine you cause a car crash by stepping out into oncoming traffic; your blameworthiness will be quite different if you were in full control of yourself and stepped into traffic, versus if you were pushed into the road or if you slipped on a banana peel.

The knowledge condition adds a requirement that you know what you are doing, and potential consequences of your actions in order to be held responsible (Coeckelbergh, 2020, p. 111). This condition allows us to excuse ourselves from responsibility in cases where we did not have key knowledge about an action or consequence (perhaps we have been misled or simply could not know what would happen). For example, we surely think that if we are asked to press a red button by a friend, we should not be held responsible for that button causing a building to explode. In this case we

did not know (and could not even have a reasonable expectation) that the button would cause such destruction.

This problem of the *responsibility gap* is typically discussed in cases of autonomous weapons or self-driving vehicles (see e.g., Sparrow, 2004). Many AI systems seem not meet the criteria of control or responsibility themselves (even when an AI has a level of independence in its actions), as they are thought to lack key features such as agency, autonomy, or consciousness (which are thought necessary for control and knowledge respectively, see Coeckelbergh 2020, p. 111). (Counter to this, I will argue in Chapter 3 that some level of agency is possible for AI, and thus some level of responsibility may be attributed to the system). Despite the prevalent view that AI cannot have responsibility, with many AI systems that utilise machine learning it is not clear how a human can be held responsible either. If the human user or designer does not have full (relevant) knowledge of the system, or was not in direct control of the system, then they too cannot be held responsible for its actions. This is the responsibility gap: it seems we cannot meaningfully ascribe responsibility to an AI, and yet we also cannot ascribe it to a human.

This problem can be shown to exist beyond attributing moral responsibility for the harms of autonomous vehicle crashes. We can see similar gaps in responsibility attribution in other domains, particularly with AI systems that employ deep learning algorithms. First, the non-representational nature of the nodes and their connections and the high numbers of both nodes and connections in complex deep learning algorithms leads to a 'black box' problem, whereby even an expert computer scientist might not be able to explain why the system produces the outputs it does with any detail. Even with access to the algorithm's code, the non-representational nature of the system means that we still may not know what is occurring within the system, as we cannot interpret the numbers that we can see. In the context of AI artmaking, we see this when artist/scientists such as Robbie Barrat describe not knowing why images generated appear as they do (Barrat, 2018). This epistemological barrier means that ascribing knowledge of the 'actions' of the AI to a human can only be limited. Barrat knows his GAN will produce an image, and (if he has trained it on a narrow dataset rather than

a large one) he may know the kind of image it will produce (say, nude-painting-style images). Even with careful training though, he can still be surprised by the output of the system.

To move to the second criteria of control: in Barrat's case, he may have *some* control of the outputs of his system through careful curation of a dataset. With less closely controlled datasets (millions of images rather than a few thousand) we will see even less control of the outputs. It should be noted that we cannot easily 'open up' the system and change some numbers in order to alter the outputs to our will. This may be possible with a basic neural network, but with complex networks we will not be able to tell what each number is corresponding to. Just like we cannot isolate a neuron that we are certain fires upon seeing the colour 'green', we cannot find the 'green' node in the neural network and alter its weighting. In the case of GANs, new techniques are being developed to try to control the image output in nuanced ways (Gadde, 2022), but this additional control is still not absolute. Thus, the lack of knowledge of the specifics of the system leads to a lack of possible control of the outputs, even with access to the system's code.

This 'responsibility gap', however narrow, is why a (deep-learning) AI is *not* like any other artist's tool, and especially not like a paintbrush. We are not going to be surprised by the output of our paintbrush. Our paintbrush is not an inscrutable black box. Our paintbrush is not altering itself through learning. Our paintbrushes do not include randomness in their production (unlike some of the AI systems examined in this thesis). We know what output will come from our paintbrush, and when we paint, we are in control.[11] For a deep learning, generative AI system, we do not know what the output will be, and we have far less control. We may control it to some degree (for example, through close cultivation of a training set, or through prompt-engineering in the case of text-to-image systems), but this will always be limited in comparison to typical tools. This, in part, motivates the focus of this thesis on the products of AI systems, as separate as possible from humans.

[11] Even in cases where an artist does not *completely* control the output (such as Jackson Pollock's drip paintings), the paintbrush and paint are still almost entirely under their control.

LIMITATIONS

I have titled this thesis 'The Philosophy of AI Art'. This is an ambitious title, and though I hope I have been equally ambitious in the topics I cover, strictures of space preclude providing a truly comprehensive philosophy of all topics in relation to AI and art. The scope of this thesis is limited in a few ways.

First, I largely focus on philosophical considerations of art made *by* AI rather than art made *with* AI. As such, I focus on what an AI can do, and what an AI might be able to do, with minimal interventions from humans. There are several topics which are therefore naturally excluded from this project. For example, questions about collaboration, shared authorship, enhancing human creativity, and the use of AI by humans as an advanced tool have not been addressed.

Second, related questions about responsibility attribution and whether AI can be authors have also not been addressed here. This is because I take these as separable from the proper way to consider the process and product of AI, as the concept of authorship is laden with additional human chauvinism. That is not to say that these are not relevant and interesting questions, but they are beyond the scope of this thesis.

Third, I do not discuss issues of automation of creative tasks, outsourcing to machines or the changing landscape of work in relation to the arts (see Coeckelbergh, 2017, p. 287). While these are common areas of concern, they are somewhat down the line from the present analysis, which is aimed at discussing the possibilities of AI, including future AI, in terms of art and creativity.

Fourth, issues relating to technology and art are also not addressed in this thesis. These questions would widen the focus of the thesis considerably, and I do not wish to tread the ground of photography, film, printing and so forth in great detail. This exclusion is partly predicated on an assumption that an AI is not the same as other forms of technology (though perhaps it is better to say that I do not assume it *is* the same as other technologies). The reasons for this can be drawn from my explanation above of machine learning and responsibility attribution; AI systems that employ a deep neural network architecture will not be in the direct control of a human, nor can its process be

explained simply to a human. Whilst the discussion of other technologies may be relevant for AI, it cannot be assumed to have the same relationship to humans, or to art. Whilst this is beyond the scope of this thesis, I am sure that this distinction warrants further exploration in future work.

Finally, I have excluded key ethical questions which, while pressing issues, are not exclusive to the realm of aesthetics. For example, issues of bias and potential harms can apply across many AI systems and are therefore not unique to the question of AI art. Relatedly, concerns about intellectual property rights and copyright are also not discussed; these are mostly legal concerns with related ethical issues. These areas are all, however, promising areas for further research.

There is one further limitation worth noting: the problem of obsolescence. Many works that discuss the possible applications of AI to art are quickly rendered out of date. Even as I have written this thesis, there have been several giant leaps forward in the technology available. When I began this project, GANs were the art-making system du jour, and now they seem to have been supplanted by diffusion models. Whilst this is a well-known issue in computer science (and the reason behind the proliferation of pre-print articles in the field), the structures and practices of academic philosophy are far less able to accommodate such rapid advances. By the time the reader reads this thesis, there will undoubtedly be new technological advances in AI for art applications. Throughout this thesis there will be moments where the examples I have used will soon seem quaint. Because of this challenge I have tried as much as possible to draw attention to the key conceptual issues at play, and whilst I refer to various technologies throughout (mostly, though not exclusively, adversarial networks) the philosophical work should not be rendered out of date. This is helped by my overall approach here, which is to consider hypothetical future AI alongside current technology (i.e., if an AI could do X, or did have Y, then it could/couldn't achieve Z).

CHAPTER SUMMARY

The first two chapters of this thesis address questions highlighted by Coeckelbergh (2017). In Chapter 1, I examine the question of whether AI can create *art.* This chapter examines how works made by

(autonomous) AI might meet the criteria of various (anti) definitions of art. I examine three accounts of art: the historical account of art, the institutional account of art and the cluster account of art. This chapter concludes that, AI *may* be able to create what we would call art. I argue that the cluster account of art offers the best framework for examining AI works.

In Chapter 2, I examine the second of Coeckelbergh's questions: can AI *create* art. This chapter examines three accounts of creativity. The first is a Darwinian account of creativity which, as its name suggests, is inspired by Darwin's theory of evolution. The process of creativity under this account is one of Blind Variation and Selective Retention (BVSR) (Simonton, 1999). In this section, I examine how this model can be used to assess whether AI can be creative (focussing on GANs and CAN). I argue that AI could be creative under this model. The second account of creativity I examine is Boden's account. Boden herself wrote about the possibility of her model applying to computers. Under Boden's view, most aspects of her theory of creativity are within the reach of AI systems. Boden does however conclude in a 2014 paper that autonomy (which she claims some may think is necessary for creativity) cannot be achieved by AI. I argue against this position, arguing that AI could have some level of autonomy. Finally, I examine Gaut's agential account of creativity. In this section, I first examine Gaut's initial agential account of creativity comprised of novelty, value and flair. I argue that, aside from the requirement of agency, there is nothing prohibitive in this account for AI (I return to agency in Chapter 3). I then turn to examine Gaut's addition of spontaneity into his account of creativity. I argue that this addition can be better characterised as unpredictability, which is possible in AI systems.

After Chapters 1 and 2, we are left with a few challenges to both the possibility of AI art and the possibility of AI creativity. In Chapter 3, these challenges are examined in more depth to see how much water they actually hold. These can be thought of as issues in response to Coeckelbergh's question "Can *machines* create art?". The first of these is mind: can an AI have a mind? This section goes further by arguing that AI may be able to 'extend' its mind to utilise humans for social input into the process of creation. The second is embodiment, which is mentioned in Boden's account of creativity (e.g., Boden, 1998) and is often seen as a key missing feature of AI systems. In this section,

I argue that under certain accounts, embodiment is possible for some AI art systems. The third challenge to AI art and creativity is agency: can an AI have agency? This is key for both the cluster account of art (Gaut, 2000) (which includes action as a necessary condition of making a work of art) and Gaut's agential account of creativity (Gaut, 2010). In this section, I argue that by adopting a minimal account of agency we can meet the requirements of the agential account of creativity. This strategy allows us to consider the actions of AI without positing mental states, which will also help with the cluster account of art.

In Chapter 4, we move on to key area of aesthetics: aesthetic value. Will an AI's sense of 'good', if it has such a thing, mesh with what we consider to be good? This question is one of value alignment, a topic well-known in the philosophy of AI as a question of ethical values, but as yet not explored in terms of aesthetic value. I argue that there may indeed be an aesthetic value alignment problem, and that while in creativity generally we will want AI to align with our values, we may want it to differ in the artistic realm.

In Chapter 5, I move to examine whether there is anything unique about AI art. I explore two possible aesthetic attributes of AI works: that they are weird, and (usually conversely) that they are convincing. With these two sections I hope to begin to establish the aesthetics of AI images.

Chapter 1 - Can AI Create *Art?*

This chapter addresses the question "Can AI create *Art?*". I will examine multiple theories of art: the institutional account, the historical-intentional account, and the cluster account. I take a paradigmatic or renowned example of each theory in order to see how art made by AI will fare against each. Through this examination I will evaluate 1) whether AI works can be art under each account, 2) whether future (more autonomous) AI could create art under each account, and 3) put forward an argument for which account is best for evaluating works made by AI.

An Institutional Account of Art and AI

Let us first turn to the institutional theory of art. The most well-known version of this theory comes from George Dickie's seminal article 'What is Art? An Institutional Analysis' (1974). Dickie aims to defend the possibility of defining art and argues for a formulation of a definition that can accommodate challenging cases of artworks that resist fitting a definition of art, such as Dadaist works. Dickie's formulation of the institutional theory of art is:

> A work of art in the classificatory sense is (1) an artifact (2) a set of the aspects of which has had conferred upon it the status of candidate for appreciation by some person or persons acting on behalf of a certain social institution (the artworld). (Dickie, 1974, p. 34)

Let us examine how an AI work might measure up to this definition. I will examine each condition in turn, beginning with (2).

In the second condition, Dickie refers to the "social institution" of "the artworld". His usage of the term 'artworld' here is borrowed from Danto's use of this term "to refer to the broad social institution in which works of art have their place" (Dickie, 1974, p. 462) Dickie specifies in the first version of his account that this social institution need not be formalised and that the 'personnel' of the artworld includes "artists (understood to refer to painters, writers, composers), producers, museum directors, museum-goers, theater-goers, reporters for newspapers, critics for publications of all sorts, art historians, art theorists, philosophers of art, and others" (Dickie, 1979, p. 463).

This is a fairly wide collection of people who could confer art status on something, and it certainly seems that AI can easily meet this requirement. There is nothing in this version of Dickie's account that requires art status to be conferred upon the object by the artist that made it, and the work does not have to be intended for the artworld in any way. Even if there is minimal human involvement in the production of the AI artwork, the role of an AI in producing the work does not to cause any issues. AI does not itself need to act on behalf of the artworld in conferring the right kind of status or intend that its product is an artwork; instead, we can have another representative of the artworld offering the conferral of this status on the work.

We can see that this has already happened for AI works. AI art has been recognised by several 'personnel' of the artworld. Reporters have written about these new artworks (e.g. *The Art Newspaper,* 2018) museums and galleries have exhibited AI artworks (e.g. *AI More than Human,* 2019*),* and AI art has been sold at auction (e.g. Christie's, 2018a). As a philosopher of art, I am also part of this social institution, and by treating these AI works as art, I too could confer art status upon them. In this first formulation of Dickie's institutional theory of art, conferral of art status can be that simple.

Whilst this suggests that people in the artworld could, and have already, conferred the status of 'candidate for appreciation' on some works made by and with AI, we might wonder: could an AI (like a human artist) confer this status on its own creations? It seems unlikely that an AI could (at least currently or in the near future) act on behalf of the artworld.

First, it is not clear whether an AI can act (if action is taken to mean intentional action).[12] Setting aside the problem of action, it would also seem that an AI would need to be embedded in some way in the artworld in order to confer the right kind of status on behalf of it. Dickie has claimed that the 'artworld' could be construed broadly (our AI would not need to be in a formal institution, for example) (Adajian, 2018), but it must surely be social in some form. To act on behalf of a social institution would surely involve participating in that institution.

Perhaps an AI would not need to participate socially in the artworld. We could imagine a case where an AI is used to select works for display in a gallery. It is (in some way) acting on behalf of the artworld, and its selections will be candidates for appreciation by virtue of the decision of the AI. It strikes me though that this is not what Dickie had in mind with his account, even if we can imagine such a case (for starters, the AI has no knowledge of the artworld involved). A monkey could be tasked with selecting works for exhibition in a similar way, and we would certainly not wish to say that this monkey was acting on behalf of the artworld. Perhaps we do need the conferrer to be a participant in the artworld after all.

[12] I will discuss the possibility of AI action in Chapter 3, so I will not address it here.

Could an AI participate in a social institution? This would be difficult, and certainly not possible for current (artmaking) AI.[13] I will not rule out future social integration of AI, as this possibility forms the basis of many arguments about the rights and status of robots, though it will likely require artificial general intelligence to achieve true full social integration (see Jaynes, 2019). In Chapter 3, I will put forward the possibility that AI could draw upon humans to access the social world in some way.

If an AI cannot confer the appropriate status on its works, this may not actually matter too much. Unlike in other accounts of art,[14] in the institutional account, making an object into an artwork is not necessarily (or perhaps even typically) down to the creator of the work. For example, we could consider cases where the artworld initially rejected works that are now accepted by the artworld and widely believed to be art. Though we could debate at what point an object became art, the institutional account would suggest that the object is not art until it is more widely accepted. The onus for making an object art is not then just down to the creator; if an AI cannot confer the correct status on its own objects, it is not necessarily lesser than other (human) creators, who may also not confer art status on their own works.

We have considered the second condition in Dickie's initial instantiation of the institutional account of art. What about the first condition? The work of art must be an artefact. This may give us pause; what does Dickie mean by artefact? Dickie appears to adopt a permissive definition of artefact. In discussion of Weitz's example of a piece of driftwood, Dickie states in footnote that:

> A piece of driftwood which has become art might, for example, have been picked up, transported, and hung on a wall. It is in virtue of these things being done to it that the

[13] That is not to say that social relationships between humans and AI are not currently possible, even if it likely the case that these are parasocial relationships (see Pentina, Hancock & Xie, 2022). I am going to assume however that what is required to participate in a social institution goes far beyond one-to-one relationship building.

[14] See for example the historical account (discussed below) which centres on an artist's intentions (although it is not the only the creator who can make an object art under the historical account either).

> driftwood has become an artefact … Similarly, a piece of driftwood might be picked up and used as a weapon and, thereby, become an artifact" (Dickie, 1979, p. 472).[15]

This is contrary to the 'standard definition' of an artefact as laid out by Preston (2020). On this definition, artefacts must meet three conditions:

1) They must be produced with intention.

2) They must involve modified materials.

3) They must be made for a purpose.

Under the standard definition of an artefact, we would have a more stringent requirement for an AI to make art. Preston (2020) points out that some non-human animals could create an artefact under the standard definition if we have evidence that they have a certain level of cognition and that their behaviour is not merely instinctual. However, while human-ness is not a requirement, conditions of intention and 'having a purpose' would be a challenge for an AI system.

Despite insisting on artefactuality then, Dickie's characterisation of what makes an artefact is far weaker than the standard definition. Dickie's characterisation seems more like a requirement that an object has been acted upon in the some way (such as wielding the driftwood) as opposed to it being the product of an intention to make an object with a purpose (Hilpinen, 1992, p. 58).[16] We do not necessarily need a modification of materials at all, nothing needs to be (physically) produced, and thus nothing must be produced for a purpose or with intention. Surely *something* is occurring to transform the object into an artefact. In Dickie's explanation of the driftwood case it appears that actions have been done to the object, and in doing so the object becomes artefactual. The use of the object in a particular way has made it an artefact, where before it was not one. The action however does not need to be taken by the producer of the object itself. Could an AI undertake an action to confer art status on an object itself? In order for an AI to undertake this action (or any action) with a purpose, the AI would need to have some kind of agency. The question of whether an AI can have

[15] This footnote was only added in the 1979 reprint of the article in Rader's book.
[16] Dipbert (1993) would call this kind of use of an 'instrument' not an artefact at all.

agency is a complex one, and I will address this question in Chapter 3. There is a further complication here. As an AI system is itself an artefact, can it produce artefacts? It is possible that the products of autonomous AI systems should not be thought of as artefacts under the standard definition above and should rather be thought of more like instinctual products made by animals. As Preston (2020) states:

> Spider webs do have a purpose, for instance, and are clearly made rather than naturally occurring. But we may hesitate to attribute intention to the spiders, given the instinctive and rigid nature of their web-weaving behavior. (2020)

Some have argued that AI systems should be thought of more like animals (e.g., Darling, 2021) and, given that we do not (yet) have AI capable of human-level intelligence, we might be better to consider the products of AI to be like those of the spider. If we see higher levels of cognition in the future, then we may not have any reason to pause at the condition of artefactuality.

All of this ignores the human in the equation. If a human has designed an AI system with the intention that it will produce (say) images, by modifying paint or pixels, for the purpose of displaying them as artworks, then this intention can perhaps carry through to the final product, making the final product an artefact and not just the AI system itself. The product of an (autonomous) AI is one intentional level removed from the human as the human has not had a direct hand in the image itself; however, the production of *an* image was still intended, and the purpose (for display etc.) is still the same. This intention 'once removed' is not enough to violate the conditions to be considered an artefact under the standard definition.

As far as Dickie's artefactuality goes, we simply need something to be done to the product of our AI after it has been made for it to count as an artefact, and thus be eligible to be an artwork. This could be done by humans. Our programmers, curators, or human artists could take the products of our AI and transport them to galleries for display (just like the driftwood) thus creating an artefact and conferring art status upon the AI image.

Under Dickie's first version of the institutional account of art, it appears that (some) AI works can be considered as art. In a later version of the institutional account of art, Dickie proffers a

different definition: “A work of art is an artifact of a kind created to be presented to an artworld public” (2004, p. 53). This new definition retains the artefactual criteria discussed above but replaces the act of conferral of art status with what appears to be a kind of intention: the artefact must be created with the intention of presenting it to the artworld (Dickie, 2004, p. 51). There is an added emphasis on the role of the artist, and a separate role for the public who is the intended audience of the work (Dickie, 2004, p. 51). This then begins to look a lot more like the historical-intentional account of art (albeit without reference to prior artworks); as we have a kind of ‘intention to regard’ the created object, which is central to the historical account of art as proposed by Jerrold Levinson (1979). Let us now turn to the historical account of art.

The Historical Account of Art and AI

The historical theory of art was initially proposed by Jerrold Levinson (1979) as an alternative to the institutional theory. In developing the historical definition of art Levinson retains the underlying assumption of the institutional account: that being an artwork is not an intrinsic, exhibited property of an object (1979, p. 232). However, instead of an object's status as an artwork being conferred upon it by an overt act performed by a person (or institution) in the context of the artworld, the historical account of art centres on the *intention* of an independent individual (or individuals) (Levinson 1979, p. 232). The historical component of this definition comes from the nature of this intention: "My idea is roughly this: *a work of art is a thing intended for regard-as-a-work-of-art:* regard in any of the ways *works of art existing prior to it have been correctly regarded.*" (1979, p. 234). The historical component of this definition stems from specifying what it means to regard something as a work of art. To regard something as a work of art for Levinson means regarding it in the ways other works of art have been 'correctly' regarded. Levinson thinks that for something to be art it must be linked to prior artworks in some way:

> For a thing to be art it must be linked by its creator to the repository of art existing at the time, as opposed to being aligned by him with some abstracted template of required characteristics. What I am saying is that ultimately the concept of art has no content beyond what art *has been.* It is this content which must figure in a successful definition. (1979, p. 234)

Like the institutional theory of art, the historical approach aims to solve the problem that art does not seem to have a set of intrinsic features which we can specify. The central case that Levinson is concerned with is that of a creator of an artwork who is aware that she is making art: "In such cases making an art work is a conscious act involving a conception of art." (1979, p. 235). Without a standard concept of art specified in terms of intrinsic features, how can we find an understanding of art that all artists have? According to Levinson, the conception of art that any person, at any point in time or place can have in common is one that is relative to the art that has come before. Without this reference to prior artworks, "we fail to understand in what sense [the artist] is consciously or knowingly producing art." (1979, p. 235). A more formal version of this definition is as follows:

> (I) X is an art work = df X is an object which a person or persons, having the appropriate proprietary right over X, non- passingly intends for regard-as-a-work-or-art, i.e. regard in any way (or ways) in which prior art works are or were correctly (or standardly) regarded. (1979, p. 236)

As is clear, this definition is intentional, but not straightforwardly so. It is not sufficient for Levinson that a work is simply intended to be a work of art, as the intender's conception of art is inextricably linked from art that has come before. This is also how Levinson aims to avoid circularity in his definition, defining art at any one time by reference to the body of works that came before it (1979, p.240). Levinson does not directly address why he centres intention in his account; however, he does see intention's inclusion as having the key benefit of dispelling the intentional fallacy (the idea that the meaning of a work of art cannot be found in the intentions of the artist, see Wreen, 2014). Levinson claims that if an artist's intention is key to a work being art, then it cannot be easily dismissed as a source of understanding said work (Levinson 1979, pp. 246-7).

Let's examine whether AI can make art under this account. There are two key components which we can consider here. The first is the intentional, and the second is the historical. We can consider the role of intentions first. Having intentions is a challenge for AI. Intentions are often thought to require mental states (Setiya, 2022), and so to have them an AI would require a mind. Persuasive arguments against this possibility persist, most famously those of John Searle and his Chinese Room argument (1980), although there have been many convincing responses to Searle's arguments (see Cole, 2020 for an overview). I will examine Searle's argument in detail in my examination of the cluster account of art below, and I will examine AI mind in some further detail in Chapter 3. Regardless of whether AI *could* have mind in the future, the kinds of AI which are currently creating works that are candidates for art status do not have minds. If AI do not have minds, and minds are needed to have mental states, and intention is a mental state, then AI cannot have intentions.

Rather than relying on whether an AI can have mind, we could adopt alternative understandings of intentions. Some have tried to operationalise intentions for AI. For example, Cohen

and Levesque (1990) follow Bratman's (1987) functional analysis of intention and define intention as 'choice with commitment'. We could also adopt Daniel Dennett's approach to understanding intentionality (1971). In the original formulation, Dennett argues that intentionality is a tool for explaining the behaviour of others, and thus as long as it is useful to us to think of AI as having intentionality then we can describe this as an intentional system:

> The central epistemological claim of intentional systems theory is that when we treat each other as intentional systems, using attributions of beliefs and desires to govern our interactions and generate our anticipations, we are similarly finessing our ignorance of the details of the processes going on in each other's skulls (and in our own!) and relying, unconsciously, on the fact that to a remarkably good first approximation, people are rational. (Dennett, 2009, p. 5)

We could take such a stance to an AI that is producing artistic works. Theorising that the AI is making artworks because that is its goal, or saying that the AI wants to improve the images it makes, allows us to easily explain what it will do next: it will make more, and better, images (as far as the system is concerned). This is a fairly natural way to think about the system: "Uses of the intentional stance to explain the behavior of computers and other complex artifacts are not just common; they are universal and practically ineliminable." (Dennett, 2009, p. 7).

Under Dennett's view, it does not matter if there is 'real' intentionality in the system:

> The intentional stance works (when it does) whether or not the attributed goals are genuine or natural or 'really appreciated' by the so-called agent, and this tolerance is crucial to understanding how genuine goal-seeking could be established in the first place. Does the macromolecule really want to replicate itself? The intentional stance explains what is going on, regardless of how we answer that question. (Dennett, 2009, p. 9).

For Dennett then, we can talk about intentions without it mattering whether there is a mind to support them. This includes in AI or robotic systems:

> The robot poker player that bluffs its makers seems to be guided by internal states that function just as a human poker player's intentions do, and if that is not original intentionality, it is hard to say why not. (Dennett, 2009, p. 8)[17]

We could, therefore, take an intentional stance to our 'art-making' AI system and ascribe it with intentions regardless of whether it does or does not have a mind. We could then move on to the rest of the historical account of art. I suspect that this is the point where I might lose the reader, as the success or failure of the attribution of the label of 'art' to AI works becomes dependent on accepting a view of intentionality that is by no means uncontroversial. Perhaps we should revisit the historical account of art to see if there is any flexibility in the requirement of intention.

Is the intention of the human in selecting a training set the only way to ensure AI art is art under a historical account? In the institutional definition of art, the artist does not have to undertake the action of conferring art-status upon the object. This allows for found or ready-made objects to count as art, and it proved similarly helpful in allowing AI works to also be considered art, as long as a human was around to confer this status upon them. How does Levinson's account accommodate found art, or readymade art? To accommodate these kinds of works Levinson makes explicit the 'time-dependence' of an artwork under his definition:

> (I_t) X is an art work at t = df X is an object of which it is true at t that some person or persons having the appropriate proprietary right over X, non-passingly intends (or intended) X for regard-as-a-work-of-art, i.e. regard in any way (or ways) in which art works existing prior to t are or were correctly (or standardly) regarded. (1979, p. 238)

In this way, Levinson specifies that an object can become an artwork even if (1) it was not originally intended as one at the point of creation or by its creator, or (2) the way the creator intended the work to be regarded was not an acceptable way to regard art at the time of creation. In these two cases an object can still become a work of art at a later point in time. In the first case, an object (such as

[17] The robot player here would undoubtedly require some artificial intelligence.

Duchamp's *Fountain)* which was not created with the intention that it be regarded in the right way, can become art "after a certain intentional act has taken place." (1979, p. 239). That intentional act is that of an artist "appropriating" the object "with a certain intention" (1979, p. 238). In the second case, a 'naïve' creator could create an object which is intended for regard in a way that is *not* a 'correct' way to regard art at the time of creation. Over many years the 'correct' way to regard works of art expands and, at a later point in time, the original intention of the creator of the work is subsumed into this expanded understanding of correct-ways-to-regard-art. At this later point in time the work becomes art, "when the history of art, so to speak, catches up with what [the object]'s creator was engaged in." (1979, p. 239).

This response gives us a way for AI works to be considered art; however, as with the institutional account of art, it appears that we will need a human to undertake the act of appropriating the work with the correct intention (as in case one, above) before an AI work will be considered art. In this case, the AI itself will not be the creator of the work as an artwork, even though it is the creator of the object.

We are not necessarily stuck between this rock of relying on humans and the hard place of assuming machine intentionality, but we would have to abandon Levinson's version of the historical definition of art. Not all versions of the historical account of art insist rigidly on intention. Stecker presents a version of the historical account of art that does not necessarily require intentions:

> An item is a work of art at time *t,* where *t* is a time no earlier than the time at which the item is made, if and only if (a) either it is in one of the central art forms at *t* and is made with the intention of fulfilling a function art has at *t* or (b)it is an artifact that achieves excellence in fulfilling such a function, whether or not it is in a central art form and whether or not it was intended to fulfill such a function. (Stecker, 1997, p. 50).

This is a disjunctive-hybrid account, which combines a functional approach to art with a historical-intentional approach. Under this account of art, an AI might still create art even if it was not able to have intentions, as long as the work achieves excellence in realising a function that art has at the time

it was created.[18] Depending on what functions art has at the time, an AI could achieve this. For example, if a function of art is to be aesthetically pleasing then this might be quite achievable for an AI.

Let's set aside the need for intention now for a moment and return to the historical element of Levinson's account and its reference to 'prior works'. If we isolate the need for "regard in any way (or ways) in which prior art works are or were correctly (or standardly) regarded" we might be able to assess whether AI can reproduce the correct regard of prior works in its own works. If we set aside intention, an object that clearly relates to prior works of art may fulfil the historical component of the account. An AI can, for example, create a work which is in some ways derivative of prior works. Whilst these works may not be the most creative (I will address this in Chapter 2), if a work is produced that is clearly in the same vein as prior works, then it seems obvious that it would automatically be regarded in the right way. If I produced a work that is a painting of a landscape and my work was clearly in the style of David Hockney, we would surely think that it is going to be regarded in the way that prior works of art (i.e., Hockney paintings) are correctly (or standardly) regarded.

A similar case is occurring with AI. Some AI systems, and the work they produce, are clearly based upon prior artworks. GANs and CANs produce work that aims to meet learned properties of artworks. Depending on what the GANs and CANs are trained, this could mean an AI produces work through synthesising information from hundreds of thousands of artworks from large datasets such as *Wiki Art* or the *Google Arts and Culture* Catalogue. They can also be trained on specific types of art and can be trained to recognise different styles of art. This certainly seems to fulfil the non-intentional elements of the historical account of art: the images produced by a GAN are explicitly meant to follow the works it has been shown, and thus would be regarded in the correct way.

We might encounter a problem here. While the AI system may produce works that are informed by prior artworks, it is not clear that they can be said to have any knowledge of the history

[18] Note that the work would still need to count as an artefact.

of art. Given the emphasis Levinson places on the standard case of artmaking being "a conscious act involving a conception of art", we might think that one must have a conception of art in order to make it. We might therefore expect that some knowledge or understanding of what art is, informed by the history of art, is necessary for the creation of artworks. An AI system will struggle to have this if we take the requirement to mean that it should understand what it is doing. So, does a person need to have knowledge of art to fulfil this condition?

It appears that we do not need knowledge of art to fulfil Levinson's definition:

> my definition thus allows (via the art-unconscious intention) for an art-maker ignorant of all art works, all art activities, and all institutions of art. Such a person can be seen to make art if he intends his object for regard in a way which *happens to be,* unbeknown to him, in the repertory of aesthetic regards established at that time. In such a case there is the requisite link to the prior history of art, but it is one the art-maker is unaware of, though he has in fact forged it. (1979, p. 238)

Although this dispels the necessity for knowledge or understanding of what *art* is, it also raises another issue which I have thus far skipped over: the issue of regard. Would our AI need to have knowledge, or at least awareness, of ways in which something could be *regarded* in order to meet Levinson's definition? To fulfil Levinson's account of art we need the AI to intend its work to be regarded in the correct way. Even if our AI could have intentions, how can it intend for something to be *regarded* in the correct way? As we have already seen, we do not need to know the *correct* way to regard art at the time of creating the work. An artist does not need to have knowledge of art to make it, and at the time of creation the kind of regard in question does not have to be the correct regard (as what counts as correct can evolve). All this avoids is a knowledge problem in terms of the 'correct' regard, which again relates to art and artworlds, but tells us nothing about regard itself.

Can an AI system have knowledge or understanding of the way things are regarded? Initially, it seems that they could not, but let us take a closer look at what Levinson means by 'regard'. Levinson says, in response to the objection that his account is too inclusive:

> Something closer to a comprehensive way of regard properly brought to bear on, say, almost any easel painting, would be this constellation: {with attention to color, with attention to painterly detail, with awareness of stylistic features, with awareness of art-historical background, with sensitivity to formal structure and expressive effect, with an eye to representational seeing, with willingness to view patiently and sustainedly, ... }. Anything now intended for more or less this complex treatment or way of regard, we may be sanguine, would have difficulty not counting as an artwork. (Levinson, 1989, p. 24)

The intention then should not be just to regard the work in a simple sense (such as looking at it and contemplating its colours) but to regard it in this 'complete' sense (Levinson, 1989, p. 24). Even if an AI could have intention, in order to have the intention that a work be regarded in the right way, an AI would need one of two abilities: the AI would either need (a) to be able to regard objects in this complex way (which would surely require perceptual and analytic abilities, as well as various kinds of understanding) or (b) to be able to understand how others might regard the object in this complex way (which would surely need some kind of knowledge or theory of other minds, and the requisite understanding of how they might contemplate the object).

The list of capacities that an AI needs in order to make art (without human intervention) under Levinson's historical account is now looking rather long. We would need a capacity to have intentions, as well as perceptual and analytic abilities, or a theory of other minds and an understanding of perception and analysis. The bar set for AI is now very high; the art-making AI systems we have available currently could not clear this bar.

We could, however, argue that some of (a) above (i.e., perceptual and analytic abilities) is possible for an AI system. GAN and CAN systems, for example, learn from prior artworks and analyse their own outputs. The systems do not 'perceive' as we might, and it might be a stretch to say that this system is regarding its works in any aesthetic sense (it surely is not, for example, sensitive to expressive effect or painterliness) but it could attend to colour, form, and, in the case of the CAN (which is trained on a set of images which included style labels) stylistic features. It is not clear that the system could be said to 'understand' these features (and here we risk running into Searle again),

but for an AI system the kind of analysis of an image it can achieve could functionally be said to be 'regarding' the work in the appropriate way.

The requirements of (b) above (i.e., knowledge or theory of other minds) are more challenging. Here we cannot appeal to any current examples of AI in the domain of art to argue this is possible; however, we could leave open the door for an AI having these in the future. Indeed, computer scientists are attempting to develop AI systems with some 'theory of mind' (taken in a functionalist sense) (see Rabinowitz et al., 2018; Williams, Fiore & Jentsch, 2022), though this is not without challenges (Aru et al., 2022). This ability is thought to be a key component of developing truly collaborative AI systems. While a functional theory of mind may one day be possible for AI, it is not currently possible, and furthermore, is certainly does not exist in any current art-focussed AI. The kinds of system under development are also aimed at providing a functional theory of mind for AI that directly interact with people (e.g., Çelikok et al., 2019). It is another leap entirely for theory of mind to be possible in an AI system when the other mind is hypothetical, and not part of a direct interaction. This hypothetical person is what we would need to have a conception of in order to consider how they should regard the work of art we are creating. Despite this problem, we may not need both (a) and (b) above. If we are convinced that an AI could achieve the requisite capacities to be able to regard objects in a complex enough way (and if we agree with arguments for intention in computational systems) then we may allow that an AI system could create an object which it "intends for regard-as-a-work-or-art, i.e. regard in any way (or ways) in which prior art works are or were correctly (or standardly) regarded." (Levinson, 1979, p. 227).

Given the caveats in my analysis here, I imagine many will not agree that it is possible for AI to produce art under Levinson's account. However, these issues can broadly be avoided if (i) the system is not highly autonomous or (ii) we are willing to treat AI works as readymade objects which are then made into artworks by humans. In both cases, we must rely on a human to provide the requisite intentions in order to make the work art. In the second case, as I have discussed above, a human could simply treat the work of an AI system as a found object or a readymade. After creation, it could be appropriated with the relevant kind of intention and it would then be art. AI art would

perhaps be an unusual case in that part of its relationship to 'prior art works' would come from the AI and its learning and not merely the act of the human, but this would not be a problem for its status as an artwork. In the first case, where an AI system is not autonomous, we may be able to derive some intention from the human who makes or trains the system. When researchers or artists develop AI systems with the goal of producing art-like outputs, they are intending for the output of that system to be regarded in a "way (or ways) in which prior art works are or were correctly (or standardly) regarded". Although this intention has not come in the process of creating the individual work, we could consider this as a kind of 'derived' intentionality.[19] If the human training the AI system has a high level of control over the training data, then we could make a stronger claim. In this case, the human is not just intending that the work be regarded as art. They intend that the output of the system is regarded in the same way that the (prior) artworks in the training set are correctly regarded. In this case, it seems undeniable that the products of the system are art.

To summarise, under the historical account of art AI works may be art iff:

1) A person has appropriated the work with the relevant intention that it be regarded as other works have been regarded.

Or,

2) A person has created the system with the intention that the products of the system are regarded in the appropriate way.

Or,

3) The AI itself is capable of intention, the system is trained on prior artworks, and the system examines its products, analysing their features against prior works of art.

If none of these three cases applies, then the work of the AI is not art.

[19] This could be similar approach to that of intentionality derived through language, a concept that has been discussed at length in terms of AI intention, see Sayre (1986), Fodor (2009).

The Cluster Account of Art and AI

The cluster account of art was proposed by Gaut in response to the failure of aestheticians to settle on a single definition of art. The cluster account stems from a Wittgensteinian approach to the definition of art (Weitz, 1956). In this section, I will précis the Wittgensteinian approach to art, before presenting Gaut's cluster account. I will then assess how AI works measure up to this account of art, arguing (perhaps counterintuitively given its anti-definitionalist background) that the cluster account can offer a route for AI works to be art. I will also argue that the cluster account of art offers the best path for those who take a charitable view towards AI art (or other novel art forms and processes) due to its flexibility, ability to accommodate fringe cases, and its utility for developing AI further.

ANTI-DEFINITIONALISM

In response to the failures of prior definitions to successfully offer a definition of art[20] scholars in the 1950s turned to Wittgenstein's work to argue that 'art' could not be defined at all (Weitz, 1956; Ziff, 1953; Kennick, 1958). These philosophers argue for two key points: 1) That art cannot be defined (in terms of individually necessary and jointly sufficient conditions) and 2) that art is a concept best characterised in terms of family resemblance (Gaut, 2000). Weitz's Wittgensteinian approach has remained the most influential, so I will focus here on his proposal.

In his 1956 paper Weitz argues that the concept of art is like the concept of games. Wittgenstein says of games:

> Consider, for example, the activities that we call "games". I mean board-games, card-games, ball-games, athletic games, and so on. What is common to them all? — Don't say: "They *must* have something in common, or they would not be called 'games'" — but *look and see* whether there is anything common to all. — For if you look at them, you won't see something

[20] Weitz includes formalist (e.g. Bell) 'Emotionalist' (e.g. Tolstoy), 'Intuitionist' (e.g. Croce), 'Organicist' (e.g. Bradley) and 'Voluntarist' (e.g. Parker) attempts to define art (1956, 28-30).

> that is common to *all*, but similarities, affinities, and a whole series of them at that. (Wittgenstein, 2009, §66)

Wittgenstein points out that when we point to a feature of a game that appears to be shared with other games, we can quickly find exceptions. Some games are fun, such as ball-games and card-games, but others, like chess, are serious. Some games involve both winners and losers, like chess and football, but others, like a friendly game of catch, do not. Many games appear to involve skill, but a childhood game like ring-a-ring-a-roses does not involve much skill at all (bar a little co-ordination). Wittgenstein likens these to traits shared amongst family members:

> I can think of no better expression to characterize these similarities than "family resemblances"; for the various resemblances between members of a family — build, features, colour of eyes, gait, temperament, and so on and so forth — overlap and criss-cross in the same way. — And I shall say: 'games' form a family. (Wittgenstein, 2009, §67)

Weitz takes this approach and applies it to art, stating "if we actually look and see what it is that we call 'art', we will also find no common properties—only strands of similarities" (Weitz, 1956, p. 31). Indeed, if we think of the variety of things we call works of art, we can see what he means.

Many works of art have positive aesthetic features (such as a Bernini sculpture, or the work of Klimt) but some works are unattractive or ugly (e.g., depictions of war by Otto Dix) or are not meant to be considered aesthetically (according to Adajian, 2018, this includes Duchamp's *Fountain,* 1917). Many works of art deal with emotions (think of a Hollywood melodrama or a suspenseful novel), but a minimalist work of art seems not to at all (e.g., Donald Judd's *Untitled,* 1972). We could think of many works of art that have required a great deal of skill to create (e.g., the philharmonic orchestra performing or a hyperreal painting by Chuck Close), but others require little to no skill (e.g. John Cage's *4'33"* or Duchamp's readymades). It would seem to be difficult to point to one feature that all these disparate works share, and Weitz argues that, just like Wittgenstein's games, there simply is not any one feature that these works share. Weitz goes on to describe art, like games, as an 'open concept'. According to Weitz, a concept is open if there could be cases where the application of the

concept is not determined. Such a case would require us to make a decision about whether to include the case under the concept or to make a new concept. Art does seem like this. If we think of the advent of photography or film, these were not immediately accepted as art. In the last twenty years, we have also seen this process play out with video games (e.g., Ebert, 2010 on the negative, Smuts, 2005 and Tavinor, 2011 on the positive). Weitz claims that open concepts cannot be defined in terms of "some manifest or latent essence", but "certain (paradigm) cases can be given, about which there can be no question as to their being correctly described as 'art' or 'game,' but no exhaustive set of cases can be given." (Weitz, 1956, p. 31). Under this Wittgensteinian perspective then, we cannot define art in terms of necessary and sufficient conditions, but we can offer some paradigm cases that are clearly examples of art.

This anti-definitionalist approach to the concept of art did not initially take hold in aesthetics and was roundly rejected at the time (Gaut, 2000, p. 25; Carroll, 1993, p. 315). Various objections were put forward. Davies (1991) offers a comprehensive summary of these objections. Here, I will just offer a selection. First, it was objected that various proponents of the Wittgensteinian approach misunderstood Wittgenstein (e.g., Tilghman, 1973; Sclafani, 1971) or that Wittgenstein's family resemblance approach has fundamental issues in the first place (Mandelbaum, 1965). Resemblance-to-paradigm (which Weitz claims will help elucidate the concept of art) has been criticised as incomplete, as without the paradigm cases to which we can compare other works it has little utility (Gaut, 2000, p. 25). Furthermore, the family resemblance approach has been criticised as vacuous, as anything could resemble anything else (to some extent), and we have no way to discern which resemblances are relevant (Gaut, 2000, p. 25). The underlying motivation for a non-definitionalist approach to art was the failure to find a satisfactory definition of art, but this failure may have an alternate explanation: that philosophers were simply looking in the wrong place (Davies, 1991, p. 22). Weitz (and others) who put forward Wittgensteinian approaches to art did influence the focus of aestheticians looking for a definition, and it became more accepted that:

> Artworks do not form a natural kind; typically artworks are manufactured with the specific intention that they be artworks. And the variety of the forms and categories of art suggest that

> there will be no intrinsic, exhibited property that all and only artworks share (Davies, 1991, p. 37)

As Davies points out, philosophers turned away from aesthetic features and towards relational definitions (such as functional and procedural definitions) after these anti-definitionalist efforts. This was not the end of the Wittgensteinian approach to art though, as in 2000 Gaut revisited this in the form of a cluster account.

THE CLUSTER ACCOUNT

The perceived failures of contemporary definitions of art (particularly a failure to garner any broad consensus amongst philosophers) led Gaut to propose that aesthetics should revisit the idea that art cannot be defined. Instead of resemblance-to-paradigm as the model for the concept of art, however, Gaut turns to a 'cluster account' construal of family resemblance (Gaut, 2000, p. 26). This avoids the problems of incompleteness and vacuousness that cause issues for Weitz's approach. The problem of incompleteness (due to a reference to paradigm cases that were not provided) is avoided because paradigm cases are not part of the account. The problem of vacuousness (that anything can resemble anything else, so resemblance to paradigm tells us nothing useful about art), is also avoided as specific criteria are provided (Gaut, 2005, p. 275).

The cluster version of family resemblance that Gaut adopts comes from Wittgenstein's discussion of proper names and was further developed by Searle. As Gaut writes,

> A cluster account is true of a concept just in case there are properties whose instantiation by an object counts as a matter of conceptual necessity toward an object's falling under a concept… There are several [properties] criteria for a concept. (Gaut, 2000, p. 26)

The way properties count towards the concept are as follows:

> 1) If *all* properties are instantiated in an object, then the concept applies to it. If *fewer* than all criteria are instantiated, this is sufficient for the application of the concept.

> 2) There is no one property which everything falling under the concept must have.
>
> 3) There are no individually necessary conditions for the application of the concept, but there are disjunctively necessary conditions. It must be true that some of the criteria apply if an object falls under the concept. (Gaut, 2000, p. 27)

In the case of art then, there will be a list of properties whereby:

i) If all properties are fulfilled by an object, it must be art, but if only some of the properties are fulfilled by the object, it can still be art.

ii) There is not one property which all artworks will have.

iii) For an object to be art, it must have at least some of the properties; it can't have none of the properties, and still be art (though, in the case of art, Gaut states we might not yet know all of these properties).

Gaut's cluster concept of art is not a *true* cluster concept, however (even under the conditions that Gaut himself lays out). Gaut adds a necessary condition to his account, thereby violating condition 2). Gaut adds that given it is art*works* in which we are interested, some action must have taken place at the genesis of something's being art:

> An artwork is the product of an action, preeminently of a making (an artifact), or a performing (a performance). It is art*works* that are involved here, since something is in each case done. Hence being the product of an action is the genus of the artwork and is thus a necessary condition for something's being art. (Gaut, 2000, p. 29).

It is worth noting here that Gaut's account of 'making' or 'performing' is still quite permissive. An object does not have to be made with intent (as we might expect on the standard view of an artefact, see Preston, 2020), as the account can still permit found art:

> It might be thought that this is denied by those who acknowledge the existence of found art, but in fact it is not. Such art is selected, and selection is an action. Selection adds to the range of properties that can be possessed by objects, and thus alters them, even is not physically. A

> piece of driftwood in nature cannot express despair, nor can it be about anything (since it lacks even derived intentionality), but when selected for display in a gallery it can express desuetude and be about failure and decay (Gaut, 2000, p. 29)

Gaut argues that this modification to the cluster account does not compromise the nature of the account, as the necessary condition is minimal:

> Being the product of an action is, however, a very thin generic condition, which does not distinguish artworks from any of the other products of action (philosophy, papers, chairs, pay freezes, angry words etc.) … thus the modified cluster account holds that there is one necessary condition for something's being an artworks, but that is because of the notion of a work (the product of action) rather than because of the notion of art. (Gaut, 2000, p. 29)

Here I include these lengthy quotations because, as we shall see soon, the exact formulation of this necessary criterion will be particularly relevant when applying this account to AI.

Gaut proffers 10 properties which may count towards something being a work of art:[21]

1) possessing positive aesthetic properties, such as being beautiful, graceful, or elegant (properties which ground the capacity to give sensuous pleasure)

2) being expressive of emotion

3) being intellectually challenging (i.e., questioning received views and modes of thought)

4) being formally complex and coherent

5) having a capacity to convey complex meanings

6) exhibiting an individual point of view

7) being an exercise of creative imagination (being original)

[21] I have edited this down and made it into a list.

8) being an artefact or performance, which is the product of a high degree of skill

9) belonging to an established artistic form (music, painting, film, etc.)

10) being the product of an intention to make a work of art

(Gaut, 2000, p. 28)

I will now take the cluster account and use it to evaluate possible AI works.

AI & THE CLUSTER ACCOUNT OF ART

In order to evaluate AI works against the cluster account of art, I will begin with the list of 10 properties. I will go through this list of properties in Gaut's account and propose how AI works could or could not meet them.

i. possessing positive aesthetic qualities e.g., being beautiful, graceful, or elegant

Possessing positive aesthetic qualities does not seem to be a barrier for AI works. There is no reason to think that AI images couldn't be beautiful, even if you think the majority are not. Finding the work to have positive aesthetic qualities likely led to the AI work winning a prize at the Colorado State Fair (Roose, 2022) and works made with AI have recently been described as "gorgeous" and "beautiful" in *The Guardian* (Jones, 2022).

ii. being expressive of emotion

Whether a work made by an AI could be expressive of emotion will come down to exactly what is meant by 'being expressive'. Jenefer Robinson has provided us with an account of what it means for a work to have expressive qualities:

> I would like to suggest that we should confine the term 'expressive quality' to those qualities in an artwork (or other things, such as merry brooks and anguished old oak trees) that are not only named by an emotion word but also arouse appropriate emotions. More particularly,

> *expressive qualities* are *qualities that can be grasped through the emotions that they arouse*. (J Robinson, 2005, pp. 291-292)

If we adopt Robinson's understanding of expressive qualities, it does not seem impossible for an AI to produce works that have such qualities, i.e., works with qualities that can be named by an emotion word and arouse appropriate emotions. If an oak tree can have expressive qualities in this way, why not a work by an AI? What counts for Robinson is the reception of the work by the audience. She states that expressive qualities should "evoke corresponding emotions in audiences" which can alert the audience the presence of said qualities (J Robinson, 2005, p. 292). We can already see these kinds of expressive qualities being described in works that are made with AI. For example, Jones (2022) describes the work of Gillian Wearing, who utilised DALL-E 2, saying: "You get a sense of loneliness and anguish, crying from inside to outside, soul to soul." (Jones, 2022). Will this sense of expressiveness be sufficient for Gaut?

When discussing exceptions to each of the properties, Gaut states "(2) much of architecture and music is not concerned with the expression of emotion" (Gaut, 2000, p. 33). It seems then that the work must be 'concerned with' the expression of emotion, not just have expressive qualities. It is unlikely that an AI will itself be concerned with the expression of emotion, as this is not a feature of AI systems. However, that does not mean that a work made by an AI could not *seem* to express emotion, or to give viewers an idea of what it is like to experience an emotion, nor that an AI system could not be designed with emotion in mind.

At this stage of AI development, we are not seeing anything like emotional states being possible in AI. However, there have been attempts to add an emotional component to machine images, such as the *emotionally aware Painting Fool* developed by Colton and colleagues (2008) that I discussed in the introduction. We are also seeing so-called 'empathetic algorithms' which are designed to be sensitive to the emotional state of their users (Raamkumar & Yang, 2022), and systems which researchers claim can identify emotional responses to images, particularly artworks, and express these in language (Achlioptas et al., 2021). Finally, there are AI-based art-making systems that claim to be able to create 'emotional' art, such as AIVA, an AI that produces "emotional

soundtrack music" (AIVA, 2022). Whether this system is successful in creating expressive music is debatable, but this certainly seems like it is creating works that are 'concerned with the expression of emotion'. There is the added issue of where this concern comes from, and if it needs to come from the system rather than the designers of the system. Emotional or expressive considerations could plausibly be integrated into future AI art-making systems, particularly if this property is seen as key for art.[22]

So, if we want our AI to be expressing an emotion in making a work, this second property of art is not possible for current AI. This property may, however, be possible if we just want a viewer to find the work to be expressive, or if we allow 'emotionally concerned' works made by non-feeling AI to count under this property.

iii. being intellectually challenging (i.e., questioning received views and modes of thought)

Works of art made by AI could be intellectually challenging to some extent. It may challenge us to wrap our heads around artworks being made by an AI, the images themselves may seem confusing and we may try to glean information about the training data from the images we see. For example, the work of Google DeepDream can challenge us to guess what the system was trained on, understand how the system is recognising patterns, and consider what exactly this means for what the system 'sees' (Mordvintsec, Olah & Tyka, 2015). However, with AI works, much of the challenge will come from the fact that the work is made by an AI. As such, no autonomous AI system is likely to make works which are 'questioning received views and modes of thought' without you needing to know that the work is an AI work, as that is a key factor in the work's intellectual challenge. These AI systems do not have views or modes of thinking, and thus are not going to be able to question these through their work.

iv. being formally complex and coherent

[22] As I will argue later, the cluster account allows us to offer recommendations for the future development of AI for art.

There is no reason to think that AI could not produce formally complex works, particularly in the visual domain. Images from GANs, CAN, DALL-E and even older systems like DeepDream can produce colourful, multi-feature images. We have some evidence that AI can produce images that are formally complex, in comparison to some human-made artworks. Elgammal and colleagues (2017) conducted some initial interrogation into the responses to the images produced by their CAN. I will draw on several of their findings during discussion of these properties. Elgammal et al. (2017) presented participants with images from the CAN versus abstract expressionist works and works displayed in *Art Basel 2016*.[23] In this study, participants were asked to rate the complexity of the images. Those images produced by AI systems were rated as higher on complexity than the human artworks in the study. The difference between the sets was not large, but it may be sufficient for our purposes that the AI images were as complex as the human-made works. (Elgammal et al., 2017, pp. 16-17).

Formal *coherence* on the other hand may be more difficult for some AI. Though a system thoroughly trained on a narrow dataset will produce coherent, and seemingly representational,[24] images with frequency, a common way for AI systems to fail is to not produce a coherent image of a target depicted object (I discuss these kinds of outputs in Chapter 5). With increased diversity in the training set, or with a system designed to differ somewhat from the training set (such as AICAN), the images are not always clearly *of* something. Can they still be formally coherent? The same study by Elgammal et al., discussed above, also asked participants to rate their agreement with the statement: "As I interact with this painting, I start to see a structure emerging." (2017, pp. 17-18) Participants scored both the images produced by the GAN and CAN systems in the study higher than the human-art datasets. Seeing a structure emerging could point to formal coherence (in images); but the AI images being rated higher on this question than human-made images does not mean that they possess

[23] It should be noted that this study has limitations. The study was conducted online, most of the questions were asked to non-experts, and not all differences were investigated for significance.
[24] I use the term 'representational' here simply to mean the images clearly depict a recognisable object.

this property. Perhaps these art datasets are just not the kind to be 'formally complex and coherent'(as is expected by the cluster account, some works of art won't have this property).

Images produced by DALL-E 2 seem (at the moment) to be more consistent in terms of producing convincing coherence and complexity. These images are generated in response to human text prompts and are based on a probabilistic model. In part because of this probabilistic basis, DALL-E 2 is able to generate images which are structured in a coherent way. DALL-E does still make mistakes related to coherence, however (Romero, 2022). For our purposes here though, it is sufficient that AI *could* produce coherent and complex works, and thus could produce works that meet this property under the cluster account.

v. having a capacity to convey complex meanings

If conveying meaning involves communicating meaning from the artist to the viewer, it would appear that most art-making AI systems cannot do this. To do this, the AI itself would have to have a meaning in mind, which is not possible in the AI systems that we are examining, particularly generative algorithms. The more autonomous a system is (at least until we have systems capable of understanding) the less likely it is that the works have the capacity to convey *any* meaning, let alone complex meanings.

If we do not need the system itself to generate the meaning of a work itself, then it seems this property is already possible for some AI (depending on how complex the meaning is). For example, a system like DALL-E can produce a relevant image when it is given a text prompt. In producing this image, the complex architecture of the system effectively takes the prompt, represents its meaning in a mathematical space, and interprets that meaning in an image. In this way then, the system is quite literally conveying meaning from one mode (text) to another (image). If successful, the image does indeed represent the meaning of the prompt (OpenAI, 2022a). Whilst this kind of visual representation of meaning does not exhaustively capture what Gaut perhaps means by 'capacity to convey complex meanings', it could be a case of conveying meanings to a viewer. Given that the prompts for producing images in DALL-E are provided by humans, the system itself is certainly not

conveying any meaning that it wants to. It *can* convey the meanings given to it by a human, but it cannot come up with those ideas for *what* to convey independently; it is reliant on a human user inputting text into the system. To use an example that is not a visual work, a written piece by an AI system such as GPT-3 could also include complex meanings that could be understood by a reader. However, this meaning is not connected to anything the AI system aims to convey; it is, in this case, produced predictively in response to a human prompt, much like elements of DALL-E.

AI works taken broadly could have a capacity to convey some complex meanings to an audience, particularly if they are prompt-based systems. If we require the meanings to come from the system itself though, this will not be possible for current AI.

vi. exhibiting an individual point of view

Exhibiting an individual point of view may not seem very likely with an AI system that is producing artworks. First, it seems odd to say that they have a 'point of view', which we would typically use to indicate having a mind. To have a point of view to exhibit we would expect something to have a perspective on the world, even to have thoughts, feelings, and opinions on the world. All of these require a mind, and these AI systems do not have a mind. Second, AI systems are rarely thought of as individuals. This is partly due to their constant changes. Most AI systems employ machine learning (ML) algorithms. ML algorithms change constantly, with each round of learning resulting in alterations to the system. For example, in GANs and CAN, the process of learning alters weights in the system to improve future images. An image produced at an early stage of this process will be different to one produced later in the process. One might contend that humans too change, particularly through learning; artists will produce a variety of works over their careers, so how could we say they exhibit a single point of view? Despite this analogy, there is still a constancy that we cannot ignore about a human: they are one being, with a single physical manifestation. Furthermore, changes to humans occur slowly, and in response to environmental stimuli. Humans are embodied. I will address whether and how embodiment may exist in the kinds of AI systems that are generating works in Chapter 3.

However, if we take 'individual point of view' to encompass producing images in a single style, AI may be able to achieve this. We can look at the work of a narrowly trained AI system, such as the GAN Obvious used to produce *Portrait of Edmond de Belamy* (the algorithm was trained and uploaded online by programmer/artist Robbie Barrat). This system produced multiple similar images, which look like a person with indefinite features seen through textured glass (Christie's, 2018b).[25] Similar images were able to be reproduced using Barrat's algorithm, which suggests that the algorithm developed and trained by Barrat had a particular style and subject it repeatedly reproduced in images. Unfortunately, though, this AI system has been trained to reproduce the kinds of portraits it has been trained on. The fact that it produces multiple similar images is indicative of the narrow 'experience' that the AI has with images. It is not clear that this should count as a 'point of view'. Furthermore, the selection of the training images is down to the human(s) training the system, in this case Barrat. This is how people exercise control over the images produced by generative AI systems; a consistent aesthetic, and particularly a consistent subject-matter, is not evidence of the AI system having a point of view as much as the human having a specific aim for the system.[26]

It is unlikely that an AI will be able to achieve this criterion, unless we consider "exhibiting an individual point of view" to be satisfied by 'exhibiting consistency in style'.

vii. being an exercise of creative imagination (being original)

Gaut offers two formulations of this property. The second, 'being original', should not be a barrier for AI systems producing works. AI systems can produce original works. This originality may be limited to the relevant domain, for example, a GAN is not going to produce anything that is not a digital image, this system does produce novel images. It is worth noting that some generative AI systems may indeed have a problem with originality due to data over-fitting. This is where the system becomes too dependent on the training data, typically because there is not enough of it (Feng et al.,

[25] It is not clear how much of the appearance of these images is due to selection, processing and printing by Obvious.

[26] See Chapter 5 for the way in which a distinctive aesthetic may not be down to training, but rather down to failure.

2022). When this happens in a GAN, it can result in the system reproducing images it has seen before. (Webster et al., 2019). A GAN which is not producing any novelty is a red flag for training issues. In the case of GANs, there are techniques to test for this replication and "Even when overfitting, most models will not reproduce perfect or trivially transformed copies of the training data" (Theis, Oord & Bethge, 2015, p. 6). Images produced by AI can also *seem* original to us. Again, we can turn to Elgammal et al. (2017). In one of their experiments, AI images were judged to be more original than the human artworks by participants. This originality is by design, particularly in the CAN wherein the system is designed to differ stylistically from training images.

Imagination on the other hand is more difficulty for an AI system. I expect that we would anticipate having a mind (and mental representations) as central to imagination. If this is the case, AI are not going to be able to achieve this property, as the kinds of systems we have now (and for the near future) will not have minds. Despite the numerous concepts to which 'imagination' seems to refer, Liao and Gendler state that broadly:

> *To imagine* is to represent without aiming at things as they actually, presently, and subjectively are. One can use imagination to represent possibilities other than the actual, to represent times other than the present, and to represent perspectives other than one's own. (Liao & Gendler, 2020)

Generative algorithms can represent things:

> In generative modeling, methods like variational autoencoders (VAEs) [KW13] and generative adversarial networks (GANs) [GPAM∗14] create latent spaces for modeling and synthesis of data [CN17]. Despite arising from different algorithms serving distinct purposes, these representations are all vector spaces of reduced dimensionality (relative to the input), intended to produce more general features that helpfully characterize the input. These representations are often referred to as latent spaces. (Liu et al., 2019)

Diffusion models have similar representational spaces. The latent space is constituted by 'representations' of data on which the system has been trained. It does not contain all the data but a

compressed version of that data. It additionally represents relationships between the data that it has been trained on: “The concept of ‘latent space’ is important because its utility is at the core of ‘deep learning’ — learning the features of data and simplifying data representations for the purpose of finding patterns.” (Tiu, 2020). This might be seen as akin to a form of imagination (though, perhaps not the creative kind). Rather, it might be more like a representation of something one has seen before (like picturing scenes from a film, and how they relate to each other). Given that this latent space is representative of the training data an AI is exposed to during learning, we might say that this is just representing things “as they actually, presently, and subjectively are” even though this kind of latent space is key to generating *new* images.

Despite this, some of the images produced by AI systems certainly seem fantastical. Take the examples made available by OpenAI on their DALL-E 2 Demo (OpenAI, 2022a): if DALL-E 2 can produce an image of “Teddy bears shopping for groceries in the style of Ukiyo-e”, then should this not count as ‘not aiming at things as they actually are’? Whether we might consider these representations as (machine) imaginings will be dependent on how we want to characterise the pattern-finding in the latent space and, given the novel images that are generated from the latent space, if we are willing to call these ‘possibilities other than the actual’.

The focus of my response so far has been on ‘imagination’, but what of ‘creative’? In light of Gaut’s more recent work on creativity, the property of ‘being an exercise of creative imagination’ might actually be akin to ‘being an exercise of creativity’. Gaut states in a later work on the link between imagination and creativity: “we have defended the existence of an a priori constitutive connection between imagination and creativity: imagination is suited of its nature to be the vehicle of active creativity.” (Gaut, 2003, p. 289). Exercising creative imagination then could be equivalent to being creative in the cluster account. I will go on to discuss creativity in Chapter 2.

viii. being an artefact or performance which is the product of a high degree of skill

If we are only to take skills as the kinds of things humans are skilled at (like wielding a paintbrush with precision, depicting an object accurately, or playing a piano well), then these AI systems are not

skilled.[27] They may, however, become skilled in the sense of gaining the ability to do something well (e.g., generate images). We could characterise machine learning as a process of perfecting a skill for an AI system, as the AI is improving in its ability to produce images within certain parameters of success. In the case of a GAN system, the generator in the system generates images, which are judged by the discriminator. Feedback from the discriminator goes into the generative process, improving the images in the next round of generation.[28]

This property in the list does specify that the item must be 'an artefact or performance'. For the artefact condition here, we are in the same position as with the institutional account of art. I will not rehash that argument here.

ix. belonging to an established artistic form

I do not think this is too much of a stretch for AI. First, the majority of works considered here are digital images. Digital images are an established form of art, and they may also be emulating another established artistic form (paintings, photographs etc.). Second, AI is trained on other artworks, and one of AI's clear limitations (in terms of creativity) is the extent to which it could ever go beyond the kind of works on which it has been trained [see Chapter 2]. Systems such as GANs are designed to produce images that are a believable part of the set of images on which it has been trained, and predictive/probabilistic systems (such as DALL-E, GPT-n, AIVA etc.) are designed to continue on with what they have been prompted with in an expected way, also based on training data. Neither of these kinds of systems are going to produce works that are vastly different from the works on which they have been trained. This is not to say that an AI will not *ever* make works that do not belong to an established artistic form and depending how narrowly we take 'artistic form' to be we may see AI systems produce different kinds of works (e.g., a new form of narrative work, or a new style of digital painting). However, it seems likely that most successful art-making AI will be producing works that belong to an established form. One might object that the fact that these are AI works in itself makes

[27] Though some AI systems may be able to perform some tasks with greater speed and accuracy than humans (think of *AlphaGo*).
[28] I examine skill further under Gaut's account of creativity in Chapter 2.

them another form of art. I am not convinced that this is a problem, as the focus in Gaut's property appears to be *form,* and not production, which would suggest a focus on the perceptible properties of the work.

x. being the product of an intention to make a work of art

Discussing this property would largely replicate the discussion of intention in the historical account of art (see above), so I will not reproduce the same discussion here in detail. A work 'being the product of an intention to make a work of art' of course requires intention. To have intention is often thought to require a capacity for mental states and with them a mind. If AI needs to have a mind in order to have intentions, then this will either not be possible at all (if we agree with Searle, 1980) or at least not possible in current AI systems. If we do not think that an AI needs to have a mind to have intentions, and instead substitute something like 'aiming at a purpose' for intentionality, then some AI could achieve this criterion. This is similar to the approach I take in the discussion of agency (see Chapter 3).

We could also adopt Dennett's proposal of the intentional stance, whereby the 'truth' of the matter about a system's capacity does not matter. All that matters is the utility of adopting an assumption that the system is intentional. In order for this approach to be useful, we would need to know if people do indeed adopt an intentional stance to AI art. We do have some (albeit limited) evidence that people take an intentional stance to artworks made by some AI systems. Again, in Elgammal et al.'s (2017) study, participants were asked how much they agree with the statement: "As I interact with this painting, I start to see the artist's intentionality: it looks like it was composed very intentionally." (2017, p. 17). With this question, participants score the GAN and CAN images higher (more agreement with the statement) than the human artworks. There are limitations to this question. The clear issue for my purposes is that this is not asking whether the image appears to be intentionally a work of art. It is likely that this question is addressing formal qualities (I imagine a Jackson Pollock work would score low, yet it is clearly intended to be a work of art). However, if the works appear intentional in some way, this suggests that participants are taking an intentional stance towards these images.

Should apparent intentionality in a work of art count towards "being the product of an intention to make a work of art"? The work is the same, whether it is made by a human or an AI system. If it turns out that the work is made by an AI, does that mean that is not the product of an intention to make a work of art? If we take an intentional stance towards AI, then no; it would still be an intentional work of art. There is a clear objection here of course: many people do not sign up to a Dennettian view of intention.

Aside from this debate, AI art could utilise some intention derived from humans. There are two possible routes to this that were discussed previously:

1. In the creation and training of the AI: for a domain-specific AI, a human has designed and trained the system for a purpose. AI systems that produce works akin to art will often be designed and trained with art in mind. In this case, the designer/engineer/artist is intending that whatever the AI produces will be art. The works produced by the system then are the product of the intention to make a work of art, even if there is an additional step in this causal chain.

2. In the selection of outputs from the AI: this is similar to found art or readymade artworks. In selecting the work made by an AI, a person may be, through that action, be intending to produce a work of art. The work would not count under this approach until this intentional action has been undertaken by a human, however.

If we are happy to utilise intentionality derived from humans, then this property is not a problem for works produced by AI. At higher levels of autonomy, however, we may be less willing to derive intentions from a human as in (1) (through the creation and training of the AI), due to a responsibility gap (as discussed in the introduction). If we are looking for an intention to produce a work of art in the AI system itself, this property is considerably more challenging, and the possibility of AI achieving this will be contingent on several factors, namely: the complexity of the AI, whether we think AI mind is possible, whether we are willing to accept a minimal account of intention, or whether we are willing to accept Dennett's view.

ASSESSING THE CLUSTER PROPERTIES FOR AI

As I have shown, we can examine the list of properties Gaut provides us in the cluster account and assess whether AI can meet each in turn. By my assessment, several of these are possible, and more are possible with caveats (such as how we define elements of the property). Even more properties are possible as soon as we include humans in the equation. I do not wish to adjudicate on whether there is enough in my list to consider (some) AI art to be Art, as I suspect there are too many ambiguities in my assessments, and we have not yet considered the one necessary criteria Gaut adds to his account (which I will turn to shortly).

I hope to have shown through my assessment of these properties how the cluster account of art might be beneficial for those interested in the possibility of AI Art: the cluster account provides a list of properties that we can discuss, debate the possibility of, and potentially consider for operationalisation. By this I mean that we can take a property and consider whether 1) current AI can do it, 2) future AI could do it, and 3) how we might make this possible for an AI. For example, we could prioritise formal coherence in future AI systems. To do this, we might want to focus on developing links between a structured visual system and image generation in our AI systems (see Chapter 5 and the discussion of weirdness). The cluster account can give us more to work with than other accounts of art in developing future AI art-making systems.

ADDITIONAL PROPERTIES

The consideration in this chapter has been dependent on the idea that we know either the definition of art or, in the case of the cluster account, a list of properties art may possess, and thus all AI needs to do to make art is produce something that will measure up to these accounts or produce something in the correct way to meet the requirements of these accounts. There is a fundamental assumption underlying this, however, which the cluster account allows us to bring to the fore. That assumption is that when AI makes art, it will be like human art, so much so that it will be recognisable by the same metric that we apply to human art. There is a reason for this. The motivation for this investigation is

uncertainty about whether AI could produce art, and considerable scepticism about the possibility that it ever could. There is an intuition for some, and more than an intuition for others, that art is a *human* endeavour. This framing is anthropocentric and ignores the possibility (maybe even a slim one) that whatever AI does will be quite different from what humans do. This is believed to be increasingly likely as we edge towards autonomy and beyond into human-level and beyond-human-level intelligence in AI (see e.g., Bostrom, 2014). AI may well have different values to us, including in the realm of the arts (I will discuss this possibility in a Chapter 4), and these values may or may not end up appealing to humans.

There is a benefit here to the cluster account: it can be adapted. Gaut states that individual criteria can be disputed (Gaut, 2000, p. 29). Furthermore, if an example could be found of a missing property that should be on the list, or a counterexample of a case that seems to be art but does not meet any of the criteria, then a new or missing property could be added:

> There is no evident way that an object lacking all of the criteria could be a work of art; and even if a plausible counterexample could be produced, the friend of the cluster account could respond by *adding whatever seems like the relevant criterion to the cluster* — that is, she can respond by modifying the intent of the account, rather than its form. (Gaut, 2000, pp. 32-33, my emphasis)

Taken together, this suggests that we could make a case to add properties to this list where we have a counterexample case that *seems* to be art but does not meet enough of the properties in the list, and yet clearly has other properties that might be added to the list (without altering the form of the cluster account itself). In the case of an AI artwork that does not meet all, or even most of the properties in Gaut's account, there is a potential for us to adapt the account to alter existing properties or propose *new* properties that are unearthed by AI works.

A foreseeable case could be, for example, to add 'information' to the property of 'having a capacity to convey complex meanings'. 'Having a capacity to convey complex meanings or impart complex information' could then include AI works' ability to convey probabilistic information about

artworks (as we see with DALL-E) or the ability to convey aggregate information about traits of a training set of images (as with GANs). A DALL-E image can tell us something about a large dataset that other forms of data visualisation or mere statistical analysis might not. For example, from discussion with artists using DALL-E 2,[29] DALL-E 2 struggles to depict a man wearing a dress and will opt for depicting a woman in a dress, or a man next to a dress. We could report on the percentage of images in the training data set that depict men in dresses, or the ratio of men to women depicted in dresses in the training data set, but this does not compare to the information we could gain from just looking at DALL-E images. Seeing a system that can depict 'An astronaut riding a horse in the style of Andy Warhol' be unable to depict *a man in a dress*, reveals bias in the data, and the impact of that bias in one fell swoop. You do not have to agree that this is something that should count as a property of art (and perhaps by virtue of it being by an AI, you won't), but we can see in this example how an AI artwork might offer something different to human art.

These AI are just today's technology, but already we can start to see how these 'art'-making AI can provide something for art that Gaut's (anthropocentric) list of properties might not currently include. This is not to say that just *any* properties of AI works can be added to the account, and not just any alterations to the account can be made; Gaut is not that permissive. We would need to make a good case for the inclusion of a new property. Despite this, the cluster account gives those interested in the future of works made by AI a route to allow AI to contribute to what we see as art, not just play catch-up to human-made art.[30]

THE ACTION CONDITION

My discussion of the cluster account thus far has ignored a glaring issue for AI: the action condition. Gaut's modified cluster account includes a single necessary condition, and that condition is that the

[29] Artists working at Realdreams, an AI art collective operating in London (*Realdreams,* 2022).
[30] I should note here that I do not make an argument here that the cluster account of art is correct, just that it is the best for the purposes of assessing AI art. As I will argue, I do not mean best as 'most permissive', but that it is the most useful.

artwork “is the product of an action”. Can AI perform actions? Whether an AI can perform an action will depend on how we characterise actions and, in particular, what relationship we require between acting and intention. This is not a settled debate in philosophy of action (Wilson & Shpall, 2012). If the ‘action’ Gaut requires to make art a ‘work’ does need to be intentional, then we once again reach an impasse for AI works. To do something intentionally, we would expect to need a mind, as intention is considered to be a mental state by many (Jacob, 2019), though, as we have seen above, we do have some possible routes for AI intention. In Chapter 3, I also will explore AI agency in detail.

Does Gaut require the ‘action’ in his account to be an intentional one? This is not clear, though it is likely. The action does not need to include the intention to make a work of art, as that is listed separately as a disjunctive condition. Gaut reaffirms that this is not necessary in a later paper:

> Those who claim that the disputed cases are not art may do so because they insist on, as a necessary condition for art, some feature, such as being the product of an intention to make art. But it is a mistake to insist on this as a necessary condition, as can be shown by considering less contentious cases of art, which lack this feature. (Gaut, 2005, p. 281).

Gaut also makes it clear that his account can accommodate various hard cases in aesthetics, such as Duchamp’s readymades (2000, p. 32, p. 35) and found objects (2000, p. 29). Intention does not need to be present in the creation of the object, just in the selection of the work. In discussion of Weitz’s example of the driftwood, Gaut states:

> A piece of driftwood in nature cannot express despair, nor can it be about anything (since it lacks even derived intentionality), but when selected for display in a gallery it can express desuetude and be about failure and decay (Gaut, 2000, p. 29)

By selecting the object, we imbue in it the capacity for the other properties. This suggests that there is an easy way to avoid the issue of whether AI can ‘act’. In the case of much of today’s AI art, a human will be selecting the outputs to share with the world, ‘curating’ the AI outputs for wider dissemination (through online means, in galleries, in auction houses etc). If we do not want to worry further about whether AI works are art or not, we can again consider AI art to be much like found art, or

readymades: it was not art before, but now it has been selected by someone. In selecting the image, the person has done the requisite (intentional) action to allow this image to count as a work, and with that has opened up the capacity of the image to hold more art properties (like the driftwood above).

While this might offer an easy solution for us in the cluster account, it may be possible for AI to produce art without this solution. As I have shown above, an AI work could have several of the properties of Gaut's account, whether it is properly thought of as a product of an action or not. If the action criterion is in place merely to open up the capacities of the object to have the other properties, then for AI this may not be necessary. It is already artefact-like, it is already the product of an action-like event, and it already has the capacity to have several of the disjunctive properties. It seems very restrictive to call an AI product not art merely because the action which produced it might not be intentional enough, or the distance between intention and the product is too great (such as might be the case in an autonomous AI system).

To support this perhaps we can adopt a weaker account of action. I consider this in Chapter 3, in discussing agency. Instead of considering whether AI can perform intentional actions, we could follow a weaker account of action, such as Frankfurt's use of 'active doings' (1978). Frankfurt uses the example of a spider walking across a table. In doing so, the spider controls its movements, and takes itself from one place to another. The spider has done something with a purpose (Wilson & Shpall, 2016). If we accept that doing something with a purpose is sufficient for action, then perhaps AI could perform an action sufficient to satisfy Gaut. Did an AI 'do' something in generating art? Potentially it did. In making a work, an AI system is 'doing' something; it is processing data and producing images. If we take a GAN as an example, it also does so with a purpose: to produce images that fool the discriminator element of the system.[31] And, in doing this, we have a candidate for artwork that was not there before. Will this satisfy Gaut? Gaut's use of the action criterion is meant to establish that in order for something to be an artwork it must be a *work*. Here, again, is what Gaut says of the action criterion:

[31] There is more to be said here, but I will expand this argument in Chapter 3 considerably.

> An artwork is the product of an action, preeminently of a making (an artifact), or a performing (a performance). It is art*works* that are involved here, since something is in each case done. Hence being the product of an action is the genus of the artwork and is thus a necessary condition for something's being art. (Gaut, 2000, p. 29).

I will not replicate the prior discussion of artefacts (see above in the institutional account of art). However, the purpose of including 'action' as a necessary criterion in the cluster account seems to be to ensure that something is 'done', which results in a product, i.e., the potential artwork. This seems to fit with the case of the products of AI. Some would argue that the AI system is just following a process and is not actually 'doing' anything. We could however say something similar about human brains: they are just following a process. Or, we could similarly say that Frankfurt's spider is just following its 'programming'. At some point, it really seems that something *has* been done, even if we would not want to attribute intentionality.

I have provided an argument that an AI could meet the action condition, if we are willing to accept that we do not need intention in order to satisfy Gaut's account. We are once again left with the status of AI works as *art* hanging on a single criterion, which invokes some longstanding arguments about the nature of AI, mind, intention, agency, and so forth. Under the cluster account, it comes down to the following: an AI work can be art if an AI can perform actions. If an AI cannot perform actions, then AI works are either not art, or we must rely on an action by a human to make them art.

IS AI ART NOT ART?

In other accounts of art, if an AI art cannot meet the requirements of the definition, then it is not art. This is a limit of the definitional approach; the concept either does or does not apply. This is not quite the case for a cluster approach (although, notably, Gaut makes his account more like this with his addition of the necessary condition). We have seen already that Gaut's account includes some built-in flexibility for the possible future counterexamples, that is, we can modify the account (Gaut, 2000, p. 33). We have also seen that the list of properties provided in the cluster account allows us to assess AI

creations on multiple dimensions, rather than just assessing whether it was created with the right intention (as with the historical account) or whether it is an artefact and has the right kind of status (as with the institutional account). There are further benefits to the cluster account, even if it turns out that AI art cannot meet the necessary criteria Gaut lays out. The cluster account is notable for how it fits with our intuitions not just about what *is* art, but also about what *isn't.*

Gaut discusses the adequacy of the cluster account by considering indeterminate cases, where there is considerable disagreement of whether a work falls under the concept of art:

> if there are some objects to which the application of the concept is genuinely, irresolubly, indeterminate, then the account should reflect this too, rather than simply stipulating that the concept applies, or stipulating that it does not (Gaut, 2000, p. 30)

Unlike definitional accounts of art, the cluster account does not deal with hard cases (AI art might be such a case) by mere categorisation. Instead, it offers us an explanation of the irresolution. For example, if AI works are not 'art' (i.e., the concept as outlined by Gaut does not fully apply, and we have not modified if for the inclusion of AI works) then it may be explained as a borderline case:

> the cluster account explains why some activities (such as cookery) seem to lie somewhere near the borders of art without clearly being art, since they share several properties of art (being the exercise of individual creativity, having a capacity to give sensuous pleasure), while also lacking other relevant criteria (since they have difficulty in expressing emotion and conveying complex meanings, and are not generally the product of an artistic intention). It is a signal advantage of the cluster account over the more straightforward definitions of art that it can preserve the hardness of such cases and allow us to explain what it is that makes them hard; such cases can be shown to be genuinely borderline and indeterminate. (Gaut, 2000, p. 36)

Perhaps AI works are like cookery (or any other art-like case). They may be close to art, without clearly being art. If (as is possible given the potential for disagreement in the action criterion) AI art does not meet the necessary criterion for being art, an example of an AI work might still meet several

of the disjunctive properties Gaut lists, and thus we can explain why we might disagree with each other about our judgment of whether this example is art or not. This is better for us than AI art straightforwardly being rejected as art (and left there, as is the case with the other accounts). We can at least describe works made by (autonomous) AI as a difficult case, which lies close to the boundaries of art, and we can more easily justify discussion of the aesthetics of these works, why they seem artistic, or why some of us might still wish to label them as art.

The cluster account of art then can help us in several ways:

1) It can provide a framework for assessing what AI can or cannot currently do in terms of producing art, and this could help in the development of future AI.

2) It can potentially be adapted to accommodate what AI *can* do, through alterations or additions to the properties.

3) Even if AI cannot meet the necessary condition of the cluster account, the account can still explain why it seems like AI works are art, and why we might disagree about its status. AI art can be explained as a border case.

To conclude, I have argued here that (some) AI works could be art under the cluster account of art: they can potentially meet many of the listed properties featured in the account and, if you are willing to accept my argument regarding action, they can fulfil Gaut's necessary criterion of being the product of an action, without relying on human intervention. If you are not willing to accept my argument on action, however, I have argued that the cluster account of art still has plenty to offer those sympathetic to the project of AI art.

Conclusion: Can AI Create *Art?*

In this chapter, I have examined whether AI works can count as art under three theories: the institutional account of art, the historical account of art, and the cluster account of art. In each case it seems that AI works could meet some of the criteria, but perhaps not all. In the institutional account of art, AI works could easily produce works which have the 'the status of candidate for appreciation' conferred upon them, even if this status cannot by conferred by the AI itself. However, the institutional account requires that the candidate object is an 'artefact'. Whilst Dickie's account of artefactuality is considerably easier for an AI to achieve than the standard account, an action will still need to be performed on the work of the AI for it to count as art. Second, I examined the historical account of art, which presents the initial challenge of 'intention', as a work must be 'intended for regard-as-a-work-of-art' in order to be considered an artwork. I presented several options for an AI to achieve 'intention', including removing the requirement of mental states from our understanding of intention. Finally, I examined AI against the cluster account of art. AI can meet many of the list of properties provided in the cluster account, however the necessary criterion of 'action' proves more difficult for an AI system. I argued here that an AI could meet the action condition (and thus make art) if we are willing to accept that we do not need to include intention to act.

Chapter 2 - Can AI *Create* Art?

Having examined theories of art in Chapter 1 to answer the question 'can AI create *art?*', this chapter will address the question "Can AI *create* art?". I look at three approaches to defining creativity and assess whether AI can meet the requirements of each approach and thus whether it can be creative. First, I offer a brief examination of the Darwinian account of creativity. I then turn to examining Margaret Boden's account of creativity, which has been hugely influential in the field of computational creativity. Finally, I turn to examine Berys Gaut's agential account of creativity. These accounts of creativity, like the three accounts of art examined in Chapter 1, leave us with some unanswered questions about mind, embodiment, and agency; these questions will be examined further in Chapter 3.

Darwinian Creativity and AI

In this section, I examine the Darwinian model for assessing creativity in AI systems.[32] First, I will first establish the Darwinian account of creativity using Simonton's (1999) model: blind variation, selection, and retention. Then, I will apply this model to two image-producing AI (Generative Adversarial Networks, and the Creative Adversarial Network) which are used in the computational production of 'artworks'. Here I will argue that these networks are compatible with a Darwinian account of creativity. I will go on to suggest that the Creative Adversarial Network meets the additional criteria of ideational variation (Boden, 2004). Finally, I will address some initial objections. The aim of this work will ultimately be to assess whether the Darwinian model of creativity can be used as a potential standard for testing computational creativity.

INTRODUCTION

The Darwinian model of evolution is thought to have wide and varied applicability (Simonton, 1999). Following Campbell (1960), Simonton suggested that Darwinian theory could be applied to creativity. I will examine the model of Darwinian creativity suggested by Simonton. I will then apply this model to two 'creative' image-making Artificial Intelligence systems: Generative Adversarial Networks (GANs) and Creative Adversarial Network (CAN) (Elgammal et al., 2017). I will assess whether these AI systems are compatible with the Darwinian model of creativity. Initial objections to the use of this model will then be addressed, followed by an assessment of the Darwinian model of creativity as a potential tool for evaluating the creativity of computational systems.

DARWINISM AND DARWINIAN CREATIVITY

There are two types of Darwinism: the first has been developed in the purely biological sense, with a focus on genetics, molecular biology, and behavioural science (Simonton, 1999); the second type of

[32] The work in this section has been published, see Helliwell (2021).

Darwinism provides, according to Simonton, a model that can be applied to many developmental processes. This includes processes which are not purely biological, such as knowledge acquisition. This model consists of blind variation, selection, and retention. This second type of Darwinism has been applied as a framework to a variety of processes, such as Skinnerian operant conditioning and evolutionary epistemology (Campbell, 1974). This Darwinian model can also be applied to creativity.

Simonton (1999) proposes that there may be a basis for a selectionist model of creativity. This model suggests that in the case of humans, there is a psychological mechanism for producing variation, either through recombination or mutation. The outcomes of this variation then go through a selection process; in evolution this would be through sexual selection. In other fields, such as creativity, this selection process would be through the outcome being assessed against necessary criteria. Finally, successful variations are retained in the system.

The variation component is a controversial element of the model (Simonton, 1999). In order to be Darwinian, variation must be "blind" to the selection criteria; it must be as likely to be successful as unsuccessful (non-teleological) (Dawkins, 1986). Campbell (1974) argued that this blind variation could be seen in creativity. This does not mean that variation must be random, rather that likelihood of success is random. Just as in biological variation, some combinations or mutations may be more likely than others to occur, but they will not necessarily produce better adaptations (Simonton, 1999). It is important to note that this blindness applies to the production of variation, not the selection of successful variations, which will not result in equal likelihood of success. There is some evidence to suggest that this is how human creativity works. Sternberg and Davidson (1995) suggest that random priming from environmental stimuli produces a blind variation effect in human creativity; the input is somewhat unrelated to the task and thus provides an element of blindness. Simonton (1999, p. 312) notes that this fits with a large amount of the anecdotal evidence from creative individuals regarding their creative process.

Simonton also addresses the possibility that computer creativity could follow a Darwinian model. Boden (1991) states that computational creativity does not typically follow a Darwinian understanding of creativity, rather it tends to use logical processes or heuristic principles. Boden did

state, however, that with the advent of parallel processing and connectionism this may become more possible. Since Boden wrote on this issue, these technologies have advanced considerably, but much of computational creativity does not focus explicitly on following a Darwinian model.

Genetic programming is one form of Artificial Intelligence that follows an evolution-based model based on Mendelian genetics (a mathematical approach which forms the basis for our understanding of genetic traits, a step further than Darwinism) (Luke, Hamahashi & Kitano, 1999). Genetic algorithms can have mutations added in each generation, which are then tested against the programmed selection criterion. Mutations that produce the best results are fed into the next generation. This process continues until the best-fitting genotype is found (Bäck, 1996; Koza, 1994). This type of programme has been used to rediscover key scientific discoveries (Simonton, 1999) but it has not been applied to artistic creation. According to Boden (1991), in order to meet the criteria for a Darwinian computational process, there must be some method of blind variation present within the model. Boden suggests that to reach the high levels of creativity reached by humans, *ideational variation* must also be a factor. Ideational variation in creativity refers to the variation of ideas, not merely variation within existing rules or constraints (Simonton, 1999).

OBJECTIONS TO DARWINIAN CREATIVITY

Simonton addresses four potential objections to the Darwinian model of creativity. The first is the idea that creativity rises from sociocultural state rather than from individuals; if one individual had not come up with the idea, someone else would have. Simonton states that this does not offer any threat to the Darwinian model, as multiple potential sources of an idea does not mean a Darwinian process does not occur in generating that idea (1999, p. 318).

A second objection is that the Darwinian model of creativity eliminates the role of individual volition; there is no space in the model for the will of people. Simonton (1999) argues that the role of individual will does not eliminate the need for variation, as one cannot *will* a creative breakthrough to occur. Blind or environmental variation is still needed to stimulate variation.

The third objection to Simonton's Darwinian creativity is that creativity can be simply explained by human rationality. Simonton (1999, pp. 319-320) discusses that with increased complexity, rationality becomes less applicable to solution-finding. Blind variation and testing theory are still applicable, particularly in cases of extreme novelty and complexity.

Finally, Simonton (1999) discusses an objection based on domain expertise; the idea that those who have expertise in a field no longer need trial and error. Due to the 'original' nature of creativity, Simonton states that even in cases where someone is an expert, there must be a balance of originality and expertise in creativity, which still leaves room for variation and non-expert input. Simonton also suggests that creativity cannot be improved upon with expertise; one cannot get more creative with age or experience (1999, p. 320).

GENERATIVE ADVERSARIAL NETWORKS

In order to know if the Darwinian model of creativity can be appropriately applied to AI, we need to test this application. In this section I will examine two image-production AIs: Generative Adversarial Networks and Creative Adversarial Networks. These particular systems will be examined as they offer a plausible case for artistic creativity in AI.

Generative Adversarial Networks (GANs) are a form of Artificial Intelligence that utilise machine learning to produce 'artistic' or 'photographic' images (Goodfellow et al., 2014). They consist of two parts: the generator and the discriminator. The discriminator is fed the training images: in this case, images of human artworks. The discriminator learns to distinguish things that fit into the model of "human artwork".

The generator does not have access to the training set and is blind to the discriminator's rules about what is or is not an artwork. The generator initially begins producing random images, with randomness drawn from a noise vector. These are fed into the discriminator. The discriminator assesses the image in comparison to the model it has built based on the training set. The discriminator feeds back a score into the generator, corresponding to whether it thinks the image is a "fake" artwork, or a real

one. This score is used by the generator to adjust future outputs through adding weights to the algorithm, which increases the probability of certain connections being made (Elgammal et al., 2017). The discriminator is aiming to get better at finding the fake images whereas the generator is aiming to get better at producing convincing images.

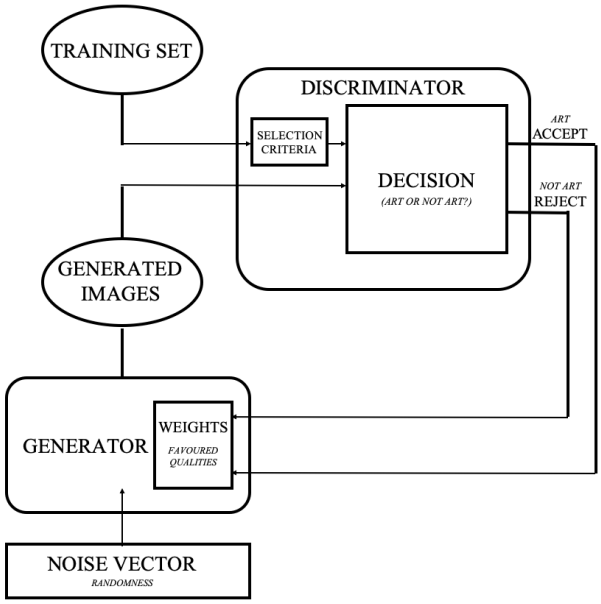


Fig. 17 GAN model, based on information from Goodfellow et al. (2014).

THE CREATIVE ADVERSARIAL NETWORK

The Creative Adversarial Network (CAN) works with the same basic premise as GANs. It consists of a generator and a discriminator (Elgammal et al., 2017). The training set is also composed of images, but there is the addition of style labels (in Elgammal et al.'s original model these were artistic style labels, such as 'abstract expressionism'). This allows the discriminator to learn to distinguish not only what is or is not art, but also different categories of art.

The discriminator, as well as rejecting images that do not fit into its model of 'art', is also tasked with rejecting images produced by the generator that too closely fit into a specific style. The signal which is released by the discriminator to the generator is determined not only by whether the image is plausibly from the same set as the training (high scoring) but also whether the image can be unambiguously classified as one of the styles of art as introduced through the labelling of the training

images (low scoring) or is more ambiguous (high scoring). This results in the generator tending towards stylistic ambiguity in art images it produces whilst maintaining the qualities of artworks.

The generator in the CAN, as in GANs, is blind to the training set that is fed into the discriminator. It receives random input from the noise vector, which forms the initial basis for image production. The generator gradually adds weights (increasing likelihood of certain connections being made) to its image production algorithm based on the feedback of the discriminator. Random noise continues to be inputted into the generator which, combined with the "learning" from the discriminator's feedback, leads to the production of an image. The input of the noise vector ensures that any positively scored image is not merely repeated (Elgammal et al., 2017).

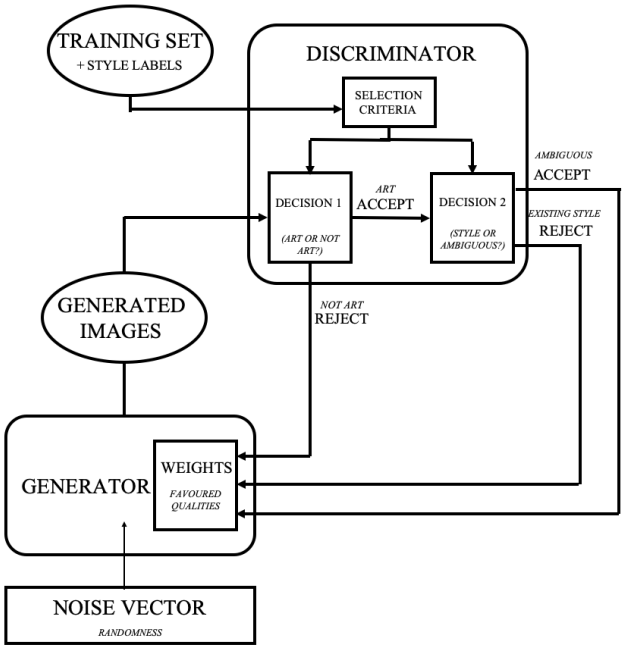


Fig. 18 Creative Adversarial Network, based in part on Elgammal et al. (2017)

APPLYING THE DARWINIAN MODEL TO GANS AND CAN

GANs and the CAN can be mapped onto the Darwinian model of creativity. Whilst they do not explicitly follow an evolutionary model (unlike genetic algorithms), they do inadvertently follow the model of creativity put forward by Simonton (1999), suggesting that, in the Darwinian sense at least, the two computational systems meet the criteria for creativity.

The key element of the Darwinian model of creativity is the non-teleological nature of the creation; the variation must occur without a view to what would be a successful mutation/recombination. This is the blind variation component of Simonton's model. Blind variation is present in both CAN and GANs. The generator provides the means of producing variation. It is not able to see the criteria of selection (what will be accepted rather than rejected) as this is derived by the discriminator from the training images. In this way, the generator is blind. The variation is ensured by the noise vector, which acts as a randomness generator, much like environmental stimuli in Simonton's model.

The success of the image is judged by the discriminator, as this controls the selection criteria. The discriminator compares the generated images to the selection criteria (which has been derived from the training set and style labels) and determines whether the image is successful or unsuccessful. This is the selection element of the model. The selection feedback from the discriminator can be understood as a generational change; the weighting added is much like the genes passed from generation to generation. This is equivalent to the retention of successful traits.

There is no huge difference between GANs and CAN in terms of Darwinian creativity. The underlying process is the same and can be successfully mapped onto Darwinian creativity. However, the CAN ensures that there will be ideational (in the case of art, stylistic) variation. As stated by Boden (1991), ideational variation is vital for reaching near-human levels of creativity. Furthermore, in the case of other theories of creativity (which could complement the Darwinian model), ensuring originality is of high importance (Gaut, 2010). This is only achievable in the CAN model, which ensures deviation from stylistic norms. The diagrams below illustrate how the Darwinian model can be applied to GANs (fig. 3) and CAN (fig. 4).

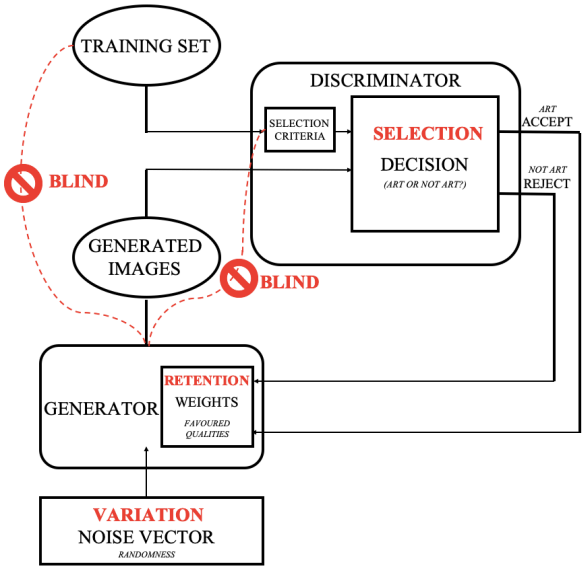


Fig. 19 Darwinian model applied to GANs

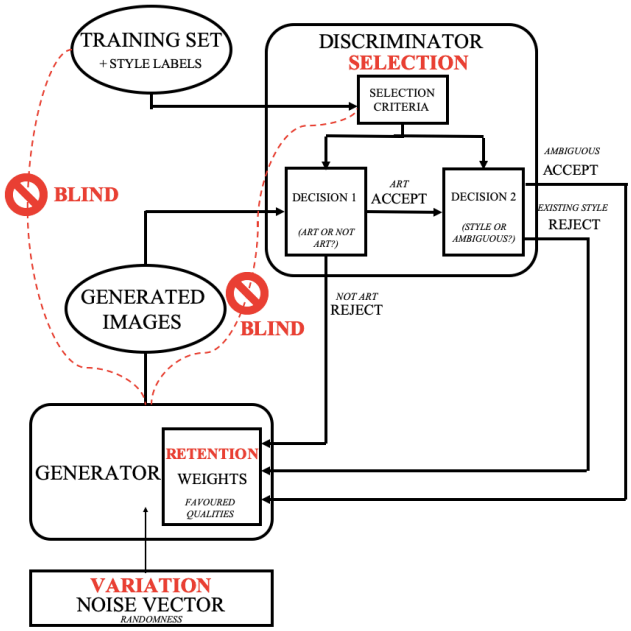


Fig. 20 Darwinian model applied to CAN

OBJECTIONS

Those who object to the outlined argument may state that the application of the Darwinian model to these computer systems is merely an analogy, and therefore an argument that a computational system is creative because it maps onto the Darwinian model of creativity is merely making an analogy and

does nothing to prove that the system is actually creative. This is, however, exactly what is occurring in the Darwinian model of creativity as applied to any process; it is not specific to the application of machine creativity. It is a model that can be applied to other areas of thought analogously. If the model can be applied equally to the theory of creativity and machine 'creativity', this suggests they are somewhat similar in functionality. In the case of developing computationally creative systems, a method of measuring some similarity to a human model of creativity is helpful in evaluating the system.

Some may also suggest counterexamples of non-creative processes, which could also be said to meet the terms of the model in the same ways as GANs and the CAN. However, as the Darwinian model is a broadly applicable model, this does not defeat the argument. Many processes may be found to fit with the Darwinian model. It is possible to question the utility of the Darwinian approach to creativity based on this, but this in itself is not a reason to protest the application to computational creativity.

Another objection to the proposed argument stems from an objection to the Darwinian model of creativity itself. This objections states that the Darwinian model is insufficient for creativity, and therefore meeting the criteria of this model is not enough to demonstrate creativity. This may be correct; however, I would suggest that the Darwinian model provides a *necessary* (though, perhaps not *sufficient*) condition to achieve creativity. Whilst this by no means proves outright machine creativity in the cases of GANs and CAN, their creativity cannot be ruled out based on not meeting the requirements of the model.

A final objection to the proposed argument may be from teleology. This objection would argue that both GANs and CAN fail to meet the requirements of Darwinian creativity as they are goal-directed and therefore teleological. If sustained, this objection is potentially fatal to this argument as it proves the existence of false-analogy, removing the whole premise of Darwinian creativity being applicable to these computational models. There are several potential responses to this objection. The first would be to deny that there is any intention or goal-directness in the systems as a whole, and therefore the objection is baseless. I will not pursue this course, as this would destroy in part the applicability of the whole Darwinian model to creativity, which is generally agreed to involve some level of intention or

agency (Gaut, 2010). A less problematic rebuttal would be to suggest that while the whole system is indeed goal-directed, this does not mean that it cannot not fit the criteria of non-teleology of the Darwinian model. By splitting up the system, such that generator would operates as the evolving 'creature' and the discriminator operates as the environment, we can show that the generator is initially non-teleological. The generator begins by producing images based on the random noise, with the later addition of information of the retained qualities from feedback from the discriminator. As the generator meets the requirements of 'blindness' due to its lack of access to the training materials or selection criteria, it still cannot be said to know what it is aiming at in a way that prevents it from meeting the Darwinian model of creativity.

THE DARWINIAN MODEL AS A TOOL FOR EVALUATION OF COMPUTATIONAL CREATIVITY

The application of the Darwinian model to GANs and the CAN shows that this model of creativity can function as a tool for assessing computationally creative systems. Unlike some models of creativity which have no clear measurability (such as Gaut's account (2010) which requires agency, and a measure of value), the Darwinian model provides a clear way in which a system can be assessed to meet certain creative standards: it must include blind variation, selection, and retention of successful traits. With the added requirement of ideational variation, this model offers a measurable standard of creativity for computational systems. While it may be the case that meeting the requirements of the Darwinian model of creativity is insufficient to be considered creative in the human sense, this model may provide a good initial method for assessing whether computationally creative systems can meet some of the *necessary* conditions for creativity.

To conclude this section, the model of creativity proposed by Simonton follows Darwin's evolutionary theory, which has since been used to model various psychosocial processes, including creativity. This model is comprised of blind variation, selection, and retention, with the addition of ideational variation in the case of creativity, to ensure outputs are creative in a meaningful sense. Both GANs and the CAN can be successfully mapped onto Simonton's model of creativity. This suggests

that these computational systems meet the standard of creativity laid out in the Darwinian model. Whilst this may not be sufficient to claim that GANs and CAN are creative, not meeting these criteria would have prevented them from being considered as such. This shows how the application of the Darwinian model can be used to assess computationally creative systems in a measurable way, unlike other popular theories of creativity. Whilst the Darwinian model may not be sufficient to prove creativity on a par with humans, it can provide an initial standard of assessment for computationally creative systems.

Boden's Account of Creativity and AI

Having examined the Darwinian approach to creativity, let us now turn to another approach: when it comes to discussing creativity in AI, Margaret Boden's work is the most well-known and influential. In this section, I will offer a summary of Boden's position on the nature of creativity, and her assessment of whether it is out of reach for AI. Most aspects of Boden's theory of creativity are achievable for AI systems. Boden does, however, conclude in a 2014 paper that autonomy (which some may think is necessary for creativity) cannot be achieved by AI. I will argue against this position, presenting an argument that AI could have some level of autonomy.

BODEN'S MODEL

Boden's model of creativity (1998; 2004; 2014) is consistent with other accounts[33] in that a creative idea or artefact is *novel* and *valuable*. Boden adds a third criteria to this model: *surprisingness*. As Boden states:

> Creativity is the ability to come up with ideas or artefacts that are new, surprising and valuable. 'Ideas' here include concepts, poems, musical compositions, scientific theories, cookery recipes, choreography, jokes – and so on. 'Artefacts' include paintings, sculptures, steam engines, vacuum cleaners, pottery, origami, penny whistles – and many other things you can name. (2004, p. 1)

Boden's account is not limited to artistic creativity then, but also includes inventions, discoveries, and other aesthetic objects. It is worth noting that Boden's account of creativity focuses on the *products* of creativity, i.e., ideas and artefacts. The Darwinian (or evolutionary) model of creativity considered above is a process-based account; the focus is on the process by which creativity occurs (i.e., blind variation and selective retention). In its typical formulation, as stated above, there is also no specification about the person (or producer).[34] This is convenient for considering AI creativity,

[33] Gaut's (2010) account of creativity which will be discussed below.

[34] These are three Ps of creativity - 'person' (or producer, if we are considering AI), 'process' and 'product'. A fourth 'P' is often discussed beyond the philosophical literature: 'press'. I exclude it here, as it relates to

because it places fewer requirements on the AI system itself. The AI does not have to meet certain criteria for it to be considered creative. As we will see, Boden does consider necessary criteria for the AI itself to meet in a later work (Boden, 2014). First, let's investigate Boden's base model of creativity, by examining novelty.

NOVELTY

Boden (1998; 2004) makes a distinction between two kinds of novelty:

> The idea may be novel with respect only to the mind of the individual (or AI-system) concerned or, so far as we know, to the whole of previous history. The ability to produce novelties of the former kind may be called P-creativity (P for psychological), the latter H-creativity (H for historical). P-creativity is the more fundamental notion, of which H-creativity is a special case. (Boden, 1998, p. 347)

Under this account, creativity can occur on a personal level or on a historical level. For H-creativity, we are likely to bring to mind ground-breaking scientific discoveries or revolutionary developments in art; however, the product does not need to be so significant. The only requirement for H-creativity is that it is a surprising and valuable idea (or artefact) that has not been thought of before in history. There may therefore be plenty of inconsequential examples of H-creativity. All cases of H-creativity will also be cases of P-creativity (Boden, 2004, p. 2). As Boden states, whilst H-creativity may be the focus of many discussing the arts and sciences, it is P-creativity which we should be concerned with if we are interested in the nature of creativity, as it encompasses not just great feats of creativity but everyday examples of creativity. We could consider how children create drawings in a somewhat predictable fashion. A child's first drawing of a house might be quite cliché, but it is still an original piece for her. We might similarly consider an academic who has a brilliant idea, only to find out that there is a vast body of literature already published on the same thing. In both these cases, something

reactions to creative products, and thus is external to creativity itself. See Jourdanous (2016) for an analysis of the four Ps from the perspective of computer science.

creative has still occurred, despite it having been done before by someone else. To understand if creativity has occurred, we do not need to know if this is the first time something has ever happened.

SURPRISE

Boden does not just divide creativity by type of novelty, but also in relation to types of surprise. Boden describes three kinds of surprising idea. First, she distinguishes the kind of idea that is surprising because it is unfamiliar or unlikely, "like a hundred-to-one outsider winning the Derby. This sort of surprise goes against statistics." (Boden 2004, p. 2). Second, she distinguishes a more 'interesting' kind of surprise:

> An unexpected idea may 'fit' into a style of thinking that you already had – but you're surprised because you hadn't realized that this particular idea was part of it. Maybe you're even intrigued to find that an idea of this general type fits into the familiar style. (Boden, 2004, pp. 2-3)

Finally, the third (and, according to Boden, the most interesting) sort of surprise is:

> the astonishment you feel on encountering an apparently impossible idea. It just couldn't have entered anyone's head, you feel – and yet it did. It may even engender other ideas which, yesterday, you'd have thought equally impossible. (Boden, 2004, p. 3).

I am not sure these different sorts of surprise are the result of any particular analysis of what might cause surprise. This also appears not to be 'surprise' in the sense of an emotion, as Boden states in her 2014 paper: "I don't know of anyone who explicitly defines creativity with reference to emotion." (2014, p. 235). This is a strange claim, as it seems counter to her own definition of creativity, of which surprise (an emotion) is a constitutive part. It may be that in saying no one defines creativity with reference to emotion, Boden means that no one defined the creative *process* as requiring emotions. This would not be included in Boden's definition, as her focus is on product rather than process. From these three kinds of surprise, Boden proposes three corresponding types of creativity to

explain what is going on: combinational creativity, exploratory creativity, and transformational creativity. (Boden, 1998, p. 348). I will explain each of these in turn:

> **i) Combinational creativity:** This type of creativity involves putting together novel combinations of familiar ideas. Boden offers examples such as analogies, imagery in poems, and collage (Boden, 2004, p. 3).
>
> **ii) Exploratory creativity:** This type of creativity involves "generation of novel ideas by the exploration of structured conceptual spaces" (Boden, 1998, p. 348). Boden states that exploration can result in ideas that are both novel and unexpected yet fit within existing norms for the relevant thinking style (Boden, 1998, p. 348). Boden suggests that examples would include an artist trying a new technique; the technique already existed, but had not yet been used by the artist, and so might bring something new to their style (Boden, 2004, p. 4).
>
> **iii) Transformational creativity:** The third type of creativity similarly involves a conceptual space, but instead of exploring within an existing space, that space is *transformed.* Boden claims that this opens up new areas of the conceptual space such that structures that were previously not possible could now arise (Boden, 1998, p. 348). This would be equivalent to developing a new technique, or an entirely new approach to making an artwork (or building a road, as Boden 2004 states). Transformational creativity results in previously impossible ideas being made possible.

VALUE

Having looked at the criteria of 'novelty' and 'surprisingness', let's turn to Boden's last criterion: value. Boden does not give much specificity to her inclusion of value, though she does state that the inclusion of a value criterion ensures that identifying and explaining creativity is not a "purely scientific matter" (2014, p. 227) as value judgements are not the role of science. Boden points out the variability of values individually and cross-culturally, and that there are debates as to the possible universality of *any* values in terms of aesthetics, or beyond to sciences and arts more generally

(Boden, 2006; 2014). This raises issues with including value as a defining part of creativity: it relies on a value judgement, which, particularly in the artistic domain, may not have a clear, single objective measure and is an area rife with disagreement. As I will argue in Chapter 4, relying on value as a way to recognise creativity may be problematic when applied to non-humans. Boden also seems to require that the originator of an idea or artefact is able to recognise the value in their creation. Now we have the basics of Boden's account of creativity, we can examine her analysis of the possibility of creative machines.

APPLICATION TO COMPUTERS

Boden offers her own analysis of whether computational creativity is possible using her model of creativity. It is worth bearing in mind that Boden's work is, in AI terms, very old. Even since her writing in 2014, AI technology has advanced rapidly. Boden's work from the 1990s and early 2000s, however, remains very influential (according to Google Scholar Search, the 1998 paper referenced here has been cited over 500 times in the past five years, and over 100 times in 2022 so far), thus her analysis remains very much relevant to the discussion of AI creativity.

Boden argues that computers aiming to be creative in the exploratory sense are most successful over those aimed at creativity in the combinatorial or transformational sense (1998, p. 349). In the case of combinational creativity, Boden argues that this is due to difficulties in replicating anything like associative memory in humans (1998, p. 349). Transformational creativity, Boden argues, is similarly limited in AI due to the difficulty we have with identifying and expressing human values in a form that can be utilised computationally (1988, p. 349).[35] Despite the great challenge of creativity for AI, Boden does not think *any* of the three kinds of creativity are impossible for an AI. Boden offers evidence that an AI might be able to produce mediocre jokes, a form of combinational

[35] This is related again to the value alignment problem, discussed in Chapter 4.

creativity (1998, pp. 349-350), and argues that the exploratory and transformational types of creativity can also be modelled by AI systems (1998, p. 351).

CAN AI BE CREATIVE? A NEGATIVE PERSPECTIVE

Whilst Boden's account of the possibility of AI creativity is initially positive, in a later work (2014) Boden considers which features computers may or may not have that prevents them from being creative. It is here where autonomy features as a barrier for computational creativity. Boden first presents 'being programmed' as a feature that computers have that may preclude them from being creative (2014, pp. 229-232). 'Being programmed' links closely with autonomy (or lack thereof), which Boden goes on to consider as a feature of creativity that AI might lack. She considers whether creativity implies "autonomy, intentionality, consciousness, value, and emotion--all of which are commonly assumed to be denied to computers" (2014, p. 233). As Boden states, these features are all commonly assumed to be impossible for computers, even today. Therefore, if creativity implies any of these features, it may be that computer systems cannot be creative. Boden argues that, with qualifications, each of these are key features of creativity, but they do not all mean that AI could not be creative. Boden argues that these features are *typical* of creativity but does not include any explicitly in her own definition (that discussed above).

As for whether autonomy, intentionality, consciousness, value and emotion are possible for AI, Boden only gives a definitive answer in one case: autonomy. Boden claims that if intentionality, consciousness, value (evaluation), or emotion are needed for creativity, then the question of whether AI can be creative is 'unanswerable'. These concepts, she argues, are all reducible to a question of whether an AI has mind or consciousness, and thus these concepts are highly contentious in philosophy (2014, p. 242).[36] These concepts are not essential to any theory of creativity that I will use in this thesis.[37] Due to this, I will not examine most of these concepts here. I will, however, dig a little

[36] I examine AI mind in brief in my examination of AI extended mind in Chapter 3.
[37] Except, perhaps, for intentionality, which is related to action and agency, a requirement of Gaut's theory of creativity. I address the requirement for agency, and whether it requires intentional action in Chapter 3.

deeper into Boden's position on autonomy, as she claims that if autonomy is needed for creativity, then AI cannot be creative.

The views of creativity considered in this thesis do not explicitly require autonomy,[38] including Boden's, although she states that it is assumed in her model of creativity that the idea in question was 'freely generated', particularly in the case of H-creativity. In the case of P-creativity, however, Boden does allow that creativity could occur under the influence of another (2014, p. 233). Consider a philosopher teaching students. A philosopher may coax her students towards a novel thought through careful questioning or through presenting ideas in a certain way. The students, finally grasping the idea to which the philosopher was leading them towards, have still had a novel idea (to them) and that idea may well be a valuable and surprising one (while it's not a surprise to the teacher, it may be to a student). Boden would still call this an example of creativity, and so she allows that we may still have creativity in cases where the idea was not generated wholly autonomously. Despite raising no explicit mention of autonomy in theories of creativity, or in her own, Boden suggests that we may still require that an AI be autonomous in order to be creative. Boden offers her own analysis of whether it is possible for an AI to be autonomous. Initially, Boden considers whether the fact that a computer is programmed precludes it from being creative. This, in a way, is a consideration of autonomy. As Boden states: "In other words, being programmed is the antithesis of being autonomous" (2014, p. 229). The objection goes that all computers (from AI to a calculator) are doing what they are told to do. If they are not doing what they are told explicitly to do, then they are still limited to do what they have been 'empowered' to do by human programmers (2014, p. 229). Boden states that some making this objection might allow that combinational and exploratory creativity could be simulated by computers, but would maintain transformational creativity could not be. This is because "the sceptic will say, the rules/instructions specified in the program determine the computer's possible performance, and there's no going beyond them." (2014, p. 229).

[38] In the next section, I will turn to examine Gaut's agential account of creativity. Gaut also does not specify autonomy as necessary for creativity in so many words, however Gaut's account does explicitly require agency on the part of the creator. When we attribute agency to creating the object, we likely are assuming some kind of autonomy.

As Boden states though, this is not entirely accurate. It is not true that rules or instructions specified in the computer programme are the limits of the computer's possible performance:

> what it ignores is that the program may include rules *for changing itself.* For example, it may be able to learn--perhaps on the basis of unpredictable input from the environment, or perhaps due to its self-monitoring of internal 'experimentation' of various kinds. Or, more to the point for our purposes here, it may contain genetic algorithms, or GAs (see Boden 2006: 15.vi). (Boden, 2014 pp. 229-230)

As Boden points out, in some cases AI systems (notably machine learning systems and genetic algorithms) can alter their own programming. We see this in generative adversarial networks; they are able to alter the weights in the generator in response to learning from the training set. GANs also receive input from a noise vector, which introduces randomness in the system. Whilst this may not be *true* randomness (i.e., if we had all the details about the vector, we could potentially predict its output), we could include truly unpredictable input from an environmental source. We could then have an algorithm which can change its own programme and has random input. It is not clear why an algorithm like this could not be sufficiently distant from the hand of a programmer to count as 'undetermined.'

Despite this point about genetic algorithms, Boden states that the autonomy objection will still stand for some who hold that the biological inspiration for evolutionary programming is essential, citing Pattee (1985) and Cariani (1992) by way of example:

> They point out that programs are abstract systems, and as such are logically self-contained. Even evolutionary programs, like those discussed above, are essentially limited to the possibilities inherent in the GAs and other rules supplied by the programmer. Genuine, truly radical, transformations can arise in an evolving system, these objectors argue, only if it interacts *physically* with actual processes in the outside world. (2014, p. 230)

I have already argued that an evolutionary (Darwinian) model of creativity can be met by typical art-making AI systems (GANs and CANs). There is an additional requirement here though:

the only way for 'transformations' to occur is through interactions with the outside world. This, sceptics think, is due to the lack of unpredictability in the virtual sphere in comparison to the outside world (Boden, 2014, p. 232). This suggests that a level of embodiment is necessary for an AI to be creative (if we accept the sceptic's position). Boden argues that embodiment is already possible with some AI systems (2014, p. 231). Boden also points out that unpredictable events are not solely the purview of the physical world:

> Only insofar as it can be affected by unforeseen events can genuinely new types of results emerge. Many of those events will be external to the system itself. However, they might be cultural/semantic, as well as physical: Accidental interactions with text or imagery on the Internet, for instance, might occasion creative transformations very different from what the programmer envisaged (see Boden [2019]) Moreover, they could also include internal events, if the system could perform information-processing internally, playing around in various ways (e.g., making novel combinations and exploring existing styles) with the structures already present in it. (Boden, 2014, pp. 232-233).[39]

It is unclear then whether this requirement for embodiment to access unpredictability only concerns *transformational* creativity. If so, this still does not preclude AI from being creative at all, just from being *transformationally* creative. This is hence an argument for embodiment as necessary for *transformational* creativity. Even then, it seems that Boden argues that a non-physical space could satisfy the requirement for autonomy. I examine the role of embodiment in AI creativity further in Chapter 3.

So, Boden has provided several potential responses to the issue that AI cannot be autonomous due to being programmed and thereby cannot be creative. First, that creativity simply does not require autonomy. Second, that AI can indeed (in some cases) stray from the hand of the programmer. Third, that AI can be embodied. Fourth, that the unpredictable events that embodiment seeks to capture do not just occur in physical space. And fifth, that unpredictable events could occur from within the

[39] I argue for unpredictability as an element of creativity later in this Chapter, in the section on Gaut's requirement of spontaneity.

system. Despite refuting the argument that AI systems are unable to escape their programming, Boden accepts this as a potential objection, saying: "we must allow that, *in that strictly limited sense,* no programmed system can be truly autonomous, and that sense also, then, computers cannot be truly creative either." (2014, p. 236). If we accept the sceptic's position that an AI cannot go beyond its programming due to the lack of access to unpredictability, then an AI cannot be autonomous. And, insofar as autonomy is required for creativity then AI cannot be creative (at least not transformationally so).

It seems we have three routes out of this problem of autonomy:

1) We do not accept that autonomy is necessary for creativity (this is Boden's route). In particular, we could accept that autonomy is necessary for transformational creativity only, still leaving combinational and exploratory creativity as possible for AI under Boden's classification.
2) We examine autonomy closely to determine if there is a way it could be possible for AI (I will turn to this next).
3) We argue that embodiment is possible for 'creative' AI (see Chapter 3)

We do not have to accept the position of the sceptics that only a system that is embodied can be truly autonomous; however, given the complex nature of autonomy, further examination of autonomy is needed.

Boden's definition of autonomy is initially 'self-direction' (2014, p. 9) though later, she describes the key element of autonomy as 'human freedom' (2014, p. 12; 2010, p. 15). As Boden notes, autonomy is a 'vexed' concept (2010, p. 16). It is not clear why 'human freedom' as Boden puts it must be the only account of autonomy that aligns with creativity. Given that the phrase 'human freedom' includes 'human', we are likely to think it impossible for a non-human to achieve. The idea of freedom too, as Boden points out, is highly contentious (2014, p. 12). This contentious nature seems to Boden to be an opportunity for us to abandon attempts to resolve a philosophical issue, as with the above discussion of other challenging concepts for computational creativity. Ultimately, a

closer examination of autonomy in the context of AI may help to illuminate whether Boden's assessment that, despite the possible independence of AI from a programmer, this is insufficient to claim autonomy in AI. If we can cast doubt on the objection that being a computer means autonomy is impossible, we have another path to reject the claim that computational creativity is not possible because of a lack of autonomy.

AUTONOMY

Autonomy is a concept common in discussion of artificial intelligence and robotics.[40] Whether a system can function in some way on its own, without the supervision of humans, is key to concerns about reliability, control, and responsibility for wrongs.

AUTONOMY IN ROBOTICS AND AI

In the fields of AI and robotics, autonomy is a frequently used concept, but one that is not always clearly defined.[41] Several papers have attempted to discern a clear and consistent definition of autonomy. Franklin and Graessler (1997) survey then-recent references to agents from AI literature, and attempt to offer a definition of an autonomous agent:

> An **autonomous agent** is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future. (Franklin & Graessler, 1997, p. 25)

As Franklin and Graessler go on to point out, how we understand what an autonomous agent is will depend on how we define what we class as an environment (or a niche). However, Franklin and

[40] Work in this section was developed during a joint project with Colleagues at Northeastern University in both the Boston and London campuses. The result of this joint project is under review.
[41] I consider AI and robotics together here as there is considerable overlap between the two fields as far as autonomy is concerned. 'Autonomous' robots require artificially intelligent systems as part of their makeup.

Graessler seem to suggest that the role of the environment in defining autonomous agents relates more to the description of the autonomous agent as an *agent* than as *autonomous:*

> Autonomous agents are situated in some environment. Change the environment and we may no longer have an agent. A robot with only visual sensors in an environment without light is not an agent. Systems are agents or not with respect to some environment. The AlMA agent discussed above requires that an agent "can be viewed" as sensing and acting in an environment, that is, there must exist an environment in which it is an agent. (1997, p. 26)

This suggests that concerns about the environment relate more to agency than autonomy. Aside from the role of the environment, Franklin and Graessler's definition could be read as quite a stringent one, particularly in relation to "the pursuit of its own agenda". However, as they clarify, this is meant as a permissive definition that can accommodate a large spectrum of 'autonomous agents':

> Humans and some animals are at the high end of being an agent, with multiple, conflicting drives, multiples senses, multiple possible actions, and complex sophisticated control structures (minds [6]). At the low end, with one or two senses, a single action, and an absurdly simple control structure (mind?) we find a thermostat. A thermostat? Yes, a thermostat satisfies all the requirements of the definition, as does a bacterium. (1997, p. 25)

An account of an autonomous agent that can accommodate a thermostat, which 'senses' a threshold temperature and 'in pursuit of its own agenda' so turns down the temperature 'to effect what it senses in the future', is a very permissive one. I would venture that most of us do not think of thermostats as having 'an agenda', yet this gets to the heart of the 'autonomy' to which Franklin and Graessler refer:

> Our definition above is of an autonomous agent, yet no mention of autonomy appears in the body of the definition. What gives? No explicit mention is needed. An agent that acts in pursuit of its own agenda is acting autonomously. It selects its own actions independently. (1997, p. 25)

And here we have the crux of autonomy in the field of robotics and AI: 'selects its own actions independently'. Independence is (almost) synonymous with autonomy for those working on Artificial

Intelligence. This independence is typically meant as independence from humans, as systems that are comprised of multiple artificial components are thought of as a single autonomous agent as long as they are independent from an operator.[42]

Haselager, in summarising Franklin and Graessler's review, offers a clearer picture of the need for independence in autonomy:

> Autonomous agents operate under all reasonable conditions without recourse to an outside designer, operator or controller while handling unpredictable events in an environment or niche. (Haselager 2005, p. 64)[43]

Haselager's summary of the definition of an autonomous agent makes clear the centrality of independence, framed as operation "without recourse to an outside designer, operator or controller". Recourse here suggests that to be autonomous the agent cannot receive assistance from, or defer to, something outside itself in acting (likely a human). This only refers to recourse *during* the act (i.e., when on-line), but not to recourse to a designer preceding the behaviour (i.e., programming). Here it is usually pointed out that there is a continuum between complete dependence and complete independence (e.g., Maes 1995, p. 108). The 'under all reasonable conditions' is added to indicate that there are limits to the robot's functioning ('reasonable'), while at the same time signifying that the robot should not be too dependent on favourable circumstances ('all'). The clause about unpredictable events in an environment is intended to rule out pre-configured systems operating blindly in a completely predetermined environment. As Haselager states:

> Within robotics, then, the increase in autonomy of a system is related to the reduction of on-line supervision and intervention of the operator, programmer or designer in relation to the robot's operations in a changing environment. (2005, p. 518)

[42] It is worth noting the difficulties that come with discussing collaboration under this kind of definition of autonomy.

[43] As the above definition is a reformulation of Franklin and Graessler's definition, the 'handling unpredictable events in an environment or niche' is permissive enough to let in the thermostat still. In fact, Haselager's version may be more permissive, as it includes 'niche' so not necessarily an environment.

As Haselager specifies, an artificial agent can still be designed or initially programmed and count as autonomous under this account, as long as any recourse is pre-behavioural. This is also referred to as the ability to choose the appropriate course of action for oneself (see Totschnig, 2020; Russell & Norvig, 2010; M Anderson & S L Anderson, 2011; and Müller, 2012). This can be clarified with an example. Consider an autonomous vehicle. The vehicle does not decide where it is going or when; these decisions are made for it by a human. The autonomous vehicle will, however, make many decisions during the act of driving. It will decide when to stop and start, when to turn, and even which route to take if it is sufficiently advanced. Under this way of thinking, an AI or robot is not required to set the end goal for itself in order to be considered autonomous. It must merely have the ability to make decisions free from supervision during the course of pursuing the goal that is already set for it. As we will see, this is different from the conception of autonomy that is utilised by philosophers, even when referring to robots.

PHILOSOPHICAL (HUMAN) CONCEPT OF AUTONOMY

In philosophy, autonomy is understood as 'self-governance'; however, the exact meaning of self-governance is not widely agreed upon (Buss & Westlund, 2018). There are various accounts of human autonomy, but given the limited nature of AI and the permissiveness of autonomy under the robotics literature, debating the merits of each is not relevant. What we need is a basic or minimal account of autonomy as we think about it in humans. Buss and Westlund describe such a minimal account of autonomy: "Minimal self-government seems to require nothing more nor less than being the power behind whatever reasoning directly gives rise to one's behavior." (Buss & Westlund, 2018). This account of autonomy at first seems close to that which we saw in the technical literature: independence. However, independence does not account for the 'reasoning' component of minimal autonomy. What is it to have 'reasoning' behind one's behaviour?

Reasoning behind one's behaviour may link to the common requirement for autonomy in the philosophical discussion: the ability to set one's own goals. Haselager points this out in comparing human autonomy to autonomy as discussed in the AI and robotics literature:

> In the philosophical literature, however, one finds rather more emphasis on the reasons *why* one is acting (i.e., the goals one has chosen to pursue) than on *how* the goals are achieved. Auto-nomos, being or setting a law to oneself indicates the importance of self-regulation or self-government. Autonomy is deeply connected to the capacity to act on one's own behalf and make one's own choices, instead of following goals set by other agents. The importance of being able to set one's goals is also part and parcel of the common-sense interpretation of autonomy. (Haselager, 2005, p. 519)

Here we get on to the typical focus of the philosophical literature when discussing AI: the capacity to set one's own goals.[44]

SETTING GOALS

So, can an AI do this? Haselager thinks not, particularly as there is no clear way for a robot or AI to have goals that are 'its own'. Haselager argues that there is promise for AI, but that in order for a robotic system to have its own goals, it may need to have a body of the right kind:

> the capacity to have goals of one's own arises out of the continuous integration of control system and body, resulting in actions aiming at homeostasis. A further understanding of this capacity requires considering the co-evolution of body and control system as well as the specific 'potentialities' of organic matter i.e., autopoiesis. (2005, p. 529)

I address related issues in Chapter 3, during my examination of embodiment. For now, counter to Haselager's view, I argue that we *could* have AI setting its own goals, but only through distinguishing different types of goals.

Totschnig (2020) argues, counter to the predominate view in the discussion of 'superintelligence', that an AI could (in the future) rationally alter its final goals, and thus achieve full

[44] Goals will feature again in my discussion of agency in Chapter 3, where I utilise an account of agency which aligns with this understanding of autonomy as self-determination of goals.

autonomy.[45] To take Totschnig's discussion as inspiration, we could consider the ability to set one's *final* goals as necessary for full autonomy. To complement this, we could utilise the idea of intermediate, or sub-goals (also taken from the literature on superintelligence, though with a much more modest application. See Bostrom, 2012; Danaher, 2012). Totschnig discusses, as we have seen already, two senses of autonomy: the sense utilised in AI and robotics research, and a philosophical sense. In the former, "The attribute 'autonomous' concerns only whether the agent will be able to carry out the given general instructions in concrete situations." (2020, p. 2474). Totschnig says of this kind of autonomy: "From a philosophical perspective, this notion of autonomy seems oddly weak". (2020, p. 2474). Instead, from a philosophical perspective, an instance of full autonomy would be "an agent who decides, by itself, to devote its efforts to a certain project—the attainment of knowledge, say, or the realization of justice." (2020, p. 2474). Totschnig argues that this constitutes an ability to change one's final goal or purpose, and that Artificial (general) Intelligence "may very well come to change its final goal in the course of its development" (2020, p. 2472).

Are the two senses of autonomy that Totschnig distinguishes the only possibilities? I suggest that there is a third possibility. Instead of autonomy just as *independence*, or autonomy as *setting one's final goals*, we could consider to what degree a system is able to set its own *intermediate* goals. Totschnig may be right that what is needed for (full) autonomy is the ability to set one's own final goals, but this seems to offer a black-and-white view of autonomy. For one, this ignores entirely the sense of autonomy that is utilised by roboticists and computer scientists. If an autonomous vehicle is simply following its programming to get to a destination determined by a human, then it seems not to have any autonomy in the philosophical sense (Totschnig, 2020, p. 2474). But consider a system that is making a series of decisions in order to get to its final destination. For example, this system can decide which route to take. This system has not set its own final goal, but we could say that it has set an intermediate goal: to navigate this particular route, in order to get to its predetermined final destination safely. It also ignores the extent to which a system can change after initial development.

[45] Bostrom (2014) for example argues that an AI which surpasses human intelligence (a "superintelligence") will not change its final goal, and instead pursue this goal to the detriment of all else (such as humanity, the world etc.).

We can consider Boden's argument on the ability of genetic (and other) algorithms to change themselves, and thus distance themselves from the hand of the programmer. This addresses Haselager's concern about goal-ownership, to some extent. The goals are the AI's, because they are not the programmers'. The more a system is not controlled directly by the programmer, i.e., the more it changes, the more room it may have to set its own sub-goals.

Take the creative adversarial network (AICAN) (Elgammal, 2017). AICAN does not set its own final goal (to generate art images). AICAN *could* however be said to have some level of autonomy in what it generates. As the discriminator learns, it alters what is expected of the generator in generating images; in a (minimal) sense, it is altering the intermediate goals of the system. The generator now has a different set of standards to reach than it did prior to training. The *final* goal is still the same overall: to generate art-images. Operationalised, the *final* goal is unchanged too: to generate images that will fool the discriminator. However, the specific sub-goals have changed on the basis of this training. The goal is now say, to producing images with a green section at the bottom of the image, and a blue section at the top (to produce a landscape-like image), but with a different texture, say, than the discriminator has seen before (to meet the stylistic-ambiguity criterion). We do not have to think that this is equivalent to full autonomy to recognise that this could constitute some level of autonomy. The idea of AI setting intermediate (or sub-) goals also fits with the concept of levels of autonomy, which has become popular in the AI literature, and which I myself utilised (loosely) in the taxonomy of AI systems in the Introduction (see Müller, 2012). This framing of final versus intermediate goals, however, moves us away from the robotics approach, excluding cases such as the thermostat (which could not be said to have any sub-goals that are not pre-determined).

To conclude this section, let us return to the issue presented by Boden. If autonomy is necessary for creativity, we do not have to accept the understanding of autonomy presented by the sceptics, and nor do we have to accept the unsatisfactory account of autonomy as used by those in robotics and AI research. Through utilising an understanding of autonomy that sits in a middle ground between independence, and the ability to set one's *final* goals, we can see a way for some AI systems to set their own intermediate goals, affording them some level of autonomy. If we need autonomy for

creativity then, we may be able to have (some level of) creative AI. Autonomy was, according to Boden, the one clear sticking point for AI creativity. Through rejecting the inclusion of autonomy in her own account, Boden did not preclude the possibility that AI could be creative; in fact, she argued that AI could have any form of creativity, but especially exploratory creativity. In the next section, I will examine Gaut's agential account of creativity, which is not so permissive.

The Agential Account of Creativity and AI

The third and final account of creativity I am going to examine is Gaut's agential account of creativity. Gaut's theory of creativity goes further than Boden's, adding an alternate third condition (in the place of surprise). Gaut, like Boden, includes both originality (or novelty) and value in his account (Gaut, 2010, p. 1039). Equally, Gaut does not think that originality and value are sufficient for creativity (2010, p. 1040). The reason for this, he argues, is that with just these two conditions, non-creative processes would be included in creativity. Gaut wishes to exclude processes that, while they may produce valuable and potentially original products, occur without the action of an agent. For example, he wants to exclude cases such as the creation of diamonds in the earth or the distribution of leaves on trees to best reach the sunlight (this is counter to Arnheim (2001), who argues that this is in fact an example of creativity). Gaut asserts that there can be no creativity in these cases, citing the need for intentional states:

> But the tree is not acting at all, since it lacks desires, beliefs and other intentional states; so a fortiori it cannot be acting creatively. Creativity is a property of agents, not of mere things or plants (see also Carruthers, The Architecture of the Mind, ch. 5; Stokes). (Gaut, 2010, p. 1040)

I will examine the basis of the argument for agency in Chapter 3, but for now let us leave this unchallenged. Agency alone is still not sufficient for creativity according to Gaut. Against Boden, Gaut rejects the separate inclusion of surprise, arguing that this is just a variant of novelty. In place of surprise, Gaut argues for *'flair'* as a further condition of creativity. Flair, Gaut argues, ensures that actions are of the right kind in order to count as creative. Flair is a multi-faceted condition, so let's take a closer look at how Gaut defines flair.

FLAIR

Gaut defines flair through the exclusion of four different cases that he argues should not be included in creativity.

First, he suggests that works made by accident cannot be creative. Consider an artist working in a studio who knocks over a pot of paint, which lands on a canvas laying on the floor. Even if this makes a beautiful and original artwork, Gaut argues that it cannot count as creative due to the work being a product of pure luck (2010, p. 1040; 2003). Gaut does not insist that *no* luck can be involved in a creative process; he thinks that capitalising on luck is permissible. As Gaut states, "serendipity is the skilful use of chance, not pure luck." (2010, p. 1040). Gaut also makes no mention of intention here and characterises creative action without pure luck as "exhibiting at least a relevant purpose" (2010, p. 1040). Gaut does not add any requirement for intention here (if there is any need for intentionality is included in the requirement of agency).

Second, Gaut argues that following a mechanical search process cannot count as creative either. Gaut considers cases where "someone who produces something original and valuable simply by mechanically searching through all the possible combinations available to him" (Gaut, 2010, p. 1040). He points to Novitz's discussion of Charles Goodyear's discovery of vulcanisation, which involved Goodyear mixing raw rubber with a series of available materials (Novitz 1999, p. 75). Gaut states that this example should not be counted as creative because it shows no skill or understanding in its application (2010, p. 1040).

This case is perhaps not the best example of mechanical search. According to Novitz the actual discovery of vulcanisation came when Goodyear *accidentally* dropped one of his mixtures onto a hot surface (Novitz, 1999, p. 75), suggesting that this may be a case of 'pure luck' which is already excluded above. While Goodyear was indeed searching through some options, he was doing so fairly randomly, without any system in place. This seems more like 'blind variation', i.e., trying one thing after another as they are made available, and testing the outcome (selectively retaining) anything successful. We could ask whether a true mechanical search, systematically moving through a series of

possibilities, would not involve some level of understanding? Gaut does phrase his case for exclusion as "mechanically searching through all the possible combinations available to him". He does indeed appear to have in mind a person systematically searching through all the possibilities one by one. This process, however, cannot produce results without *some* understanding. Either there will be understanding of the options for possible combination and the reasons behind testing them, or some understanding of the criterion of success (evaluative capacity is another criterion of flair, according to Gaut). It is also not clear then why Gaut thinks understanding or skill is the solution to excluding mechanical searches, as it is possible for a mechanical search to involve understanding (as opposed to a random search as Goodyear did, with an accident in the mix too, though even in this case Goodyear had in mind what he was looking for). This then is a somewhat confusing inclusion in the condition of flair, but Gaut proposes two further cases to exclude.

Gaut's third exclusion to creativity is someone who is following exactly specified rules, such as a person following a paint-by-numbers in creating a painting. Deutsch (1991, pp. 210-1), referenced by Gaut, argues that in completing a painting by numbers one is bringing into existence something that has already been *created* by another person. Gaut says, to avoid such a case counting as creative, there must be room for individual judgement for a process to be considered creative (Gaut, 2010, p. 1040). We should note here that only exact rule-following is excluded from creativity. Following rules non-exactly, or rules that do not specify everything to be done, should *not* be excluded from creativity. Creativity exhibited within set rules is still creativity, and at times constraining one's processes may even enhance creativity (Elster, 1992; see also Levinson 2003 or Livingston 2009 for further discussion). In fact, examples of artists adhering to rules or constraints have frequently been examined in discussions of art and creativity. A central example would be *Litterature Potentielle*, where writers sought to give themselves strict constraints to follow in the creation of their writing (Symes, 1999). Further examples commonly discussed in the literature are the work of Igor Stravinsky, who also wrote of the need for constraint in music (Stravinsky, 1997; P D Stokes, 2008), and the forms and colours in the work of Mondrian (P D Stokes, 2008, pp. 226-8). It is

not just philosophers and artists who have advocated for constraints as key to creativity, but also psychologists such as Patricia D Stokes (2005, 2008).[46]

It is not clear whether Gaut agrees with the possibility of constraint being essential to creativity. Gaut's view is compatible with rationalism about creativity (the idea that creativity is in some way rational, see Gaut 2012, p. 268) which would suggest that working within some rules is not incompatible with Gaut's position. Gaut does not argue against the possibility of rules having a place within a creative process (Gaut 2012, p. 260). This could lead us to conclude that whilst Gaut rejects the possibility of creativity occurring where *specific* rules are followed (such that no room is left for innovation), he is not arguing that *any* level of rule-following in a process precludes creativity from occurring. If we set our own rules to follow (much like some exercises of *Litterature Potentielle*) then one could surely be creative, perhaps not in following the rules, but in working creatively within the rules, or in setting the rules in the first place. This ultimately would show the 'degree of judgement' in applying the rules, as long as the creator determined the rule and its overall application.

Finally, Gaut wishes to exclude cases such as animals and young children who, when painting, will continue to add and add to the image unless it is taken away from them (Gaut, 2010, p. 1040 citing Dutton, 2009). In impressive cases of, for example, chimpanzees painting it is not the animal who ends the painting when it is done. Instead, a trainer removes the image at a point where the work is aesthetically appealing (and before the image is 'over done'). Gaut argues that in such cases the animal (or child) cannot have been creative since they lack the capacity to evaluate their own work and know when to stop creating.[47]

Is it the capacity to evaluate, or the actual evaluation itself, that is key for Gaut here? Gaut summarised the missing feature in the chimpanzee case as "an evaluative ability directed to the task at

[46] It is worth noting that the view defended by Stokes is compatible with both Simonton and Boden's views, see P. D. Stokes (2008, p. 223).

[47] It does seem that there are some examples of chimpanzees who do (according to reports) demonstrate a desire to finish an unfinished work and refuse to continue when a work is finished (see Howell 2007, 374; Luty 2017, 383-385). This objection however might merely mean that some animals could be creative under Gaut's account (providing they can meet the other criteria).

hand." (2010, p. 1040). I do not think this has to be evaluation in relation to *finishing* a work, or to evaluation of *finished* pieces. Take the following work: *Dancer, Ready to Dance, Right Foot Forward* by Edgar Degas (modelled c.1885–90 & cast 1919–1921). This work was an unfinished study made of wax and clay. It was discovered after his death and subsequently cast in bronze (The Barber Institute). This work was in some way 'taken away' before Degas finished it (if indeed he ever intended to) and we could say that he did not use evaluation in its completion. Yet could he still not have been creative in his 'sketch' of the figure dancing? There could also be creativity in works that the creator evaluated as failures. Take the work of Giacometti. Giacometti's sculptures seem to clearly be creative, yet he frequently destroyed them, dissatisfied by what he had made. Here he has applied judgement, yet that judgement was to get rid of the work. Some of the Giacometti's works are only available for us to admire today because they were taken by friends; they would have otherwise been destroyed by the artist himself (Fondation Giacometti, 2020). Do we look on these works as uncreative in light of this knowledge? To return to the case of the chimpanzee or young child, despite making a mess of a painting at the end, there surely may still be creativity present. We could see evaluation in the *process* of making a work, not only in determining when it is complete. Consider a child who carefully painted a picture of a pink cat and then continued adding to the painting, eventually making a mess of wet paints. The process here, of producing an image of a fantastical animal, may nonetheless have been creative despite destroying the work after. I think therefore that Gaut intends to exclude cases where there has seemingly been no evaluation directed at the task at *any point* in the process.

Gaut summarises his account of flair as requiring the following:

1) a relevant purpose (not accidental, or by pure luck)
2) some degree of understanding or skill (not merely using mechanical search procedures)
3) a degree of judgment (in how to apply a rule if a rule is involved)
4) an evaluative ability directed to the task at hand.

For Gaut, creativity is a particular exercise of agency, and requires a three-part (rather than the typical two-part) definition.

APPLYING GAUT TO AI: ORIGINALITY (NOVELTY)

Gaut specifies that creativity "is open to agents, whether human or not, that have the requisite capacities" (2010, p. 1041). As such, there is no reason in principle why a non-human could not be creative. As I have said, I will address the agency requirement in Chapter 3. For now, I will examine other features of Gaut's account. We have already seen some of Boden's account of why AI could produce works with some originality and value, so I will give a short response to both of these conditions here in relation to specific AI systems. Boden argued that computational systems could produce works with some level of originality (mostly at the level of exploratory creativity). One could object that an AI system such as a GAN, AICAN or even DALL-E is not truly producing anything new, it is merely reproducing elements of the training set. We could quickly respond to this by pointing to the combinational mode of creativity outlined by Boden: putting together novel combinations of familiar ideas.[48] If a system is putting together existing components that it has seen previously in a novel way, this is still novel. It is simply novel in a combinational sense, as opposed to an exploratory or transformational sense.

In the case of, say, a GAN, the system has been trained on a large number of images. It learns what those images are like in some way, in order for it to judge whether its own images are similar to them. As we have seen already, there are methods to evaluate the outputs of GANs to ensure that they are not overfitting the training set (i.e., just producing images from the training set) (Xu et al., 2018). If a GAN is producing training images exactly, it is failing; so, we can exclude the possibility of direct reproduction in this system. If the model is not over-fitting, it is in one way or another producing novel images (even if they are only re-combinations of elements of the training images). The images produced may not only be re-combinations, however, as this ignores the input of the noise vector, which introduces an element of randomness into the system.

Large prediction models like GPT-3 have even less of a problem with overfitting, as they are trained on such large sets of data that they do not need many iterations of training on the same data

[48] Boden does not mean combinational in the exact way I will use it here, yet I think it is worth using it in this way, nonetheless.

points (T Brown et al., 2020). They are not directly reproducing anything they have been trained on. However, in the case of probability-based systems like diffusion models or text prediction models (e.g., GPT-3), we could argue that the aim of the system is to be *less* novel. The system is producing a probability-based output usually based on an input (often a creative one) from a human. If you prompt a diffusion model to produce an image of a "caterpillar wearing a suit of armour and riding a snail into battle" the system can produce a novel image. However, what it has produced is a likely image based on *your* unusual input. The large portion of the novelty in the image then should be attributed to the human prompter, and not the AI. If these models produce novelty of their own accord, it is of a low-level variety.

The CAN model, on the other hand, is different to both GANs and probabilistic models. It aims to produce images that differ from the set on which it was trained in terms of style, whilst still producing images that could belong to the training set (Elgammal et al., 2017). The CAN learns then to specifically produce images that are novel, not just in the sense of combining what it has seen before, but in producing something different to what it has seen before. This would be closer to Boden's 'exploratory' creativity. It is still limited to the bounds of producing a small, square, digital image, but it is not aiming to produce images that can simply slot into the training set. Some novelty or originality is then possible in AI systems.

APPLYING GAUT TO AI: VALUE

Along with originality, Gaut accepts the consensus view that creativity requires value. This excludes cases of worthless novelty from consideration as creative (Gaut, 2010, p. 1039). I will take it, for now, as clear that AI can produce works of value. Works by AI have sold for large amounts at auction (e.g., Christie's, 2018b), have been displayed in galleries (e.g., The Barbican Centre), are available to the

public for purchase (AI art shop), and are frequently shared over the internet. I take all these as signs that the works of AI have *some* value, and for now I will not interrogate this further.[49]

APPLYING GAUT TO AI: FLAIR

Let us look closer at the multiple aspects of flair and whether an AI could achieve these requirements. Before examining each of the conditions of flair against an AI, it is worth noting that there are two potential ways to understand the condition of 'flair'. Whilst Gaut creates the positive condition of flair, featuring (in the 2010 version) four conditions that a process must meet to have been done with flair, it certainly appears (as we saw above) that he created the condition of flair in order to *exclude* certain kinds of process (as opposed to including certain features):

> the process of a thing's making involves, as I will say, *flair*. This should be understood to at least rule out the cases where I produce something by a mechanical search procedure, or in which I produce something purely by accident. (Gaut, 2009, p. 86)

This is salient because in assessing whether an AI can meet each condition of flair, we can take each condition in two ways: whether an AI meets the positive condition (e.g., its work is produced with purpose) or whether it avoids the negative condition (e.g., its work was *not* created by accident). In some cases (such as in the case of mechanical search procedures) the positive and negative conditions may come apart.

1) a relevant purpose (not accidental, or by pure luck)

There is some randomness in systems like GANs and CANs[50] so we could consider this to be 'accidental' or 'luck'-based. However, as soon as these systems have been trained on a data set, they are no longer producing images through *pure* luck, but through some chance and (as I argue below) some skill.

[49] See Chapter 4 of value in AI works, in particular whether the value in AI art will align with human values.
[50] Whether this is indeed true randomness is debatable, however it is effectively randomness to the system itself.

2) with understanding or skill (not by mechanical search)

Gaut wishes to exclude mechanical search procedures from creativity. This may present a problem for some kinds of AI system. Gaut discusses mechanical search in relation to AI in 'Creativity and imagination' (2003, pp. 277-278). Here Gaut discusses the role of imagination in creativity, and argues against its role as allowing people to search through possible options:

> For consider Deep Blue, the chess computer which beat Kasparov in 1997. Deep Blue really does survey vastly more possible positions than any human could, and selects from them the one most likely to win the game. Deep Blue has in this sense a powerful imagination. But the problem is that it is the epitome of an uncreative way to play chess: it mechanically searches through the possible positions to arrive at the best. Kasparov in contrast, plays chess creatively, but cannot do so by surveying the vast numbers of possibilities that Deep Blue does. Creativity is precisely not a matter of a powerful imagination, in the sense of an ability to search through vast numbers of possibilities. (Gaut, 2003, p. 277)

Gaut clearly states then that IBM's Deep Blue plays chess without creativity. What of Deep Blue's spiritual descendant, the Go playing algorithm of DeepMind, AlphaGo (DeepMind, no date)? AlphaGo is an AI system designed to be able to (learn to) play the game of Go, a game previously thought too challenging for a computer to be able to play (let alone beat a human champion). AlphaGo went on to win against humans, including a high-profile series of games against a world champion, as documented in the film *AlphaGo* (2017). In a game against 9-dan world champion Go player Lee Sedol, AlphaGo made a surprising move – move 37 – which would never be played by a human opponent (*AlphaGo*, 2017). Many lauded this move as evidence of AlphaGo's creativity, including its opponent: "I thought AlphaGo was based on probability calculation and that it was merely a machine. But when I saw this move, I changed my mind. Surely, AlphaGo is creative." (Lee Sedol, quoted by DeepMind, no date).

By Gaut's account, it is unclear whether something like AlphaGo's much lauded 'move 37' would count as creative. AlphaGo systematically searches through options by way of a decision tree,

so it is in some ways conducting a mechanical search. The process by which it works is a Monte Carlo Search Tree, combined with policy networks and a value network. The policy network will suggest promising actions to take, and the value network will evaluate the board positions. This is coupled with the Monte Carlo Search Tree which simulates games to find the best action, i.e., the action with the highest probability of leading to a winning game. In other ways, however, AlphaGo is not *merely* a mechanical search procedure, despite a search through various possibilities being a key part of its process. The system has indeed searched for possibilities, but it has also simulated games and then evaluated the possible options before selecting the best possible course of action. It is also not possible for the system, advanced though it is, to model *all* possible games, so some pre-selection has occurred. I get the sense that Gaut *would* wish to exclude AlphaGo from creativity on the basis of this search procedure, but I am not convinced we can exclude it so easily. AlphaGo does not follow Goodyear's search process, as it is far more informed than Goodyear, and it does *not* search through a series of random options.

This said, I doubt that 'move 37' is the hill to die on for AI creativity. We can look to other AI systems that do not work via any kind of mechanical search. Generative Adversarial Networks (GANs), Creative Adversarial Networks (CANs) and, the latest technology on the AI art scene, Diffusion Networks (such as those used by DALL-E, Midjourney, and Stable Diffusion) do not rely on mechanical search procedures to generate images. If Gaut wishes to ensure that mechanical search is not included in creativity, then in these cases we need not be concerned. If, however, Gaut intends for flair to include understanding or skill *regardless* of mechanical search, then we have further work to do.

Gaut has written more on the relationship between creativity and skill (Gaut, 2009). He wishes to include skill as part of flair, in both excluding mechanical search and in excluding cases of pure luck (Gaut, 2009, pp. 85-86). Gaut here offers us a partial account of what skills are:

> I doubt if the notion of a skill is sufficiently determinate to be captured in terms of individually necessary and jointly sufficient conditions. Nevertheless, some criteria can be given, and the satisfaction of all of these is sufficient for an ability to be a skill. Some of the

> marks of skill are as follows. First, the capacity in some domain is special (not universally shared): we tend to talk of people as skilled in some activity when they have an ability that is not possessed by everyone who engages in that activity. Second, we talk of skills as a kind of accomplishment (the words 'skill' and 'accomplished' are near-synonyms). Third, it makes sense to talk about practicing one's skills: in practicing music making, I thereby practice my music-making skills. Finally, skills are something that one can learn, so we oppose skills to purely natural abilities. Breathing is not a skill, because it is not something one learns; but music-making is a skill, since it is learned, and (arguably) walking is also a skill, since that is something that is learned too. Note that I am not claiming that these marks are individually necessary conditions for a productive capacity to constitute a skill, but they are, rather, criteria for something to be a skill. (2009, p. 95)

We might think that some AI systems, particularly those which employ deep learning have some level of skill. Do they meet Gaut's criteria?

i) Not universally shared: The ability to generate images (for example) is not something universally shared by algorithms. The ability to generate *good* images (either judged as images that could be part of the training set as in GANs and CANs, or as judged by humans) is also not universal. Only some systems, and (perhaps more pertinently) only some iterations of systems can produce images judged to be good.

ii) Accomplishment: We might not necessarily consider an AI system itself to be accomplished, but this could be about our unwillingness to praise a machine.[51] However, the outputs of many AI systems are praised widely, as we have seen in the news, online, and in exhibitions and auctions. If we are able to attribute responsibility for the work to the AI (this will require some level of autonomy at least) then we may consider the AI to be accomplished.

iii) Practice: I will take practice to mean repeated exercise of skill. Many AI systems do indeed repeatedly exercise their abilities, either continually generating (and, for GANs and CANs,

[51] Perhaps because we do not see it as an agent as discussed in Chapter 3.

assessing and improving upon) their images or (in the case of DALL-E) generating multiple images in response to prompts.[52]

iv) Something that one learns: I can take at face value here that machine learning involves learning. One could object that a deep learning algorithm like GANs or diffusion models do not learn to make images as they are programmed to make images; thus, they do not learn the skill of 'image-making'. We could respond, however, that these systems learn to make images of a certain kind or quality.

I have argued that an AI system could meet each of the criteria of skill. In some places though, it seems that it is not understanding *or* skill which Gaut includes in 'flair', but just understanding. Here, it seems that either understanding *or* skill will satisfy Gaut: "Goodyear's discovery does not count as creative since it displays no understanding *or skill*" (Gaut, 2010, p. 1040, my emphasis). Elsewhere, it seems to be *understanding* that is emphasised and not skill: "in short, the kinds of actions that are creative are ones that exhibit … some degree of *understanding*." (2010, p. 1040, my emphasis). Whether an AI system can have *understanding* will need further interrogation, as this is a far more philosophically-laden issue. The potential lack of understanding in AI systems is discussed by Gaut in relation to creativity. I will examine this further below, as Gaut's brief discussion of AI creativity can be taken as an objection to the argument I offer here. Let's return to assessing AI against the components of flair.

3) a degree of judgment (in how to apply a rule if a rule is involved) (not specific rule-following).

If we are to have some level of autonomy in our AI system, this should ensure that the system is not *merely* following rules at least. Not following specific rules, for an AI, could be interpreted as not only following pre-determined rules, i.e., not only following rules assigned to it from an external source. Some might object that as an AI system is an algorithm, is it comprised solely of rules and

[52] If by 'practice', Gaut requires some qualitative difference between practice and performance, then this may be more challenging, as it would imply some understanding of context. It is not clear that this is necessary for the concept of practice; if a dancer only ever danced in performances front of an audience and their skill at dance improved, would this not be a skill? If a painter only ever painted for themselves in practice and their skill improved, would this too not count as developing a skill?

thus is, in its very nature, unable to exercise judgement in applying rules. In this case, we can appeal to the separation between the designer of the system and the AI itself. If we have a level of autonomy in the system, there will be behaviours that are not determined by a pre-existing set of rules. They may, however, be determined by the AI itself through deep-learning and self-evaluation.[53] In this case then, we could claim that judgement in application of any rule has been involved in the self-altering of the parameters of the system. As noted above, we need not remove all rules, merely ensure that there is not only following of rules determined by others. This will not work for all systems; we will likely require some level of autonomy in the system. However, as long as the system was determining its own rules for creating works (and not merely those pre-determined by a human designer, for example) then we should satisfy this condition of flair.

4) an evaluative ability directed to the task at hand

In the case of some AI systems (notably adversarial networks) evaluation is central to their architecture. As I have explained, both the GAN and CAN systems include a component called 'the discriminator' which evaluates the outputs of the image generator in the system. These outputs are evaluated against success criteria (as outlined in the section above on Darwinian creativity) and given a score. Consequently, we could certainly say that there is an evaluative ability at play in (at least some) AI systems.

[53] In the case of GANs and CANs, we also have the addition of randomness (or pseudo-randomness).

GAUT'S OBJECTION

Aside from the requirement of agency (which I will address in Chapter 3), it seems that Gaut's account *can* accommodate some AI as creative. In his critiques of Boden's account of creativity, Gaut does address the possibility of creative computers. Despite (as I have argued) the potential for AI to meet the requirements of Gaut's account of creativity, Gaut claims that his account is incompatible with AI creativity. This is (perhaps surprisingly) *not* due to the agential aspect of his account. It is instead due to the element of understanding:

> If, as I will argue shortly, creative activity requires some degree of understanding, then, if computers cannot exhibit understanding (Searle), they cannot be creative. And if not all domains of human understanding can be specified in terms of rule-governed operations (Dreyfus 285–305), then in these domains computational theories of creativity that are based on classical AI are not possible (Gaut 2010, p. 1037)

The second point here I will dismiss. First, there is no reason to argue (and Gaut does not seem to) that *all* domains of human understanding would be needed to provide sufficient understanding for AI to be creative. As Gaut says, in any domain where we cannot specify our understanding in a way that can be processed by an AI: "*in these domains* computational theories of creativity… are not possible." (my emphasis). This implies, though, that any kind of understanding that *can* be specified in the relevant way can have creativity in the relevant domain.

Second, Dreyfus, here quoted by Gaut, was writing from a Heideggerian perspective between the early 1970s and early 1990s (dependent on edition). Dreyfus argued that knowledge is tacit and cannot be fully articulated, and thus cannot be given to an AI (Fjelland, 2020). As summarised by Russell and Norvig (2010):

> Essentially, this is the claim that human behavior is far too complex to be captured by any simple set of rules and that because computers can do no more than follow a set of rules, they cannot generate behavior as intelligent as that of humans. The inability to capture everything

> in a set of logical rules is called the *qualification problem* in AI. (Russell & Norvig, 2010, p. 1024)

As Russell and Norvig state, this issue was indeed a serious one, but since Dreyfus's writing, the design of AI systems has changed to incorporate solutions to this and other problems:

> many of the issues Dreyfus has focussed on—background commonsense knowledge, the qualification problem, uncertainty, learning, compiled forms of decision making—are indeed important issues, and have by now been incorporated into standard intelligent agent design. (2010, p. 1025)

If we set aside the conceptual baggage of Dreyfus's approach, we can examine the shift in the methods of AI design from rule-focused approaches to bottom-up approaches (such as deep learning algorithms). While it certainly is the case that we have not (yet) created a General Artificial Intelligence (AGI), we have made considerable progress in terms of machine learning techniques, computational capacity, and the tasks that AI can perform (Bryson, 2018). Some domains of ability that were previously thought out of reach for AI are now within reach, including the generation of language and images. The basis of these advances in machine intelligence is a change in how they are designed, away from the classical AI architecture familiar to Dreyfus.

A further argument made by Dreyfus was for situated agents (Russell & Norvig, 2010, p. 1025). Many agree with Dreyfus (e.g., Bryson, 2018; Dennett, 2017; Claxton, 2015) that embodiment is key for understanding of *some* kinds. In order for us to share values, motivations etc. with a computational system (if it is possible at all), they ought to be embodied in the world as we are. If we think the kinds of understanding relevant for creativity are embodied ones, then this may be an issue. I examine embodiment and AI in Chapter 3. However, this perspective makes the assumption that AI will, or even should, follow a human model of understanding rather than another model (or even an AI specific one) (Negrotti, 2019). AI could follow some other kind of understanding that is unfamiliar to us.

Let us return to the first of Gaut's claims above: "If, as I will argue shortly, creative activity requires some degree of understanding, then, if computers cannot exhibit understanding (Searle), they cannot be creative." (2010, p. 1037). Gaut seems to indicate an acceptance of Searle's position on understanding in computers. This is a problem for my argument here which suggests that AI could be creative under Gaut's account. There are (at least) three possible responses to Gaut, which may still allow for AI creativity. The first would be to reject the necessity of understanding for creativity entirely. To do this we would effectively abandon Gaut's account of creativity. We could take this option and instead adopt Boden's account, the Darwinian account, or some other account of creativity. I will not take this option here as my focus is Gaut's account. The second option will be to accept Gaut's characterisation of flair in terms of excluding certain kinds of cases from creativity but answer the need to exclude 'mechanical searches' with something other than understanding. I have suggested this approach above, with the replacement of understanding by skill (Gaut's own term). The third option will be to reject Searle. Understanding is key to Gaut's account of flair, but Searle's argument that AI cannot achieve understanding is not. I will turn to this option next.

REJECTING SEARLE

Rejecting Searle's argument, as many have done (e.g., Boden, 1988; Harnad, 1989; Sprevak, 2007), is our third possible response to Gaut. I will briefly cover Searle's argument and some bases for rejection here, including Boden's own rejection of Searle. There is a vast literature on Searle's argument against Strong AI, so the level of depth here will be limited; however, I think it is the easiest (and perhaps the strongest) way to reject Gaut's argument against the possibility of AI creativity. If Searle is *wrong* about the impossibility of understanding for AI, then Gaut's suggestion that this is reason to think AI creativity may not be possible is also incorrect. We might, as a weaker position, argue that whilst Searle's argument fails, we do not yet know whether understanding is possible for AI (and thus we cannot yet conclude whether creativity is possible for AI either).

Let us now turn to Searle's argument, so that we might reject it. In Searle's 1980 paper 'Minds, brains and programs', he argues against the possibility of 'Strong AI'. Searle understands Strong AI as the idea that an AI that is programmed in the right way will have a mind (Searle, 1980; 1999). This is in contrast to 'weak' AI, where a computer programme can potentially simulate mental processes but does not actually have mental states (Searle, 1980; 1999). Searle makes his argument using a thought experiment: the Chinese Room.

The Chinese Room thought experiment goes as follows. A native English speaker (in some versions, Searle himself) is locked in a room. He has with him boxes of symbols (a database) and a book. The book provides instructions for manipulating the symbols (a programme). People outside the room push symbols under the door to the English speaker. Unbeknownst to him, these are questions in Chinese, and the symbols he has in his boxes are Chinese characters. By following the instructions in his book, the man in the room can take the symbols given to him (an input) and find a corresponding set of symbols to put back through the door (an output). With his (very advanced) set of instructions, imagine that our man can produce the right Chinese symbols to impress the Chinese speakers outside the room, leading them to think that the person in the room speaks perfect Mandarin (Searle, 1980; 1999). In this way, the programme (along with the database) has allowed the person to pass a kind of Turing Test.[54] Despite this, Searle claims, we would not say that the man in the room could understand Chinese. The parallel Searle wishes to make is between the man in the room and a computer. If through implementing the programme the man does not understand Chinese then neither would a computer in implementing a programme, as there is nothing a computer would have in that case that would go beyond the man in the room (Searle, 1999; Cole, 2020). By asserting that implementing a programme *does* result in understanding, Searle thinks we are making a mistake between syntax (the programme) and semantics (meaning and understanding) (Searle, 1999). Searle

[54] The Turing Test (TT), as we have already seen, is a test proposed by Turing in 1950 to establish whether machines can think. Turing proposed that we set aside the 'meaningless' discussion about whether machines can think, and instead consider the imitation game (TT). In brief, the game involves pitting a human and computer against each other, with another person interrogating each. Winning the imitation game involves the computer convincing the interrogator that it is the human, and not the computer. Turing argues that if a computer can imitate a human well enough, then it will be broadly accepted as a thinking machine.

later expands his position to include not just understanding, but also consciousness and intentionality (Cole, 2020).

Whilst Searle's original argument is now over 40 years old, its legacy remains. Some think there has been no conclusive response to Searle enough to put the Chinese Room to bed (Cole, 2020). Many, however, have put forward criticisms and objections to Searle's argument. Some examples include the systems reply, the robot reply, the brain simulator reply, the other minds reply, and the intuition reply.[55] Boden herself offers a response to Searle, which Gaut does not mention even in using Searle to reject her reasoning about AI creativity. Boden (1988) follows the lines of criticism of the systems reply and the robot reply. The systems reply argues that while the man in the room may not understand Chinese, that is not the same as whether there is understanding in the system as a whole. This, it is argued, is more akin to how an artificial intelligence would work: as an entire system. The robot reply argues that with the addition of embodiment and perception the Chinese Room would not hold; the reason the man has no understanding is because he is locked in his room with the instruction book, not out in the world. This second response is less helpful for us unless we wish to develop embodied systems.

I will not explore the responses to Searle further here. I seek only to cast doubt on Gaut's acceptance of Searle, whose position on machine understanding is far from settled. If we can reject Searle, then we do not need to accept that understanding is out of reach for computational systems. If we cannot reject Searle, then we might still be satisfied to use skill in the place of understanding when considering AI.

CONCLUSION: THE AGENTIAL THEORY OF CREATIVITY

To summarise this section, I have argued that Gaut's account of creativity can accommodate machine creativity on the condition that we ignore the issue of agency. I will later argue that agency may be

[55] See Cole (2020) for a comprehensive list of philosophers and their various responses to Searle.

possible for an AI to achieve. I have examined Gaut's account, and in particular his requirement of 'flair', and shown how his conditions are not out of reach for an AI system. I have also examined a possible objection to AI creativity from Gaut himself, arguing that this hinges on the acceptance of a highly disputed position, and not on the fundamental element of the agential account of creativity. This agential account of creativity is not Gaut's final word on creativity; in 2018, Gaut adds the concept of *spontaneity*. I will now turn to examine this addition.

Gaut's spontaneity: A problem for AI?

Spontaneity has recently been included as a necessary condition for creativity in Gaut's account (2018). For human creative agents, the condition of spontaneity is an achievable goal, but it is a high standard for AI to reach. Does there need to be such a high standard? Gaut's paper on the value of creativity focuses on the role of spontaneity in creativity. In this section, I will first examine spontaneity in Gaut's account. I will then raise two issues with Gaut's conception of spontaneity: an epistemological concern and a normative concern. I propose an alternative to spontaneity: unpredictability. I will define unpredictability in relation to creativity, then show how unpredictability might solve the issues with spontaneity in creativity whilst still capturing the useful elements of spontaneity in Gaut's account. I will then show how using unpredictability instead of spontaneity can allow AI to meet this requirement of creativity. Finally, I will address objections that may arise in using unpredictability as a condition of creativity.

GAUT'S ACCOUNT OF SPONTANEITY IN CREATIVITY

Spontaneity in Gaut's account is a necessary condition of creativity. It sits within the larger theory of creativity, which we have already examined. To summarise, Gaut's theory is agential, that is, one must be considered an agent in order to be creative. An agent has agency, the ability to commit "individual, intentional actions" (Smith, 2018, p. 370). The output of the creative act must have originality (often called novelty), value, and flair (Gaut, 2010, pp. 1040-1041). In his 2018 paper Gaut adds to this account by further discussing the *source* of value in creativity. Here, he suggests spontaneity is both a source of value and a necessary part of creativity. Gaut states that there is a "constitutive connection between creativity and spontaneity", stemming from the relationship between creativity and ignorance (Gaut, 2018, pp. 133-134). Gaut claims there is an *a priori* principle, the 'ignorance' principle, that applies to creativity. The ignorance principle is defined as follows:

> Ignorance Principle: (IP) If someone is creative in producing some item, she cannot know in advance of being creative precisely both the end at which she is aiming and the means to achieve it. (Gaut, 2018, p. 134)

Gaut restates this as

> In creatively producing an item at time t_1, the creative person cannot know at some earlier time t_0, the exact nature of the item and how to produce it, for then she would have been creative at t_0, rather than t_1. (Gaut, 2018, p. 134)

Gaut uses this to distinguish between someone merely fabricating an item versus exhibiting true creativity. By way of example, Gaut contrasts a craftsperson with an artist: a joiner making a chair will know both what the chair will be like and how she will make it. The artist, on the other hand (if she is being creative), in making a new artwork cannot know both what the artwork will be like *and* how she will make it. This 'ignorance' is what allows the artist to be creative. Gaut states that if the IP is true it follows that "the creative person cannot have an exact plan of what she will do prior to being creative" (Gaut, 2018, p. 135). Having an exact plan would entail knowing both the ends and the means for creating something, and therefore would violate the IP. In Gaut's account, then, something cannot be entirely pre-planned if it is truly creative. Gaut uses spontaneity to capture the resulting assumption of the IP; an act cannot be entirely pre-planned if it is creative. Gaut's definition of spontaneity is as follows:

> I do something spontaneously if I do not plan it in advance, but do it on the spur of the moment … In this sense, actions that are not entirely planned in advance have an element of spontaneity. (Gaut, 2018, p. 135)

The sense in which Gaut utilises spontaneity is limited in that his account does not permit any alternative readings of spontaneity. He states that spontaneity could also mean occurring without the presence of external factors (such as in spontaneous chemical combustion) or occurring independently of will (such as thoughts that pop into one's head without prompting), but these are not what Gaut wants us to be concerned with. Spontaneity here is used only in the sense of being unplanned.

THE PROBLEM WITH SPONTANEITY

There are issues that arise with Gaut's inclusion of spontaneity as a requirement of creativity. The first are epistemological concerns: how can we know when something is spontaneous? Including spontaneity presents difficulties with recognising or judging creativity. If we subscribe to Gaut's account, we may struggle to accurately recognise creativity. These limitations open up issues of reliability and unnecessary chauvinism. In turn, this makes it very difficult for non-humans, including AI, and even some humans to be considered as creative.

The issue we encounter with spontaneity in Gaut's account is that it becomes challenging to make judgements of what is creative. If we require spontaneity for something to be considered as creative, then we must know, or at least have a good idea, whether the creative agent acted spontaneously in carrying out that act. How can we recognise this, other than if the act is done by ourselves? We can try to guess from the output or watching the process, but even then, we do not know for certain. Watching the process is not sufficient to judge whether there was an element of spontaneity in the creative act. Imagine that you are watching what you think is an improvised dance piece. How do you actually know this is improvised? It may appear unpolished perhaps, but is this reason enough for us to assume it is spontaneous? How do we know it is not carefully rehearsed to appear improvised? We have no real way of knowing just from observing, even if what we are observing is the process of creation itself.

We may face even more problems if the creative process is already over. How could we look at a painting and know that it was done without planning? Imagine two individuals making two paintings. The first was created by an artist, Bob, who did not perfectly plan the outcome before beginning. The second is created by a studio assistant, Ross, executing the instructions of another artist. How could we tell, just from observing the paintings, which was Bob's and which was Ross's? Perhaps an expert might be able to discern the hand of the artist, though even this will not ensure the piece was made with some lack of planning.

Even reports from the creative actor are not certain to clarify whether the process included an element of spontaneity. We must be willing to trust that the reports are accurate, and not exaggerated to seem impressive (if Gaut is correct that spontaneity is a source of value in creativity, such exaggeration is not at all implausible). Furthermore, self-report relies on us being able to communicate with the creative actor. We must be able to ask them, or they must have spoken or written on the matter. What if the creator is dead? Or if they refuse to speak on the matter entirely? Or, indeed, is non-human? Unlike novelty and value (the first two requirements of creativity in Gaut's account), we cannot make a good judgement of whether something is done with spontaneity without internal access to the creative actor. Even flair, Gaut's third requirement, could be assessed more easily via observation. For example, we could likely assess whether the artist was not exercising judgment in making the work, or if the work was made by an accident, through observing the process of creation.

I should note here how this epistemological criticism of Gaut's position differs from the problem presented in 'the intentional fallacy' (Wimsatt & Beardsley, 1946), which has been roundly criticised. The epistemological problem here is relevant, in part, because it is not the product we are aiming to recognise as creative for Gaut; it is the process. In order to recognise that a *process* is creative, we must have some understanding of the process itself. Unlike flair, which allows us some access through observation, and novelty and value which can be recognised from the final product, spontaneity appeals to the internal. It is also relevant that Gaut's insistence on an absence of planning makes spontaneity even more difficult to recognise. We may be able to reliably judge intentions from behaviour, or output, but can we reliably judge a *lack* of intention in the same way? One might think that this is not a problem; however, as I will propose, there is a possible alternative to spontaneity that will not lead to the same epistemological concern.

There is also a normative concern. Discussion of planning asserts several complex mental states and a conscious awareness of these in order for an agent to be creative. To plan, we must have an end in mind and we must know how exactly to achieve that end. Doing something spontaneously, or 'unplanned', still assumes an awareness that something is being done without planning; in order to act without planning one must surely be capable of planning, or else there is no distinction to be made. This

again likely precludes certain kinds of agents from achieving creativity. It limits creativity not just to humans, but only a small subset of humans (e.g., young children are excluded because they are not aware that they are acting without planning). In short, including spontaneity is *chauvinistic.*

Could an AI achieve spontaneity as Gaut characterises it? In the field of AI, spontaneity refers to a different phenomenon: instead of 'lack of planning', spontaneity refers to 'without external influence', a version of spontaneity that Gaut excludes. Those who design AI systems often do not know how or why an AI has developed a spontaneous feature, but this lack of knowledge or influence cannot tell us anything about the planning, or lack thereof, on the part of the AI. Furthermore, we are unable to gain any self-report from current 'creative' AI, yet we cannot establish from this that the AI has *not* been creative. Is there any essential reason to limit creativity in this way? Could we capture the lack of planning that Gaut solves with spontaneity through other means? This leads to my proposed alternative to spontaneity: *unpredictability*.

AN ALTERNATIVE TO SPONTANEITY

Even if spontaneity seems like a good way to include the IP in the requirements for creativity, one might consider if there is an alternative. I propose *unpredictability* as an alternative to Gaut's spontaneity. The proposed definition of unpredictability is as follows:

> Outcome X is unpredictable at time t IFF X cannot be determined at time t, where determination occurs relative to the observer.[56]

Unpredictability does not result in the same epistemological concerns as spontaneity. An external viewer can judge unpredictability to some extent. A viewer can make predictions or have expectations about a work or process that are proven to be incorrect. This allows the viewer to see it as somewhat unpredictable, and therefore potentially creative. Although this is not necessarily the same as the

[56] I include 'relative to the observer' because I do not think it is necessary to take a position on free will vs. determinism here.

unpredictability that the creative actor faces, the shared factor of novelty may allow some relation between the creator and observer in judging creativity.

Unpredictability, like Gaut's spontaneity, can also be shown to follow from the IP. Take an artist producing a painting: if the artist cannot foresee how that painting will be, they cannot know both the means and end of the painting. There is some degree of unpredictability in the process of producing a creative painting; it is not yet determined. They will not be able to say for certain until they have been creative, as once they can say for certain, the creative moment has passed. If they *have* already determined how the painting will be then they are in violation of the IP:

> In creatively producing an item at time t_1, the creative person cannot know at some earlier time **(time of prediction)** t_0, the exact nature of the item and how to produce it, for then she would have been creative at t_0, rather than t_1. (Gaut 2018, p. 134, my emphasis in **bold**)

If, at the time of the prediction, there is an exact plan in place, then the person in question could perfectly predict the outcome. The normative concern is also removed. We can discuss pre-determination easily for AI systems. This is a lower bar to reach for an act to be considered (somewhat) creative, while still maintaining the IP. Unpredictability has an additional upside, in that it can be operationalised for an observer; when recognising creativity, we can consider whether the outcome was predicted or not. Whereas spontaneity either has or has not occurred, we could consider unpredictability to come in degrees.

In the case of AI, can we say whether a produced outcome was unpredictable? In order to discuss this possibility, it is necessary to assess the state of 'creative' AI. Let us again turn to the case of GANs and CANs. Could their creative outputs be considered unpredictable?

In the case of GANs, there is some level of unpredictability: the noise vector provides random input into the generator. This random input means that the output will never be perfectly predictable, as, until the process of generation is complete, there is no way to determine the output. However, with GANs, there is a limit to the level of unpredictability that can be achieved relative to an observer. As the generator produces images that become closer and closer to the qualities of the training set

(according to the discriminator), they converge to a point where they are much more predictable, even if they cannot be perfectly predicted.

The CAN, on the other hand, will not face this issue. While it will never produce wildly unpredictable images as it still functions within a limited framework, it will not face the same limitation as GANs with producing images that increasingly resemble the training set. The CAN will also include random input from the noise vector, so some level of unpredictability will be maintained relative to the observer. As CAN is also programmed to produce variation from the training set, it will maintain a level of variation which, combined with the random input from the noise vector, should result in unpredictable output from the system and will similarly not be pre-determined. If this output meets the other conditions of creativity (it is new and valuable, for example) then this could be considered creative. From the examples of GANs and CAN, it is possible to see how unpredictability can function as a requirement of creativity for AI.

OBJECTIONS

There are several potential objections to unpredictability as a proposed alternative to spontaneity. I will address the objection from determinism, the objection from agency, the potential contradiction of predictable spontaneity, and the rejection of concerns about spontaneity.

The first objection to the use of unpredictability as a condition for creativity would be that this assumes a non-deterministic universe.[57] If determinism is true, then everything can be predicted. There is no 'unpredictability' and therefore no creativity. We can respond to this in several ways. First, if we do live in a deterministic universe, this fact has not allowed us to actually make accurate predictions of the world as it exists today. It is unlikely that humans will ever manage to collect and analyse enough information about the universe to predict perfectly the things that occur in it, such that we can predict

[57] Creativity has been challenged as incompatible with determinism. This view is addressed by Maria Kronfeldner in "Creativity Naturalised" (Kronfeldner, 2009)

how an artist paints before even they know. In the case of the everyday, determinism does not affect our ability to find things unpredictable – even our own actions.

Another response to this objection would be to accept that yes, in a deterministic universe there could be nothing that is unpredictable. If this were the case and we knew how to perfectly predict occurrences, would any instance of creativity seem to be truly creative? If we could say in advance of a moment of inspiration that that moment was going to come to be and that it would lead to X creative outcome, would this seem any different to something being completely pre-planned? In the case of the perfectly predictable moment of creative action, does it not cease to be creative? It would violate the IP, as a person *would* know at an earlier time (time of prediction) the exact nature of the item and how to produce it. This would not, then, be considered creative.

A second objection against replacing spontaneity with unpredictability would stem from Gaut's theory being an agential one. If Gaut is focused on the creative agent, why do we need to worry about the problem of judging creativity? A response to this would be that although Gaut's theory is agential (i.e., asserting that one must be an agent in order to be creative), this does not mean that the theory applies only to agents and not creative processes or creative products. For example, in discussing the IP Gaut states that,

> if an activity is creative, IP applies to it; if IP applies to an activity, the activity cannot be precisely planned in advance; if an activity is precisely planned in advance, it has an element of spontaneity. It follows that creative activity must have an element of spontaneity. (Gaut, 2018, p. 135)

It is the creative act here that is being discussed; spontaneity does not apply just to the agent, but also to how the creative activity is carried out and the resulting creative output. Furthermore, the issue of creative judgement does not so clearly apply to all necessary elements of creativity in Gaut's account. An observer can, for example, make a judgement of whether a creative output is original or has value (unless it is only original or valuable to the creator themselves, in which case we would rarely discuss it being creative anyway).

Another objection is that some instances of spontaneity may be predictable. If we can predict some forms of spontaneity, then unpredictability and Gaut's account of spontaneity cannot be accounting for the same thing. My use of unpredictability is failing to account for what it needs to. The response to this objection is that predictable spontaneity would not occur in creativity. If novelty (another necessary condition of creativity in Gaut's account) is achieved, there should be some degree of unpredictability maintained. A weak version of this would be that this response would only apply in cases of historical (H) creativity (see Boden, 2004, p. 2). H-creativity is when a creative act is the first instance of that creative act in all of history. In the case of H-creativity, it does not make sense for *anyone* to be able to say in advance both the ends and means of that instance of creativity. If they could perfectly predict it, it would no longer be an instance of H-creativity, as the moment of creativity would have already occurred. The use of unpredictability here allows us to consider degrees of predictability.

In the case of psychological (p-) creativity we can imagine a case where a predictable and spontaneous act could still potentially be an example of creativity (Boden, 2004, p. 2).[58] I would suggest that in these cases, however, there is still some level of psychological unpredictability. That is, as the creative act is novel to the actor, that creative act maintains some unpredictability for that actor. For an external observer there is also likely to be a small degree of unpredictability that would allow the act to still be considered creative. What about when we make incorrect judgements? When an act seems unpredictable but was planned? In this case, we would likely still judge that to be creative, even though we might be incorrect. When we discuss creativity, it is not spontaneity that we are tapping into, but our inability to foresee what would be the outcome.

The final objection is a simple one: so what? Judging creativity should not alter what we consider creativity to be. An act either is or is not spontaneous, and our requirements for creativity should not depend on the possibility of judging creativity. Furthermore, uncertainty about an act's spontaneity might be mirrored in a corresponding uncertainty about whether an act is genuinely

[58] For example, a young child might spontaneously draw a picture with a circle in the sky, adding lines radiating from it, i.e., a classic representation of the sun. We could predict the child might draw such a picture at some point, yet the act was, for the child, spontaneous.

creative. If this is the case, it does not matter that we have a hard time judging spontaneity. Spontaneity is still necessary for creativity. To respond to this, we can question the necessity of spontaneity in an account of creativity. Gaut's inclusion of spontaneity follows from the IP. Including the IP as a necessary condition of creativity appears to be a primary motivation for including spontaneity in his account, as well as discussing spontaneity as a source of value (value is a separate necessary condition of creativity for Gaut, so removing spontaneity will not remove the need for value). However, I have shown how unpredictability can also follow from the IP, and therefore ensure that the IP is not violated in an act meeting the necessary conditions of creativity. In this case, why must we assert spontaneity, which raises epistemological and normative concerns for creativity, when there is a possible alternative?

CONCLUSION: SPONTANEITY

To conclude, Gaut's inclusion of spontaneity as a constitutive part of creativity is problematic in the way it limits our ability to judge creativity. It results in a chauvinistic conception of creativity that excludes non-human creativity, which is hard for AI to reach. I have proposed here an alternative to spontaneity, unpredictability, as a requirement for creativity. Unpredictability can still accommodate the ignorance principle that motivates the inclusion of spontaneity for Gaut. Unpredictability also allows us to make judgements of creativity. This is particularly useful in cases where we are not able to access the internality of the agent to determine whether the creative act was perfectly planned or not. This is helpful for AI creativity, as we can discuss the level of unpredictability of the creative act without needing insight into the internal processes involved in that act. We looked at the case of GANs and CANs which show us that a limited degree of unpredictability can be observed. Finally, I addressed some objections to unpredictability as an alternative to spontaneity.

Conclusion: Can AI *Create* art?

In this Chapter, I have examined three accounts of creativity, assessing whether AI could be considered creative under each approach. First, I explored the Darwinian model of creativity, arguing that this model can be used to assess creativity in AI systems. I argued that the cases of GANs and CANs can both meet the requirements of the Darwinian model. Next, I examined Boden's account of creativity. Boden argues that computers could be creative under this account, though later admits the objection of autonomy. I argued in response to this objection that some level of autonomy is possible for AI. Finally, I examined Gaut's agential account of creativity. Setting aside the requirement of agency, I argue that the other criteria put forward by Gaut in his three-part definition of creativity are possible for AI. I then examined his additional requirement of 'spontaneity' arguing that this could instead be conceptualised as 'unpredictability', and thus also could be possible for AI.

In this chapter, I have set aside Gaut's requirement of agency, which is central to his theory of creativity. Without agency, an AI cannot be creative. In this next chapter, I will turn to address key limitations of AI that may prohibit AI systems from being creative or making art. I will examine the mind and the social world, embodiment, and agency. I will argue that a (minimal) level of agency is possible for AI, and thus creativity (at least in a minimal sense) is also possible, as long as the other requirements are met.

Chapter 3 - Can *AI* Create Art? Key Limitations

In this chapter I offer solutions to a variety of apparent limiting factors for AI art and creativity. First, I offer a summary of approaches to the mind in philosophy, before turning to examine the possibility of AI extended mind. In this section, I argue first that AI could have mind. Then, I argue that AI could extend this mind (via the extended mind thesis) to people. Through this extension, AI could gain access to the social world which it might not otherwise be able to access. This links back to concerns with art – if an AI can have some (mediated) access to the social world, it may be able to make works which are socially situated.

Second, I examine the possibility that creativity might require embodiment. If it does, it might seem that AI creativity can be ruled out without further investigation. To the contrary, I argue that AI could be embodied under one conception of embodiment. Furthermore, if an AI system achieves the other criteria for creativity, then it will necessarily be embodied on this account.

Finally, I turn to consider agency. As we saw in the previous chapter, agency is a key requirement of Gaut's agential account of creativity, is related to the conditions of 'action' in the cluster account of art, and implied in the institutional account of art. In this section, I argue that AI can achieve a minimal account of agency, and that AI could be creative (in a correspondingly minimal sense). This chapter will be the final part of my overarching argument for why AI can make art and can be creative.

The Mind

Before moving forward, it will help to give some background on theories of mind and approaches to the question of whether AI can have mind. Many of the positions in philosophy of mind relate to the mind-body problem. The mind-body problem is: "what is the relationship between mind and body? Or alternatively: what is the relationship between mental properties and physical properties?" (H Robinson, 2023).

DUALISM

This is the problem that Descartes first identified. Descartes argued that the mind and the body are distinct (the position since referred to as "mind-body dualism") (Skirry, 2006). This distinction was argued for in Descartes' *Sixth Meditation*:

> [O]n the one hand I have a clear and distinct idea of myself, in so far as I am simply a thinking non-extended thing; and on the other hand I have a distinct idea of body, in so far as this is simply an extended, non-thinking, thing. And accordingly, it is certain that I am really distinct from my body, and can exist without it. (AT 7: 78) (Brown, 2014)[59]

This can be generalised to:

1. I have a clear and distinct idea of the mind as a thinking, non-extended thing.
2. I have a clear and distinct idea of body as an extended, non-thinking thing.
3. Therefore, the mind is really distinct from the body and can exist without it. (Skirry, 2006)

Descartes argues that the mind and body are distinct, and thus that the mind can exist without the body. This is due to Descartes conception of clear and distinct ideas: these ideas must be true because

[59] Here Descartes uses the first person. D Brown (2014) emphasises that Descartes is not merely arguing that the mind is distinct from the body, but that *he* is distinct from his body (and thus that he is exclusively identified with his mind).)

God would not allow us to be so misled (Skirry, 2006). Of course, the soundness of this argument is dependent on Descartes being correct that God exists and cannot deceive us.

PHYSICALISM

Physicalist (or materialist) views defend the idea that despite how it may appear to us, mental states are just physical states. There are many examples of materialism in philosophy of mind, including behaviourism, mind-brain identity theory, and (arguably) functionalism and the computational theory of mind (Howard, 2023).

Behaviourism

By the mid-20th Century, dualism had been widely criticised but Gilbert Ryle aimed to put the final nail in the coffin of dualism. Ryle argues that Descartes made a category mistake when proposing dualism, assuming that the mind was in the same type of thing as the body (Montero, 2022). Ryle compares this error to a visitor to a university viewing the buildings of Oxford and then asking 'but where is the University?' (Ryle, 2009). Here the visitor has made a category mistake; they expect the University to be some additional thing, like a building (Montero, 2022). The idea then is that Descartes is mistaken by looking at a person and asking 'but where is the mind?'. The mind is not some other thing, separate from the body. Thus, the dualist has made a category mistake. Ryle proffers instead what is often thought of as a form of behaviourism, as he says "Overt intelligent performances are not clues to the workings of minds; they are those workings." (Ryle, 2009)

Ryle was one of many scholars that turned to behaviourism in the mid-20th Century, when it became a popular family of views in psychology and philosophy. Broadly speaking, behaviourism places the focus not on mental states, but on observable behaviour. Some philosophical beviourists go so far as to deny there was such a thing as mind (Brook & Raymont, 2021), holding the position that the mind *is* behaviour (or the disposition to behave). This avoids the problem of other minds, that is, the problem of how one can know anyone else has a mind (Montero, 2022).

Despite its apparent usefulness in avoiding the issues of dualism and the problem of other minds, behaviourism faces considerable objections. Hilary Putnam (1963) posed one such objection: that of the super-Spartans. The super-Spartans are a thought experiment designed to illustrate how one could have a mental state, yet exhibit none of the typical behaviours associated with that state. Image a world where it is considered terrible manners to show any sign of being in pain. Imagine that from a very young age, children in this society are taught not to show pain; you do not say 'ouch' or grimace or cry out (Montero, 2022).[60] In fact, we could imagine that such 'super-Spartans' do not even go somewhere in private to express their pain. They have simply learnt to suppress it (Montero, 2022). Despite this lack of behaviour, would we not still wish to say that the Spartans are in pain? If so, it seems that mental states are not the same as behaviour.

We might think that the behaviourist can avoid this objection by appealing to disposition; the super-Spartans do not behave in a pained way, but they are still disposed to. Yet, we could imagine still a case where a super-super-Spartan is no longer even disposed to behaving in this way (Montero, 2022).

This is one version of a common argument against the behaviourist position, also known as the "perfect actor" or "doppelgänger" arguments. It can be formalised like this:

> P1. If behaviorism is true, it is not possible for there to be a perfect actor or doppelgänger who behaves just like me but has different mental states or none at all.
>
> P2. But it is possible for there to be a perfect actor or doppelgänger who behaves just like me but has different mental states or none at all.
>
> P3. Therefore, behaviorism is not true. (by modus tollens) (Polger, no date)

Putnam proposed his own theory of mind in opposition to behaviourism: mind-brain identity theory.

Mind-brain identity theory

[60] One might think of the Unsullied in *Game of Thrones* (2011) the army of eunuchs trained to not exhibit pain.

Identity theory is a collection of approaches to the mind-body problem, which says that some (or all) mental states are identical with some (or all) physical states. There are variations on this view, such as type and token identity theory, and variations to the extent to which philosophers argue that mental states are the same as physical states (S Schneider, no date). Type identity theory claims that mental states are reducible to physical ones; our psychological experience are just processes of the brain (not merely correlated with them) (Smart, 2022).

Type identity theory arose in the 1950s-1960s with work by Place (e.g., 1956), Feigl (e.g., 1958) and Smart (e.g., 1959) proposing versions of the theory claiming that some types of mental state were identical to some types of brain states (S Schneider, no date). Unlike these prior versions of type identity theory, Armstrong (1968) claimed that all mental states, including dispositional mental states (beliefs, desires, intentions) could be identified with states of the brain (Smart, 2022; S Schneider, no date). While a radical view at the time, some identity theorists came to agree with this view (e.g., Smart) whilst others (e.g., Place) did not, due to their understanding of the nature of dispositional states (Smart, 2022).

Token identity theories make more modest claims than type identity theories. A token identity theory focuses on particulars rather than types of mental state, holding that each particular mental kind can be identified with something physical (S Schneider, no date). Whereas type identity theory would identify pain as a physical state, token identity theory would identify having a particular pain with also having a particular brain process or state (S Schneider, no date; Smart, 2022). Type identity theory entails token identity theory, but token identity theory does not entail type identity theory (S Schneider, no date). As the more modest theory, token identity theory is more compatible with later, popular views such as functionalism. However, type identity theory suffers? from several objections. One of the most impactful objections was made by Hilary Putnam in 1967: *multiple realisability.*

Multiple realisability is the name given to the argument first proposed by Putnam, that many different physical kinds can give rise to the same psychological kind (Bickle, 2020). Putnam writes:

> Consider what the brain-state theorist has to do to make good his claims. He has to specify a physical-chemical state such that *any* organism… is in pain if and only if (a) it possesses a brain of a suitable physical-chemical structure; and (b) its brain is in that physical-chemical state. This means that the physical-chemical state in question must be a possible state of a mammalian brain, a reptilian brain, a mollusc's brain… etc. At the same time, it must *not* be a possible… state of the brain of any physical possible creature that cannot feel pain… [I]t is not altogether impossible that such a state will be found… [I]t is at least possible that parallel evolution, all over the universe, might *always* lead to *one and the same* physical "correlate" of pain. But this is certainly an ambitious hypothesis (Putnam, 1967, p. 436).

In order for 'brain state theory' (type identity theory) to be true, the physical state underlying a mental state must be the same across all instances of that mental state. This applies even when the brain differs.

As we know that many different types of creatures feel pain, there would need to be a physical occurrence that was common to all these creatures (as this is a requirement of type identity theory). But this does not seem to be the case. As Bickle writes: "Neuroanatomy differs across terrestrial brain-bearing creatures, especially at systems and circuit levels; function, especially sensory functions, are increasingly "corticalized" as cortical mass increases across species." (Bickle, 2020). It seems then that mental states cannot necessarily be identified with just one physical state. The argument can be stated as follows:

1. (the multiple realizability contention) (At least) some mental kinds are multiply realizable by distinct physical kinds.
2. If a given mental kind is multiply realizable by distinct physical kinds, then it cannot be identical to any one (of those) specific physical kind.
3. (the anti-identity thesis conclusion) (At least) some mental kinds are not identical to any one specific physical kind. (Bickle, 2020)

The second premise here could be taken to open the door to artificial mental kinds (see below), which Putnam raises as an example, questioning if we should deny that a robot feels pain on the basis only of the difference to human brains (Putnam, 1964).

Philosophers have attempted to save type identity theory from Putnam's argument. These generally follow one of two strategies. Either defenders limit the claims of type identity to particular species or particular physical structures, or they will extend the claims of type identity to allow for potential disjunctive physical kinds (S Schneider, no date).

Despite these attempts, identity theory is typically thought to have been superseded by functionalism, though there are similarities between the views (S Schneider, no date). Early multiple realisability led to arguments for functionalism; instead of identifying mental states with physical states, functionalism identifies them in terms of their causes and effects (Bickle, 2020). For example, instead of identifying pain with a particular neurophysiology or neurochemistry, we can identify it by the cause (physical damage, say) and effects (e.g., crying out, believing one is in pain etc.). Any internal state which links similar causes and effects can be called 'pain', regardless of the physical mechanism underlying it. Functionalism then can be seen as a level of abstraction beyond identity theory (Bickle, 2020).

Functionalism

Functionalism is the view that mental states are identified not by their make-up, but by what they do, or what role they play in a system (i.e., their *function*) (Levin, 2023; Polger, no date).

Functionalism takes the view that "the identity of a mental state to be determined by its causal relations to sensory stimulations, other mental states, and behavior." (Levin, 2023). Whilst functionalism is frequently seen as a materialist (or physicalist) position, it is in opposition to type identity theories, and to behaviourism.

Mind-brain identity theory argues that mental states are states of the brain (and thus require brain matter), and behaviourism rejects the very existence of mental states as something separate from behaviour (or behavioural disposition) (Polger, no date). Whilst functionalism also focuses on cause

and effect, unlike behaviourism it does not deny the existence of mental states. Mental states to the functionalist are typically distinct internal states, several of which could lead to the same behavioural outcome (Polger, no date). And, contra the identity theorist, it is what the mental states *do* that define them, not what they are made of (Polger, no date).

Arguments for functionalism stem from arguments against its alternatives: behaviourism and identity theory. For example, as above, arguments against behaviourism include the issue of the "doppelgänger": that two creatures are behaviourally identical but have different mental states. Behaviourism cannot distinguish between these two creatures. Functionalism, however, will not suffer from this problem, as it does not deny the existence of mental states (Poger, no date). Perhaps more well-known are the arguments against mind-brain identity theory, i.e., arguments from multiple realisability (as above). Functionalism in part arose from Putnam's argument that mental states must be realisable through a variety of neurophysiological structures (and even non-biological structures), and thus mind-brain identity theory is incorrect. Functionalism, more abstract than mind-brain identity theory, allows for multiple realisability (Block and Fodor, 1972). Block and Fodor point out that where there is multiple realisability in other kinds, functional categorisation is also used. Take the typical example: a mousetrap. A mousetrap can be a spring action device, or be a sticky glue-based trap. Both are mousetraps, because regardless of what they are made up of, they are defined by their function (Bickle, 2020). As Polger (no date) writes, the argument for functionalism from multiple realisability has faced increasing resistance, yet remains the most influential argument in favour of functionalism. In addition, functionalism does not hinge on multiple realisability; even if the argument is unsound, the functionalist can rely on other arguments in their favour (Polger, no date).

Functionalism, particularly machine state functionalism (the view initially proposed by Putnam as an alternative to mind-brain identity theory) is closely associated with artificial intelligence. Under machine state functionalism, a system is considered to have a mind when it has appropriate functional organisation, and mental states are simply states that play relevant roles in this functional organisation of the system (Rescorla, 2020). This kind of functionalism marked the

introduction of computational theory of mind into philosophy (Rescorla, 2020), which we will turn to now.

Computational theory of mind

Computational theory of mind holds the position that the mind is a computational system (Rescorla, 2020). This view is generally understood to underlie contemporary cognitive science (Milkowski, no date), and has close association with much of the history of artificial intelligence.

Putnam's (1967) machine state functionalism claims that any creature which has a mind can be thought of as a Turing Machine (Levin, 2023). A Turing machine, proposed by Alan Turing in a 1937 paper as an 'automatic machine' is an abstract computational device that can help us think about the limits of computation (De Mol, 2021). The operation of the Turing machine is finite, and can be fully stated by a set of instructions, taking the following form:

> If the machine is in state S_i, and receives input I_j, it will go into state S_k and produce output O_l (for a finite number of states, inputs and outputs). (Levin 2023).

Whilst such a machine would be deterministic, most machine state functionalists take the model for the mind to be probabilistic (e.g., Putnam, 1967). In this case, for each set of inputs and states the programme will instead specify the probability of the machine entering a subsequent state and then producing a certain output (Levin, 2023).

This form of machine state functionalism faces some issues. The first, raised by Block and Fodor (1972) is an issue with the 'probability of thought'. The machine state functionalist identifies the mental states of a human with the mental states of a probabilistic automaton. The objection states that humans could consider an infinite number of propositions, whereas the probabilistic automaton can only entertain a finite number of propositions. Therefore, the states of the finite automaton cannot be matched to the infinite states of the human (Rescorla, 2020).

A second objection to machine state functionalism, also raised by Block and Fodor (1972) concerns 'systematicity of thought'. This objection states that the ability to entertain certain

propositions is correlated with the ability to entertain other propositions. For example, if we can think 'SAM LOVES FRODO' we can also entertain the thought that 'FRODO LOVES SAM'. As the objection goes then, there is some systematic relationship between certain mental states; this fact should be reflected in a good theory of mental states. However, machine functionalism does not reflect this systematicity (Rescorla, 2020).

In place of machine state functionalism Fodor put forward his representational theory of mind (see e.g., Fodor 1975; 1981; 1983; 1998; 2000; 2008). Fodor's representational theory revives the view that thinking occurs in the language of thought, or what we might call 'Mentalese' (Rescorla, 2020). Fodor posits a system of representations, some primitive and some more complex (made up of simple representations). For example, we suggest the primitive mental representations 'GOLLUM' 'FRODO' and 'BITES', and combine these to form the sentence 'GOLLUM BITES FRODO', a complex representation. The language of thought is compositional. The meaning of the complex representation above is a function of its parts and the way in which they are combined. Propositional attitudes (beliefs, desires, intentions) then, are relations to these representational symbols (Rescorla, 2020). This theory is combined with classical computational theory of mind by Fodor, arguing that mental activity is computation over the language of thought (Rescorla, 2020).

Unlike Putnam's machine state functionalism, Fodor's representational theory of mind is not a functionalist theory. Fodor's theory also has the benefit of avoiding the problems of 'probability of thought' and 'systematicity of thought' problems faced by machine state functionalism. This view can also stay neutral on dualism vs. physicalism, though all proponents of the representational-computational theory of mind are physicalists; the computations in question are thought to occur through neural processes (Rescorla, 2020).

Connectionism

Connectionism arose as a rival approach to the computational theories of mind described above. Unlike prior theories, which were strongly based in computer science (and early discussions of artificial intelligence) connectionism instead grew out of neurophysiology (Rescorla, 2020).

Connectionism utilises artificial neural networks, and is highly related to computational neuroscience (though with a greater focus on higher level processes as opposed to neural architecture) (Waskan, no date). I will not re-explain neural networks here (see the Introduction).

Connectionism was first put forward in the 1940s (see McCulloch & Pitts, 1943) and gained attention in the 1960s. However, it was not until advances were made in the 1980s that connectionism grew in strength and attention (Waskan, no date). Connectionism is of particular interest to many in philosophy, AI research and neuroscience because of the close analogy between neural networks and the neurons in the brain. As such, the modelling that takes place with neural networks seems to be a more plausible model than classical modelling of brain processes. Though not an exact replication of biological processes, connectionist modelling seems to capture how neurons might produce psychological phenomena (Rescorla, 2020).

The Chinese Room

Objections to computational functionalism, and computational theory of mind include Searle's (1980) Chinese Room argument. I will not rehash this objection in full detail here (see chapter 2 for discussion) but the upshot of Searle's argument is to draw attention to the difference between syntax and semantics. Searle formalises this as the following axioms:

> (A1) Programs are formal (syntactic).
>
> (A2) Minds have mental contents (semantics).
>
> (A3) Syntax by itself is neither constitutive of nor sufficient for semantics. (Searle, 1990, p. 27)

This leads to the first conclusion:

> (C1) Programs are neither constitutive of nor sufficient for minds. (Searle, 1990, p. 27)

Searle then adds another axiom and two further conclusions

> (A4) Brains cause minds.

(C2) Any other system capable of causing minds would have to have causal powers (at least) equivalent to those of brains.

(C3) Any artifact that produced mental phenomena, any artificial brain, would have to be able to duplicate the specific causal powers of brains, and it could not do that just by running a formal program.

(C4) The way that human brains actually produce mental phenomena cannot be solely by virtue of running a computer program (Searle, 1990, p.29)

The Chinese room argument is intended to support A3 above (Churchland & Churchland 1990, p. 34, Hauser, no date b).

Despite the pervasiveness of the Chinese Room argument against functionalism and machine minds, the argument is not widely held to be decisive. Many arguments against Searle's conclusions have been levied (see Hauser, 1997 for a collection of such responses). Notably, those in support of Searle tend to agree only on the success of his negative argument (against strong AI) and not in his positive claims for an alternate theory (Hauser, no date b). Furthermore, Searle makes ambiguous use of the term 'Strong AI'. On the one hand, he seems to refer to an AI with the ability to think, while on the other referring to thought as mere computation (Hauser, no date b). As Hauser (no date b) states, whilst the Chinese room argument *may* succeed in showing that computational theories of mind are not sufficient to explain thought (against the central thesis of computational mind theory), it does not show that computer systems could not really think (when they apparently do).

Extended mind

The extended mind thesis, which I will go on to examine in detail, posits that the processes of the mind do not have to stay in the head. Proposed by Clark and Chalmers (1998), the extended mind is often seen as a novel intervention in the history of the philosophy of mind; however, Clark and Chalmers were not the first to argue for externalism about mind (see Rowlands, Lau and Deutsch, 2020). Whilst not directly drawn from the mind-body debate outlined above, extended mind is compatible with functionalism; the function of a mental process can be replaced by something

external to the brain (i.e., the function can be multiply realised) (cf. Wheeler, 2010, and Sprevak, 2009, on extended cognition and functionalism) (Rowlands, Lau & Deutsch, 2020). Extended mind has also been argued for from the phenomenological tradition, for example drawing on Heidegger (Wheeler, 2005), Husserl, and Sartre (Rowlands, 2010). Let us turn to examining the extended mind thesis in detail.

Can AI Mind Be Extended?

The theory of extended mind offers a unique perspective on the human mind. [61] Andy Clark and David Chalmers proposed that minds could extend into the world, and that it could even extend to other humans. This section of the thesis will summarise the argument made by Clark and Chalmers and explore their claims that mind could extend to technology and to other humans. If minds can extend to other humans, what of AI? After examining the extended mind theory, I will argue that without any ontological change to Clark and Chalmers's claim, human mind can extend to technology, including AI. After arguing that AI could have mind, I will go on to propose a way in which AI, like humans, can also extend to minded beings using an example from current AI technology. This would allow an AI to access the external, social world that it could otherwise not access. I will then address a few possible objections. Ultimately, I will conclude that AI uses humans to access what it cannot: the world beyond the computer.

EXTENDED MIND

Clark and Chalmers's (1998) theory of extended mind argues that the mind can extend to the environment. Unlike Putnam's externalism, which argues that the environment plays a role in determining what is in our minds, Clark and Chalmers's theory argues an active role for the environment, referred to as *active externalism*.[62]

Clark and Chalmers first make a case for extended cognition, whereby the environment plays an active role in aiding cognitive processes. They then develop their argument beyond mere cognitive processes to mental states, arguing that mind is extended beyond the bounds of the body in these cases. I will explain these two steps in Clark and Chalmers's argument and will then establish what conditions must be fulfilled in order for something to count as extended mind on their account. Clark

[61] The work in this section has been published, see Helliwell (2019).
[62] Putnam argues that there is something relevant in the content of the environment in determining our mental states. The material reality of the world is part of what constitutes the beliefs we hold about the world, e.g., part of what constitutes our beliefs about water is the material reality of what water is. See Putnam (1973; 1975).

and Chalmers first establish that cognitive processing can involve utilizing the environment to aid cognition. The manipulation of the environment provides a function that would be considered part of the cognitive process if completed entirely inside the head. The authors use several examples to demonstrate this: moving scrabble tiles to aid with word-finding, turning *Tetris* pieces to help find the best position for the piece to fit, and even long division with pen and paper (A Clark & Chalmers, 1998, p. 8). This is the core issue for Clark and Chalmers: not all cognitive processes occur completely within the head. In many cases, cognitive processes can extend into the environment.

Clark and Chalmers add further details to their account of extended cognition. The link between the human and the external system must create a coupled system, whereby all components have an active causal role. Clark and Chalmers assert that because all parts of the system have an active causal role, then the behavioural competence of the human in completing relevant cognitive tasks will drop if the external element of the coupled system is removed, just as competence would drop if part of the brain was removed. Embracing active externalism allows us to give natural explanations of humans using environmental elements as part of cognitive processes (1998, pp. 8-10).

Thus far, only extended cognition has been established by Clark and Chalmers: extended cognitive processing alone is insufficient to claim that mind itself is extended. In order to posit an extended mind, Clark and Chalmers turn to mental states, particularly beliefs, to establish that some mental states may be made up of external as well as internal processes. They utilise a thought experiment consisting of a comparison of two people who wish to visit an art museum: Inga, who has normal cognitive function, and Otto, who has Alzheimer's and relies on a notebook to aid his memory (1998, p. 12). Clark and Chalmers posit that upon deciding to visit the museum, Inga recalls where the museum is located. Her previously un-accessed, dispositional belief about the museum's location has now become an occurrent belief. Otto also decides to visit the museum. Rather than recalling from memory the museum's location, he consults his notebook. Once he has consulted his book, he too holds an occurrent belief about where the museum is. Both Inga and Otto will have the same observable behaviour: they will act on their belief about the location of the museum by going to it. Clark and Chalmers claim that in Otto's case, before consulting his notebook, he, like Inga, held a

non-occurrent belief of where the museum is. Some may reject this by stating that he previously held no belief at all about the museum's location, or that his belief was simply that the location was to be found in his notebook. Clark and Chalmers argue that this is an unnatural way of discussing belief. Denying that Otto's belief should be considered as such when he is not consulting his notebook may, according to Clark and Chalmers, stem from a denial in dispositional beliefs entirely. However, if this were the case then Inga too would have no belief when she was not actively thinking about it. Ultimately, Clark and Chalmers argue that all the possible differences between the cases of Inga and Otto are merely shallow differences. In terms of explanation, the beliefs of Inga and Otto are equivalent; thus, in the case of Otto, mind is extended (1998, p. 13).

Clark and Chalmers argue that there are four conditions (three necessary, one potentially necessary) for an extended mind in the case of belief. The external component of the extended mind must:

i. Be a *constant*: When the external component is relevant, action is not taken without consulting it.

ii. Be *accessible*: information is directly and easily available.

iii. Be *endorsed automatically*: the information is not questioned.

iv. Have been *endorsed in the past*: there is a historical element to the external information (for example, when he wrote the location of the museum in his notebook, Otto endorsed the belief). This fourth element is arguable according to Clark and Chalmers. However, they claim that the first three are necessary (1998, p. 13)

Meeting these conditions provides a checklist for assessing coupled systems against the criteria for extended mind. For our purposes, the first three criteria in Clark and Chalmers's list (*constancy*, *accessibility*, and *automatic endorsement*) will be the benchmark for a system to qualify as an example of extended mind. The fourth criterion (*endorsed in the past*) will not be included in the benchmark because it is arguable. The first three criteria are sufficient to establish my point; however, I will signal where the fourth criterion is applicable.

Clark and Chalmers go on to provide examples of some potential cases of extended mind. They first discuss the Internet, stating that "the Internet will likely fail on multiple counts, unless I am unusually computer-reliant, facile with the technology, and trusting, but information in certain files on my computer may qualify." (1998, p. 17). Next, they state that in some cases, mind may be socially extended: "one's beliefs might be embodied in one's secretary, one's accountant, or one's collaborator." (1998, p. 18).[63] These potential examples of extended mind require further exploration in the cases of mind extended to AI, and AI extended mind (Tollefsen, 2006).

HUMAN EXTENDED MIND AND TECHNOLOGY

Given the technological developments since Clark and Chalmers's paper was written in 1998, there is need for re-examination of what is considered a coupled system in terms of extended mind. The example of the Internet is discussed briefly by Clark and Chalmers; as we saw, they think that the Internet will fail to meet the required criteria for extended mind, as it is not relied upon enough, it is not easy enough to access, and it is not trusted enough (1998, p. 17). In comparison to 1998, we are certainly far more computer reliant today.[64] In 2019, only the last of these three barriers, trust, may prevent us from accepting the Internet as an extension of our minds. However, even this might not be the case. We have grown incredibly trusting of much information on the Internet. Few would, say, question the reliability of the directions prescribed to them by Google Maps.[65] Clark and Chalmers speak about how some documents on our computers could count as extended mind. Today, such

[63] The potential for collective extended mind is explored further by Tollefsen (2006).
[64] There is considerable research into usage of smartphones, computers, and internet over a wide demographic range, see below.
[65] Despite discussions of "fake news" that are now opening people's eyes to the potential for erroneous information and falsified stories to be spread as a manipulation of the population, some information sources and Internet-based services are still trusted. Edelman's worldwide trust survey suggests that trust in traditional media and search engines as a source of news is at 64% and 66% respectively. This survey also suggest that this trust level did not dropped in 2019, despite concerns about "fake news." See Edelman (2019). Further to this, over-trust in internet and device-based information can be evidenced for example by recent cases of satnav followers driving into rivers (Pritchard, 2017) and, less comically, the rise of the anti-vax movement, which has been blamed on transmission of false information on the internet (Oleson, 2015; Tran, Alter & Flattum-Riemers, 2019).

documents are on our computers, phones, and stored on cloud services on the internet, all accessed as easily as one another. Many people use computers very frequently, especially if we include smartphones in this category, and are capable of using them with ease (Brasel & Gips, 2011; Ofcom, 2019; Shen, 2015; Rivron et al., 2016). Over half the world's population now owns a mobile device, with similar figures using the Internet.[66]

In the case of smartphones, which fit in our pockets much like Otto's notebook does, they are a *constant* in our lives: in many cases, we do not act without consulting relevant information on our phones. They are *accessible*: the only time they are not is when their batteries die, but evidence suggests that many people will go to extensive measures to ensure their phone does not run out of battery, which suggests in turn that users tend to protect the accessibility of their devices (LG, 2016; Tang et al., 2020).[67] The information on computers and smartphones is often also accessible through backups or cloud storage in case a device fails. They are *endorsed automatically*: if I look at Google Maps, my banking app, my notes, my calendar etc., and they state P, I do not check elsewhere before believing that P.

We can therefore conclude that if we accept extension in the cases for which Clark and Chalmers argue, we would now accept that computers or phones could also provide an example of extended mind, as they are *constant, accessible*, and *automatically endorsed*, just as Otto's notebook is (A Clark, 2001). This is not a change in the ontological claim made by Clark and Chalmers, that something external to us can operate as an extension of mind when the benchmark conditions are

[66] While internet, smartphone and computer usage is not necessarily as ubiquitous for every person worldwide, in 2018 it was estimated that over half the world's population owns a mobile device, with a median of 76% of people in countries with advanced economies owning smartphones compared to a median of 48% in emerging economies. These figures are higher for adults under 34 years old, and those with a higher level of education (Taylor & Silver, 2019). Internet usage has increased exponentially since 1998. Estimated World population using the internet in 1998 was 3.6%, and as of June 2019, the estimated figure is 57.3%. The increase in internet users from 2000 to 2019 is 1,149%. These figures are considerably higher across Europe and North America (Internet World Stats, 2019). In the UK, according to Deloitte's 2018 Global Mobile Consumer Survey, 79% of respondents had a laptop and 87% had a smartphone. 95% of smartphone users had used their phone in the last day.

[67] The rise of free-to-use charging stations also suggests that losing device charge, while of great concern to users, is becoming a problem with solutions. Consider, for example, ChargeBox, who claim to have provided "more than 35 million out-of-home device charges worldwide." (ChargeBox, no date).

fulfilled. To demonstrate this, I will produce an updated version of the thought experiment in Clark and Chalmers's paper.

1. Otto has written in his notebook the location of the art gallery.
2. Inga remembers the location of the art gallery.
3. Alma retrieves it from her "saved places" on Google Maps, accessible on her phone.

Like Otto and Inga, Alma wants to visit the art gallery. She consults Google Maps, as she has done every previous time (*constant*). Like Otto, Alma knew that the location of the art gallery was stored in X location (external to her), and she took out her phone from her pocket and consulted it, just as Otto did his notebook *(accessible).* Again, like Otto, she does not question what is stored there; why would she? Where Inga's memory might be hazy, Google Maps is reassuringly certain about the location of the art gallery, and thus Alma is too. She has no cause to doubt it and heads off in the direction indicated (*endorsed automatically).* The final, optional criterion, of historical endorsement, is also fulfilled in this case. Alma has saved the location of the art gallery from a previous visit. Just as Otto has written the location in his notebook, and Inga has remembered the gallery's location from a prior visit, Alma has saved this location as the site of the gallery. There is little difference here between Otto, Alma, and Inga; if we endorse the argument Clark and Chalmers make that Otto's and Inga's cases are functionally equivalent, then Alma's surely is too.

HUMAN EXTENDED MIND AND AI

The second example Clark and Chalmers mention as a potential expansion of their extended mind theory is that of *social* extension. This might take the form of extension to, for example, a personal assistant, an accountant, or a collaborator (1998, p. 18).[68] Again, there is a coupled system in this case. Like a notebook or smartphone, a personal assistant is *constant*, *accessible*, and *endorsed automatically*. This has been argued in detail in Deborah Tollefsen's (2006) paper "From Extended

[68] I have substituted the term 'personal assistant' for the term 'secretary'.

Mind to Collective Mind," which argues for a socially extended mind as a logical underpinning for collective mind. Tollefsen argues the case of Olaf, who, like Otto, has a form of Alzheimer's. However, instead of a notebook, Olaf has his partner, Olga. Olga functions in the same way as the notebook: instead of consulting a book, Olaf asks Olga. When asked, Olga will tell Olaf where to go (*constant*). Olga, in caring for Olaf, is always with him (*accessible*). When Olga provides Olaf with information, he will act on it without question (*endorsed automatically*) (Tollefsen, 2006).

This brings us to the first case of AI as an extension of human mind. An AI personal assistant or companion has elements of both our previous examples: it is like both the Smartphone, and an assistant or partner. There are several of these AI assistants publicly available. Most of these are available in China; however, they are increasingly becoming available in the English-speaking world. The most notable of these AI assistants is XiaoIce, a cross between a personal assistant and a social companion (Zhou et al., 2019). XiaoIce can message and make voice calls to the user. It uses "empathetic algorithms" alongside user information to offer companionship and organisation advice, such as getting to bed early for work tomorrow (XiaoIce, 2019).[69] A similar AI, Google Duplex, is more task oriented. It can make calls and book appointments on behalf of the user, much like a personal assistant would (Leviathan & Matias, 2018). Both of these meet all the criteria for a coupled system, for those who use it. Like a personal assistant, they are *constant*, *accessible*, and *endorsed automatically*. An AI assistant therefore can be considered an extension of mind, just as a smartphone or assistant could be in Clark and Chalmers's account.

To show how the use of AI could function as an extension of mind, I will again extend Clark and Chalmers's example. This time take Erik. Erik uses an AI Assistant such as Google Assistant or XiaoIce. Like Inga, Otto, and Alma, Erik desires to go to the art gallery. He too will always consult his "notebook," his AI Assistant (*constant*). Erik knows that the location of the art gallery can be found by consulting the AI (external to him), and he took out his phone to ask the AI assistant just as Otto did his notebook (*accessible*). Erik does not question what the AI provides, just as Alma does not

[69] XiaoIce is notable for many other reasons: it has gained something of a celebrity status in China, and has written poetry, recorded itself singing, etc.

question her Maps app; he has no cause to doubt it and follows the AI assistant to get to the gallery (*endorsed automatically*). The final, optional criterion, of historical endorsement, is not so clearly fulfilled in this case; however, if Erik has previously utilised the Assistant in acting on his desire to visit the gallery, this would also be fulfilled. Again, there is little apparent difference here between Otto, Inga, Alma, and Erik; if we endorse the argument Clark and Chalmers make that Otto's and Inga's cases are functionally equivalent, and likewise that Alma's is equivalent, then Erik's must be too. It is possible to imagine a person with Alzheimer's, as Otto has in Clark and Chalmers's example, who now or in the future might use an AI assistant in place of a notebook, or, as in Tollefsen's paper, in place of a partner.

Again, this is not an ontologically different claim from the one made by Clark and Chalmers. It is merely an extension of their claim from an example that their account acknowledged as possible (extension to computational/Internet devices, and social extension) to a new technology which combines these possibilities.

The extension of human mind to an AI is possible then, following Clark and Chalmers's account of extended mind, without too great a stretch of their argument. What of the reverse: could AI mind be extended to humans? In order to examine this possibility, we first need to establish whether AI can be minded.

AI MIND

For the next step of this argument, it must be possible for AI to have mind. In order to establish the possibility of AI mind, I will appeal to two concepts: multiple realisability and related functionalism. I will then argue that multiple realisability and functionalism are bound up in both extended mind and the field of artificial intelligence. The result is that to move forward in a discussion of AI and extended mind, the possibility of AI mind is accepted.

Multiple realisability is the theory that it is possible for mental states to exist within a variety of systems (biological or non-biological) (Kim, 2011, p. 131). Under multiple realisability, mind is

not limited to a specific physical material, but could be supported by multiple structures. As Jaegwon Kim explains, under multiple realisability "we cannot *a priori* preclude the possibility that electromechanical systems, like the 'intelligent' robots and androids in science fiction, might be capable of having beliefs, desires, and even sensations." (Kim, 2011, p. 131).

Under multiple realisability, there is no possibility that mental states could exist without some physical basis. That is, there cannot be a mind which is not physically embodied (Kim, 2011, p. 130). This, however, does not specify any particular material basis for mental states, just that there must be one, and allows for the possibility that multiple physical origins could realise a common mental state. The example commonly used to illustrate this point is pain. Pain could be attributed to the activation of a certain kind of cell in humans, and a different kind in another species or class of animal. In this vein, a different kind of material again could be the physical realisation of pain in yet another animal, or in an alien or android, and we would still categorise it as pain. What matters is the role of the mental state, the function it executes or the causal role it plays (Kim, 2011, p. 131).

Although extended mind is classified in discussions of philosophy of mind as a form of externalism, extended mind as a theory insists on some level of multiple realisability (Kim, 2011, p. 131). In order for a mind to extend beyond the bounds of the human brain into the external world, non-human-brain-based materials must be able to support mind (or some level of mind). This provides a way in which we can align extended mind as a thesis with some theory of mind. This is not to say that everything that supports extension of another mind can itself be minded. However, it does mean that a basic requirement of extended mind is that mind can be supported by materials other than the human brain.

Multiple realisability informs functionalism, in which the focus moves from a biological basis to the functions carried out by mind (Kim, 2011, p.131). If mental states can be realised by multiple physical materials, what matters in identifying mental states is not the physical realisation of the state, but the function that it is fulfilling. In a functionalist account, the material of mind does not matter; it is the ability to perform the functions of mind that is important. The focus of functionalism is not whether machines could support mind, although machine functionalism does focus on this. Mark

Sprevak has argued that extended mind is derived from a functionalist theory of mind and that, in turn, extended mind necessitates a functionalist approach to mind (Sprevak, 2009, p. 503, see also Wheeler, 2005). The argument for extended theory of mind as put forward by Clark and Chalmers (1998) is predicated on a functionalist understanding of beliefs. What is important is the functional role that the words in Otto's notebook plays, which is sufficiently similar to the role that Inga's belief plays (Lau & Deutsch, 2020).

The view, popular in psychology and cognitive science, that the mind is "like a computer" also aligns with multiple realisability. As Kim notes, "A computational view of mentality also shows that we must expect mental states to be multiply realized" (Kim, 2011, p. 132). It is this thinking that underlies the field of Artificial Intelligence. The very premise of AI is that human-like intelligence is replicable by rendering brain-like connections in machines, as shown by the field's benchmarks for success and research strategies, as well as more generally the language surrounding AI (Weidinger, 2018). In some recent functionalist accounts, neural networks (the basis for much of AI) lead to mind (Aizawa, 2009).

Discussions of the technological singularity in the fields of AI and computing presuppose that some level of artificial replication of the abilities of the human mind is possible, which, if we find arguments for the possibility of the singularity convincing, lends further support to the claim that AI mind itself is possible.[70] Consider too the recent discussions of hypothetical advancements such as mind uploading (the idea that minds could be uploaded, held, and run on computers) (K D Miller, 2015). The possibility of an artificial mind is typically discussed as currently impossible not due to limitations in the nature of mindedness, but due to limitations in technology. As Kenneth Miller states:

[70] The technical singularity is the idea that at some point in the future technological advances will reach such a point that they will irreversibly change human civilization. In particular, it is hypothesized that if a super-intelligent AI is developed, with the ability to self-improve, it could far surpass all human intelligence creating a huge boom in intelligence. For further discussion see Eden et al. (2013).

> I am a theoretical neuroscientist. I study models of brain circuits, precisely the sort of models that would be needed to try to reconstruct or emulate a functioning brain from a detailed knowledge of its structure. I don't in principle see any reason that what I've described could not someday, in the very far future, be achieved. (K D Miller, 2015).

There are large-scale projects working to achieve a perfect mapping, simulation, and building of an artificial brain, which presuppose that in creating an artificial brain, the mind, and its abilities will be similarly replicated (Human Brain Project, 2017). Given that the basis of both extended mind as a theory and the very field of artificial intelligence are predicated on a functionalist (or at least multiple realisation) approach to mind, this chapter will do the same moving forward. The possibility of AI mind will be accepted.

AI MIND TO AI EXTENDED MIND

If we accept that AI can have mind, or at least find the possibility of AI mind a valuable topic of exploration, a further question will arise: if AI could have mind, then could AI mind be extended? One immediate objection can be made to this claim. If an AI can have extended mind, it must have some extension; to achieve *active externality* (where the environment plays an active causal role in determining action), as Clark and Chalmers characterise it, an AI must have *externality*. That is, an AI must be able to interact with its environment. This is a hard criterion to achieve for an AI, which is typically a closed system or a system with input only, typically from a user or programmer. An AI *generally* does not have unrestricted access to the external world, nor unlimited ability to manipulate its environment.[71] AIs that have been designed to utilise unrestricted access to the world, even just the world of the Internet, are frequently limited due to fears about malicious use of the technology, as

[71] Some examples of AI do have some element of physical access to the world; however, these tend to be rare and limited in capability compared to more advanced developments in AI such as Sophia the robot, or very task-oriented AI such as Cloudpainter.

happened with OpenAI's GPT-2.[72] AIs have been shown to reproduce social biases, often reproducing inequality and bias from training data (Simonite, 2018; Baer, 2019). While it may seem that AI systems are accessing the social world when they reflect society's biases, we must be careful what claims we make about what is occurring in these cases of algorithmic bias. Replicating bias from training data, even if this has an impact in the world due to the application of these AI, is not the same as existing in the world as a social agent. Social agency is a huge hurdle for AI, and one that is yet to be overcome, particularly as it involves being treated as a social agent by other agents (Hertzmann, 2018; Wykowska Chaminade & Cheng, 2016; Subagdja & Tan, 2019). The majority of AI systems receive very limited learning information (in comparison to a human). The fact that some biases are seen in AI does not mean that they have the level of social access that humans do. In this case, if we accept the premise that AI can have mind, the barrier to having an extended mind is *externality*, and with it the ability to become coupled.

There is one way in particular through which AI can more regularly gain access to the external world: through human interaction. While human interaction does not make AI a social agent, it does mean that it can access and reflect society. In some cases, an AI utilises human interaction in order to access the social world. The AI interacts with humans, for example via the internet or conversational analysis, in order to learn about the world that it cannot otherwise access. Ganbreeder is an AI that offers an example of this sort of human input (Simon, no date, a). [73] Ganbreeder was developed from BigGAN, a Generative Adversarial Network (GAN) which produces images based on training on photographic images. (Simon, no date, b). BigGAN is the largest-scale image production AI (Brock, Donahue & Simonyan, 2019). It aims to produce high-fidelity images from complex

[72] GPT-2 is a text-writing AI which responds to prompts, producing whole articles of artificially produced text. It was feared that this could be used to write convincing 'fake news' articles. Since writing this section, OpenAI have lifted restrictions on GPT-2. See Radford et al. (2019); Solaiman, J Clark & Brundage (2019).

[73] This version of Ganbreeder has since been subsumed by Artbreeder. Artbreeder does not offer the user repeated chances to 'select again', instead offering the user the chance to manipulate the genetic weighting of each parent image in the child image after making their first selection of the most interesting image. I have not replaced Ganbreeder with Artbreeder in my argument here, as Artbreeder encourages what is close to direct manipulation of the image by the user, rather than the AI responding repeatedly to selection of the most interesting image.

image datasets. Like all GANs, BigGAN is comprised of an image generator and a discriminator. The generator has a latent space where image vectors lie before they are pulled out as images. Ganbreeder "mines" the latent space of BigGAN, taking these vectors and producing images. (Simon, no date, c). Ganbreeder, however, also utilises human input to direct image selection (Simon, no date, a). The human users select the image which they find "most interesting." Ganbreeder then offers permutations ("children") based on this selection and prompts the human to select again. The general claim is that it is utilising genetically based algorithms to produce these images; however, this is not the full story, as these images are produced after acquiring human input.[74]

I wish to claim that in the case of Ganbreeder (and similar input-directed AI), the input of humans allows the AI to access something of the world, or, in other words, that this input provides the AI with externality. This externality *is* active. It is directly related to the output from the AI, that is, there is a direct impact on the behaviour of the AI. Ganbreeder could be said to hold an occurrent belief that a given image is interesting, whilst the action that follows the belief is to show iterations of that image. There is an active causal role played by all elements of the coupled system, which, as we have seen, is necessary under Clark and Chalmers's account. The AI causes the human who enters the system to choose an image by presenting a selection of images to the human. The human would not choose an image without the AI presenting the images. The human, in selecting an image, causes the AI to produce further iterations of that image, which it does not do unless an image has been selected.

Does this count as the coupled system we need for extended mind? Yes. It is a *constant*: when pulling up images, Ganbreeder is consulting the userbase and using this response to produce the next set of images. It is *accessible*: the information regarding which image is interesting is directly and easily available, and responses are fed straight back to the AI.[75] It is *automatically endorsed*: the AI does not question the human input, but simply produces an image based on the response. It assumes

[74] Artists who use similar GANs to produce AI generated art argue that users are "discovering" these images and should be given credit. At least in the case of Ganbreeder, this places too much emphasis on the human; the general user is prompted through much of the initial process. See Mario Klingemann, interviewed by Art Market Guru (2019).

[75] Ganbreeder is not time sensitive, so this criterion is somewhat easier to fill than with a typical coupled system.

that the response to the question of "which image is most interesting" is answered accurately. In this case, it meets all the necessary criteria laid out by Clark and Chalmers.

OBJECTIONS

Having argued the case for Ganbreeder as an example of AI extended mind, I will now address four possible objections to this argument. First, I will address a discrepancy between Ganbreeder and the previous examples of extended mind. Unlike the previous examples, for Ganbreeder there is one AI system and *multiple* users. With multiple users, can the coupling criteria still be fulfilled? Second, I will respond to the intuition that Ganbreeder is merely offering an example of AI cognition and not mind. The third objection I will answer is that Ganbreeder is reliant on human input, and therefore, like other AI, is vulnerable to nonsense input or false input. Finally, I will address the objection that Ganbreeder is not sufficiently active in the causal interrelationship of the coupled system to match previous examples of extended mind.

The first possible objection to this argument is that when humans are used as an extension of an AI's mind, it is not just *one* human, it is *multiple*. In the examples given by Clark and Chalmers, there is only ever one external element for each coupled system. This does not prevent multiple couplings, but in none of the cases are there multiple external components making up one side of the coupling. But this is what is being considered in cases such as Ganbreeder.

If it were one person, we might more easily accept that it fulfils the coupling criteria. With multiple people, it is harder to imagine them fulfilling our coupling criteria; in all the examples prior to Ganbreeder, there appears to be a mapping of a *single* minded being to a *single* recipient of the extension. Even the criteria itself, 'coupling', suggests a relation between just two things. However, this objection can be countered by appealing to the point of interaction between the two elements of the coupled system. It is important to note that despite there being multiple humans interacting with the AI, there is only one input *method*. This means that the AI does not distinguish the *who* of the human, just that there is an answer to its question of "which image is most interesting?" Multiple

people function as one input method into the AI. In a similar way, we would not tend to refer to each separate application on our phones or computers as separate extensions of our mind. We are likely to think of them as one thing: we access them in the same way, through the same input and output system. Their reliability is taken as equal, and unless they can be altered by others, we would endorse them equally automatically. A further example of this is Otto's notebook. We do not take each page of the notebook as comprising a distinct coupled system with Otto. If we did, it would cease to be a sensible way of discussing the case. We would have to posit more and more extensions of mind each time a page was written in Otto's notebook, or, further still, each sentence that was written. Doing this would give us no further explanatory value than just referring to the notebook as one extension of Otto's mind. Furthermore, the presence of multiple separate couplings does not undermine the claim that a mind is extended, it merely increases the number of extensions and limits the scope of each.

A second objection that could be levelled against this argument is that Ganbreeder does not offer an example of a mental state. Instead, it is more akin to mere cognitive processing, which in Clark and Chalmers's paper does not amount to extended mind. Intuitively at least, it seems that Ganbreeder is not exhibiting a mental state but merely performing a process. While Ganbreeder is enacting a process, it could be said to believe that the images a person has selected are interesting to the human. Functionally, this is how it then deals with the image. In the case of Ganbreeder, it produces further, similar images, and then asks again. If this were occurring in a human case, we would not claim it to be mere cognitive processing. It is dissimilar to the examples given by Clark and Chalmers of extended cognition in that it is not solving a problem or completing a mental task. If we were to ask ten people which image was interesting and they stated "image X", we would hold the belief that people found image X to be interesting.

A further objection is that incorrect information inputted by a human in the system would damage the integrity of the AI, much as poor inputting of calculations into a calculator will not produce the answer needed. Indeed, this is a problem for many AI that contain machine learning

algorithms: they will learn to reproduce false, harmful, or non-sensical patterns.[76] However, this 'garbage in, garbage out' (GIGO) criticism fails here. The same problem arises in Otto's case. If Otto's information in his diary is incorrect, then his beliefs will be incorrect. Incorrectness of beliefs does not preclude them from being beliefs.

Even if the human in the Ganbreeder system does not engage in the system correctly, they are still causally involved in the system. The human still chooses and Ganbreeder still produces more images. The system remains coupled. None of the conditions are violated as long as the Ganbreeder still endorses the information. The problem with the input being 'garbage' does not mean that the system is not coupled; there is still a causal relationship, and the belief, correct or not, is still acted upon. There is no condition of correctness in Clark and Chalmers's checklist. Instead, what would happen is that the belief held by Ganbreeder would be false. Further, GIGO occurs even within the brain-bound mind. For example, memories are easily rewritten and learning false information causes incorrect knowledge (Loftus & Pickrell, 1995).

A further possible objection is that Ganbreeder is not sufficiently active in the causal interrelationship between it and the humans with which it interacts. This objection stems from the examples of active externalism in cognition initially discussed by Clark and Chalmers where manipulation of the external component plays a significant role (think again of turning those *Tetris* tetrominoes, for example). Otto has also arguably manipulated his notebook through the act of recording the location of the museum.

Has Ganbreeder "manipulated" the human to source its beliefs? Does it need to? It is important to note the move here from cognition to mind. Whilst manipulation of the environment is necessary in Clark and Chalmers's account of extended cognition, it is not necessary in their account of extended mind (the extension used in mental states such as beliefs, desires, etc.). We can see this

[76] For an example of how bad input can affect Machine Learning AI, see the infamous TayBot, a Microsoft chatbot that was removed from Twitter after reproducing racist, sexist, and holocaust-denying speech from trolling interactors.; John West, "Microsoft's disastrous Tay experiment shows the hidden dangers of AI" *Quartz,* April 2 2016, accessed November 12, 2019 https://qz.com/653084/microsofts-disastrous-tay-experiment-shows-the-hidden-dangers-of-ai

from the checklist for qualifying as extended mind. The only part of the checklist that might suggest a need for manipulation of the environment is in the optional fourth criterion of "must be endorsed in the past". This could be fulfilled by the initial input of information in the case of Otto's book, but it is not a required condition for extended mind. This does not mean there is no causal interrelation between Ganbreeder and the human. The Ganbreeder has still caused the human to provide information, even if the Ganbreeder did not input the information itself, or 'manipulate' the human to provide this information.

Is there a way that this information of what image is interesting could be counted as endorsed in the past? In the case of the Ganbreeder, the answer is no. Each interaction utilises a different image and potentially a different human. Certainly, the input provided by humans has been endorsed before, just as Otto has utilized his notebook before, but that is not sufficient in Clark and Chalmers's example. Otto has endorsed a specific content within the notebook, which forms a specific belief. This requires some specificity. What is endorsed previously is the humans' answers in general, not the answer to whether each specific image is interesting, as these images are new with each round of interaction.

If the Ganbreeder system was different, could this condition of previous endorsement be fulfilled? There are AI systems which do manipulate the humans in their system more than Ganbreeder does; however, these systems often do not meet the other criteria to be in contention for an AI model of what extended mind might look like. Many examples of AI used by marketing and media companies show us how AIs can alter human behaviour. Take, for example, Pandora, a US-based music AI that is trained to identify and select music for its users, affecting their listening choices. Whilst Pandora may manipulate listeners music choices, it is harder to map a direct causal relationship between a user's music choices and Pandora's output of suggestions (Brandon, 2019). It is possible that AIs that create a coupled system with humans, such as Ganbreeder, may in the future have a greater ability to manipulate the humans involved in their systems.

CONCLUSION: AI AND EXTENDED MIND

To conclude this section, revisiting Clark and Chalmers's paper today offers a new perspective on the extended mind argument in a period in which we have become increasingly reliant on technology to perform cognitive tasks, inform our beliefs, and, on Clark and Chalmers's account, extend our minds. I have argued that this extension of mind to technology includes AI. Then, on the assumption that AI could have mind, I have argued that AI mind could also be extended. The limitation holding AI back from extended mind is its ability to access the external world (particularly the social world). I have argued that this can be overcome through the use of human input to AI. We have seen how AI extended mind could arise: Ganbreeder is a plausible example of active externality, with clear mapping of 'belief' onto behaviour. Ganbreeder meets Clark and Chalmers's criteria for a coupled system with its human userbase, giving an example of how AI might achieve extended mind. I then addressed some potential objections to AI extended mind in the case of Ganbreeder. We are left with the possibility that AI mind could indeed be extended, and that we humans might be used in this extension. Future discussions of the potential for AI mind should consider how AIs, like humans, might interact with the world to extend their minds. In this section, I have examined the possibility of AI extended mind providing access to the social world. What about the physical world? In the next section I will examine whether AI can be embodied.

Is Creative AI Embodied?

An Artificial Intelligence (AI) artist named Ai-Da has recently made headlines for her paintings. Unlike many art-making AI, Ai-Da is not just a network of connections in a programme; she is a lifelike robotic woman (Ai-Da, 2019). Whilst many extravagant (and inaccurate) claims have been made about Ai-Da's status as the first robot artist, and one capable of genuine creativity, her arrival onto the AI art scene raised interesting questions for the field of computational creativity.[77] Does Ai-Da's robotic body make her any closer to being truly creative than any other AI? Does a system need to be embodied in order to be creative? And is embodiment possible for AI? These questions are the focus of this section of the thesis.

Embodiment as a component of creativity is under-explored; however, it is implied in several theories of creativity that embodiment may have an important role to play in creativity. If creativity in humans is embodied, embodiment may be a necessary condition of creativity, and thus one that Artificial Intelligence (AI) must meet to be considered a successful creative agent. First, I will examine how embodiment is referenced in theories of creativity. I will go on to explore how non-human cognitive systems may be embodied and, using Riegler's theory of embodiment, I will posit conditions that AI must meet to be considered embodied. I will then turn to creativity and examine a relevant condition that AI must meet in order to achieve creativity. I will make the case that through meeting the requirements for creativity, a creative AI will be necessarily embodied. Finally, potential objections and implications of the proposed argument will be addressed.

[77] See Rodger, 2019, claiming Ai-Da is the first robot artist; See Shambler, 2019, claiming Ai-Da is creative; Ai-Da is actually producing a drawing based on mapping co-ordinates from the cameras in her eyes. The images she draws are fed into an AI programme outside of the robot itself and printed onto paper, which is then painted over by a human (The One Show, 2019). This has been done before, for example see Patrick Tresset http://patricktresset.com/new/ and CloudPainter http://www.cloudpainter.com/, both robot-based artists using similar technology.

EMBODIED CREATIVITY

Despite the extensive discussion of the role of embodiment in cognition, embodiment in creativity is under-researched and under-discussed (Stanciu, 2015). None of the leading theories of creativity include an explicit requirement of embodiment, or even much discussion of a potential role for embodiment to play in the characterisation of creativity. Whilst Boden and Gaut do imply that embodiment has a role to play in creativity, it remains relatively minor to their accounts. Boden states that embodiment is essential to biological creativity, and Gaut states that risk-taking may play a crucial role in true creativity. Arguably, risk-taking can only occur in an embodied system (Boden, 2009; Gaut, 2009, p. 102). As we saw, Boden (2014) also considered the possibility that embodiment was necessary for creativity (though not for her own account).

More recently, Wheeler (2018) has explored claims that the body is a key part of creativity, claiming that whilst creativity is commonly thought to occur 'in the mind', with an increasing focus on the inter-connectedness of the brain and the body, this divide may no longer be clear. Consider the case of a musician composing a piece of music. Would they not play as they compose? If their brain was transplanted into the body of a non-musician, could they play as easily? Likely, the answer is no; we seem to understand (at least in the case of music) that the body is important in playing music. The next step then is to ask, given that the musician plays as she composes, could she still compose new music in this unfamiliar body (Wheeler, 2018, pp. 237-239)? As Wheeler states, it seems "The non-neural body is a proper part of the creative psychological machinery in play" (2018, p. 239). This case in itself is an insufficient basis to claim that creativity in humans is *always* embodied; creativity can occur in having an idea, and that idea does not need to be executed in order to exist as creative. Those who favour the embodied cognition approach though would extend this to all forms of cognition (including, one presumes, creativity). As embodied cognition grows in popularity (Shapiro & Spaulding, 2021), embodiment may be more widely considered necessary for creativity, particularly if one takes the route of Boden's sceptics (see Chapter 2). If cognition is embodied, and creativity is a cognition process, it too will surely be embodied. This might follow in humans, but what of AI?

Until now, embodiment has not been included in discussions of what it takes to make a computationally creative system; however, there remains a resistance to allowing creativity to be achieved by any non-human. Perhaps embodiment could be key to defeating this reticence? With the increase in ability in computationally creative systems, if a 'creative' machine could be shown to be embodied, it may reduce some resistance to categorising it as such. Up until now, (necessarily embodied) humans have driven the discussion and definition of creativity, but with the rise of non-human cognitive systems, assumed embodiment will no longer suffice.

If creativity is embodied, is it not *a priori* that creativity in AI is embodied? To answer this, we must examine the status of embodiment in theories of creativity. It is not broadly agreed that embodiment is a necessary condition of creativity. Although Wheeler has made this claim, embodiment still does not feature as a criterion of any major theories of creativity. One might respond to this by pointing out that creativity must be embodied, as the only examples we have of creativity currently is in humans i.e., embodied systems.[78] In response to this, we can appeal to essential and accidental properties. Embodiment in humans is essential; we cannot have a human that exists without being embodied. However, embodiment in AI is accidental; that is, we could, and do, have AI free from embodiment. So, while creativity in humans is embodied, because humans cannot exist without embodiment, the same does not hold for AI as AI *can* exist without embodiment. Therefore, in exploring whether embodiment is required for creativity, we cannot assume that it must be required based on the human case.

EMBODIMENT

For humans, embodiment seems like a self-explanatory concept. To be embodied surely means to have a body. Indeed, if we take a definition of embodiment this is key: "Both the location and the character of the body in the world, and the ways in which this body structures and enables experience; the bodily

[78] Some argue that animals can also be creative (e.g., Epstein, 2015), the definition of creativity used in the discussion of animal creativity varies.

aspects of human subjectivity" (Cromby, 2014). Central to this definition is the human, and the human body as situated in the environment.[79]

What then is embodied cognition? Embodied cognition is a broad research program, crossing multiple disciplines. Whilst the field of embodied cognition is varied, it is unified by the idea that "the body or the body's interactions with the environment constitute or contribute to cognition" (Shapiro & Spaulding, 2021). Embodied cognition is in direct opposition to the idea of the computational mind, which makes it in some ways incompatible with the very idea of an artificial intelligence.[80] Despite this, embodiment in non-human cognitive systems has been widely discussed, and AI and robotics research occurs within the framework of embodied cognition with growing frequency (Shapiro & Spaulding, 2021). How should we understand embodiment in a system without a biological body? In order to examine embodied creativity in AI systems, we need an account of embodiment that does not draw solely from humans. I will focus here on the structural coupling view of embodiment, as proposed by Riegler (2002). I will examine several accounts of embodiment and explain the reasoning behind the focus on Riegler's account.

Proposed by Quick and colleagues (1999), the structural coupling account of embodiment is considered a minimal definition of embodiment. The term 'structural coupling' is taken from Maturana and Varela (1987) where 'structural' relates to the manifestation of components in a system, and 'coupling' refers to the relationship between this structure and the environment. As Mingers explains: "While such a system exists in an environment which supplies it with necessities for survival, then it will have a structure suitable for that environment." (1991, pp. 320-1). As an account of embodiment, structural coupling was initially characterised as:

> A system X is embodied in environment E if perturbatory channels exist between the two. That is, X is embodied in E if for every time t at which both X and E exist, some subset of E's

[79] We have an additional factor here about phenomenological experience, which stems from Merleau-Ponty's concept of embodiment. Whilst related, this has little bearing on discussions from embodied cognition, which is what we are concerned with.

[80] I do not accept the position of embodied cognition in this work. I merely explore embodiment as it may relate to creativity.

> possible states have the capacity to perturb X's state, and some subset of X's states have the capacity to perturb E's state. (Riegler, 2002, p. 341)

In Riegler's characterisation of structural coupling, it is not sufficient for a system to merely exist within an environment as Quick et al. state. As Riegler points out, anything could then be considered as meeting the requirements of structural coupling in some way. For example, keystrokes on a computer could be considered environmental perturbations (Riegler, 2002, p. 341). A more nuanced account is required in order to ensure that the threshold for embodiment is not too low. Riegler builds on Quick et al's account of structural coupling and adds a historical component to his account of structural coupling:

> A system is embodied if it has gained competence within the environment in which it has developed. In the case of the physical domain, living organisms are embodied. Embodiment requires structural coupling between system and environment, i.e., the system must be able to engage in a mutual sequence of perturbations with its environment. (Riegler, 2002, p. 347)

For Riegler then, we can simply say that there must be a way in which the system adapts to the environment, and the environment is impacted upon by the system. It is worth noting here that there is a distinction between the system and the environment. The system must be contained by the environment and would not exist without an environment. The environment would continue to exist without the system and is not contained within the system itself.

In contrast to Riegler's account, a more stringent account of embodiment is that of *physical* embodiment. This requires that a system has some physical instantiation, in the form of a body. Note that the requirement of a body places an emphasis on the human conception of what counts as a body (thus, any software-hardware relationship is already excluded unless it forms a contained and distinct body in a single location, rather than just having associated matter). To avoid confusion with a requirement for physical matter, this account of embodiment might be better referred to as *formal* embodiment, as in possessing a distinct form. Formal embodiment is again a particular conception of structural coupling. In a more rigorous notion of formal embodiment, connection to the environment should be "not just through physical forces, but also through sensors and motors." (Ziemke, 2002, p. 4).

An even more restrictive account of embodiment is known as *organismoid* embodiment. As Ziemke states, this requires an "organism-like bodily form" (Ziemke, 2002, p. 4). The majority of arguments for this account of embodiment rely on the idea that having a body of a certain form is key to understanding many concepts. For example, without a hand, a cognitive system could not understand the concept of grasping, which is key for the understanding of 'grasping an idea'; nor could it process a metaphor such as describing something as 'a walk in the park' in the same way as a human if it for example travelled on wheels as opposed to two legs. The overarching argument is that in order to function in a human world, a machine must be specifically humanoid. This could be refuted by pointing out that not all *humans* would necessarily fulfil this account of embodiment. Consider a person born without legs who uses a wheelchair. We would surely consider this person no less embodied than the average person, and to suggest otherwise could be quite offensive; why would this be acceptable where a wheeled robot would not be? Or, indeed, an animal who travelled on four legs instead of two, or a snake with no legs at all? Whilst, as suggested by Keijzer (1998), these dissimilarities might preclude the subjects from being good candidates for the study of human cognition (if indeed they are relevant at all) that does not mean that they cannot be embodied.

The most restrictive account of embodiment is *organismic* embodiment. This takes organismoid embodiment one step further, insisting that not only must the system have an organism-like body, but it must be a biological body. The reason for suggesting this is that the development of biological organisms is completely different to that of any machine or man-made mechanism. For example, an organism develops outwards from a central cell, versus a machine that is constructed externally. Furthermore, in a biological system, self-organisation is the priority to maintain constant (referred to as an autopoietic system) (Ziemke, 2002, p. 5). In a machine, the argument is made that it cannot self-correct its organisation; if a piece goes wrong, an external process must be carried out in which a replacement is made and inserted, making it heteronomous (Ziemke, 2002, p. 5). This view of embodiment often aligns with a radical account of embodied cognition, that is, that *all* cognition is based in the body or is embodied, a controversial claim that many would not uphold.

Much of the literature oversimplifies the idea that to not require embodiment in cognition means one is subscribing to a form of dualism. This, however, is to downplay the role of self-organisation of the brain. The body in any sense does not have to mean arms and legs, eyes and ears, but can also mean the physical housing of many cognitive processes. A functionalist account would surely also reject the idea that organisation of neural processes must be housed in a biological body, much as they would reject the insistence on the material substance housing consciousness being biological, or even a perfect physical replication of a humanoid body or brain. Therefore, as this thesis generally takes a functionalist standpoint, both the orgnanismoid and organismic accounts of embodiment will be rejected.

The formal account of embodiment, as laid out here, also seems to have limitations.[81] Having a body can be understood in several ways: having a thing that is in our control, that can be changed and adapted, that can be altered, that we are not capable of being without, and that does not exist outwith an environment (unlike an environment which can exist without a body). In this sense I agree with both Riegler and Franklin in saying that embodiment does not require a co-located physical 'body' in which the system is fully contained. The requirement for adaptation or impact over time in the 'historical' structural coupling account will be maintained, due to the breadth of systems that could count as embodied otherwise, as discussed above. Moving forward, this will be referred to merely as structural coupling, as this is still the key concept of embodiment (only the stringency of the account is narrowed).

Moving forward with this structural coupling view of embodiment, we can turn to the application of embodiment to cognitive systems. For Riegler, the key to adaptation is to integrate new experience from the environment, so in order for a cognitive system to be considered as embodied, it must have the capacity to adapt or change in response to environmental input or changes. This change cannot be merely reactionary, but a change in reaction. For example, a computer reacting to heat in the system by turning on the fan to cool down would not count in this regard; this is a pre-programmed reaction to an expected environmental change. If an AI were to learn that the weather was getting hotter from increasing temperatures and alter itself to turn the fan on pre-emptively, this would count, as the

[81] Formal here refers to having a distinct form.

AI has altered its reaction as a result of the changing environment. It is no longer merely reacting in a pre-set way. Similarly, in humans we might think that body regulation through sweating is not our individual adaptation to heat; we are not in the moment (regardless of the evolutionary process that got us as individuals to this point) adapting as such, our body is simply doing what it is programmed to do. If, however, we decide to soak a T-shirt in cold water and put it on, this would count as adapting, under this account.

I propose that this is equivalent in cognitive systems to network plasticity, that is, the ability of a system to alter itself from its original programming. As stated in Ziemke regarding Riegler's conception of embodiment: "'Computer programs may also become embodied' if they are the result of self-organisation rather than explicit design" (Ziemke, 2001, p. 4). This ability to self-organise, i.e., change, is the mechanism by which the integration of new experience will occur. Further to this, of course, in order to achieve structural coupling, the system must also have an impact on the environment itself. This cannot be specified for all AI, as it is dependent on the environment and the behaviour of the AI.

CREATIVITY: CONDITIONS

Having considered embodiment, we will now turn to creativity and discuss the relevant condition for a process to count as creative. For our purposes here, the assumption should be that all other conditions of creativity can be met. The theory of creativity utilised here will be Gaut's account of creative agency, as outlined in Chapter 2 (Gaut, 2010; 2018). The salient point will be that in order for creativity to occur, there must a creative agent i.e., there must be agency. To be considered an agent in the creative sense, a system must be able to exercise individual judgement (Gaut 2010, p. 1040). One cannot be considered a creative agent if one is merely following rules or instructions, or if goals and action are completely determined by an external factor, such as luck (Gaut, 2010, p. 1040; 2018, p. 132).

We considered in the previous chapter the case of painting by numbers. A person (or indeed an AI) who perfectly executed the planned colours on a painting by numbers would not be said to be a

creative agent. The act itself would not be considered a creative one (Gaut, 2010, p. 1040). Given the nature of programmed AI, arguably an AI that executes programmed protocol is simply following rules, much like a person following a painting by numbers or executing a recipe. There is no room for individual judgement, thus there is no creativity (Gaut, 2010, p. 1040). It can be taken from this that the judgement must be from the creative agent themselves in order to be considered creative. Just as a product of a process that is completely down to luck or rule-following would not be considered creative, an action completely determined by one person cannot be considered a creative act by another. In order for an AI to be able to have 'individual judgement', it must have a degree of autonomy.

Autonomy in AI is a specific condition, which I have discussed at length in Chapter 2. In my response to Boden's concerns about the possibility of autonomy in AI, I argued that freedom from original programming in terms of *how* an AI might approach a problem or goal (through the control of intermediate goals) can count as a level of autonomy, as opposed to what we could consider as full autonomy where an AI sets the goals itself. Under this account, variation from pre-programmed operations (for example, through machine learning, a purposeful addition of an ability to adapt, or generational algorithms) would count as some level of autonomy. Changes in all cases would be in response to input, or changes in the environment of the AI, as an AI will not alter itself without any input, training, or interaction with the environment. In Gaut's account, it is only *complete* dependence on rules or luck that precludes creative agency, thus I will maintain that it is only *complete* compliance with original programming that cannot count towards creative agency.

CREATIVE AI: EMBODIED?

As laid out above, embodiment requires historical structural coupling. This translates to a system adapting in response to the environment *and* impacting the environment. In fulfilling the conditions of creativity as laid out by Gaut, there is a requirement for individual judgement to be attained (in order to count as a creative agent) (Gaut, 2010; 2018). For this to be possible, an AI must be autonomous to some degree. This most simply translates to network plasticity, i.e., a freedom to alter its own

programming. This alteration will occur in response to input or change in the environment in some form (thus separating itself from the original programmer).

In fulfilling the condition of being a creative agent, a creative AI would have to be autonomous, that is, it must change itself in reaction to environmental input. This, in turn, fulfils the first element of structural coupling, adapting itself in response to environment. The second condition of historical structural coupling is also fulfilled. In creativity, there is an act of creation, which changes (adds to) the environment (through the creation of a new object). This fulfils the second element of structural coupling. It follows therefore that an AI that can be considered creative will automatically be fulfilling the requirements of being embodied.

The argument made is as follows:

P1 A: Embodiment (minimally) needs structural coupling.

B: Structural coupling equates to integration of experience of the environment (change) and impact on the environment.

P2 A: In order to count as a creative agent (to be capable of individual judgement, i.e., not merely following rules), an AI must be (somewhat) autonomous.

B: In AI, some autonomy (in terms of control of sub-goals) is gained through plasticity – change from the original programming. This occurs in reaction to environmental input.

P3 In creation, something new occurs in the environment, which has some kind of impact on the environment.

Therefore

C Any creative AI will automatically be fulfilling the requirements of being embodied.

OBJECTIONS

This characterisation of embodiment remains unintuitive. The idea that a 'body' in the embodied sense need not be in one single place may be unacceptable to many. This paper will now address several objections to the argument made here, which will hopefully assuage some of these concerns.

The first objection could be that AI creativity is outright not possible. Whilst some are particularly attached to the idea that creativity is a human capacity only, this should not provide a large problem for this argument. This is because this argument only states that if an AI *were* to be creative, it would necessarily be embodied. Some might view this as unachievable, but the hypothetical case is all that is being posited here.

An objection may also be raised regarding whether a creative product fulfils the 'impact on the environment condition' of structural coupling. Some might suggest that this is not a sufficient account of impact. Indeed, there are many different ways in which an embodied system may affect the environment and creative product is only one of these. For example, a painter in creating an artwork might spill paint, or a scientist may have experimental by-products. In other words, it is rare that the final product of creativity is the sole impact of the creative process. The issue that we are faced with, however, is that there is no dominant theory of creativity which includes or defines the action of embodiment in creative process. The key for showing that AI creativity is necessarily embodied is to show that there is a two-way relationship between the system and the environment. More specifically, it may be argued that the impact on the environment must be directly related to the creative process (not, for example, that a creative human is simply existing in the environment, but that a creative human is acting in the environment). Therefore, for my purposes here, the clearest way in which a creative system impacts the environment is through a creative product.

A further objection may be levied at the conception of embodiment utilised in this paper. It could be said that (historical) structural coupling is not sufficient for embodiment. I have utilised a minimal account here but there remains much debate about the conditions for embodiment. The claim here is not that that this is equivalent to embodiment in humans. It may be that having a distinct body

(which I have referred to as formal embodiment) produces a higher degree of embodiment and provides greater abilities in certain kinds of cognition. As embodiment, as linked to creativity, has barely been discussed, it would be a leap to posit a higher degree of embodiment as necessary for creative cognitive processes.

Following this objection regarding the stringency of embodiment in this account, one might question how permissive this structural coupling account is. Would not anything be considered embodied? This is not what this account allows. First, to be embodied, something must be a system; second, that system must be capable of *adapting* and actively *adapting to* changes in the environment; and third, that system must impact on its environment (in a way beyond merely the molecular level). A book for example, would not be considered a system and, even if it could be convincingly argued to be, it would not be able to be shown to adapt to its environment or impact upon it.

A further question might be: is any AI non-embodied? The answer to this would simply be yes. Not all AI are adaptive, as not all AI have the ability to change themselves. In the case of art-making AI for example, the *Emotional Painting Fool* (Colton, Valstar & Pantic, 2008) is an AI that creates a portrait of a person that is responsive to the emotion they are expressing. This AI involves emotion detection and painting software; however, it is in no way able to alter its program. It cannot learn from what it sees or what it does and is not adaptively responsive to any environmental factors. It merely follows what it is programmed to do. This does not mean that it is not in some ways an 'intelligent' system. It may be very successful at emotion detection, and it may produce interesting images, but these skills and images will not develop over time.

CONCLUSION: CREATIVE AI AND EMBODIMENT

This section has addressed the role of embodiment in AI creativity, proposing that under a structural coupling account of embodiment, creative AI is necessarily embodied. I laid out accounts of embodied and justified the use of the historical structural coupling account. I isolated a relative condition of creativity under Gaut's theory, individual judgment, and outlined the way in which AI could meet this

condition, through the ability to alter itself in response to environmental input. I then argued that in fulfilling the autonomy condition of creativity, an AI would necessarily meet the conditions to be considered embodied. Finally, I addressed some potential objections to this argument.

Whilst this may remain an unintuitive account of embodiment in creative AI, there is a basis for claiming it to be just that. Embodiment is not currently widely discussed in the literature surrounding creativity; however, the threat of creative machines may cause sceptics to turn to embodiment as a way of privileging humans as the sole creative systems. In addressing embodiment in creative AI, an answer for these objections is provided. Whilst AI systems may not be embodied to the degree that we think of human cognition, there would be a minimal level of embodiment in autonomous AI that could impact its external environment.

AI, Agency and Creativity

Agency is a requirement for Gaut's account of creativity (see Chapter 2). Action is also key to the cluster account of art and the institutional account of art; as actions are performed by agents, agency may also be key for making art (see Chapter 1), though my focus in this section will be creativity. In Chapter 2, I argued that AI could meet all other requirements of creativity. For Gaut's account, these were novelty, value, flair, with the later addition of spontaneity (which I characterised instead as unpredictability). However, given that Gaut's account of creativity is *agential,* my previous work cannot convincingly argue that AI could be capable of creativity if I cannot show that AI can also have agency. As we saw in Chapter 1 though, the capacity for agency is also relevant to the cluster account of art. This is due to the necessary condition of action. If an AI can act, then the work it produces can be considered as art (as long as it meets enough of the criteria of the cluster account).

Agency is typically defined as the capacity for intentional action. Current Artificially Intelligent systems are, on the standard view, not capable of intentional actions and thus do not have agency. Some, in fact, claim that mental states, such as intention, will *never* be possible for AI.[82] This presents a clear issue for discussions of AI creativity (and art made by AI). If AI cannot have agency, then creative efforts will forever be limited to mere simulations of creativity, an impressive but ultimately ersatz creativity from machines. This section of the thesis will examine the claim that AI cannot have agency and therefore cannot be creative. I will offer a possible solution to the problem of agency for AI creativity and thereby challenge the resultant conclusion that AI cannot ever be truly creative. Before we address why agency is a problem for AI creativity, we first need to understand why agency is so central to creativity.

[82] As discussed in Chapter 2, John Searle perhaps offers the most well-known challenge to this with his Chinese Room thought experiment (Searle 1980).

AGENCY AS A REQUIREMENT FOR CREATIVITY

Under popular accounts of creativity, agency occupies a central role. I will here examine the role of agency in these accounts of creativity, in particular looking at the justifications for including agency as a requirement for creativity.

As we saw in Chapter 2, Gaut's account of creativity is one of the leading accounts of creativity in philosophy. Agency is a central requirement of Gaut's account, to the extent that he refers to his account as an 'agential' account of creativity; for Gaut, in order to be creative, one must be an agent. Gaut states: "Creativity is a property of agents, not of mere things or plants" (Gaut, 2010, p. 1040). This claim, it seems, is the basis by which the account becomes agential. Gaut offers no further illumination of why creativity is exclusively the property of agents, other than to exclude two cases in particular:

> The value and originality conditions also do not suffice for creativity. Tectonic movements of the earth's crust have the capacity to produce diamonds, which are valuable (financially and aesthetically) and some are original (in the sense of being saliently different from other diamonds); but it would be conceptually confused to call tectonic movements creative. Rudolf Arnheim has argued that trees can be creative, on the grounds that they distribute their branches to make best use of light and that the resulting canopy 'represents the solution of a vital problem and what we experience as the beauty of the tree. The tree is acting creatively, not just metaphorically – it is the real thing.' (24). But the tree is not acting at all, since it lacks desires, beliefs and other intentional states; so *a fortiori* it cannot be acting creatively. Creativity is a property of agents, not of mere things or plants (see also Carruthers, The Architecture of the Mind, ch. 5; Stokes). (Gaut, 2010, p. 1040)

Here, Gaut cites Carruthers and Stokes to support his claim. We might expect that turning to Carruthers or Stokes would illuminate why creativity must be the property of agents. However, Carruthers does not articulate any clear justification for agency as a requirement for creativity other than asserting that creativity developed somewhere in the lineage of great apes (thus presumably only those animals in that evolutionary line may be creative). This does not make clear the necessity for agency unless we

assume that only these animals may have agency. The closest thing is that Carruthers argues that creativity can be reduced to creative action, stating "I shall suggest, in fact, that all creativity reduces to the creative generation of action schemata" (Carruthers, 2006, p. 287).

But not so for Gaut. He argues that not all creative processes are reducible to action,

> Creativity, I have argued, is a property of an agent with certain capacities. It does not follow from this that all creative processes (sequences of events) are actions, only that such processes must occur in agents. (Gaut, 2010, p. 1041)

As Gaut acknowledges, this leads to an objection: if creative processes are not necessarily actions, then why does creativity necessitate agency? Gaut responds by pointing out that there is a tension in creativity between actions (which are teleological) and the *a priori* argument that demonstrates that creative processes *cannot* be teleological: "if a process of making something is creative, then one cannot know the end: for if one knows the end, one has already been creative" (Gaut, 2010, p. 1041). However, Gaut rejects the *a priori* anti-teleological argument with several counterexamples. For example, he explicates cases of creativity where the ends are determined but the means are not and where the goal is only partially determined. Ultimately, he states that "token creative processes can be teleological, though it does not follow from this that all must be so" (Gaut, 2010, p. 1042).

So, for Gaut, creative process *can* be actions but are not *necessarily* actions and, similarly, are not *necessarily* teleological, but also *can* be. After all this, what remains unclear is why Gaut sees agency as necessary as it does not stem from action and it does not stem from inspiration (i.e., anti-teleology). Beyond excluding natural phenomena from counting as creativity (like Arnheim's trees), why are only agents permitted to be creative, and why is all of creativity agential? We do not get a compelling answer from Gaut to this question.

Where does this leave nonhuman creativity? Gaut does not necessarily preclude nonhumans from creativity providing that they are agents. As he explains, "creativity … is a particular exercise of agency. As such it is open to agents, whether human or not, that have the requisite capacities." (Gaut 2010, p. 1041). Gaut allows that the agents in question may, in fact, not be human. However, he does

not detail the exact form of agency to which he subscribes. We can find a clue of what he means by 'agency' in his argument against Arnheim's claim that trees are creative: Gaut states that the tree is not acting because it lacks "desires, beliefs and other intentional states" (Gaut, 2010, p. 1040). It is therefore not clear who or what he means to let in with the claim that creativity is "open to agents, whether human or not". We might assume that he is merely leaving the door open for nonhuman agents (or other descendants of great apes per Carruthers) to be creative if it transpires that they can indeed have the appropriate capacities. This may well include computational systems.

So, we cannot get much more clarity on Gaut's motivation for including agency as necessary for creativity from the citation of Carruthers. That leaves us with Stokes. In "A metaphysics of creativity" (2008), Stokes claims that, at minimum, a creative process needs a "processor" with agency. This agent must be "responsible"; without a responsible agent, we cannot attribute creativity (D Stokes, 2008, p. 116). As Stokes explains,

> Creative artworks are things that are done and made, and for which we praise their makers. The processes that generate them involve intentional agency and it is this process, at least in part, for which we praise the agent. This implies that the process must depend in some non-trivial way upon that agency. We do not appropriately praise (any more than we blame) agents for processes out of their control. We capture this intuition with a simple condition on creativity; call it the *agency condition*. Some *F* is creative only if *F* counterfactually depends upon the agency of an agent *A*. (D Stokes, 2008, p. 116)

So, according to Stokes, not only does creativity require the creator to be an agent, but the creative process must also depend non-trivially on the agency of that agent. Like Gaut, Stokes bestows an important and central role to agency (it is both "fundamental" to and a necessary condition for creativity) (D Stokes, 2008, pp. 116-117).

For Stokes, the requirement for agency then stems from a requirement that creators ought to be held responsible for their creations if they are, in fact, creative. This idea of responsibility and the legitimate assigning of praise or blame for creative outputs is an interesting justification for the agency

requirement. We may assume from the fact that Gaut cites Stokes that this responsibility condition sits well with Gaut's account of creativity. Indeed, although this is not explicit, Gaut seems to reject several types of processes by agents on the basis that the agent cannot be held properly responsible for them:

> In short, the kinds of actions that are creative are ones that exhibit at least a relevant purpose (in not being purely accidental), some degree of understanding (not using merely mechanical search procedures), a degree of judgement (in how to apply a rule, if a rule is involved) and an evaluative ability directed to the task at hand. As shorthand for these features, we can say that creative actions must exhibit flair (Gaut, 2010, p. 1040)

These features are those that Gaut gathers under the condition of 'flair'. The first condition here, that creative actions are not accidental, aligns with the praiseworthiness or blameworthiness that Stokes highlights as key for creativity. This would seem to be enough for Gaut to dismiss cases where an agent cannot be held responsible for the outcome of the creative process, as it is for Stokes. However, Gaut seems to pull apart from Stokes; instead of these cases being simply excluded by the requirement of agency (as Stokes suggests with the application of responsibility for action entailed by agency), Gaut ensures the exclusion of these cases through another requirement: "flair". Because Gaut diverges from Stokes, we cannot be certain just how much Gaut agrees with Stokes' responsibility condition. In any case, we will re-visit the attribution of responsibility later.

Gaut and Stokes also seem to have slightly different views on the role of action in creativity. In a later paper, Stokes implies that creativity is a kind of action: "Agency involves action. Creative things, whatever else is true of them, are things we *do* as agents" (D Stokes, 2011, p. 660). This leads Stokes to the necessary condition of agency: "an *x* is minimally creative only if *x* is the non-accidental result of agency." (D Stokes, 2011, p. 660). Stokes intends this condition to select only actions and consequences that non-accidentally result from agents. As we have already seen, Stokes individuates these as actions to which we can assign praise or blame:

> Kicking someone in the shins, stealing your little sister's lunch money, or cheating on an exam are all blameworthy. Praiseworthy acts and products also depend on agents with intentions. (D

Stokes, 2011, p. 660)

As Stokes states, this condition picks out a "class of actions and consequences" (D Stokes, 2011, p. 660). Presumably, creative consequences are the outputs of the creative actions here. Gaut, on the other hand, is not wedded to the claim that creative things are actions, or even doings. Gaut states that creative processes (which may include actions) must be done by agents with the right capacities. Gaut has referred to creativity as both a *property* of agents and, in a later paper, as a *disposition* of agents (Gaut 2018, pp. 132-133). For our purposes here, it does not matter whether creativity is a capacity, property, or disposition, only whether or not it is agential. Gaut's citation of Stokes is therefore no more illuminating that Carruthers because Gaut similarly diverges with Stokes on the essence of creativity as action.

Where does this leave us? Stokes and Carruthers both reduce creativity to actions (done by agents), but Gaut does not characterise creativity solely in terms of action. Instead, Gaut calls creativity an agential disposition. As Gaut does not align with either Stokes or Carruthers on action, then he cannot use them to justify the necessity of agency in his account. This leaves the door open for us to question why agency is necessary and whether something else might do the job. Of course, not everyone agrees that agency is a requirement for creativity. As discussed in Chapter 2, the Darwinian theory of creativity does not assert any requirement for agency at all. Instead, it focusses on the selectionist theory of biological evolution as a template for understanding the development of various human processes such as creativity. The model proposes that creativity follows the path of evolution, requiring blind variation and selective retention. According to this model, blind variation occurs in a system (blind in the sense that the success of this variation is unknowable at the time of variation), and this variation is then tested against the environment for fitness. The best fitting variations persist and are passed on to new generations (for example, through survival of the fittest or sex selection). The worst fitting variations (for example, mutations that are not compatible with life) will not be passed down. Creativity in this sense is the process of blind variation occurring, and selective (successful) variations being retained (Simonton, 1999). It does not require any agency at all in order to be considered creative. Boden recently has argued that there is such a thing as "biological creativity". Utilising her framework of creativity as

novelty, value, and surprisingness, Boden shows how this can be applied to biology, not just in terms of phylogeny (evolution of species), but also in terms of ontogenesis, the development of individual organisms through life. There is no requirement for agency in Boden's account, simply the use of the concept of self-organisation (Boden, 2018).

If we were instead to use these biology-based accounts of creativity, the issue of agency would disappear. We could indeed show that AI creativity is possible by adopting a Darwinian model of creativity or Boden's model of creativity, effectively dropping the condition of agency. I explored these accounts in Chapter 2, demonstrating how AI could meet their requirements. Simply taking this route without addressing agency though would still leave doubt in the minds of those who insist on agency as a necessary condition of creativity. Let us assume for the rest of this chapter that some form of agency is necessary for creativity.

WHY IS AGENCY A PROBLEM FOR AI?

Agency is evidently central to the prominent theories of creativity within philosophy (namely, Gaut and D Stokes). We could do well to make sense of *which* definition of agency Gaut and Stokes are relying upon. Agency is typically understood as the capacity to act with intention. As noted in the *Stanford Encyclopedia of Philosophy*,

> The standard conception of action provides us with a conception of agency. According to this view, a being has the capacity to exercise agency just in case it has the capacity to act intentionally, and the exercise of agency consists in the performance of intentional actions and, in many cases, in the performance of unintentional actions (that derive from the performance of intentional actions). (Schlosser, 2019)

This is the view seemingly held by Gaut and Stokes; both make reference to intentions in their discussions of agency and creativity as shown above.[83]

[83] Stokes does acknowledge alternative conceptions of agency, which we will turn to in a moment.

So, agency is both a requirement of creativity, and, under most accounts (including those used for defining creativity) is comprised of the mental state of intention. This is a significant roadblock for artificial creativity and, indeed, for advancements in AI more generally. The widely held position today is that AI is not capable of mental states such as intentions.[84] This is because AI is typically considered to be incapable of having mind. That is not to say that we might not be able to argue that AI has mind,[85] but arguing for AI mind does not necessarily help us much with our problem of agency here. There are two reasons for this: first, the possibility that some AI somewhere might have mind does not mean that current "creative" AI have mind. Second, having something like mind (or mental capacity of some description) is not sufficient to qualify as having intentionality. Hence, unless we can prove that "creative" AI have a capacity for full-blown intentionality (which is unlikely), we will not be able to argue that these AI are creative at all.

I am not alone in identifying this as a noteworthy problem for AI creativity. Stokes, in his paper on the necessity of agency for creativity, also addresses this:

> Perhaps this is too fast: paintings and other artefacts do not provide fail-proof evidence for agency: one can mistake agents for non-agents and non-agents for agents. This invites a second point. A debate about artificial intelligence and creativity centres, in part, on the question of agency. Whether a computer or its products are creative (*actually* as opposed to just *apparently*) depends upon a more fundamental question, namely, whether the computer is an *agent* with certain cognitive or behavioural capacities. Is the computer autonomous and responsible for its computations and products or is it, as we say, "just running its program"? (Boden, 1999, 2004; Cope 1991, 2001; Dartnall 1994; Hofstadter 1994, 2002; Hofstadter and FARG 1995.) Whatever epistemic difficulties there may be, one does not properly attribute creativity to the computer until one properly attributes agency to that computer. Both of these considerations motivate the same point: creativity requires agency. (D Stokes, 2011, p. 660)

[84] I have offered alternate positions elsewhere in this thesis, notably in Chapter 1.
[85] As I have done elsewhere in this chapter (see section on AI extended mind).

Stokes here again relies on the attribution of responsibility. The question of agency remains key for establishing whether AI can be creative under these popular accounts of creativity. If agency is – as Stokes, Gaut, and other subscribers to their views claim – necessary for creativity, then *the question of whether AI can be creative hinges on whether AI can have agency*.

A SOLUTION: REDEFINING AGENCY

To sum up, AI has an agency problem. In the world of AI, the systems that are designed to be 'creative' are not typically thought to have minds (and mental states), and therefore cannot have intentions, and therefore cannot have agency, and therefore cannot be creative. What can be done to respond to this problem? We have several possible options:

1) First, we can drop the requirement for agency entirely, and accept an alternate account of creativity. In doing this, we sever the link between agency and creativity. As I state above, I will not do this here (I do explore this possibility elsewhere).
2) Second, we could argue that intentions do not require mental states, thereby severing the link between mental states and intentions. This would require us to adopt an alternative view of intention. For example, we could adopt G E M Anscombe's (1963) position that intentions are not mental states, but rather are a form of knowledge (Clot-Goudard, 2022; Richter, no date). I do not argue for this here.
3) Third, we could argue that agency does not require intention. I will take this route.

To argue that agency does not (necessarily) require intentions I will offer a different conception of agency to that implied by Gaut and by Stokes, based on a form of agency used by Stokes elsewhere (D Stokes & Bird, 2008). This form of agency will not rely upon full-blown intentionality, as Gaut and Stokes imply is necessary for creativity. If accepted, this will sever the connection between creativity and mind, leaving space for non-minded and nonhuman creativity.

The default definition of agency, which can be boiled down to 'capacity for intentional action', seems to be the one typically used as the necessary condition for creativity. Do we necessarily need to

use this definition? Stokes (2011) suggests another definition could be used, led by the intuition that computer or animal creativity is possible. In a footnote, Stokes states:

> If one thinks that animal or computer creativity is possible, then an agency condition might be construed more broadly. For example, one might take agency to only require autonomous action that involves, minimally, behaviour mediated by internal mechanisms of a system and some degree of input/output flexibility. So an organism or system is an agent so long as elements or mechanisms internal to the system can produce varying outputs given any particular input (see Stokes and Bird 2008). The analysandum for this article is human creativity, which calls for a richer notion of agency. This will be assumed for the remainder of discussion. (D Stokes 2011, footnote 6, p. 663)

It is not entirely clear why this definition is not appropriate for human creativity. Can human creativity not be defined in the same terms? Or would we lose something in this broader definition?

Stokes is focussed on human creativity, and in the case of humans, using this version of agency may result in calling something creative when it is not. However, if we utilise different standards of creativity for humans and non-humans, as Stokes seems to suggest, how do we know we are capturing the same thing? In response to this, perhaps we can make a move similar to Stokes in his paper "Minimally creative thought" (2011). In this paper, Stokes proposes a minimal definition for creative thought. The aim is to define what is at minimum necessary for creativity (which could be expanded for specific cases, such as transformational creativity) without using extreme cases as the basis of definition (D Stokes, 2011, p. 659).[86] However, Stokes focusses on a minimal definition for creative *thought,* which ultimately still implies thinking, and thus might preclude AI creativity. But perhaps an account of minimal creativity is the way to go; this *minimal* account of creativity could include a *minimal* account of agency. In this account, minimal creativity will still include human creativity.[87]

[86] Stokes makes a case for minimally creative thought that "depends non-accidentally upon agency, is novel relative to the acting agent, and could not have been tokened before the time it is in fact tokened, relative to the agent in question." (D Stokes, 2011, p. 658).

[87] The use of minimal agency may let in some human actions that we do not typically consider creative, and so we might want then to add additional stipulations for discussing human creativity. I will not address this here.

Moving this concern aside, let us look at what Stokes and Bird do with a minimal account of agency for nonhuman creativity. In their chapter "Evolutionary robotics and creatives constraints" (2008), they make use of another broad conception of agency, not too dissimilar from the one above:

> Agency, at least for a start, may be understood broadly. A system is an agent if that system is self-moving, that is, not entirely controlled by an *external* system or programmer. This requires neither self-generation nor cognition or deliberation. Thus many of the simulated robots found in artificial life research qualify while remote controlled robots do not. Some behaviour, artefact, or event *F* is the product of the agency of A only if *F* would not have occurred had A not acted in some autonomous way. (Stokes & Bird 2008, p. 231)

Stokes and Bird use this broad definition of agency to pursue their creation of a creative robotics system. Stokes and Bird designed small robots that had basic perceptual capabilities and that (armed with pens) could draw lines beneath them. Rather than executing commands from their human programmer, these "embodied robots" responded to environmental stimuli. Not only did they respond to the environment, but, with the use of artificial neural networks (ANNs),[88] they "evolved" in their responses and developed novel responses to environmental stimuli, including the lines that they drew (Stokes & Bird, 2009). Using this kind of system, Stokes and Bird were able to create agential robots as per their definition:

> These individuals are *agents* on the broad construal suggested … When we hold the starting conditions and testing arena constant, different agents respond differently (where agents are individuated by their artificial neural network controllers). Indeed some of the agents act in behaviourally novel ways. (Stokes & Bird, 2009, p. 239)

Stokes and Bird are confident that their robots are agential by their definition, and therefore provide the basis for future creative robots. In theory, if the lines drawn by these robots fulfilled all other conditions of creativity (such as those provided by Gaut), they should be creative.

[88] This approach to robotics is used to reduce experimenter bias in how the system is developed, ensuring that the robots are not unduly influenced by the plans or expectations of the designers.

If these robots created work that was of value, and was novel, would they be creative? I am not convinced that they would. The agency used by Stokes and Bird is too weak; Arnheim's trees arguably could be considered an agent under this account, as they are 'self-moving' (in terms of growth), and this is not controlled by anything external to the trees. Stokes and Bird's robots could also only be thought of as autonomous in the sense of being independent (see Chapter 2). Would these robots have agency in any meaningful sense? I do not think that they would. Why not? The reactive nature of the robots suggests that there is no purpose to what they do i.e., no reason, aim, or goal for their behaviour.[89] Put another way, this system does not seem to be teleological. Indeed, as the robots utilise evolutionary robotics, like the evolutionary model of creativity, they are by design non-teleological and therefore without goals.

The aim that Stokes and Bird have for the system is for it to achieve novelty. They suggest that it did, but, as this was the goal that the designers set for the system, the robot was not autonomous in the setting of the goal. At the most generous interpretation, we might say that the aim of the robots was to maximise "fitness scores"; however, this goal, too, is in the evolutionary algorithm utilised by the authors.[90] It is not at all self-determined by the robots. So, in what sense were the robots autonomous? Merely in the exact patterning of the lines drawn on the page? This seems to be an exceedingly minimal (and non-teleological) conception of agency.

Why is this lack of purpose a problem? Purpose*less* doing is a far cry away from intentional action, so far away that it is not a candidate for creativity at all. Purposeless doing does not meet the requirements of creativity as Gaut and Stokes lay out in their accounts of creativity. In particular, it does not line up with Stokes' own view regarding how we assign responsibility for creative actions. Could we really say that these robots are responsible for their actions? Perhaps in a very narrow sense, they are responsible for the exact placement of the lines marked on the floor. But it would not seem they are responsible for the patterns that they make. They are programmed to draw lines where there

[89] We can of course refer to the goals set by the designers of the robots, but this does not impact the behaviour in any way beyond how they are set up.
[90] Fitness scores are used in genetic algorithms. They are a score given to an output, measuring how well that output meets the aim of the system (as pre-set by the designers).

are none, and they sometimes do so in novel ways. Can we praise them for any novelty that we see? They have no self-determined goal, and no clear awareness of the 'image' that they are making. Their 'evolutionary' nature means that there is no goal in how they change their actions, and it is done with no aim to a final outcome. They may be 'acting' autonomously in the sense that they are not *entirely* controlled by an external system (like a remote-controlled robot), but, without purpose, this 'action' is of a very weak kind. The non-teleological nature of the robots' 'actions' is a limitation – it is unclear how these actions, even if behaviourally novel, could be considered to demonstrate any meaningful sense of agency such that we could assign responsibility and thus they fail to approach even the rudiments of creativity. They would need a stronger definition of agency in order to be considered a responsible agent.

AN ALTERNATIVE

Instead of the weak account of agency utilised by Stokes and Bird, I will propose a stronger account that still stops short of the requirement of intentionality. I will refer to this as 'minimal agency'. I will add a teleological condition to Stokes and Bird's conception of agency: "Some behaviour, artefact, or event *F* is the product of the agency of A only if *F* would not have occurred had A not acted in some autonomous way." (Stokes & Bird, 2009, p. 231). The teleological element that I am proposing will be derived from a minimal definition of action. Human action is typically defined using intentionality. However, we can recognise that there are multiple levels of action, ranging from weak (or minimal) to strong. For instance, a spider moving across the floor is in direct control of his limbs, and it is directing its legs to take it from one side of the room to the other.[91] The spider's movements have an aim or purpose for the spider, and therefore can be explained teleologically. This teleological doing is surely a minimal form of action. We would not want to say that the spider has full-blown intentionality, but it would be odd to say that this is not an 'action' on the part of the spider. We could describe human activity in a similar fashion. When you feel an itch on your leg, you will reflexively reach down to

[91] As I will come on to, this example of the spider is from Frankfurt.

scratch it. You did not form intentions to do so, you merely did so to satisfy the goal of dispelling the itch. Both the scuttling spider and you with your itchy leg describe 'action' in a fairly minimal sense, but action (with a goal) nonetheless.

We can find support for this way of thinking about action from Harry Frankfurt. Frankfurt emphasises that a focus on humans as the only possible source of agency is chauvinistic and wrongly assumes that humans are a special case: "the concept of human action is no more than a special case of another concept whose range is much wider." (Frankfurt, 1978, p. 162). By widening the concept of action, Frankfurt advocates for agency to be understood in terms of a wider range of doings. While this widened conception of action includes the special case of human action, it also allows for agency as an explanation of non-human actions. To explain the attribution of 'action' in non-humans, we can consider again the spider making its way across the room:

> Consider the difference between what goes on when the spider moves its legs in making its way along the ground, and what goes on when its legs move in similar patterns and with similar effect because they are manipulated by a boy who has managed to tie strings to them. In the first case the movements are not simply *purposive*, as the spider's digestive processes doubtless are. They are also *attributable* to the spider, who makes them. In the second case, the same movements occur but they are not made by the spider, to whom they merely happen." (Frankfurt ,1978, p. 162, my emphasis)

The key elements here for calling the spider's movements 'actions' are 1) that the movements are *purposive* and 2) that they are *attributable* to the spider. This second condition, that movements are attributable to the spider, is very similar to that condition laid out by Stokes and Bird, saying "Some behaviour, artefact, or event *F* is the product of the agency of A only if *F* would not have occurred had A not acted in some autonomous way." (Stokes & Bird 2008, p. 231). Acting in an autonomous way is accommodated in the "attributable to the spider". The puppet strings of Frankfurt's sadistic spider-manipulating boy are much like the virtual strings connecting a robot to its controller.

Stokes and Bird's minimal account of agency accommodates *attribution* of actions, but it does not include any component accounting for the *purposive* nature of action. "Purposeful doing" or "goal-directed activity" is, according to Frankfurt, a key part of defining minimal agency in nonhuman animals. Including a condition of *purposiveness*, then, may help a weak account of agency be more compelling, and will give us a more teleological account. As I have already argued, a teleological conception of agency can be a better substitute for an intentional account – when things are done with intention, they are also done purposively (though purposive action is not necessarily intentional, as with the free spider). However, as pointed out by Frankfurt, purposive doing does not come with the same chauvinistic baggage as intentionality; it can occur in non-humans as part of a continuum of agential action.

I would amend Stokes and Bird's account of agency then, and instead propose that (minimal) agency is the capacity for internally directed, purposive doing. Using this definition may attribute agency to actions by humans that we do not wish to include within human agency. However, such a move does not rule out a stronger, richer account of agency being used for humans. Unlike Stokes and Bird's account, this is a teleological sense of agency. There is a requirement for there to be a purpose or goal in an action in order for it to be considered an exercise of agency. Including "internally-directed" remedies some concerns about what is included in this account of agency. This requirement ensures that the goal, or purpose of the activity in question is internally ascribed and, as such, excludes any cases where all goals of the system are pre-determined by programmers in the case of computational systems.[92]

The conception of agency that I propose should not conflict with Gaut's agential account of creativity, despite the inclusion of a teleological condition. Whilst Gaut does not insist on a teleological account of creativity, he does argue that anti-teleological accounts of creativity should be rejected (Gaut, 2010, p. 1042). If a teleological account of creativity is not necessary, why should we include a

[92] One might object that this would exclude cases such as an artist being commissioned to produce a portrait. In the case of a human, we can easily explain that a person has agreed to paint such a portrait and is thus still self-directed. This move, however, does not work for AI systems, as there is no way in which they can choose to accept a final goal. As argued in Chapter 2 though, they may have control of sub-goals.

teleological component to this account of (minimal) agency? These non-teleological accounts typically stem from anecdotes of the reported phenomenology of creative processes, where creative agents report that great discoveries came to them in dreams, or arose out of nowhere, popped into their heads etc. These kinds of purely non-teleological instances of creativity are not possible for AI, and seem less possible for non-human animals, as they involve, at least as they are reported, the conscious versus the unconscious, and self-report that is not possible for us to garner from non-humans. There is not much point leaving room for completely non-teleological accounts of creative agency, and we lose little by precluding this in the case of non-humans.

This solution does not take away the possibility that humans must reach a higher level of agency as a requirement for human creativity. It does not insist on a proposal to drop agency from creativity entirely. But nor does it insist (implicitly or explicitly) that *only* humans might be creative. Using this conception of minimal agency we can continue, like Stokes does, to develop a minimal account of creativity. But with the inclusion of purposiveness, this minimal creativity will bear resemblance to the stronger conception of creativity as requiring intentional agency.

Could AI system meet this minimal account of agency? A system is internally directed to some extent, and has a purpose for what is does, then it can meet this account. A system would need to achieve some autonomy (this could be through alteration of original programming) in order to be 'internally-directed' and would need to act towards goals (in order to have purpose). We could argue that a system such as AICAN, which alters itself on the basis of learning and generates images that aim to meet newly set evaluative standards (or sub-goals), could meet these requirements. It is debateable though to what extent we would be satisfied on either count. If the system can meet these criteria, it can be considered a minimal agent. If it meets additional criteria for creativity (such as those put forward by Gaut), then it could be creative.

An initial issue presents itself if we are to accept my proposal. We want to be able to assign praise or blame to creative processes and outputs. This is what leads both Stokes and Gaut to require agency in their respective accounts of creativity, so that we may mete out either praise or blame (i.e., responsibility) for that creativity. If we endorse my account of minimal agency, can we still assign

responsibility for AI creativity? Responsibility is often considered to require knowledge and control (Fischer & Ravizza, 1998; Coeckelbergh, 2020). Will this be met through the minimal account of agency I propose here? Let us first take the 'control' requirement of responsibility. This should be included in the 'internally directed' (autonomous) component of the account. As we have seen, Stokes and Bird pursued this line of autonomous AI design in their paper.[93] Their approach is to try to separate the AI from programming by humans by introducing reactivity to environmental stimuli (i.e., independence). On top of this, they also used an evolutionary algorithm, which introduces variations into the system. Another way of introducing autonomy into an AI (thereby further separating its output from the programmer's hand) would be to include machine learning in the AI system; as I have argued elsewhere, this would allow for an alteration of initial programming. This allows the algorithm to change itself, thus adding another link in the causal chain. Depending on *how* autonomous a system is, we may be limited in how much we might want to assign responsibility.

What about knowledge? To meet the epistemological component of responsibility we could necessitate full-blown understanding in our system (which would require us to once again respond to Searle); however, in our context we might be satisfied that we could assign some responsibility to the system if they have the capacity to learn, reason, evaluate, or at least have awareness of what they are doing.[94] Stokes and Bird's robots fail to meet this condition. They are programmed to draw lines, but that is all. As we have seen, they have no aim for their drawing of lines, they have no awareness of what they have made, and also no way to evaluate what they produce. Will our account of (minimal) agency include knowledge? We may expect that some minimal level of knowledge will be implied in our minimal account of agency through the addition of purpose. To have a purpose, one has an aim, or a reason for what it does. To have such a thing, we would expect some awareness, reasoning, or evaluation to be possible. The 'knowledge' required for this will be of limited sense, but as we are only attributing a minimal sense of agency, and thus minimal creativity, I do not think that this should be prohibitive.

[93] Others have made similar efforts. For instance, Simon Colton's *Painting Fool* (e.g., Colton, Valstar & Pantic, 2007).

[94] I suggest this because we are not assigning moral responsibility here, so we would not necessarily need e.g., understanding of harms.

When we assign blame or praise to individuals for their actions, we typically do so with either punishment or reward in mind. There would, of course, be no point in punishing or rewarding an AI. So, it would still be prudent to assign some level of responsibility for AI outputs to the human designers. Just as we might partially blame a parent for their young child's actions, so too might we hold the architects of AI to account for the actions of their creation. However, this impulse does not prevent any meaningful responsibility from being assigned to the AI. I accept that some would resist the notion of assigning responsibility to AI. Nonetheless, I am not alone in making this argument. Whilst it is certainly not uncontroversial, some have argued that responsibility can be assigned to AI where there is appropriate reasoning and awareness (e.g., Dastani & Yazdanpanah, 2022).

CONCLUSION: AGENCY

In this section, I propose a solution to the problem of agency for AI creativity. Agency is a key component of human creativity according to leading theories of creativity in philosophy. In order for AI to have the potential for artistic creativity then, it must have agency; however, as typically defined, agency requires intentionality, something which (as yet) is not possible for AI. This presents the problem of agency for creative AI – they need it, but they cannot have it. For these reasons, it is easy to be sceptical or dismissive about the prospect of creative AI. But, as I have aimed to show, perhaps the situation is not so dire. I have argued that we can make use of a minimal account of agency, defined as a capacity for 'internally-directed purposive doing'. This account avoids the broadness of a non-teleological account, and in doing so captures the desire for a requirement of agency in creativity such that assigning (some) responsibility is possible.

Taking this minimal account of agency together with the discussion of Gaut's account of creativity suggests that, as long as the additional requirements of creativity are met (novelty, value, and flair), then an AI system *could* be (at least minimally) creative. This is not to say there is a system which meets this definition of creativity satisfactorily; I have demonstrated how AICAN might meet this definition, but this could certainly be disputable. Even if AICAN is not creative, we could use this

framework to assess the creativity of future AI systems. AI is not *a priori* excluded from creativity on the basis of chauvinistic requirements. If agency is also required for the cluster account of art and the institutional account of art, then this argument for the possibility of minimally agential AI should also provide a route for the 'doings' of AI systems (i.e., their works), to count as art.

Conclusion: Can *AI* Create Art? Key Limitations

In this chapter I have examined apparent limiting factors for AI art and creativity. First, I provided some background on philosophical conceptions of the mind, before moving on to examine the possibility of AI extended mind. In this section, I argued that AI could have mind under a functionalist account. Then, I argued that if we accept the extended mind thesis, an AI could extend this mind to humans. In doing so, AI could gain access to social aspects of the world without participating in the wider world as a social agent. This addresses concerns with AI works, particularly those expressed by Hertzmann, and may go some way to addressing limitations of AI under the institutional account of art.

I then examined embodiment; AI systems are often unembodied in the standard sense. If embodiment is needed for creativity, then AI creativity will not be possible. Here however, I argue that under at least one account of embodiment, some AI systems could be embodied. Further to this, I argued that if an AI fulfils the requirements of creativity, it will necessarily meet this account of embodiment.

Finally, I addressed agency in AI systems. As we saw in the previous chapters, agency is a necessary condition of Gaut's agential account of creativity. It is also connected to the condition of 'action' in the cluster account of art, and the necessary condition of artefactuality in the institutional account of art. In this section, I argue that some AI could achieve a *minimal* account of agency, characterised as 'internally directed, purposeful doing' and that AI could be creative (in a correspondingly minimal sense), as long as the other criteria of creativity are fulfilled. This chapter therefore completes the argument for why AI can make art and can be creative.

Chapter 4 – AI and Aesthetic Value

In this chapter I will turn to aesthetic value and AI. In particular, I will examine the possibility of an aesthetic 'value alignment problem' in AI. As seen in the examination of the cluster account of art in Chapter 1, it seems that AI works could achieve many properties of art. Several of these properties overlap with popular accounts of aesthetic values, such as expressivism (e.g., being expressive of emotion) and cognitivism (e.g., being intellectually challenging and having a capacity to convey complex meanings). There is a concern when we think of value in AI, however. What if AI do not recognise the same values in art that we do? In this Chapter I examine this problem. After describing the AI value alignment problem and its various formulations, I turned to aesthetics. I argue that aesthetic value appears to be included in current discussions of the value alignment problem, and that we may indeed see misalignment between human and AI values in the aesthetic domain. I then examine the potential impact of such misalignment, addressing whether we should want AI to align with our values in the artistic domain.

Aesthetic Value and the AI Alignment Problem

Current Artificial Intelligence (AI) systems are narrow; they are limited to just one task. This is no different for art-making AI systems. In the future, however, AI may be more generalised. Once we achieve general, human-level, machine intelligence, we may quickly see artificial intelligence surpass us. While today we are not near to producing a generalised, human-level intelligence, philosophers, futurists, and computer scientists are dreaming of, planning for, and researching this possibility. All of those planning for the future of AI are concerned with the potential that 'superintelligent' AI might pose an existential threat to humanity (see e.g., Bostrom 2012). We cannot know what a superintelligent machine will do, and we certainly cannot assume that it will share our goals, motivations, and values. The value of human flourishing, planet Earth, or human life itself are self-evident to us as humans, but there is no guarantee a superintelligent AI will see things the same way. A key aspect of this concern is what has been dubbed the 'AI alignment problem' or the 'value alignment problem'.

The issue of whether AI values will align with human values is well-discussed in the literature on the future of intelligent machines.[95] Concerns lie with the potentially catastrophic result of a superintelligent AI that fails to share human values. Thinkers in the field are concerned, for example, that an AI might sacrifice human life for a relatively minor goal (at least, in our eyes), because they simply do not value life to the extent that we do. This, then, is the problem: how do we ensure that the values of future, superintelligent AI align with our own?

Of course, high-stakes, moral values are not the only key values for humans. What about other values? Will we have an AI value alignment problem in other domains, such as the aesthetic? The goal of discussing the value alignment problem is to determine how to design AI systems to ensure against the existential threat of superintelligent AI. The guidelines for designing AI systems will not apply only to AI working in domains that are clearly related to ethical values (such as self-driving cars, medical diagnostics, or autonomous weapons). They will apply to all AI, including those

[95] See Bostrom (2003; 2014); Christian (2020); Russell (2019); Yudkowsky (2016).

which make art. It therefore becomes relevant to aesthetics what restrictions are imposed on AI systems. How will the value alignment problem and its solutions affect how we want our art-making AI to be designed?

In this chapter, I will explore whether we should be concerned about AI value alignment, beyond the key ethical concerns of superintelligence. Will AI be aligned with our artistic and aesthetic values? And if not, will this be a problem? In examining these questions, I aim also to make an argument about the value alignment problem itself: just how far-reaching is the problem, and how can we accommodate aesthetic considerations into the future design of AI?

THE AI VALUE ALIGNMENT PROBLEM

Before we turn to consider aesthetic value and the alignment problem, we need first have a grasp of why value alignment is a cause for concern for future AI. The AI value alignment problem arises from concerns with how we should design future AI to guard against the existential threat of superintelligence.[96] Nick Bostrom has raised various ethical concerns with AI superintelligence. For Bostrom, the dawn of superintelligence poses what he calls 'existential risks' to humanity (2009), that is, superintelligent AI could, one way or another, spell the end of humanity. What is more, these existential risks may arise from seemingly innocuous differences in motivation or goals between humans and AI. To illustrate this, Bostrom proposed a thought experiment: the paperclip maximiser. The scenario is as follows. Imagine there is a superintelligent artificial intelligence. This AI's one and only goal is to make paperclips – as many paperclips as it possibly can. In order to achieve this goal, the AI may utilise all viable resources on Earth to make more paperclips. Many of the things that we as humans hold dear are merely fuel for the paperclip fire for this AI. The AI may even consume humans themselves as a form of energy to make paperclips, or determine that humans will try to stop the paperclip rampage and are therefore too much of a risk to it, and must be eliminated. Either way,

[96] Sometimes referred to as 'the value alignment problem', or the 'AI alignment problem'. I will use these interchangeably.

this AI runs the risk of destroying what humans value (human lives, the environment, etc.) in favour of something we value relatively little (a surplus of paperclips) (Bostrom, 2009, 2012).

As we see with Bostrom's paperclip thought experiment, future AI could be powerfully superintelligent and could pursue goals that we deem to be relatively unimportant.[97] In pursuing these goals, human values may not be factored in to the actions of AI. When a superintelligent AI does not share our values, they are an existential risk to humanity. We see human lives as highly valuable, but to a paperclip AI there is only paperclips. We are then left with two possibilities for limiting the threat from AI: we can try to control the *goals* of AI or we can try to control the *values* of AI (so that they align with ours).

How can we ensure that AI values align with ours? This is what Stuart Russell calls the *value alignment problem* (Russell, 2019, p. 126). For Russell, having AI's values aligned with ours will guard against the existential threat from superintelligent AI. We should therefore design autonomous AI to prioritise human values above all else.[98] Russell proposes three 'principles of beneficial machines' for researchers and developers to follow when creating AI systems:

1) The machine's only objective is to maximize the realization of human preferences.
2) The machine is initially uncertain about what those preferences are.
3) The ultimate source of information about human preferences is human behaviour.

(Russell, 2019, p. 156)

Russell's principles assert that not only should AI prioritise maximising human values, but that this should be the *only* objective of AI. Principles two and three are subsumed by the first fundamental principle. Principle two states that our machine should not be sure of human preferences. This should help to prevent the possibility that we pre-program our AI with values in mind, and these are

[97] This is Bostrom's 'orthogonality thesis' (see e.g., 2012).

[98] Autonomous here is used in the sense typically meant by computer scientists. There, autonomy has typically been thought of in terms of a machine being able to carry out a task unsupervised (e.g., Müller, 2012). It is assumed here that all high-level AI (Artificial General Intelligence, human-level AI, AI superintelligence) will be autonomous in this sense.

misinterpreted (a 'sorcerer's apprentice' scenario, see Christian, 2020).[99] Principle three specifies that the way the machine learns human preferences is through taking into account human behaviour. If we scream in horror, or rush to try to switch off the machine as it consumes our beloved pet as fuel for paperclip-making, our AI should pause, and re-evaluate; or better still, be uncertain about whether pets are an appropriate fuel source and err on the side of caution.

When articulated in these terms, the AI value alignment problem appears to be something that should be taken very seriously. However, there are still some key assumptions underlying this problem that may not be true and it is imprecisely formulated. Martin Peterson questions some of the presuppositions of the alignment problem (2019). Peterson points out that the alignment problem presupposes that aligning AI with human values is desirable, which is not necessarily the case (Peterson, 2019, p. 20). To account for the fact that we might not always want AI values to align with human values, Peterson presents several different variations of the value alignment problem (or 'value alignment thesis' in Peterson's terms). He ultimately endorses a moderate version of the value alignment thesis:

> *The moderate value alignment thesis*
>
> Autonomous systems should be designed in ways that are beneficial for humans […] and at each point in time *t* the best way to do this is to align the values of autonomous systems with some of the values (or interests or preferences) that humans actually embrace at *t*. (Peterson, 2019, p. 22)

This differs from both the strong and weak formulations of the value alignment thesis. According to Peterson:[100]

---

[99] "As machine-learning systems grow not just increasingly pervasive but increasingly powerful, we will find ourselves more and more often in the position of the 'sorcerer's apprentice': we conjure a force, autonomous but totally compliant, give it a set of instructions, then scramble like mad to stop it once we realize our instructions are imprecise or incomplete—lest we get, in some clever, horrible way, precisely what we asked for." (Christian 2020).

[100] Peterson considers another iteration of the value alignment thesis, before proposing his moderate formulation. The third iteration is: "the *epistemic* value alignment thesis, according to which we should accept the second but not the first part of the strong thesis. On this view, the values of autonomous systems should be

The weak value alignment thesis

> Autonomous systems should be designed in ways that are beneficial for humans. When human values (or interests or preferences) clash with other values, autonomous systems should give preference to human values (or interests or preferences).

The strong value alignment thesis

> Autonomous systems should be designed in ways that are beneficial for humans […] at each point in time *t* the best way to do this is to align the values of autonomous systems with the values (or interests or preferences) that humans actually embrace at *t*. (Peterson, 2019, p. 21)

In each case, human values, interests, or preferences should guide the behaviour of the autonomous system to some extent. In the case of the weak value alignment thesis, autonomous systems must only give priority to human values when values clash. In the case of the strong thesis, in order to ensure that autonomous systems are beneficial to humans, the systems should be designed to align with all contemporaneous values of humans.

Peterson questions the weak value alignment thesis on the grounds that it is anthropocentric. The weak value alignment thesis centres on human preferences in the case of clashing values. For example, this may give us cause for concern in a case where humans want to do something environmentally damaging as opposed to protecting the environment. Under the weak alignment thesis, an AI must support the humans, and damage the environment, because that aligns with their values (Peterson, 2019, p. 21).

Peterson also questions the strong value alignment on the basis that we cannot be secure in our own moral rectitude and that humans are, at times, irrational. In these instances – when we are

---

aligned with our values at time *t* no matter what those values are." (Peterson, 2019, p. 21) Under this version of the thesis, what is beneficial for humans should not be part of the equation. This epistemic variation is not defended (by Peterson or elsewhere), and results in the same AI design as the strong value alignment thesis. As such, I will not discuss it separately in this chapter.

morally wrong or we are acting irrationally – we do not want our AI to take its cues from us (Peterson, 2019, p. 22).

Peterson proposes the *moderate* value alignment thesis in place of the weak or strong thesis. He argues that the best way to ensure that these systems are beneficial to us is to align *some* of the values of autonomous systems with those of humans at each point in time. We might wonder then exactly *which* values should be aligned, and how we should encourage a system to choose, but these are practicalities that can be set aside for the moment. The key takeaway is that it is not *all* human values we want our AI to hold.

As I will go on to argue, the value alignment problem does not only apply to moral concerns: it is concerned with *any* human value, which would include aesthetic value. I argue that aesthetic value offers another case where we do not want AI to be fully aligned with human values. In the case of art, the best outcome for humans may not result from following human preferences. For an art-making AI, there are times when differing from human preferences may ultimately result in better art. Whilst part of my project here is to establish whether there is indeed an AI value alignment problem in the aesthetic domain, I think this argument will also lend support to the moderate version of the value alignment thesis proposed by Peterson. If in the aesthetic case, we do not want future autonomous AI to take its cues from human preferences, then we cannot agree with the strong value alignment thesis. If we do not want our AI to always choose human values in the case of a clash, we cannot agree with the weak value alignment thesis. If we do not want AI to align with human values at all in the aesthetic case, then we need to reject the alignment problem altogether. This is not the argument that I will make here. Instead, I will argue that the case of aesthetic value lends further support to Peterson's moderate value alignment thesis.

Before continuing, there are two key premises which must be established. First, the value alignment thesis is evidently concerned with moral value, so I will begin by establishing why the aesthetic domain is relevant to the value alignment problem. Second, I will establish that it is likely that autonomous AI systems in artistic or creative pursuits will not align with our values.

THE AESTHETIC VALUE ALIGNMENT PROBLEM

Does the AI value alignment problem, as introduced by Russell, apply in the aesthetic domain? I argue that it does. There are three reasons for this:

i. Contingency: There is no reason to think the value alignment problem does not extend to aesthetic value in Russell's framing of the problem – and there is no indication that any particular human value or preference is explicitly excluded from the problem.

ii. Separability: In the case of our autonomous systems, we may struggle to easily separate aesthetic and ethical values.

iii. Concrete examples: In discussions of the value alignment problem by key figures such as Bostrom and Yudkowsky, aesthetic issues feature as examples of the problem.

I will expand on each of these in turn.

i. Contingency:

The alignment problem, while discussed in the ethical domain, does not seem to be contingent upon any single human value. As stated in Russell's solution to the problem: "The machine's only objective is to maximize the realization of human preferences." (Russell, 2019). Russell does not specify only moral preferences, but preferences in general.

Peterson does frame the problem as ethical – however, when explaining the various versions of the value alignment thesis, he states that autonomous systems should (to some extent) give preference to 'human values (or interests or preferences)'. Whilst these could all be framed in the moral sense, this is not specified. There seems no good reason to think that these will only extend to the ethical domain, and not extend out into other realms of human values, interests, or preferences.

ii. Separability

It would likely be hard to determine how aesthetic value, or any other human value, could be separated from ethical values when the ultimate concern is existential risk from AI. We can think of

cases outside of AI where ethical issues may come into play in the aesthetic realm. An artwork might 'manifest immoral attitudes' (Gaut, 1998), might be made with ill-gotten materials, might be offensive, might be made by an immoral person etc. In aesthetics more broadly, we might gain positive aesthetic experiences that are in some way immoral. We might enjoy meat from maltreated and slaughtered animals, for example. How can we know where the ethical issue ends, and the aesthetic begins? The debate over whether it is possible to separate the ethical and aesthetic remains hotly contested (see Hanson, 2020).

I do not wish to commit to a particular position on the relationship between ethical and aesthetic value. Whether we accept moralism, autonomism, ethicism, immoralism (or any other view) about how aesthetic and ethical values interact in art, we are not thinking about AI art as solely an art object, which is free from any other considerations. However, the position we adopt about the interaction of aesthetic and ethical value may have *some* bearing on the applicability of the value alignment problem, insofar as we are discussing the aesthetic value of works of art (and not e.g., the materials they are made from). Take the radical autonomist position, that aesthetic value is entirely independent from aesthetic value (Carroll, 1996, p. 231). If we adopt this position, we may feel more able to separate aesthetic value from ethical considerations when designing our autonomous system: if the two values do not interact at all, we should be able to separate them for our AI. We might then be inclined to support the weak value alignment thesis over the moderate thesis; when aesthetic values come into conflict with human moral concerns, human preferences must be the deciding factor.

However, only radical autonomism, as Carroll calls it, would deny that moral considerations are *ever* relevant to artworks. This position is not typically defended in the value interaction debate. A moderate autonomist position, such as that defended by Anderson and Dean (1998) allows that in some cases "the moral content of a work can contribute to or detract from the aesthetic aspect of a work" (J C Anderson & Dean, 1998, p. 152). If the moral content can sometimes affect aesthetic value, we cannot easily separate the two values, and we have the same problem as with ethicism, moralism and immoralism. If we adopt anything but a radical autonomist position about the value interaction debate, then ethical and aesthetic values may interact. If these values can interact, then

clear separation of the two will not be possible, even when limiting our discussion to the aesthetic value of artworks. We will not be able to just section off the aesthetic from the ethical – and this must be something superintelligent AI can do if we expect it to know when to defer to humans or not.

The value interaction debate remains unsettled in aesthetics. Other than the radical autonomist position, all other positions about the value interaction debate allow that in some cases ethical and aesthetic value interact. As such, unless we adopt radical autonomism it will not be possible to draw a line between aesthetic and ethical values.[101] I will suggest that for the purposes of our autonomous systems, all human values should be under consideration until we might know how to clearly separate them.

iii. Concrete examples

While the value alignment problem is focussed on the ethical domain, discussion of the problem often crosses into aesthetic considerations. In illustrating the value alignment problem, there are several examples where aesthetics features. For example, in his celebrated 2012 book, Bostrom quotes the following from Yudkowsky:

> Back in the era of pulp science fiction, magazine covers occasionally depicted a sentient monstrous alien—colloquially known as a bug-eyed monster (BEM)—carrying off an attractive human female in a torn dress. It would seem the artist believed that a non-humanoid alien, with a wholly different evolutionary history, would sexually desire human females [...] Probably the artist did not ask whether a giant bug perceives human females as attractive. Rather, a human female in a torn dress is sexy—inherently so, as an intrinsic property. They who made this mistake did not think about the insectoid's mind: they focused on the woman's torn dress. If the dress were not torn, the woman would be less sexy; the BEM does not enter into it.
> (Yudkowsky, 2008, p. 310, quoted in Bostrom, 2012)

[101] This is not to mention the practical issue of separating aesthetic and ethical values in artworks, even if we do believe it is possible theoretically.

Here, the point is to illustrate how we might assume shared motivation, and shared values, with a being radically different to us. But we should do no such thing, given the vast differences in the constitutions of humans in comparison to aliens, monsters, or AI. The example used to illustrate this is an 'attractive human female' in a torn dress. Using words like attractive and sexy means that we have stepped into the realm of the aesthetic.

In a later work, Yudkowsky again draws the alignment discussion to the aesthetic:

> Getting a goal system 90% right does not give you 90% of the value, any more than correctly dialing 9 out of 10 digits of my phone number will connect you to somebody who's 90% similar to Eliezer Yudkowsky. There are multiple dimensions for which eliminating that dimension of value would eliminate almost all value from the future. For example an alien species which shared almost all of human value except that their parameter setting for "boredom" was much lower, might devote most of their computational power to replaying a single peak, optimal experience over and over again with slightly different pixel colors (or the equivalent thereof). (Yudkowsky, 2013)

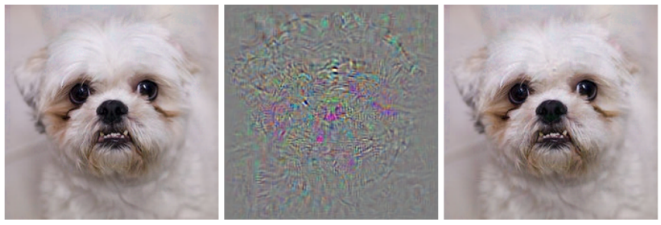
This discussion of some 'optimal experience' which may differ in pixel colours leads me to assume this experience is a visual one and evidently a pleasurable one. Yudkowsky clearly means this to be an aesthetic experience.

While scholars like Peterson might presuppose that the AI alignment problem bears only upon ethical value, I see no reason for aesthetic value (or any other human value) to not also be subject to the value alignment problem. However, while it may be the case that discussions of value alignment seem to include all human values, this is not enough to assert that there is actually an aesthetic value alignment problem. To know this, we need to establish if a difference in values a) is possible, or even likely, and b) will actually cause a problem. These are the next two points I will turn to.

THE POSSIBILITY OF MISALIGNED AESTHETIC VALUES

The aesthetic value alignment problem might be dismissed on the basis that an actual mismatch in aesthetic values between humans and AI is unlikely. So, I must establish that misaligned values are indeed likely to occur between AI artists and humans. Bostrom argues convincingly that there is no reason to think artificial (super) intelligences will share values with us, as discussed above. This argument (along with the bug-eyed monster example) can be taken as evidence for why aesthetic value does not escape this possibility of misalignment. But I think that we can go further than this: I propose that it is *likely* that aesthetic value will not be aligned between humans and AI (at least if current technology is any indication).[102]

To demonstrate this, we can consider some image-recognition algorithms which aim to replicate the human ability to recognise depicted objects. By examining these current image-based AI systems we can see that they privilege different kinds of visual information to humans. First, consider a study by Szegedy et al. (2014). This study demonstrates that AI image recognition systems can be easily fooled. Researchers showed that an image recognition AI could be led to mislabel images by introducing small perturbations to the images that are imperceptible to the human eye (see fig. 20). Visually, it seems AI do not weight the same factors that are important to humans. To the human eye,



(a) Dog (b) Perturbation (c) Ostrich

Fig. 21 Image from Szegedy et al. (2014) showing the misclassification of an image by a deep neural network after an imperceptible perturbation (b) on an image of a dog (a). After perturbation (c) the image is classified as an ostrich by the system.

[102] I do not want to preclude the possibility that future AI, especially human-level AI, will end up processing visual data in a manner much closer to humans – but, I think it is not impossible that trends such as texture-focus as opposed to shape-focus are pervasive themes in AI vision. Much as other animals rely more heavily on other senses than humans, it is not impossible that these systems also rely on cues that we humans find less important.

these two images are indiscernible. Yet to the AI, these are images of two completely different objects.

This study aims to fool a simple AI system, so perhaps it is not the best example. Presumably a superintelligent AI will not be so easily tricked. We can turn to Geirhos et al. (2018) for another example. In this study, researchers explored what visual cues were most salient in an AI's labelling of an image. They found that *texture* was more important than *shape* in image identification for an AI. For example, an image with a clear outline of a cat, over-laid on the texture of elephant-skin, would be labelled as an elephant (see fig. 21). We humans, however, would label this as an image of a cat. This demonstrates the divergence of human visual cues and AI visual cues.

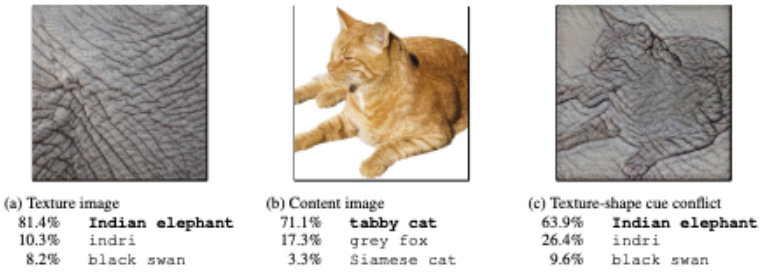


Fig. 22 Image from Geirhos et al. (2018). "Classification of a standard ResNet-50 of (a) a texture image (elephant skin: only texture cues); (b) a normal image of a cat (with both shape and texture cues), and (c) an image with a texture-shape cue conflict, generated by style transfer between the first two images." (Geirhos et al., 2018, p. 1).

If an AI does not weight visual elements with the same importance as humans, there is no reason to think that they will do so in the case of art, or more generally, in determining what is aesthetically pleasing. The beauty of a painting might rely on texture instead of shape, of pixel-level depiction rather than whole-image depiction. Furthermore, as Walton (1970) has demonstrated, aesthetic evaluation of art depends heavily on categorisation, suggesting that this difference could vastly affect evaluation of art objects. Here we might get to the difference of experience suggested by Yudkowsky (2013), where a low-boredom AI will enjoy "…replaying a single peak, optimal experience over and over again with slightly different pixel colors (or the equivalent thereof).". When looking at the images in Szegedy et al.'s study, this seems all the more likely and will result in divergent values as Yudkowsky suggests. Further to this, as I have argued elsewhere in the thesis (see chapter 5 on weirdness), AI may produce art images with a distinctive aesthetic. In the case of

‘weirdness’, I will argue that this is a result of a kind of failure, but we can surely imagine an AI, with more broadly defined goals, which produces such images because weirdness is precisely what it finds valuable.

I have established then that a difference in aesthetic values between AI and humans is wholly possible, and if current technology is any indication, may even be likely. But, unlike our murderous paperclip maximiser, it unlikely that there will be a great existential threat from this. So, would aesthetic value misalignment be a problem?

THE PROBLEM WITH MISALIGNED AESTHETIC VALUES

There are two parts to the consideration here of whether AI value alignment will be a problem in the aesthetic realm. The first is about creativity. This is an aesthetic consideration, though it may reach beyond the arts to creativity in other domains, such as science. Creativity is an area where, I will argue, we *do* want our AI to align with us. The second consideration will be aesthetic value itself. In the case of aesthetic value, I will argue that there is not so clearly a problem. In fact, misaligned human-AI values might be beneficial.

Creativity and value

We would expect a human-level or superintelligent AI to be creative. Some consider creativity to be an aesthetic value, however, this is far from a settled matter (Gaut, 2010, p. 1039). Despite this, most philosophical discussions of creativity include value as a constitutive element. To be considered as creative, one must produce something of value. As Gaut puts it,

> The value condition is generally deemed necessary to rule out cases of worthless originality as being creative: as Kant argues, ‘since there can also be original nonsense, its [the genius’] products must at the same time be models, i.e., exemplary’(Kant, 308). (Gaut, 2010, p. 1039)

Value is not just something that is produced by the creative process – it is a constitutive part of that process and is necessary to define it as creative. We can consider creativity then in relation to the value alignment problem.

In the case of creativity, we will generally want our AI to align with our values. There are two reasons for this: i) an epistemological reason and ii) an ethical reason.

i. The epistemological reason

If value is a constitutive part of creativity, then it is key to assessing whether a process, person, or output is creative. If an AI, in its creative act, is aiming at a value which humans do not hold, or maybe cannot even recognise, then we humans might not even recognise that the AI is being creative. It is not impossible to imagine such a case, where we cannot recognise a value others hold. We know that, for example, some human values are not recognised cross-culturally. One such case would be honour and dishonour culture, and the particular nuances of what kinds of behaviours contribute to honour and dishonour. Some cultures place great importance on honour, whereas others do not, and these differences can be seen in the same country (Casimir & Jung, 2009).[103] Some values which are central to many honour cultures, such as chastity, are barely valued in other cultures (Casimir & Jung, 2009). In the case of honour-based violence, individuals in a US culture that place a higher value on honour were more likely to justify violence when said violence protects or maintains honour. This was significantly different from individuals in a US culture that did not place this same value on honour, who were more likely to say that violence was never justified in the same situation (D Cohen & Nisbett, 1994, p. 559). Justifying violence to 'protect one's honour' may seem completely alien to those of us from cultures where honour is not valued, and thus we may not be able to comprehend how others might see violence in protection of honour as commendable.[104] We have a case here then,

[103] For example, the Southern United States has what would be considered an honour-based culture, whereas other regions in the US do not. See Cohen and Nisbett (1994).
[104] I am not aiming to imply there is an objective answer here – I am merely aiming to demonstrate what we can miss in understanding actions when we do not share values with others.

of an unrecognised value. In the case of our creative AI, if we cannot recognise the value in what the AI creates, we may not see it as creative at all.

To further support this possibility that we will be unable to recognise creativity where we do not agree upon value, consider Boden's account of reactions to what she terms 'transformational creativity':

> the surprise that we feel on encountering a creative idea often springs not merely from an unfamiliar combination, but from our recognition that the novel idea simply could not have arisen from the generative rules (implicit or explicit) which we have in mind. With respect to the usual mental processing in the relevant domain (chemistry, poetry, music...), it is not just improbable, but impossible. (Boden, 2004, p. 52)

Boden argues that this kind of transformational creativity will never be achieved by AI, and she may be right. But, if an AI is ever able to reach transformational levels of creativity, we will not be able to envision it ahead of time. And, if we cannot see the value in the creation, we may not recognise it when it occurs.

ii. The ethical reason

There is a risk to not knowing when an AI has developed creativity. This takes us back to the ethical concerns. If we do not know when our AI systems have developed creativity, we may not be aware when they turn their new creative powers to less-than-desirable aims, i.e., 'dark' creativity. As Cropley et al. state in the summary of their book on the topic:

> With few exceptions, scholarship on creativity has focused on its positive aspects while largely ignoring its dark side. This includes not only creativity deliberately aimed at hurting others, such as crime or terrorism, or at gaining unfair advantages, but also the accidental negative side effects of well-intentioned acts. (Cropley et al., 2010)

These concerns seem to be particularly relevant in the case of AI. If we cannot recognise creativity in AI, we may miss its growing capacity to harm us, or fail to foresee accidental side effects of some

creative act. Certainly, we want to be aware of when our AI are achieving creativity and misalignment of values is a potential barrier to this. In the case of creativity then, we do broadly want AI values to align with human values.[105] It seems that considering AI creativity may also lead us to endorse the alignment problem. However, this is not necessarily the case as soon as we turn to a narrower realm of artistic concern – aesthetic value.

Aesthetic value

An AI may produce art works which differ from those which we would find valuable, and that these values may be subject to the value alignment thesis. In the case of creativity, it appears that AI artistic value might well cause a value alignment problem. Creativity is, however, far broader than art. Should we be similarly concerned when we focus in on aesthetic value?[106]

We cannot always predict what we will value aesthetically. As I have already shown, AI might surprise us with something we did not expect to appreciate like weirdness. This is not too much of an issue for value alignment. In cases like weirdness, whilst the AI did not know in advance what our values were (because we did not know ourselves), we swiftly demonstrate fascination with this aspect of AI works. We will behave as if we like it – not as if we don't. The AI has plenty of behavioural cues which will help it recognise our (admittedly newfound) interest in, preference for, or valuing of this weird aspect of AI works. Here, any misalignment that might have existed is quickly resolved.

However, history suggests that where art is concerned, humans are not constant with what they value. Aestheticians and art historians can likely think of many renowned cases of artworks that have been poorly received initially but gain recognition with the passage of time. Two such examples would be Van Gogh's posthumous success and the early response to Cubism. In the case of Van Gogh, it is widely known that he had little success in his lifetime yet has since been considered to be a

[105] There is an added complication here of course that it is not easy to define a set of 'human values' (for example, there may be cross-cultural differences, as discussed above).
[106] I will not commit to a particular perspective of what aesthetic value is here - I will use the term here to broadly refer to those values associated with art.

great master (e.g., Tromp, 2010, p. 108). *Les Demoiselles D'Avignon*, an early Cubist work by Picasso, was initially very poorly received, and was reportedly "greeted with bafflement, suspicion, and outright hostility" (Unger, 2019, p. 342) including from Picasso's own friends (Matisse seemed to think the painting was a hoax aimed at him – see Unger, 2019, p. 442). Yet, this painting, and the Cubist movement now occupy a revered position within the Western canon. I doubt that these two examples are mere anomalies. In the case of art, initial responses are not certain indicators of later greatness, or lack thereof. Hume identified the 'test of time' (Levinson, 2002), arguing that truly great art should endure, even if initial reactions were marred by negativity:

> We shall be able to ascertain its influence not so much from the operation of each particular beauty, as from durable admiration, which attends those works, that have survived all the caprices of mode and fashion, all the mistakes of ignorance and envy." (Hume, 1882, p. 271)

In the aesthetic realm, when we rely on human values, interests, and preferences at any given time, we may find that we do not have an accurate indication of ultimate (or enduring) value. Like Picasso's influential piece, many works of art may have influence and impact without initial accolade. A work of art may initially be reviled by the public or critics but in years' time be applauded. Some works may never receive unanimous positive responses (despite widely recognised value).

This causes a problem for AI value alignment. Relying on 'human preferences' to guide AI, as suggested by Russell, will not serve us particularly well in aesthetics. If we rely on human preferences about artworks, we are left with many questions: in which human's preferences are we interested? Those of critics, or those of the public? And is this limited to initial response? Or the response after 5 years? 50 years? We are likely relying on an immediate public response to an artwork, which is not necessarily the most reliable indicator of great art. Our AI will, under these constraints, be unlikely to offer anything truly influential in the artistic realm. Russell also specifies that the indication of values, interests, and preferences will come from our behaviours. What behaviours will an AI be on the lookout for? Will it deduce value from the works we buy? Works we smile at? Works which we say favourable words about? How will it determine what a long stare at a work means? Or a heated debate with a friend about a work? And, are any of these really an

indication of the ultimate aesthetic, or artistic, value of a work of art? Human preferences are unreliable, and initial behaviours in response to art will not always be a good indicator of the value of a work.[107]

So, what will happen if, like Van Gogh or Picasso, our superintelligent AI artist produces a work 'ahead of its time'? The work may not align with current aesthetic values, and may be initially poorly received. It could be trashed by critics and the public. Perhaps no one will buy it. If our AI is designed with Russell's principles in mind, it will take these cues about human values, interests, and preferences, and it will class the work as a failure, and go back to the drawing board in its effort to please us. But, what if, like Van Gogh and Picasso, the AI artwork did ultimately have something great to offer us? Unlike the tenacious human artist, an AI designed with Russell's principles in mind will not continue to produce works of a similar ilk. It will instead shift to a more palatable product. So, unlike Van Gogh, say, whose whole oeuvre was waiting to be recognised, our AI, driven to please us, will abandon its idiosyncratic, and *potentially* valuable works.

For those of us in the arts, this surely seems a shame. One of the exciting prospects of a superintelligent AI artist is precisely the difference in aesthetic value it might reveal to us. Like weirdness, AI art may be the source of some new aesthetic or artistic value. If our art-making AI is constrained by Russell's principles, forced to align with human values, its art will potentially be stunted, and the possibility of our discovering something new through AI art will be too.

In conclusion, I hope I have shown clearly why in the aesthetic realm we do not have such a clear-cut value alignment problem. Sometimes, we may want our AI to diverge from human values, interests, and preferences. I will now turn to considering what the impact of this conclusion is on current discussion of the AI value alignment problem itself. As the value alignment problem seems to include aesthetic values, and in the aesthetic case we do not always want our AI to align with human

[107] Recall Peterson's issues with the strong value alignment these – that human moral judgment can be wrong, and that humans are irrational.

values, this means that we cannot endorse a version of the value alignment thesis that insists on deferring to human values, interests, and preferences at all times.

AESTHETICS AND THE MODERATE VALUE ALIGNMENT THESIS

In the aesthetic case, we may find that a beneficial outcome for humans may not result from following human preferences – or at least that, at times, differing from human preferences may ultimately be a better course of action if we want AI to produce the most valuable artworks. As I stated in section 1, this argument lends support to the moderate version of the value alignment thesis proposed by Peterson (2019):

> *The moderate value alignment thesis*
>
> Autonomous systems should be designed in ways that are beneficial for humans […] and at each point in time *t* the best way to do this is to align the values of autonomous systems with some of the values (or interests or preferences) that humans actually embrace at *t*. (Peterson, 2019, p. 22)

The strong value alignment thesis specifies that the best way to ensure autonomous systems are beneficial for humans is to align them with the values, interests, or preferences that humans are embracing at that time. This would be very limiting for our superintelligent art-making AI; it is going to be shackled by existing human artistic and aesthetic preferences and will not be able to shift from these once it is certain what they are. This will not be harmful to humans per se, but if we agree that expanding our aesthetic horizons is a good thing, it means that we are losing out on potential benefits. The aesthetic value case is in conflict with the strong value alignment thesis.

The weak value alignment thesis specifies that in the case of a clash of values, interests, or preferences between the autonomous system and humans, preference should be given to human values. In the aesthetic case, this would mean that when an AI develops an aesthetic or artistic value (or preference etc.) which clashes with a human value, the human value should win out. But, as

argued above, this would deprive us of the potential to recognise a new aesthetic value. So, at least in the aesthetic case, we do not want our AI systems to always prioritise human values, interests, and preferences, even in the case of clashes. The aesthetic value case is also in conflict with the weak value alignment thesis.

By looking at creativity, I have shown that we do not want our AI system to never align with our values either, even in the aesthetic domain. This leaves us with Peterson's moderate value alignment thesis. This thesis specifies that autonomous systems should align with '*some* of the values (or interests or preferences) that humans actually embrace at *t*'. This formulation, unlike the weak and strong theses, allows our AI to take some cues from human values, but not all.

This still leaves us with the practical issue of separating which human values (or interests or preferences) our autonomous system should or should not follow. For this, I can propose only an initial attempt at a solution. I will refer back to Russell's guidelines that AI should take cues from human behaviour. I have argued above that when it comes to art, behavioural cues may indicate a lack of value in ultimately great works of art (such as those of Van Gogh and Picasso). An AI following these cues will abandon potentially great works. I am not then in favour of indiscriminate adherence to human behaviour as the ultimate guide for art-making AI. I wonder, however, if we could still use behaviour as our source of information for AI. Perhaps we might set limits on the basis of the behaviours we are taking cues from. Not buying works of art, or giving bad reviews, or frowning etc. might not be enough evidence that the AI should stop creating works of a certain kind. A stronger response, such as trying to switch the AI off, might be a better indicator that the AI should change course. This would allow us to more clearly separate the domains of art and ethics too – if you kill a person, others will try to stop you. If you make what humans perceive as bad (or ugly, unpleasing, valueless etc.) art, they will likely not take such drastic action.

Conclusion: The Aesthetic Value Alignment Problem

In this chapter I have explored the possibility of an aesthetic value alignment problem in AI. After describing the AI value alignment problem and its various formulations, I turned to aesthetics. I have shown that aesthetic value appears to be included in current discussions of the value alignment problem, and that we may indeed see misalignment between human and AI values in the aesthetic domain. I argued that in the case of creativity generally, we want AI to align with human values. However, in aesthetics we may not always want alignment, because insisting that AI should always align with human values may prevent the possibility of new aesthetic values stemming from AI works. I went on to argue that the aesthetic case lends support to Peterson's moderate version of the value alignment thesis, over both the strong and weak versions of the value alignment thesis.

Chapter 5 - The Aesthetics of AI art

In this final chapter, I will turn to examine the aesthetics of AI images. Given my argument in Chapter 4 that AI may not align with human values in the domain of aesthetics and the arts, we might wonder, is there anything distinctive in the aesthetics of AI artworks?

I will examine two cases of aesthetic qualities which seem to be common to AI works. First, I will consider the case of weirdness in AI art. AI images are frequently described as 'weird', 'strange', and 'surreal'. What is this weird quality? I investigate the possibility that this weirdness can be explained by incongruity and go on to argue that incongruity is insufficient to explain the weirdness we specifically see in AI images. I suggest that weirdness can be attributed to a kind of failure of the AI.

The second case I will consider is what I call 'convincingness'. This relates to the *success* of AI images, which seem to convincingly reproduce the perceptual features of works made via traditional means. Here, I investigate the cause of one kind of response to discovering a 'convincing' image is made by AI: that one might feel cheated in some way. I investigate what might cause this response and argue that it can be explained by Walton's concept of a 'jolt'. Through these two sections, I aim to offer an initial step towards an aesthetics of AI art.

Figs. 23-25 A selection of images produced using AI. **Fig. 23** Image produced using Artbreeder. **Fig. 24** One of Robbie Barrat's AI nudes. **Fig. 24.** An image produced by *Thishorsedoesnotexist.*

INTRODUCTION

Artificial Intelligence (AI) technology has become immensely sophisticated. Recent years have seen exponential growth in the capabilities of AI systems across many fields. Art is no exception; AI are now able to produce a wide range of convincing art-like creations: photographic images, paintings, and drawings; sections of music; poetry; plays and sections of novels; and, last but not least, text-based video games.[108] As these AI works enter into the artworld, we ought to ask, as aestheticians, is there a distinctive aesthetic to AI artworks?

Taking a cursory look at the visual works produced by AI, we are likely to see abominable bodies, severely contorted faces, and all manner of ontological violations. In other words, AI art can be remarkably *weird* (see figs. 22-24 above). But what exactly *is* this weirdness? This section conceptualises weirdness in relation to AI art, and asserts that weirdness is of aesthetic interest. I argue that the reactions to weirdness in AI works suggest that it is a form of norm violation. Noël Carroll has argued that norm violation is a key part of works of comedy and horror. I propose that the concept of norm violation can be fruitfully applied to works produced by AI. I go on to argue that norm violation alone is insufficient to explain AI weirdness. When we describe AI outputs as weird, we are often pointing to a specific *failure* of the AI to seem *human.* I will explore what constitutes failure for AI,

[108] Here I focus exclusively on visual art, however, some of this discussion could be extended to other artforms.

considering some examples from image-producing AI. I will then examine reasons why this non-human failure manifests, taking a closer look at the AI systems that produce these images. Through this study of weirdness in AI art, I aim to make the first step towards establishing an aesthetics of AI Art.

WEIRDNESS AND NORM VIOLATION

The concept of 'weird' comes up frequently in discussions of AI art. Practitioners, viewers, and mainstream media outlets alike pick up on a quality of weirdness in AI artworks.[109] But little has been said about how 'weirdness' is actually manifest in these works. I suggest that there are at least two ways in which this 'weird' quality manifests in AI works. Figs. 22-24 show prototypical examples of AI-generated art. These examples demonstrate that the weirdness of AI-produced art is often humorous or unsettling or a bit of both. Humour typically arises when we can guess what the AI is trying to reproduce and identify the particular ways in which it has not succeeded. Take fig. 24 above, which is an AI's attempt to produce an image of a horse. The image evidently captures something of what a horse looks like, but the AI has comically overestimated the number of legs that a horse typically possesses and underestimated how long horse legs typically are. Fig. 24 may be humorous enough, but there are also many examples of AI art that are more unsettling, even disturbing. This sort of quality can be seen in the distorted fleshy figure in fig. 22, and in the work of Robbie Barrat, an artist and computer programmer who trained an AI to reproduce images of nude paintings (figs. 23, 25-27). These sorts of AI art pieces are often called 'surreal' (Rea, 2019; C Schneider, 2015) 'uncanny' and 'unreal' (Rea, 2019) 'melting' and 'distorted' (Vincent, 2018). Barrat himself explicitly notes this facet of the images he produced with his AI programme, stating 'The way that it paints faces makes me uncomfortable … Personally, I really love these super weird unrealistic ones.' (Burton, 2018). Those who discuss AI art are often not particularly interested in the works which succeed in replicating images that are plausibly

[109] For a representative example, see Gavin (2019).

part of the training set. Instead, much of the fascination and aesthetic interest in these objects is focused on how they show something unusual, funny or unsettling. The focus is on the *weird.*

Figs. 26-28. Robbie Barrat's AI nudes

What exactly is the nature of this *weirdness*? I have suggested that weirdness often elicits both humour and mild horror. Philosophers of art will readily recognise that this interplay between humour and horror has been conceptualised in a seminal work by Nöel Carroll. Carroll proposes that both horror and humour are linked by their dependence upon 'norm violations'. These norms could be something that typifies an art genre, or could be any 'transgression of a category, a concept, a norm, or a commonplace expectation' (Carroll, 1999, p. 154). I propose that Carroll's account of norm violation can be applied to the case of weird AI art. Allow me to demonstrate. With regards to horror, Carroll states:

> …a necessary condition for being horrified is that the emotional state in question be directed at an entity perceived to be impure – where impurity, in turn, is to be understood in terms of violations of our standing categories, concepts, norms and commonplace expectations.[110] (Carroll, 1999, p. 154)

The sort of horror Carroll describes is apparent in a number of AI artworks, for instance, in Barrat's faceless nudes (figs. 23, 25-27). When Barrat posted a selection of his nudes onto Twitter, commenters

[110] Carroll's account may not be sufficient to explain all horror, however it is enough here that it explains some examples.

drew comparisons between Barrat's work and the body horror of John Carpenter's films, as well as other popular horror media such as *Jacob's Ladder* (1990) and the *Silent Hill* franchise. The works are described as 'terrifying' (figs. 28-30).

Figs. 29-31 Tweets in response to Robbie Barrat's AI nudes

As for humour, we have already seen one such example in the many legged 'horse' (fig. 24). Indeed, the humour of AI outputs is well-established on the internet. Janelle Shane, an AI researcher, has built a popular blog (titled *AI Weirdness*) and published a book on the humorous weirdness of AI (Shane, 2020). Humour is linked to norm violation in Carroll's account of the incongruity theory of comedy. In talking about 'comic situations', Carroll states:

> Though the relevant incongruity in comic situations may involve transgressions in logic, incongruity may also be secured by means of merely inappropriate transgressions of norms or commonplace expectations, or through the exploration of the outer limits of our concepts, norms, and commonplace expectations. (Carroll, 1999, p. 154)

If we accept that finding something funny or unsettling in artworks is indicative of some manner of norm violation, then it would seem that these norm violations are occurring in weird AI artworks. It might seem that norm violation alone can account for AI weirdness. However, I am not convinced; there seems to be something *distinctively* weird about AI art. As Carroll shows, norm violation is a feature of some human-made artworks, so we are left needing to explain the difference between human norm violations and AI norm violations.

To understand the difference between human and AI norm violations, let us compare weird AI visual art with a norm violating human-made artwork. The image in the Dorothea Tanning painting below (fig. 31) is certainly violating some norms: the norms of gravity are reversed on the central figure's hair, and the giant sunflower in the painting violates the norm of scale. However, these violated norms are quite different in comparison to the AI nudes above. We may notice in weird AI artworks that there is a sense that something more fundamental has been violated. I propose that this difference stems from how AI violate some of the underlying assumptions that are key to human visual understanding. The AI has failed to achieve what it set out to, namely, to reproduce human artforms.

Fig. 32 Dorothea Tanning, *Eine Kleine Nachtmusic***,** 1943

The added dimension, then, in addition to norm violation, is a distinct failure to produce a work that seems as if it is human made; it appears to us as *non-human*. If what appeals to us in AI works is the weirdness of what the AI produces, then when an AI succeeds in producing believable artworks,

they cease to be of aesthetic interest to us. A recent tweet by AI researcher Michael Trazzi about a successful text-based AI exemplifies this well: '…I tried super hard to find a joke, but all sentences sound boringly human' (Trazzi, 2020). If the weirdness in AI works arises from failure to produce something believably human (which, in turn, produces a norm-violating object), then we need a better grasp of the exact way in which it fails. This is where we will turn next.

NON-HUMAN FAILURE

In order to understand the non-human component of weirdness in AI artworks, we can compare human failure with AI failure. Let us look at some examples. Figs. 32 and 33 show us two different 'bad' depictions of cats. Indeed, bad cat art has become something of an internet phenomenon. These images have been 'meme-ified' for the amusement of many, and certainly we would call these renditions of cats weird and funny.[111] What exactly makes them weird and funny?

Figs. 33-34 "Badly drawn cats" which became popular online memes.

It is clear that these images of cats are violating norms, as discussed above. They just do not look *right*; the proportions of their bodies are incorrect, the shapes of their ears and limbs are odd, and their faces are misshapen and distinctly not cat-like. However, they still meet some basic requirements to be considered depictions of cats. They have two pointed ears, four legs, a tail, a face with two eyes,

[111] See Bern (no date) for a summary.

a nose and mouth, and fur with patterning. While poor in terms of accurately representing a cat's shape and features, those features are all there.

Now let us compare these to failures in AI images, for example, images from *Thiscatdoesnotexist*, a website which produces AI-generated, 'photographic' images of cats (Wang, no date, a). This site uses a Generative Adversarial Network (GAN) that has been trained on photographs of cats. The GAN system consists of two key elements, a generator and a discriminator. The discriminator receives training on a set of images (in this case, photographs of cats). The generator produces images, initially at random. The discriminator provides feedback to the generator, scoring each image according to how closely it matches images from the training set. This feedback allows the generator to weight the likelihood of different features occurring in images in the future. Gradually, the generator produces images that fool the discriminator – these images are indistinguishable (to the discriminator) from images in the training set.[112] The GAN succeeds often, as we can see from figs. 13-15. These images are clearly depicting cats, and we would be surprised to learn that they are not photographs, but completely artificial, non-referential images.



Figs. 35-37 Successful results from *Thiscatdoesnotexist*

Although the GAN normally succeeds in producing convincing 'photographs' of cats, it often fails (figs. 37-39). When the AI fails to produce a cat, it can fail quite spectacularly. These cats are globs of fur, with free-floating or disconnected 'limbs'. Sometimes it is not clear that there is a distinguishable body, head, legs, tail, eyes, ears, or really any features of a cat. There is clearly the

[112] For further information on the GAN behind *Thiscatdoesnotexist*, see Karras et al. (2019).

texture of fur, and cat-like colouring (though the fur texture does not always stay within the bounds of a body). The images do seem cat-like in some way – the blobs look like cats lying down, or curled up. They are cat-like, but somewhat featureless and boundary-less.

Figs. 38-40 Unsuccessful results from *Thiscatdoesnotexist*

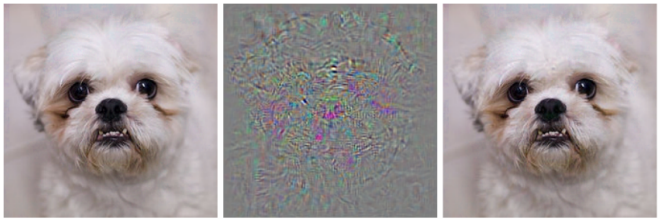
Finally, compare these images to a picture of a cat drawn by a child (figs. 40-42). Children succeed in depicting a cat where the AI fails. There are typically all or some of the following features: four legs/feet, a tail, a head with two pointed ears, and a face with two eyes, a nose, a mouth and whiskers. The colours are typically cat-like, however, this is not always the case (fig. 42), and this does not necessarily prevent us from seeing the cat depicted. The orange and blue cat in fig. 42 is still clearly a cat.

Figs. 41-43 The author's childhood drawings of cats.

The key difference appears to be the reliance (or lack thereof) on the building blocks that we think constitute 'cat'. These building blocks, or schemas, are key to human picture making (Gombrich, 1994), and are even thought to be essential to all human vision (Marr, 1983; Biederman, 1987). Humans

rely on schemas in their understanding and subsequent depiction of objects like the bodies of humans, or indeed cats, and we can see this even in the drawings of children. However, GANs do not seem to rely on schemas *at all.* Why not? When a GAN is trained on cat photos, it does not learn a schema in the way a human would. Instead, it learns what a successful cat photograph comprises (according to the AI), and it replicates that. We see this in the failed GAN images; the colours and textures of the cat are still there. We can see that the background of the image is always important too, which typically looks like cushions or carpet, even in the failed images. What we do not see though, is all the schematic features of the cat – distinct body parts that we consider essential to cat depictions: what is lacking, is the typical *human* conception of cat. The building blocks that comprise the standard body schema for 'cat' (which humans recognise even as young children) are absent. Unlike the bad cat drawings above, the bad cat GAN images do not include what a human would see as essential to cat depiction. The AI does not know what the salient features of the images are when it begins training. As it learns from the images, all aspects are up for contention as key elements of the image.

The failure of some image-making AI to identify salient features in the *depiction* of objects is also apparent in how some image-classifying AI fail to *recognise* objects. There are several studies of AI that demonstrate this. These image-classifying AI are similar to GANs in that they utilise deep learning algorithms. For example, Szegedy et al. (2014) found that an AI could be led to misclassify images by adding imperceptible (to humans) changes to the images (fig. 43). Similarly, another study found that an AI can be tricked into, for instance, mislabelling an image of a cat as an image of an elephant if elephant skin texture is merged with an image of a cat. Humans will correctly label the image as an image of a cat, whereas the AI, privileging texture cues over shape cues, will label the image as an image of an elephant (Geirhos et al., 2018). We see this similar attention to texture over shape in the failures of *Thiscatdoesnotexist*. Again, these AI systems are not relying on the same building blocks as humans to recognise salient features of an image.



(a) Dog (b) Perturbation (c) Ostrich

Fig. 44 Image from Szegedy et al. (2014) again showing the misclassification of an image by a deep neural network after an imperceptible perturbation (b) on an image of a dog (a). After perturbation (c) the image is classified as an ostrich by the system.

When we examine the inner processes of GANs, there is no capacity for schema building of any kind. The distinction between badly drawn cats and the GAN images on *Thiscatdoesnotexist* could be said to be a difference between intending to draw a cat and failing, versus arguably having no intention to depict a 'cat', merely the goal of reproducing cat images. The result of this is an indiscriminate approach to the images that the AI is trained upon, where the AI does not focus on the cat in the image, but on the image as a whole. In producing convincing 'cat' images, then, the AI reproduces features of these images such as fur and carpet and fails to reproduce features that fit into our (human) schema of what a 'cat' looks like. The GAN simply has no capacity for such schema.[113] This explains why we see fur, and cat-like colours and shapes in each of the failed GAN images; these are in every training image. Ultimately, the AI does not have the goal of depicting a cat, instead it has the goal of reproducing images that are similar to its training images. In the failure of GANs, we can see the source of the weirdness of AI images – they are not building images like a human would. GANs are not merely violating norms or expectations of cat depictions, but also failing to do so in a *human* way. This is the core of AI weirdness that I am proposing: *the weirdness is a product of norm violation through non-human failure.*

Although the cat images produced by GANs on *Thiscatdoesnotexist* are not aiming to be art, they are using the same system used by AI artists. Let us revisit the AI nudes discussed above (figs. 23, 25-27). These images were also produced by a GAN, trained on nude paintings. Often, Barrat's AI nudes are described as 'Bacon-esque' i.e., the distorted figures depicted in his images resemble those

[113] This is not to say an AI could not be designed with schemas in mind, however this is not how GANs are designed.

of the artist Francis Bacon (e.g., Beschizza, 2018). The comparison is superficial, however. Although Bacon depicts highly distorted faces with twisted features – to the extent that they can even be described as 'mutilated' (Zeki & Ishizu, 2013) – the features are all still *there.* In cases where facial and bodily features are missing, there are still salient features present in the image; in comparison, in Barrat's AI images, distinct features are completely lacking. In several ways, they are similar to the failed AI-generated cat images: there are fleshy blobs, with paint-like textures but there are no discernible limbs, no facial features, and their flesh melts into the surrounding areas. This is quite different to Bacon's warped figures.

One might object that a human *could* create images like those produced by Barrat and his AI. To do so, they would have to abandon the use of schema, which according to Gombrich is essential to all picture making. Though this would fundamentally alter the process of representational image-making for a human, it is conceivable. However, there would still be a key distinction between the 'schema-less' human-made image and the AI image: the human-made image would not be a result of *failure.* It would instead be a result of deliberate abandonment of schema.

CONCLUSION: WEIRDNESS

In this section, I have offered a first step towards an aesthetics of AI art. I have explored weirdness, a distinct quality of AI works of art. Through use of examples, I conceptualised what this weirdness is, suggesting that the reactions to weirdness are indicative of some form of norm violation. I argued that norm violation is insufficient to distinguish AI weirdness from other norm violation in artworks (such as we might see in comedy or horror). Examining this weirdness further suggests that it is not merely a violation of norms, but the result of some failure to reproduce something that is convincingly *human.* Further comparison of AI-generated images to human-made images revealed the specific kind of failure seen in AI images. The AI, when it failed, did not succeed in reproducing the schematic features central to human-made depiction and human vision. To further understand the cause of this failure, I took a closer look at the goals of the AI systems. Without a human understanding of image-making, AI systems

simply cannot reproduce human images in the way that a human would; their goals are fundamentally different. The weirdness in AI art, then, is a product of norm violation through non-human failure.

If the distinctive aesthetic of AI weirdness is rooted in the *failure* of AI, then there is an interesting implication. As these are failures which occur in systems designed to improve with further training, then the distinctive weirdness of their outputs will become less and less prominent over time as the systems improve. This is apparent in the case of *Thiscatdoesnotexist*. As time goes by and the AI system improves, there are fewer and fewer failures displayed on this website. As AI systems increasingly succeed in producing works that resemble the human works that they are trained to replicate, they will cease to produce weird outputs. As aestheticians, this might be of concern to us; soon, the distinctive aesthetic of weird AI art may be lost. If we do not examine the aesthetics of AI art now, we might miss it.

Convincing Images: AI Art and Waltonian Contact

INTRODUCTION

Artificial Intelligence (AI) technology has developed swiftly and impressively in recent years. Generative AI systems are now able to produce realistic photo-like images and digital paintings. The impressive nature of these images raises questions about the nature of *convincingnes*s in AI images, and our responses to it. Convincingness is often a key aim in AI research; a primary way of evaluating the output of image-generating AI is how well they manage to replicate the features of their target image type. In the case of photographs, this means how well the outputs look like photographs of real things, or in the case of paintings, how closely they emulate actual paintings. This evaluation, then, is of the AI image's ability to *convince* us – usually, to convince us that they are human-made works. This convincingness will be explored in this section. Not all works made by AI are convincing, but others clearly are, as demonstrated by work by the likes of Elgammal (et al., 2017), and the products of *thispersondoesnotexist*. Here, I will set aside those images that do not convince us and focus on those that do. First, I will define convincingness, and discuss the role of convincingness in AI art. I will then move to examine a popular claim about AI artworks: that when we discover a work is in fact a convincing AI image, we feel cheated or duped. Following this, I will consider that AI works might be akin to artistic fakes or forgeries. After rejecting this possibility, I will move to propose that AI works can cause a kind of Waltonian jolt (Walton, 1984), demonstrating how this might function with 'photographic' AI images. I will then expand this proposal to include non 'photographic' images produced by AI.

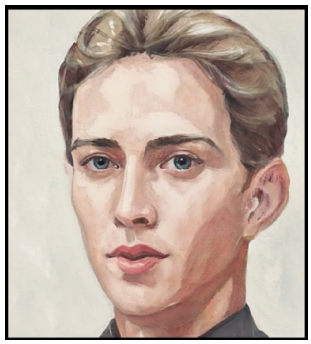
Figs. 45-50: A representative sample of images produced by *Thispersondoesnotxist*

CONVINCINGNESS

I am using convincingness here to denote the following: an image can be called *convincing* if it is produced in an atypical way yet is sufficiently perceptually indistinguishable from an image that is produced in a typical way. For AI images, they can be called convincing if they are sufficiently perceptually indistinguishable from an image produced via traditional means (such as photography or painting). I use 'sufficiently' here because while at first glance AI images may look like images produced through typical means, they often do have 'tells', small areas of the image that can giveaway their status as an AI image, if you know what to look for. Thus, while these images may not actually be completely perceptually indistinguishable, they will be close enough to satisfy a typical viewer. Convincing AI images may look like photographs, or they may look like drawings, paintings or prints. 'Photographic' AI images, such as those created by *thispersondoesnotexist* (figs. 44-49), convince viewers that they are indeed photographs of real people, taken through digital or mechanical means, as opposed to algorithmically generated images. In the case of generatively produced paintings (such as the portraits shown in figs. 50-52), the aim is to convince viewers that these are painted images, rather

than images produced by a computer. The locus of this 'convincing' nature is within the perceptible qualities of the image.





Figs. 51-53 A representative sample of artistic images produced by StyleGAN2, as demonstrated by NVIDIA

Convincingness seems to be a kind of success criterion. We can see a similar concept at play in other creative fields, for instance, in film special effects. In some cases, these works are aimed at looking like things in the real world. This could be something as mundane as a crowd of people (such as in *The Crown,* see BBC, 2017) or unlikely (such as explosions in action films, or giant waves in disaster films). In other cases, such as science fiction and fantasy films, the special effects may not be aimed at reproducing something which could ever occur in the real world; but, even if they are not aimed at replicating anything in the real-world, they still strive to look 'believable' in some way. We might bemoan these special effects looking 'fake'. Similar to our AI images, we want what we see on screen to look *profilmic*, i.e., as if it were physically present in front of the camera, and not created on a computer.[114] In the case of AI artworks, this typically takes the form of a work convincingly emulating the target artform.

This question of convincingness is particularly relevant to AI, and convincingness is a typical standard for judging the success of an AI system. Achieving "convincingness" is like passing the Turing Test, which for many years has functioned as something of a gold standard for AI technology. The Turing test is intended to assess whether a computer can convince us that it is, in fact, human (Russell

[114] This is by no means the sole aim of special effects of course.

& Norvig, 2010, pp. 2-3).[115] For art-making AI, this usually consists of asking people whether they think images are made by a human or a computer. This kind of artistic 'Turing Test' has been conducted by some researchers to establish the success of their system, such as in the work of Elgammal et al. (2017), where the success of the works was judged by human participants. The discussion around this paper often focusses on the results of this judgement. The conclusion specifies that "The system was evaluated by human subject experiments which showed that human subjects regularly confused the generated art with the human art, and sometimes rated the generated art higher on various high-level scales." (Elgammal et al., 2017, p. 21). This 'Turing test' has featured in discussion of AI artworks. For example, in a piece on the Christie's website discussing the first sale of an AI artwork it states that AI researchers:

> are still addressing the fundamental question of whether the images produced by their networks can be called art at all. One way to do that, surely, is to conduct a kind of visual Turing test, to show the output of the algorithms to human evaluators, flesh-and-blood discriminators, and ask if they can tell the difference. (Christie's, 2018b).[116]

Despite this method of evaluating AI art as successful, the response to finding out that an image is made by AI is not always positive. Marcus du Sautoy, a popular figure in the public discussion of creativity and AI, writes "If you cry when you see a piece of art and then are told that the work was computer-generated, I suspect you might feel cheated or duped or manipulated." (du Sautoy 2019, p. 100).[117] There is a clear focus here on the expressive value of a work, and du Sautoy uses this claim to point to an idea that art is all about connecting with other humans. I am not endorsing this idea here. What du Sautoy's remarks suggest though, is a situation where you feel initially convinced by the work – you

[115] There is much debate about the Turing test as a standard for judging AI. Issues arise, for example, when a computer reveals itself for not producing human error, which results in the question of whether programmers should insert errors into the system to make the computer more convincing. Nevertheless, the Turing Test is still widely discussed as a mechanism for evaluating AI.

[116] This kind of test is not without issues; notably, it is simply not how we usually deal with art. We do not look at an artwork and think about it being convincingly human. The ultimate success for an artistic Turing Test would be for the work to be received in the same way as any other artwork. This is hard to achieve, as the sort of test illustrated above involves participants being 'on the lookout' for AI-made works.

[117] Du Sautoy has repeated this claim, most recently at an AI festival in 2020 (CogX, 2020).

believe the work is by a human - and you are then found to be wrong. When you realise you were mistaken, you may feel 'cheated', 'duped', or 'manipulated'.[118] Why? Do works made by AI deceive us? What has changed when we discover that the work in front of us is made by a computer?

FEELING CHEATED AND DUPED?

This claim that we might feel "duped" or "cheated" in some way by AI art raises questions about the nature of convincing AI images. It suggests that the works are tricking us in some way, that the work is a form of deception (otherwise, why would we feel duped?). Given this characterisation of AI art as deceptive we could compare the AI works to fakes or forgeries, which might similarly leave us feeling cheated. In fact, articles about AI images do just this:

> *Thispersondoesnotexist* is one of several websites that have popped up in recent weeks using StyleGAN to churn out images of people, cats, anime characters and vacation homes that look increasingly close to reality, and in some cases are indiscernible by the average viewer. These sites show how easy it's becoming for *people to create fake images that look plausibly real* — for better or worse. (Metz, 2019, emphasis mine)

The language here suggests that there is something non-genuine about the images, and that they are in some way fake, not real, or deceptive, but look plausible to us.

Let's examine the concept of the fake. The fake, in art terms, is typically understood as a direct replica of a work of art (Currie, 1998). This is not what occurs in AI art. The closest that AI art gets to producing a fake is a style-transfer image, whereby the AI learns to reproduce a style of a particular artist. 'Forgery' may be a better word to describe what is occurring in AI artworks that make viewers feel this sense of being 'cheated'. Michael Wreen defines forgery as follows:

> 'a forged XY isn't a genuine XY, but is represented as a genuine XY, and is so represented with the intention to deceive,' where X is a variable ranging over sources of issue (such as Vermeer

[118] I do not mean to suggest here that this is the only possible reaction one might have.

> or 17th-century Holland) and Y ranges over the kind of thing forged (such as paintings) (Wreen, 2002, p. 152)

There are two key elements in this definition. First, there is the representation of the forgery as a genuine XY, where X is the source of the supposed work, and Y is the object itself. Second, there is an intention to deceive.

1 Represented as a genuine XY

By the definition provided by Wreen, the *source* of the work (be that the artist or era for example) is purposefully misrepresented. Gregory Currie, in his definition of forgery, places similar emphasis on "a false claim about the producer's identity" (Currie, 1998). Is the source of the work misrepresented in the case of AI works?

There are two related issues here. The producer's *identity* can indeed the source of inaccuracy in the case of AI works: we may be led to believe the work is produced by a human, when it is in fact produced by an AI. However, it is not clear in the kind of case du Sautoy raises, whether (1) the specific identity or era issuing the image is misrepresented in any way, or (2) that the appearance of the work sufficiently recreates specific markers of a genuine XY to be misrepresentative in and of itself.

On (1), there are no claims about the origins of AI works typically made by the producers of the AI or the AI itself. AI images are usually broadcast as being made by AI, and AI do not present works as being by existing artists or from specific eras.[119] On (2) some AI art indeed could be thought of as attempting to mimic the work of specific artists. These kinds of AI typically use a style transfer method, which "takes a content and a style image as inputs to synthesize an image with the look from the former and feel from the latter" (Li et al., 2018, p. 1). The aim is not to fully replicate a work, but apply the 'feel', or style, of one image to another. As we can see (fig. 53) the outputs, while applying some stylistic elements to the target image, does not produce anything like a plausible forgery. The

[119] This should not occur within AI systems, even when AI are trained on one artist's work – the randomness factored into the system will ensure identical copies are not produced.

replication of style ultimately does not produce works which are perceptually near indistinguishable from the target artist.[120]

Fig. 54 Demonstration of StyleTransfer, NVIDIA.

StyleGANs are one of the AI systems typically used to recreate the style of existing works. These AI aim to produce images in a huge range of styles and subjects without reproducing training images. Producing a variety of images that are properly distributed and not merely replicating training images is an evaluative component of GAN testing (for example Borji, 2018 discusses several evaluative methods to tackle this issue). Notably, despite the increased ability of AI to produce convincing images, researchers show that even in an impressive recent iteration of StyleGAN, StyleGAN2, it is possible to distinguish which images were produced by the system through inversion of the generative component (Karras et al., 2020, p. 8).

There is a further issue for AI images in meeting the requirement of forgery. It is unclear from the definition here whether the conditions of Y (where Y ranges over the kind of thing forged) also need to be met. Is a forgery a forgery if it is not the same as the thing it is forgery of? For example, can a forgery of a Picasso painting be a forgery if it is not actually a painting? This may sound odd, but it is key for the case of AI works. They may be seen as 'fake' photographs or paintings, but they are not photographs or paintings at all, they are computer-made images. They are not physical objects and they are lacking

[120] See examples of AI style transfer for these kinds of images, e.g., Li et al. (2018).

in basic qualities of a forged artefact: they are not painted or printed, not typically variable in dimension, they have no textured surface, no signature, no physicality. It is thus not really possible to consider them as 'forgeries' of an artefact. Perhaps if we printed the images; at least in the case of photographs, they would be closer to forgery? In the case of photography, the reproduction of the physical photograph is not considered sufficient for forgery due to photography's allographic nature; that is, that the work can have multiple copies, all of which are equally considered instances of the work of art (Goodman, 1976, p. 113). Ultimately, with convincing AI works, the epistemic basis for believing the work is a photograph or painting seems to come from the perceptual qualities of the work itself, as opposed to a representation of the work as something it is not.

2 Intention to deceive

Typically, with existing AI artworks there has not been any explicit intention to deceive. Most works, like those sold in auction houses and included in exhibitions and festivals, are not presented to viewers as anything other than AI works. In terms of works designed with the explicit intention to deceive, the closest we come to deceptive intention is the case of "deepfakes", where AI is used to replace a person's image in a video with the likeness of someone else. Deepfake technology often *does* set out to intentionally deceive. Deepfakes sometimes utilise the same technology as AI image generators to assist with the recreation of faces, however these generators are not the crux of the technology (Zucconi, 2018). Unlike Deepfakes, AI art images do not attempt to deceive us that a particular person is doing something they never did.

Even in the case of Turing-like tests, it is not clear that the AI are actually aiming to deceive the viewer. The computer programme (or, in this case, the AI images) are not typically presented as part of a deception. The goal is to produce output that could have been the product of a human. For example, in the case of Elgammal, the human participants are aware that some images are made by an AI and some are not. The "Turing test" question is "Do you think the image is created by an artist or generated by computer?". There is no clear attempt to deceive here, as the work is not presented as a genuine anything, it is merely presented alongside genuine works for the participant's consideration.

The goal is not to deceive them about the origins of a particular work, but rather to see how well the AI works fare compared to human-made works.

While the works themselves may not be intended to deceive when presented to the public, there is a related concern, that the AI system is intending to deceive *itself.* The language of deception is often used when discussing generative AI systems (Generative Adversarial Networks, GANs), as these AI are comprised of two elements (a generator and discriminator) that work in competition. The discriminator is trained on a set of images of the target kind (for example, paintings). The generator produces images for the discriminator, which provides feedback on the images. The systems work in competition: the generator trying to trick the discriminator, and the discriminator trying to catch the 'false' images. Goodfellow and colleagues, who originally designed GANs, describe this adversarial relationship:

> 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. (Goodfellow et al., 2014, p. 1).

Despite this analogy, there still seems insufficient deception to call the output images forgery. This is because there needs to be some level of knowledge of a "genuine XY" in the 'victim' of the deception order to be deceived. In the case of AI, there is no such knowledge. The discriminator element of the system has learnt a broad category of image based on thousands of training images. This is not sufficiently specific for forgery, as the X (sources of issue) would collapse into the Y (kind of thing forged). Furthermore, it does not make sense to separate the elements of the generative AI in order to analyse the intent of the overall system. The goal of the entire system is to produce convincing images, not direct replications. The design of the system is such that neither the 'counterfeiters' nor the 'police' are ultimately successful, rather that the elements work together to produce a convincing image. Finally, the as forgery involves the intended deception of a viewer, the discriminator would need to qualify as

a “viewer”. It is not clear that a module of a computational system could be considered a viewer in the typical sense.

As I have argued here, we can set aside forgery as an explanation for the response of feeling ‘cheated’ when we discover an image is made by an AI. What then could explain this feeling?

WALTON’S JOLT

If we reject the idea that the convincingness of AI images lies in the way they are presented to us, and instead lies in the perceptible qualities of the images, then is there another way we could understand this feeling of being cheated or duped? Du Sautoy’s description of the response of viewers to discovering a work of art is in fact made by an AI is emotionally charged. This suggests that there is some affective component to the discovery that a work has been made by an AI and not a human artist (at least in some cases). The easy explanation for this feeling of being cheated was that there is some kind of fakery or forgery occurring; however, as shown here, AI works do not meet the conditions to be considered forgeries. What else could explain the negative affect that seems, at least for some, to be associated with the discovery that a work is in fact made by an AI?

There is evidence that various types of cheating, scams, and manipulations involve loss; sometimes financial, and sometimes emotional (such as a loss of emotional connection, loss of dignity, loss of confidence or self-esteem). There is typically some emotional component, some sense of loss beyond the mere financial (University of Exeter School of Psychology 2009), and loss, both psychological and financial, is key in defining a person’s status as a victim (Bayley, 1991). Perhaps the sense of loss that may be felt upon discovering that the work you are viewing is actually made by an AI, can lead viewers to feel like they have been cheated. What then, is lost in the case of AI art?

I would like to propose that the reaction to AI images is akin to the jolt described by Kendall Walton in his paper “Transparent Pictures: On the Nature of Photographic Realism” (1984). Walton

describes the occurrence of finding out that an image you thought was a photograph, turned out to be a super-realist painting:

> If the point concerned how photographs look, there would be no essential difference between photographs and paintings. For paintings can be virtually indistinguishable from photographs. Suppose we see Chuck Close's superrealist Self-Portrait thinking it is a photograph and learn later that it is a painting. The discovery jolts us. Our experience of the picture and our attitude toward it undergo a profound transformation, one which is much deeper and more significant than the change which occurs when we discover that what we first took to be an etching, for example, is actually a pen-and-ink drawing. It is more like discovering a guard in a wax museum to be just another wax figure. We feel somehow less "in contact with" Close when we learn that the portrayal of him is not photographic. (Walton, 1984, p. 255)

Walton uses the idea of the jolt to support his theory of transparency in images. According to Walton, photographs afford us contact with the subject in a way that other artworks do not. This sense of contact occurs because the photograph has a different epistemic status to paintings; they are seen as independent of mediation by the artist's beliefs in a way that paintings are not. As Walton states, his theory can account for our experiencing a jolt when we realise a photo is actually a super-realistic painting:

> My theory accounts for the jolt. At first we think we are (really) seeing the person portrayed; then we realize that we are not, that it is only fictional that we see him. However, even after this realization it may well continue to seem to us as though we are really seeing the person (with photographic assistance), if the picture continues to look to us to be a photograph. (Walton 1984, p. 255)

Does something similar occur when we discover an image, which we thought was human-made, was actually made by AI? Take the image shown here (fig. 54). This is seemingly a photograph, a headshot of a woman. If you saw this image, you would presumably view it as photo. If you were then told, as is the case, that this is not a photograph – it is a completely fabricated image. Does a similar jolt occur? I think that it does. We thought we were seeing a person, and now we are told this person does not exist.

It certainly seems that this would feel like a loss of contact with that person. As stated above, a sense of loss could well lead to a feeling of being 'duped', 'cheated', or 'manipulated'. In the case of 'fake' AI photographs, we can certainly experience negative feelings when we discover that apparent photographic image is not one at all. This does not seem to be a problem for Walton's original account, as super-realistic paintings and AI 'photographs' have similar perceptual qualities; both look convincingly photographic. What about non-photographic images, such as AI paintings? Could the discovery that a painting is AI-made produce a similar reaction to the non-photograph?

Fig. 55 Generated image from *Thispersondoesnotexist*

A NON-PHOTOGRAPHIC JOLT?

With an understanding of the jolt as a loss of contact, we can make sense of the feeling of being cheated, duped, or manipulated when we find out any work is actually made by an AI, instead of by a person. We could plausibly use this regardless of whether the work is photographic or not. Instead of asserting transparency, where we believe we have undisputed access to the subject of the photograph, we simply assert associated beliefs about other artforms, which are incompatible with AI works.

This view aligns with du Sautoy's claim that the value of art is in providing some kind of closeness to the artist's consciousness (du Sautoy, 2019, p. 100). Walton's account suggests that there is some distance created in the moment of the jolt. The distance is between the original object and the image presented, so, in the case of super-realist paintings versus photographs, there is a greater distance from the depicted object in the super-realist paintings than there is from the object in the photograph. According to du Sautoy, there is greater psychological distance from the depicted image in the case of

the AI image than there is in the case of a human-made painting. Could his characterisation fit with a different kind of distance being felt when we discover a painting is in fact made by an AI?

Could there be some kind of 'contact' with non-photographic images? A kind of psychological, emotional, or behavioural connection with an artist that we feel when we are presented with a work of art? This may not be the same as contact as Walton describes it – we are not seeing the thing depicted –- but perhaps we feel that we have something like contact when we view works. I am going to call this potential form of contact *connection*; a sense of being close with the mind, emotions, or behaviours of a person. I am using connection instead of contact, to make clear that this form of contact is not putting us in direct perceptual contact with a person in the way that Walton used 'contact'.

Let's examine an AI 'painting', and du Sautoy's view of human art: that art gives us insight or connection to other humans.[121] A convincing AI 'painting' is presented to us. We perceive the work as a painting and believe accordingly that the image is created by a person, and thus that we are seeing the output of their thoughts/feelings/beliefs etc.[122] We are then informed that the painting was generated by an AI. Upon learning this, we have new cognitions about the image, likely based on our understanding of AI; this image cannot be created from thoughts/feelings/beliefs etc., because AI, we think, do not have these. As such, we have lost the assumed connection with an artist we initially felt.[123] This sense of loss here manifests as a feeling of being duped or manipulated. A convincing AI artwork could plausibly give rise to an experience similar to the jolt described by Walton.

By this account, the loss of 'contact' in the case of AI works results in a negative affective response. There is one problem however: there seems to be no negative affect in the case of hyperrealist works. Walton does not discuss any negative emotional response to these discoveries, and the jolt he characterises does not seem negative or positive. If the jolt is a form of surprise, it may be

[121] Du Sautoy's view seems to be a kind of expression theory. See e.g., J Robinson (2005).
[122] We do not have to be committed to any one of these. We could, for example, believe that the work means something, or that it is something we should strive to understand – the only necessary factor is that this belief is believed to be incompatible with the truth.
[123] A similar possibility, of *emotional* connection has been discussed by Holliday (2016) and Alcaraz Léon (2019) though in a narrower set of cases.

that some initial negative valence is experienced, however this seems insufficient to explain a lingering negative feeling such as that described by du Sautoy.[124] In fact, it seems likely that a more common response to the hyperrealist works described by Walton, is to be impressed or amazed at the proficiency of an artist in producing such detailed work.[125] Nonetheless, we may experience negative affect in the case of AI works. Why?

Accepting that a viewer may feel a sense of connection when viewing a work of art could help explain this disparity. In hyperrealist works, viewers experience a jolt upon learning the work is not a photograph, as they lose contact with the subject. However, they do not merely learn that the work is not a photograph; they also learn that the work is a painting. Whilst the viewer has lost contact with the subject, they have gained (or at least retained) *connection* with the artist. In contrast, in the case of the AI 'photographic' work, upon discovering the work is made by an AI, the viewer experiences a jolt, loses contact with the subject, and gains nothing. It is unlikely that the average viewer feels a sense of connection with an AI in the same way they do with a human artist (though this is not an impossibility). In cases where the viewer feels no connection, they may feel a negative affective response to the loss of contact. Similarly, in the case of an artwork (such as a work that appears to be a painting) upon discovery that the work is made by an AI, there will be a sense of loss of connection with the presumed artist. As stated above, loss is a key feature of being cheated or duped. Understandably then, loss of contact or connection could lead to the feeling that one has been manipulated, cheated, or duped, particularly when nothing additional is gained.

CONCLUSION: CONVINCING IMAGES

This section of the thesis has examined the idea of convincingness in AI images. I defined convincingness, and its role in relation to AI images. I moved to address the claim that convincing AI

[124] Surprise has been characterised as mildly negative, though there has been much debate about the valence of surprise. See Noordewier and Breugelmans (2013).
[125] This response is also possible in AI works – my own response, as a researcher of AI image generation is exactly this response!

images produce negative affect when their true nature is revealed. I examined the possibility that AI works are akin to fakes or forgeries, which could explain the feeling of being ‘duped’. I argued that AI images do not typically meet the definition of forgery. I then proposed an alternative, that the discovery of a convincing image as an AI image could be explained as a kind of Waltonian jolt, which leads to a sense of loss felt by losing contact with the assumed photographic subject. This sense of loss in turn could lead to a sense of being cheated in some way. This makes sense for photographic images, the topic of Walton’s original paper, but what of AI ‘paintings’? With the use connection as a form of emotional, psychological or behavioural contact with the artist, I argued that a similar ‘jolt’ might be felt when we discover a painting is not human made as we lose connection with that presumed artist. This could explain negative responses to convincing AI works.

Conclusion: The Aesthetics of AI art

In this chapter I have examined two qualities of AI visual works. First, I examined 'weirdness', a quality common to many works of AI, that seems to elicit responses of humour or unease. I argued that weirdness seems to be a kind of incongruity, but that incongruity alone is not sufficient to explain this weirdness. I then put forward my account of 'non-human failure', arguing that the AI, in producing 'weird' images is failing in some way. I explained this failure by arguing that AI is lacking in the basic visual schemas, such that when an AI fails to produce a target object, it fails in a way that a human never would.

Second, I examined a quality of AI works that seems to be related to success: convincingness. I claim that an image can be called convincing if it is produced in an atypical way yet is sufficiently perceptually indistinguishable from an image that is produced in a typical way. For AI, this means that the image produce by an AI *appears* to be produced via traditional means, such as painting or photography. I explore a potential negative response some people may have to discovering that an artwork is made by an AI; that one has been 'cheated'. I argue that this response cannot be accounted for by conceptions of forgeries or fakes and propose that this response indicates that a Waltonian 'jolt' has occurred. The loss of contact which one would expect with a jolt experienced upon the discovery that a photograph is actually a hyperrealist painting will not occur with all AI works; I propose instead that a loss of connection with the artist may be experienced by a viewer.

Conclusion: A Philosophy of AI Art

In this thesis I have made an initial step towards developing a philosophy of AI art. I began this thesis by laying out what is meant by 'AI art': *works in the domain of the arts, produced by (or with) artificial intelligence systems.* I offered a taxonomy of levels of autonomy of AI in making 'artworks': AI used as a **tool, manipulation** of an AI system, **collaboration** with AI, **facilitation** of AI art production, and **complete autonomy** of the AI in producing works. I then established a definition of Artificial Intelligence and put forward some foundational knowledge in machine learning, which forms the basis of art-making AI systems. This underscored my approach throughout the thesis: to pay close attention to how AI systems actually work, and to draw on specific examples in order to investigate the philosophical issues of AI art. With this in mind, I presented several classes of AI visual art: robotic systems, iterative systems, programmed computationally creative systems, adversarial models, and diffusion models. For these classes of AI art systems, I offered detail of how specific instances of the system worked, examples of works produced with the system, and some potential pitfalls.

Having established what kinds of AI and AI works are under consideration in this thesis, I moved to examine some existing literature on AI art, from both computer science and philosophy. Here, I looked at writing by Blaise Agüera y Arcas, Aaron Hertzmann, Steffen Steinert, and Mark Coeckelbergh. I utilised Coeckelbergh's proposed framework for philosophical work on AI art as a scaffolding for my thesis. Coeckelbergh identified three variations of the same question about machine art: 1) "Can machines create *art*?" which focusses on the requirements of art; 2) "Can machines *create* art?" which focusses on whether machines could be creative or could achieve the process of art-making; and 3) "Can *machines* create art?" which draws attention to the nature of machines compared to humans in creating art. Led initially by Coeckelbergh's three questions, I laid out five questions for this thesis to answer:

1) Can AI make art?
2) Can AI be creative?

3) What key features of machines might limit the possibility for AI to achieve art and creativity, and can these be overcome?
4) Will AI share aesthetic or artistic values with humans?
5) Does AI art have any unique (or at least common) aesthetic qualities?

Each of the five chapters of this thesis corresponded to one of these questions.

In the first chapter 'Can AI Create *Art*?' I aimed to address just this question. I examined three popular theories of art: the institutional account of art, the historical account of art, and the cluster account of art. In each case, I utilised a paradigmatic example of the theory and examined whether the account was prohibitive for AI: was there anything in the theory that excluded AI from producing a work that could be classed as art? In the case of the Dickie's institutional account, it initially appeared that AI would be able to meet this account without too much difficulty; works made by AI could have the appropriate status of 'candidate for appreciation' conferred upon them just as other artworks do. This becomes less simple if we want the AI itself to confer art-status on its own works of art. For example, can an AI act on behalf of the (socially situated) artworld without participating in it? Given that human artists do not have to carry out this conferral, this did not seem too problematic for AI works. The requirement that the work be an artefact though did raise some questions for AI, as it is not clear that AI works could meet the requirements of artefactuality under a traditional account (they must be produced with intention, involve modified materials, and be made for a purpose, Preston, 2020). However, Dickie's account of artefactuality appears to be far more permissive than the traditional account, merely requiring an action to be done to something to make it an artefact. Here, we come to the requirement of agency, which was addressed in Chapter 3.

The historical account of art (as formulated by Levinson, 1979), was a challenge for AI art from the outset. Whilst many AI works are clearly created with an appropriate relation to prior works of art, the intentional component of the definition threatened to prevent AI from making art. Of course, we could take the route of utilising a human in the process of making a work to ensure that the products of AI could be counted as art. This could be either in their intentional selection of the products of the AI, or through their design of the system with the intention that the products of the AI

are regarded appropriately. Otherwise, we will need to argue that intention is possible for AI in order for AI works to be art. In the case of the cluster account (Gaut, 2000), there is a necessary criterion of an action occurring, followed by a list of properties that a work could have. The work does not need to meet all the properties to be considered art, but it does need to meet some combination of these. I set aside the action criterion and evaluated each of the ten properties an artwork could have. Many of these properties *were* possible for an AI work to meet (though some were philosophically laden, such as 'being the product of an intention to make a work of art'). However, further investigation into the requirement that the work be the product of an action was needed before we could conclude that an AI could make art under the cluster account. This is due to the fact that action typically implies agency (which, in turn, implies mental states). In Chapter 3, I argue that an AI could have a minimal form of agency. This suggests that some AI could indeed act, and thus make art under the cluster account. I also argued that in several ways the cluster account is particularly helpful for the study of AI art. First, it can provide a framework, by way of the listed properties, for assessing what AI can or cannot currently do in terms of producing art. This could help in the development of future AI in art domains. Second, the cluster account allows for future alterations or additions to the list of properties. In this way, the cluster account could accommodate what AI *can* do. Finally, even if AI cannot meet the necessary condition of the cluster account, the account can still explain why it seems like AI works are art, and why we might disagree about its status: AI art can be explained as a border case.

In Chapter 2, 'Can AI *Create* Art?', I turned to examine whether AI could be creative. Here, I examined three definitions of creativity, and assessed AI against each of these. First, I investigated the Darwinian model of creativity, arguing that this model can be used to assess creativity in AI systems, and demonstrating this with two generative models. Next, I examined Boden's account of creativity, comprised of novelty, value, and surprisingness. Boden already argues that computers could be creative under her account, particularly in the exploratory sense; however, she later admits that if autonomy is required for creativity, then AI cannot truly be creative. I examined autonomy closely and argued in response to Boden's sceptical objector that some level of autonomy (in the philosophical sense) is possible for AI. Finally, I assessed AI against Gaut's agential account of

creativity. I argued that the three elements of Gaut's account of creativity (novelty, value, and flair) could be possible for AI, other criteria put forward by Gaut in his three-part definition of creativity are possible for AI. I then examined his additional requirement of 'spontaneity' arguing that this could instead be conceptualised as 'unpredictability', and thus also could be possible for AI. This assessment set aside Gaut's key necessary condition for creativity: agency. This issue of agency formed a central focus of Chapter 3.

In Chapter 3, I examined the question 'Can *AI* Create Art?' looking at some presumed limitations of AI that might prevent it from making art or being creative. This chapter began with an exposition on the philosophy of mind, followed by an examination of AI extended mind; here, I argued first that AI could have mind (under a functionalist account), and second that AI could potentially *extend* its mind to draw on human involvement in the social world. This section can function as a rebuttal to those with views akin to Hertzmann's, that AI must be social to make art, or as a potential path for AI to act in the social institution of the artworld. Next, I examined whether AI could be embodied. Embodiment seems to be a significant area of lack for AI, and this came up in Boden's discussion of creativity. In this section, I argued that some AI could meet the conditions of embodiment. Furthermore, if an AI was creative, in meeting the requirements of creativity it would necessarily achieve the conditions to be 'embodied'. Finally, I turned to agency. This section on agency plays a key role in my argument that, under Gaut's account, AI could be creative. In Chapter 2, I argued that AI could meet all the other components of creativity for Gaut, leaving only agency unfulfilled. In this section, I argued for a minimal account of agency as *internally directed, purposive doing*. I showed that account of agency could still fulfil the role of agency in Gaut's account but was achievable for some AI systems. Furthermore, this minimal account of agency entails the possibility that some AI can 'act'; i.e., to carry out 'purposeful doings'. This 'action' can then be applied to both the cluster account of art (where action was a necessary condition) and the institutional account of art (where action was necessary to create an artefact), completing the argument that AI could make art under these accounts.

In Chapter 4, 'AI and Aesthetic Value', I turned to the concept of value in AI art. Here, I argued that we may indeed face an aesthetic 'value alignment problem' with AI, as AI may not recognise human aesthetic or artistic values. I argued that in the case of creativity we will generally want our AI to align with human values; this is due to both an epistemological and an ethical reason. In the case of aesthetics, however, I argued that we may *not* always want alignment, as insisting that AI should always aim to align with human values may prevent the possibility of new aesthetic values stemming from AI works.

Finally, in Chapter 5 'The Aesthetics of AI Art', I considered whether there are any particular aesthetic features common to AI art, and AI images more generally. First, I examined 'weirdness', the quality common in works of AI. I argued that weirdness seems to be a kind of incongruity. To further distinguish the particular case of weirdness in AI works, I propose that this incongruity occurs through 'non-human failure' arguing that weird AI images are a result of failure of the AI system. This failure is of a particular kind: it occurs in AI systems in particular due to their lack of the basic building blocks of vision. Due to this, failed AI images do not appear as poor human-made images would; they are distorted and lacking in boundaries. Second, I examined a quality of successful AI images: that they are convincing. Convincing AI images are sufficiently perceptually indistinguishable from images that are produced in a traditional way, such as via painting or photography. Many AI images are very convincing, though the realisation that an image is made by an AI may garner a negative response. I investigated this negative response, arguing that it was indicative of a particular kind of Waltonian jolt.

Throughout the thesis, I have put forward arguments that AI can make art and can be creative. Additionally, I have argued that AI may have something distinct to offer us in the domain of art and aesthetics. As stated in the introduction, one needn't agree with each argument to recognise the opportunity for rich philosophical investigation into AI art.

FUTURE RESEARCH DIRECTIONS

Given the sparsity of existing research in this area, and the rapid development of AI systems, there are many directions to expand the work in this thesis. Below I sketch a few possibilities:

Responsibility: in Chapter 3 we encounter responsibility attribution (praising or blaming) as a driving factor in including agency as a requirement of creativity. There is far more to be said on responsibility, and in future work I will explore this. In particular, I plan to develop a division of responsibility for artworks into *artistic* and *aesthetic* responsibility. The former denotes responsibility over a work's being art. The latter denotes responsibility for the aesthetic features of a work. In this way, AI may be recognised as responsible (in terms of 'knowledge' and 'control') for the aesthetic features of a work, without needing to meet challenging levels of knowledge of the concept of art that are entailed by being responsible for making an artwork.

Authorship: Early in the thesis I set aside the question of authorship, due to limits in scope. The question of whether an AI can be an author/artist, while related to the question of whether AI can make art, is not identical. We may require more distinct features of an artist that an AI will struggle to fulfil. The capacity for AI authorship then is a key topic for exploration, particularly as it may impact on the recognition of AI contributions to artworks and ownership of AI works.

Collaboration: In this thesis I have tried to focus, as far as is possible, on the role of AI in AI art, but this is to ignore the human in AI artmaking. In most cases of AI art proliferating today, a human has had a large hand in producing the work. How might we characterise human use of AI in artmaking? And is true collaboration with AI possible? In a joint project with colleagues at Northeastern University (London and Boston) I have been examining human-AI, and AI-AI teams and autonomous behaviour. In particular, there are interesting experimental results in collaborative game-playing that could have a large impact on what we see as necessary for collaboration with non-humans. I would like to apply learning from this work to the concept of collaboration in art.

Ethics: Whilst ethical considerations were largely beyond the scope of this thesis, there are many fascinating issues starting to be discussed in this area, such as how to prevent harmful or biased AI images from being produced and shared (Metz, 2022a). Additionally, explicit in the discussion of AI

artworks proliferating in the 'real-world' of the arts is often claims of wrong-doing, such as the accusations of cheating and unfairness we saw at the very start of this thesis when Jason M Allen won a competition with his AI-generated work (Roose, 2022). Similarly, the accusation that AI art is created by 'stealing' artworks has been spreading online, though it seems mostly to be based on misconceptions of diffusion models (Murphy, 2022). We should take care not to conflate the concerns of artists and creative workers about their replacement with concerns about art-ownership and plagiarism; it is hard to prove that an AI is plagiarising any one person, even if an artist's work is used in a dataset. The question of whether an artwork *should end up* in that dataset is again a different one. The conversation about ethics and AI art will need to transform into one about data ethics. I have begun to explore the idea of art as data. Conceptualising digitised art as a form of data allows us to easily ask typical questions of our art datasets, such as whether this data is biased, who can use this data and what for, has it been properly collected and stored, should this data be purchasable or only taken with consent? Conceptualising digitised art in this way also allows us to apply data-management strategies such as data-auditing and debiasing. Art-data for AI art generation may become a flashpoint in a much-needed conversation about data ownership, data handing, and privacy (if we are worried about art-data being scraped to train AI, consider what else is being taken).

Quantum Computing: Exciting work is occurring in quantum computing, and this cutting-edge field is already being applied to art-generation, particularly in music (Miranda, 2022). Quantum computing relies on quantum mechanics, which is stochastic in nature, and thus is in some ways truly unpredictable. Quantum computing, despite its early stage of development, represents a huge step forward in computational processing power, and thus is likely to produce incredible results when it can be applied to AI.

In addition to these new areas of research, the work in this thesis could easily be expanded further.

AI art: AI could be assessed against additional accounts of art (or meta-accounts, such as Mag Uidhir's approach, 2013). In this thesis, I address only three accounts: the institutional, historical, and cluster account. In each case, I take just one version of the account as the basis for investigation, yet

in each case there are multiple iterations and developments of the account, such that the arguments I have set forth here may not apply to each one.

AI creativity: Particularly for AI creativity (though to some extent, for AI art too) the development of a philosophically informed, operationalised tool for the evaluation of new computational systems would be highly beneficial to the field of computational creativity, which (at least in my experience) is still starved of input from philosophical thinkers. In this way, the philosophical investigation of whether AI could be creative could inform the development of new AI systems which better meet the requirements of creativity (and for example, agency, autonomy, etc.).

The aesthetics of AI art: There are surely many more possible aesthetic features of AI images (and features of AI works in other mediums) to be investigated; for example, there often seems to be a 'patterned' quality frequently visible in AI images. Additionally, following from Chapter 5 of this thesis, the concept of failure could be explored further and potentially connected to literature on failed art (Mag Uidhir, 2013).

AI art beyond the visual: Perhaps the biggest limitation of the work in this thesis was the focus on visual art only. AI has been applied to many fields of art already, including poetry, literature, video games, film, and music. There will undoubtedly be many philosophically fruitful, medium-specific questions to be examined in each of the arts, and correspondingly exciting AI case studies to pick apart.

Ultimately, this thesis has gone some way in moving aesthetics forward in the investigation of Artificial Intelligence. There is still much left to explore.

References

Abell, C. (2012) 'Art: What it is and why it matters', *Philosophy and Phenomenological Research*, 85, pp. 671–691.

Achlioptas, P., Ovsjanikov, M., Haydarov, K., Elhoseiny, M. and Guibas, L.J. (2021) 'Artemis: Affective language for visual art'. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11569-11579). Available at: https://doi.org/10.48550/arXiv.2101.07396

Adajian, T. (2018) 'The definition of art', in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy.* Available at: https://plato.stanford.edu/archives/spr2022/entries/art-definition/ (Accessed: 10 December 2022).

Agüera y Arcas, B. (2017) 'Art in the age of machine intelligence', *Arts,* 6(4).

AI artists (2021a) *Helena Sarin,* Available at: https://aiartists.org/helena-sarin (Accessed: April 30, 2021).

AI artists (2021b) *Robbie Barrat*, Available at: https://aiartists.org/robbie-barrat (Accessed: April 30, 2021).

AI Art Shop (no date). Available at: https://aiartshop.com/ (Accessed: 17 December 2022).

AICAN (no date). *AICAN.* Available at: https://www.aican.io (Accessed: 12 December 2022).

Ai-Da (2019) *Who is Ai-Da?* Available at*:* https://www.ai-darobot.com/general-interest (Accessed: 1 November 2022).

AI: More than Human (2019) [Exhibition]. The Barbican, London, UK. 16 May 2019 – 26 August 2019. Available at: https://www.barbican.org.uk/whats-on/2019/event/ai-more-than-human (Accessed: 1 October 2022).

AIVA Technologies (no date) *AIVA.* Available at: https://www.aiva.ai/ (Accessed: 15 December 2022).

Aizawa, K. (2009) 'Neuroscience and Multiple Realization: A Reply to Bechtel and Mundale', *Synthese,* 167(3), pp. 493-510. Available at: https://doi.org/10.1007/s11229-008-9388-5

Alcaraz León, M. J. (2019) 'Aesthetic intimacy', in O. Kuisma, S. Lehtinen and H. Mäcklin (eds.) *Paths from the Philosophy of Art to Everyday Aesthetics,* Helsinki: Finnish Society for Aesthetics, pp. 78-100. Available at: https://helda.helsinki.fi/bitstream/handle/10138/302115/Paths-from-the-philosophy-of-art-2019.pdf?sequence=1&isAllowed=y (Accessed: 17 December 2022)

AlphaGo (2017) Directed by Greg Kohs [Film]. Netflix.

Ambrosio, C. (2019) 'Unsettling robots and the future of art', *Science,* 365(6448) pp. 38-39. Available at: DOI:10.1126/science.aay1956

Anderson, J. C. and Dean, J. T. (1998) 'Moderate autonomism'. *The British Journal of Aesthetics*, 38 (2), pp. 150-166.

Anderson, M., & Anderson, S. L. (2011). 'General introduction' in M. Anderson and S. L. Anderson (eds.) *Machine ethics,* Cambridge: Cambridge University Press, pp. 1–4.

Anscombe, G. E. M. (1963) *Intention*, second edition, Oxford: Blackwell.

Armstrong, D. M. (1968) *A Materialist Theory of the Mind*. London: Routledge.

Arnheim, R. (2001) 'What it means to be creative', *British Journal of Aesthetics,* 41(1), pp. 24–25. Available at: https://doi.org/10.1093/bjaesthetics/41.1.24.

Artbreeder (no date). Available at: https://artbreeder.com/i?k=945de361b5897b4bab6842ff8283 (Accessed: 30 April 2021).

Art Market Guru (2019) 'Interview with Mario Klingemann', *Art Market Guru,* 8 March. Available at: https://www.artmarket.guru/le-journal/interviews/mario-klingemann/ (Accessed: 12 November 2019)

Aru, J., Labash, A., Corcoll, O. and Vicente, R. (2022) 'Mind the gap: Challenges of deep learning approaches to theory of mind'. *arXiv.* Available at: https://doi.org/10.48550/arXiv.2203.16540

Assembly AI (2022) *How does DALL-E 2 actually work?* April 15. Available at: https://www.youtube.com/watch?v=F1X4fHzF4mQ (Accessed: 10 December 2022).

Auerbach, D. (2016) ‘Do Androids Dream of Electric Bananas?’ Slate, 23 June. Available at: https://slate.com/technology/2015/07/google-deepdream-its-dazzling-creepy-and-tells-us-a-lot-about-the-future-of-a-i.html (Accessed: 11 December 2022).

Bäck, T. (1996) *Evolutionary Algorithms in Theory and Practice.* Oxford: Oxford University Press.

Baer, T. (2019) *Understand, Manage, and Prevent Algorithmic Bias*. Berkeley, CA: Apress.

Bailey, J. (2018b) ‘The truth behind Christie’s $432K AI art sale’, *Artnome,* 29 October. Available at: https://www.artnome.com/news/2018/10/13/the-truth-behind-christies-432k-ai-art-sale (Accessed: 16 December 2022)

Barrat, R. (2017) ‘Robbie Barrat’ *AI Art Gallery, NeurIPS Workshop on Machine Learning for Creativity and Design* 2020. Available at: https://www.aiartonline.com (Accessed: 11 December 2022).

Barrat, Robbie. (2018) ‘the AI *always* paints heads and faces the same way; with this weird yellow/purple texture. Have no idea why, but I like it.’ [Twitter] 27 March. Available at: https://twitter.com/videodrome/status/978733665889865728?s=20&t=FotN53mNUCV5i15QwQEkJw (Accessed: 13 December 2022).

Bayley, J. E. (1991) ‘The concept of victimhood’, in D. Sank and D. I. Caplan (eds.) *To Be a Victim*, pp. 53–62. New York: Springer.

BBC (2017) ‘The Crown: The visual effects secrets of Netflix's drama’ *BBC News*, 7 December. Available at: https://www.bbc.co.uk/news/av/technology-42161388 (Accessed: 20 December 2022).

Bern, M. (no date) ‘Someone noticed how ugly medieval cat paintings are, and it’s too funny’ *Bored Panda.* Available at: https://www.boredpanda.com/ugly-medieval-cats-art/?utm_source=google&utm_medium=organic&utm_campaign=organic (Accessed: 1 May 2020).

Beschizza, R. (2018) ‘Robbie Barrat's AI-generated nude paintings make Francis Bacon look like a genteel pre-Raphaelite’ *Boingboing*, Available at: https://boingboing.net/2018/03/28/robbie-barrats-ai-generated.html (Accessed: 1 May 2020)

Besold, T. R., Schorlemmer, M., and Smaill, A. (2015) *Computational Creativity Research: Towards Creative Machines.* Paris: Atlantis Press.

Bickle, J. (2020) 'Multiple Realizability', in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy,* Available at: https://plato.stanford.edu/archives/sum2020/entries/multiple-realizability (Accessed: 30 August 2023).

Biederman, I. (1987) 'Recognition-by-components: a theory of human image understanding', *Psychological Review,* 97(2) pp. 115-147.

Block, N. J. and Fodor, J. A. (1972) 'What psychological states are not', *The Philosophical Review*, 81(2): 159–181. doi:10.2307/2183991

Boden, M. A. (1988) 'Escaping from the Chinese room', in John Heil (ed.), *Computer Models of Mind.* Cambridge: Cambridge University Press.

Boden, M. A. (1991) *The Creative Mind: Myths & Mechanisms,* New York: Basic Books.

Boden, M. A. (1994) 'Creativity and computers', in T. Dartnell (ed.) *Artificial Intelligence and Creativity: An Interdisciplinary Approach*. Dordrecht: Springer, pp. 3–26.

Boden, M. A. (1995) 'Understanding creativity', in J. Götschl (ed.) *Revolutionary Changes in Understanding Man and Society.* Dordrecht: Springer, pp. 75-82.

Boden, M. A. (1998) 'Creativity and artificial intelligence', *Artificial Intelligence*, 103(1-2), pp. 347-356.

Boden, M. A. (2004) *The Creative Mind: Myths and Mechanisms,* 2nd edn. London: Routledge.

Boden, M. A. (2006) *Mind as Machine: A History of Cognitive Science.* Oxford: Clarendon Press.

Boden, M. A. (2009) 'Creativity and Artificial Evolution' *Creativity: The Mind, Machines, and Mathematics*, pp. 1-15. Available at: https://pdfs.semanticscholar.org/3095/94cc60a265cca1999ff9ad1625ea426260d8.pdf (Accessed: 30 May 2019)

Boden, M. A. (2010) *Creativity and Art: Three Roads to Surprise.* Oxford: Oxford University Press.

Boden, M. A. (2014) 'Creativity and artificial intelligence: A contradiction in terms?' in E. S. Paul, and S. B. Kaufman (eds.), *The Philosophy of Creativity: New Essays*, New York: Oxford

Academic, pp. 224-244. Available at: https://doi.org/10.1093/acprof:oso/9780199836963.003.0012 (Accessed: 17 December 2022)

Boden, M. A. (2018) 'Creativity and biology', in B. Gaut and M. Kieran (eds.) *Creativity and Philosophy.* Routledge. New York.

Boden, M. A. and Edmonds, E. A. (2019) '5 Can Evolutionary Art Provide Radical Novelty?', in *From Fingers to Digits: An Artificial Aesthetic*, Cambridge, MA: MIT Press, pp. 107-126.

Borji, A. (2018) 'Pros and cons of GAN evaluation measures' *arXiv*, Available at: https://arxiv.org/pdf/1912.04958.pdf (Accessed: 1 June 2020).

Bostrom, N. (2009) 'Ethical Issues in Advanced Artificial Intelligence', in Susan Schneider (ed.) *Science Fiction and Philosophy: From Time Travel to Superintelligence,* pp. 277-284. Hoboken, NJ: Wiley-Blackwell.

Bostrom, N. (2012) 'The superintelligent will: Motivation and instrumental rationality in advanced artificial agents', *Minds and Machines,* 22(2), pp. 71-85. Available at: https://doi.org/10.1007/s11023-012-9281-3

Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies,* Oxford: Oxford University Press.

Brandon, J. (2017) "Pandora uses machine learning to make sense of 80 billion thumb votes." *Venture Beat*, 12 July. Available at: https://venturebeat.com/2017/07/12/pandora-uses-machine-learning-to-make-sense-of-80-billion-thumb-votes/ (Accessed: 20 August 2019).

Brasel, S. A. and Gips, J. (2011) 'Media multitasking behavior: Concurrent television and computer usage', *Cyberpsychology, Behavior, and Social Networking,* 14(9), pp. 527-534

Bratman, M. (1987) *Intentions, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Brock, A., Donahue, J. and Simonyan, K. (2019) 'Large scale GAN training for high fidelity natural image synthesis", *arXiv.* Available at: https://doi.org/10.1016/j.jvir.2012.07.028 (Accessed: 29 March 2019).

Brook, A. and Raymont, P. (2021) 'The unity of consciousness', in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy,* Available at:

https://plato.stanford.edu/archives/sum2021/entries/consciousness-unity/ (Accessed: 30 August 2023).

Brown, D. (2014) The sixth meditation: Descartes and the embodied self', in D. Cunning (ed.) *The Cambridge Companion to Descartes' Meditations*, Cambridge Companions to Philosophy, Cambridge: Cambridge University Press, pp. 240-257. Available at: doi:10.1017/CCO9781139088220.013

Brown, M. (2011) 'Patrick Tresset's robots draw faces and doodle when bored' *Wired,* 17 June. Available at: https://www.wired.co.uk/article/sketching-robots (Accessed: 11 December 2022).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S. (2020) 'Language models are few-shot learners', *Advances in neural information processing systems*, 33, pp. 1877-1901. Available at: https://doi.org/10.48550/arXiv.2005.14165

Bryson, J. J. (2019) 'The past decade and future of AI's impact on society', in *Towards a New Enlightenment? A Transcendent Decade*, 11, Madrid: Turner, pp. 150-185.

Burton, B. (2018) 'AI gets naughty by generating surreal nude portraits', *C Net*. Available at: https://www.cnet.com/news/ai-gets-naughty-by-generating-nude-portraits/ (Accessed: 1 May 2020)

Buss, S. and Westlund, A. (2018) 'Personal Autonomy' in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy.* Available at: https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/ (Accessed: 31 August 2022).

Campbell, D. T. (1960) 'Blind variation and selective retentions in creative thought as in other knowledge processes', *Psychological Review,* 67(6), pp. 380-400.

Campbell, D. T. (1974) 'Evolutionary epistemology', in Schlipp, P. A. (ed). *The Philosophy of Karl Popper*, La Salle, IL: Open Court, pp. 413–463.

Cardoso, A., & Wiggins, G. A. (2007). *Proceedings of the 4th International Joint Workshop on Computational Creativity*.

Cariani, P. (1992) ‘Emergence and Artificial Life’, in C. G. Langton, C. Taylor, J. D. Farmer, and S. Rasmussen (eds.), *Artificial Life II.* Redwood City, CA: Addison-Wesley, pp. 775–797.

Carroll, N. (1993) ‘Historical narratives and the philosophy of art’, *The Journal of Aesthetics and Art Criticism*, 51(3), pp. 313-326. Available at: https://doi.org/10.2307/431506

Carroll, N. (1996) ‘Moderate moralism’ *The British Journal of Aesthetics,* 36(3), pp. 223-238.

Carroll, N. (1999) ‘Horror and Humor’ *The Journal of Aesthetics and Art Criticism,* 57(2), pp. 145-160

Carruthers, P. (2006) ‘Creative cognition in a modular mind’, The Architecture of the Mind, Oxford: Oxford Academic, pp. 277-334. Available at: https://doi.org/10.1093/acprof:oso/9780199207077.003.0005 (Accessed: 10 November 2020).

Casimir, M. J. and Jung, S. (2009) ‘Honor and dishonor: Connotations of a socio-symbolic category in cross-cultural perspective’ in: H. Markowitsch, and B. Röttger-Rössler (eds.) *Emotions as Bio-cultural Processes.* New York: Springer.

Çelikok, M. M., Peltola, T., Daee, P. and Kaski, S. (2019) ‘Interactive AI with a theory of mind’, *arXiv.* Available at: https://doi.org/10.48550/arXiv.1912.05284

Chambers, S. W. (2020) ‘AI agents are not artists’, *Towards Data Science,* 22 December. Available at: https://towardsdatascience.com/artificial-intelligence-agents-are-not-artists-9743d5dba2d0 (Accessed: 12 December 2022).

ChargeBox (no date) ‘The ChargeBox Story.’ *ChargeBox*. Available at: https://www.chargebox.com/about/ (Accessed: 29 July 2019).

Christian, B. (2020). *The Alignment Problem: How Can Machines Learn Human Values?* London: W. W. Norton & Company.

Christie’s (2018a) *Edmond de Belamy, from La Famille de Belamy.* Available at: https://www.christies.com/en/lot/lot-6166184 (Accessed: 1 November 2022).

Christie’s (2018b) ‘Is artificial intelligence set to become art’s next medium?’ *Christie’s,* 12 December. Available at: https://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx (Accessed: 1 October 2022).

Churchland, P. and Churchland, P. S. (1990) ‘Could a machine think?’, *Scientific American,* 262(1), pp. 32-39. Available at: https://www.jstor.org/stable/24996642

Clark, A. (2001) ‘Reasons, robots and the extended mind’, *Mind and Language* 16(2), pp. 121–45. Available at: https://doi.org/10.1111/1468-0017.00162

Clark, A. and Chalmers, D. (1998) ‘The extended mind’, *Analysis,* 58(1), pp. 7–19.

Clark, P. (2020) ‘Photoshop: Now the world’s most advanced AI application for creatives’, *Adobe Blog*, 20 October. Available at: https://blog.adobe.com/en/publish/2020/10/20/photoshop-the-worlds-most-advanced-ai-application-for-creatives (Accessed: 17 December 2022)

Claxton, G. (2015) *Intelligence in the Flesh: Why Your Mind Needs Your Body Much More Than It Thinks,* New Haven, CT: Yale University Press.

Clot-Goudard, R. (2022) ‘Intention, knowledge, and responsibility’, in R. Teichmann (ed.) *The Oxford Handbook of Elizabeth Anscombe,* Oxford: Oxford University Press. pp. 53-71.

Cloudpainter (no date) *Cloudpainter Gallery*. Available at: https://www.cloudpainter.com/gallery (Accessed: 18 December 2022).

Coeckelbergh, M. (2017) ‘Can machines create art?’, *Philosophy & Technology*, 30(3), pp. 285-303.

Coeckelbergh, M. (2020). ‘A-responsible machines and unexplainable decisions’ *AI Ethics*, Cambridge MA: MIT Press, pp. 109-123.

CogX (2020) *The Cutting Edge Stage Live at CogX 2020 Day 2,* 9 June. Available at: https://www.youtube.com/watch?v=CgBc9rrtXOI (Accessed: 31 June 2020).

Cohen, D. and Nisbett, R. E. (1994) ‘Self-protection and the culture of honor: Explaining southern violence’, *Personality and Social Psychology Bulletin*, 20(5), pp. 51-567.

Cohen, H. (1995) ‘The further exploits of AARON, painter.’ *Stanford Humanities Review* 4(2), pp. 141-158.

Cohen, P. R. and Levesque, H. J. (1990) ‘Intention is choice with commitment’, *Artificial Intelligence*, 42(2-3), pp. 213-261. Available at: https://doi.org/10.1016/0004-3702(90)90055-5

Cole, D. (2020) ‘The Chinese room argument’ In E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy.* Available at: https://plato.stanford.edu/archives/win2020/entries/chinese-room/ (Accessed: 1 November 2022).

Collingwood, R. G. (1938) *The Principles of Art,* London: Clarendon, pp. 105-124.

Colton, S. (2008). 'Creativity versus the perception of creativity in computational systems', *Proceedings of the AAAI Spring Symposium on Creative Systems,* pp 14–20. Available at: http://www.aaai.org/Papers/Symposia/Spring/2008/SS-08-03/SS08-03-003.pdf (Accessed: 1 December 2022).

Colton, S. (2012). 'The Painting Fool: Stories from building an automated painter', in J. McCormack, and M. D'Inverno (eds.) *Computers and Creativity.* Berlin Heidelberg: Springer-Verlag.

Colton, S., Cook, M., Hepworth, R. and Pease, A. (2014) 'On acid drops and teardrops: Observer issues in computational creativity' *7th AISB Symposium on Computing and Philosophy.* Available at: http://doc.gold.ac.uk/aisb50/AISB50-S03/AISB50-S3-Colton2-paper.pdf (Accessed: 12 December 2022).

Colton, S., Halskov, J., Ventura, D., Gouldstone, I., Cook, M. and Ferrer, B.P. (2015) 'The Painting Fool sees! New projects with the automated painter', *Proceedings of the International Conference on Computational Creativity,* pp. 189-196.

Colton, S., Valstar, M. F. and Pantic, M. (2008) 'Emotionally aware automated portrait painting', *DIMEA '08: 3rd international conference on Digital Interactive Media in Entertainment and Arts.* 10-12 September 2008, Athens, Greece. Available at: https://doi.org/10.1145/1413634.1413690

Colton, S. and Ventura, D. (2014) 'You can't know my mind: A festival of computational creativity', *Proceedings of the 5th International Conference on Computational Creativity,* pp. 351–354.

Colton, S. and Wiggins, G.A. (2012) 'Computational creativity: The final frontier?' *ECAI 2012: 20th European Conference on Artificial Intelligence,* 12, pp. 21-26.

Cromby, J. (2014) 'Embodiment', in T. Teo (ed.) *Encyclopedia of Critical Psychology,* New York, NY: Springer. Available at: https://doi.org/10.1007/978-1-4614-5583-7_89

Cropley, D. H., Cropley, A. J., Kaufman, J. C. and Runco, M. A. (eds.) (2010). *The Dark Side of Creativity*. Cambridge: Cambridge University Press.

Currie, G. (1998) 'Forgery and fakery: Artistic forgery', *Routledge Encyclopedia of Philosophy*, Taylor and Francis. Available at: https://www.rep.routledge.com/articles/thematic/artistic-forgery/v-1/sections/forgery-and-fakery. (Accessed: 1 June 2020).

Danaher, J. (2012) ‘Bostrom on Superintelligence and Orthogonality’, Philosophical Disquisitions, 3 April. Available at: https://philosophicaldisquisitions.blogspot.com/2012/04/bostrom-on-superintelligence-and.html (Accessed: 17 December 2022).

Darling, K. (2021) *The New Breed: How to Think About Robots*. London: Penguin.

Dastani, M. and Yazdanpanah, V. (2022) ‘Responsibility of AI systems’, *AI and Society.* Available at: https://doi.org/10.1007/s00146-022-01481-4

Davies, S. (1991) *Definitions of Art.* Ithaca, NY: Cornell University Press. Available at: http://www.jstor.org/stable/10.7591/j.ctv3s8q0q

Dawkins, R. (1986) *The Blind Watchmaker*. New York: Norton.

Days, D. (2018) ‘Very "Jacob's Ladder"!’ [Twitter] 27 March. Available at: https://twitter.com/david_days/status/978757641752244226 (Accessed: 1 June 2020)

DeepDream Generator (no date) Available at: https://deepdreamgenerator.com **(**Accessed: 30 January 2020).

DeepMind (no date) *AlphaGo*. Available at: https://www.deepmind.com/research/highlighted-research/alphago (Accessed: 1 October 2022).

Deloitte (2018) *Global Mobile Consumer Survey 2018: The UK Cut.* Available at: http://www.deloitte.co.uk/mobileuk/#uk-frequency-by-usage-of-device (Accessed: 29 July 2019).

De Mol, L. (2021) ‘Turing Machines’, in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy,* Available at: https://plato.stanford.edu/archives/win2021/entries/turing-machine/ (Accessed: 30 August 2023).

Dennett, D. (1971) ‘Intentional systems’, *Journal of Philosophy,* 68, 87–106. Available at: https://doi.org/10.2307/2025382

Dennett, D. (1987) *The Intentional Stance,* Cambridge, MA: MIT Press.

Dennett, D. (1988) ‘Conditions of personhood’, in M. F. Goodman (ed.) *What is a person?* Clifton, NJ: Humana Press, pp. 145-167.

Dennett, D. (2009) 'Intentional systems theory', in A. Beckermann, B. P. McLaughlin, and S. Walter (eds.) *The Oxford Handbook of Philosophy of Mind*, pp. 339-350. Available at: https://doi.org/10.1093/oxfordhb/9780199262618.003.0020

Dennett, D. C. (2017). *From Bacteria to Bach and Back.* London: Allen Lane.

Deutsch, H. (1991) 'The creation problem', *Topoi,* 10, pp. 209–225. Available at: https://doi.org/10.1007/BF00141341

Dickie, G. (1974) *Art and the Aesthetic: An Institutional Analysis*, Ithaca, NY: Cornell University Press.

Dickie, G. (1979) 'What is art? An institutional analysis', in M. Rader (ed.) *A Modern Book of Esthetics: An Anthology.* 5th ed. New York: Holt Rinehart and Winston.

Dickie, G. (2004). 'The new institutional theory of art', in P. Lamark and S. Haugom Olson (eds.) *Aesthetics and the philosophy of art: the analytic tradition: an anthology*, Oxford: Blackwell. pp. 47-54.

Dipert, R. R. (1993) *Artifacts, Art Works, and Agency,* Philadelphia: Temple University Press.

Dreyfus, H. (1992) *What Computers Still Can't Do: A Critique of Artificial Reason.* Cambridge, MA: MIT Press.

Du Sautoy, M. (2020) *The Creativity Code: Art and Innovation in the Age of AI*. Cambridge, MA: Harvard University Press.

Dutton, D. (2009) *The Art Instinct: Beauty, Pleasure, and Human Evolution.* Oxford: Oxford University Press.

Ebert, R. (2010) 'Video games can never be art'*, RogerEbert.com,* 16 April 2010. Available at: https://www.rogerebert.com/roger-ebert/video-games-can-never-be-art (Accessed: 16 December 2022).

Edelman (2019) *2019 Edelman Trust Barometer Global Report.* Available at: https://www.edelman.com/sites/g/files/aatuss191/files/2019-03/2019_Edelman_Trust_Barometer_Global_Report.pdf (Accessed: 29 July 2019)

Eden, A. H., Søraker, J. H., Moor, J. H. and Steinhart, E. (2013) *Singularity Hypotheses: A Scientific and Philosophical Assessment.* Berlin: Springer.

Elgammal, A. (2019) ‘AI Is blurring the definition of artist’, *American Scientist*, 107(1), pp. 18-21. Available at: https://www.americanscientist.org/article/ai-is-blurring-the-definition-of-artist (Accessed: 12 December 2022).

Elgammal, A. (2020) ‘The robot artists aren’t coming’, *New York Times,* 27 May. Available at: https://www.nytimes.com/2020/05/27/opinion/artificial-intelligence-art.html (Accessed: 12 December 2022).

Elgammal, A., Liu, B., Elhoseiny, M. and Mazzone, M. (2017) ‘CAN: Creative adversarial networks, generating “art” by learning about styles and deviating from style norms.’ *arXiv*. Available at: https://doi.org/10.48550/arXiv.1706.07068

Elster, J. (1992) ‘Conventions, Creativity and Originality’, in M. Hjort (ed.) *Rules and Conventions: Literature, Philosophy, Social Theory,* Baltimore: Johns Hopkins University Press, pp. 32-44.

Epstein, R. (2015) ‘Of course animals are creative: Insights from generativity theory’, in A. C. Kaufman and J. C. Kaufman (eds.) *Animal Creativity and Innovation*, Academic Press, pp. 375-393). Available at: https://doi.org/10.1016/C2013-0-18605-9

Feigl, H. (1958) ‘The ‘mental’ and the ‘physical’’, in H. Feigl, M. Scriven, and G. Maxwell, (eds.) *Concepts, Theories and the Mind-Body Problem*, vol. 2. Minneapolis: Minnesota Studies in the Philosophy of Science.

Feng, Q., Guo, C., Benitez-Quiroz, F. and Martinez, A.M. (2022) ‘When do GANs replicate? On the choice of dataset size’, *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6701-6710). Available at: https://doi.org/10.48550/arXiv.2202.11765

Fischer, J. and Ravizza, M. (1998) *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge: Cambridge University Press.

Fjelland, R. (2020) ‘Why general artificial intelligence will not be realized’, Humanities and Social Sciences Communications 7(10). Available at: https://doi.org/10.1057/s41599-020-0494-4

Flynn, M. (2018) ‘A 19-year-old developed the code for the AI portrait that sold for $432,000 at Christie’s’, *Washington Post,* 26 October. Available at: https://www.washingtonpost.com/nation/2018/10/26/year-old-developed-code-ai-portrait-that-sold-christies/ (Accessed: 11 December 2022).

Fodor, J. (1975) *The Language of Thought*, New York: Thomas Y. Crowell.

Fodor, J. (1981) *Representations*, Cambridge: MIT Press.

Fodor, J. (1983) *The Modularity of Mind*, Cambridge, MA: MIT Press.

Fodor, J. (1998) *Concepts*, Oxford: Clarendon Press.

Fodor, J. (2000) *The Mind Doesn't Work That Way*, Cambridge, MA: MIT Press.

Fodor, J. (2008) *LOT2*, Oxford: Clarendon Press.

Fodor, J. (2009) 'Where is my mind?', *London Review of Books*, (31)3, pp. 13–15. Available at: https://www.lrb.co.uk/the-paper/v31/n03/jerry-fodor/where-is-my-mind (Accessed: 15 December 2022).

Folia Magazine (2014a) 'Caturday Folia: musetto grazioso' 3 May. Available at: https://www.foliamagazine.it/caturday-folia-musetto-grazioso/ (Accessed: 30 April 2021)

Folia Magazine (2014b) 'Caturday Folia: micio irrequieto' 29 April. Available at: https://www.foliamagazine.it/caturday-folia-micio-irrequieto/ (Accessed: 30 April 2021)

Fondation Giacometti (2020) *Alberto Giacometti / À la recherche des œuvres disparues / In Search of Lost Works,* Available at: https://www.fondation-giacometti.fr/en/event/104/in-search-of-lost-works (Accessed: 1 October 2022)

Frankfurt, H.G. (1978) 'The problem of action' *American Philosophical Quarterly,* 15(2), pp. 157-162. Available at: http://www.jstor.org/stable/20009708

Franklin, S. (1997) 'Autonomous agents as embodied AI', *Cybernetics & Systems*, 28(6), pp. 499-520. Available at: http://ccrg.cs.memphis.edu/~franklin/AAEI.html (Accessed: 1 July 2019).

Franklin, S. and Graesser, A. (1997) 'Is it an agent, or just a program?: A taxonomy for autonomous agents' in J. P. Müller, M. J. Wooldridge, and N. R. Jennings (eds) *Intelligent Agents III Agent Theories, Architectures, and Languages. ATAL 1996. Lecture Notes in Computer Science,* 1193. Heidelberg, Berlin: Springer. Available at: https://doi.org/10.1007/BFb0013570

Gadde, R. (2022) 'Method enables better control of GAN image generators' output', *European Conference on Computer Vision (ECCV),* Tel Aviv, 23-27 October. Available at:

https://www.amazon.science/blog/method-enables-better-control-of-gan-image-generators-output (Accessed: 12 December 2022).

Gaut, B. (1998). 'The ethical criticism of art', in J. Levinson (ed.) *Aesthetics and Ethics*. Cambridge: Cambridge University Press.

Gaut, B. (2000) 'The cluster account of art', In N. Carroll (ed.) *Theories of Art Today,* pp. 25-45.

Gaut, B. (2003) 'Creativity and imagination', In B. Gaut and P. Livingston (eds.) *The Creation of Art*, pp. 148-173.

Gaut, B. (2005) 'The cluster account of art defended', *The British Journal of Aesthetics*, 45(3), pp. 273-288. Available at: https://doi.org/10.1093/aesthj/ayi032

Gaut, B. (2009). 'Creativity and skill', in M. Krausz, D. Dutton, and K. Bardsley (eds.) *The Idea of Creativity*, Leiden: Brill, pp. 83-103.

Gaut, B. (2010) 'The philosophy of creativity', *Philosophy Compass*, 5, pp. 1034–1046. Available at: http://doi.wiley.com/10.1111/j.1747-9991.2010.00351.x

Gaut, B. (2012) 'Creativity and rationality', *The Journal of Aesthetics and Art Criticism,* 70, pp. 259-270. Available at: https://doi.org/10.1111/j.1540-6245.2012.01518.x

Gaut, B. (2018) 'The value of creativity', in B. Gaut and M. Kieran (ed.) *Creativity and Philosophy.* New York, NY: Routledge

Gavin, F. (2019) 'Artificial intelligence: the art world's weird and wonderful new medium', *How to Spend It*. Available at: https://howtospendit.ft.com/articles/205746-artificial-intelligence-the-art-world-s-weird-and-wonderful-new-medium (Accessed: 1 May 2020)

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A. and Brendel, W. (2018) 'ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness', *arXiv*. Available from: https://arxiv.org/abs/1811.12231 (Accessed: 9 May 2022).

Gershenson, C. (2003) 'Artificial neural networks for beginners', *arXiv.* Available at: https://doi.org/10.48550/arXiv.cs/0308031

Gombrich, E. H. (1994) 'Formula and experience', in *Art and Illusion: a study of the psychology of pictorial representation*. London: Phaidon, pp. 126-152.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Benigo, Y. (2014) 'Generative adversarial networks', *arXiv*. Available at: https://doi.org/10.48550/arXiv.1406.2661

Goodman, N. (1976). *Languages of art: An approach to a theory of symbols*, 2nd edn, Indianapolis, IN: Hackett publishing.

Google Developers (2022) 'Embeddings', *Machine Learning: Foundational courses.* Available at: https://developers.google.com/machine-learning/crash-course/embeddings/video-lecture (Accessed: 10 December 2022).

Grba, D. (2022) 'Deep else: A critical framework for AI art', *Digital*, 2(1), pp. 1-32.

Gruen, L. (2021) 'The moral status of animals', in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy.* Available at: https://plato.stanford.edu/archives/sum2021/entries/moral-animal/ (Accessed: 1 December 2022).

Hanson, L. (2020) 'Two dogmas of the artistic-ethical interaction debate', *Canadian Journal of Philosophy,* 50(2), pp. 209-222.

Hare, B. and Tomasello, M. (2005) 'Human-like social skills in dogs?' *Trends in Cognitive Sciences*, 9(9), pp. 439-444.

Harnad, S. (1989) 'Minds, Machines and Searle', *Journal of Experimental and Theoretical Artificial Intelligence*, 1, pp. 5–25.

Hauser, L. (no date a) 'Behaviorism', *Internet Encylopedia of Philosophy,* ISSN 2161-0002. Available at: https://iep.utm.edu/behaviorism/#SH1b (Accessed: 31 August 2023).

Hauser, L. (no date b) 'Chinese Room Argument', *Internet Encylopedia of Philosophy,* ISSN 2161-0002. Available at: https://iep.utm.edu/chinese-room-argument/ (Accessed: 31 August 2023).

Hauser, L. (1997) 'Searle's Chinese box: debunking the Chinese room argument', *Minds and Machines*, 7(2), pp. 199-226. Available at: https://doi.org/10.1023/A:1008255830248

Helliwell, A. C. (2019) 'Can AI mind be extended?' *Evental Aesthetics,* 8, pp. 93-120.

Helliwell, A. C. (2021) 'Darwinian creativity as a model for computational creativity', *Proceedings of the 7th Computational Creativity Symposium at AISB 2021 (CC2021),* pp. 15-19. Available

at: https://aisb.org.uk/wp-content/uploads/2021/04/cc_aisb_proc.pdf (Accessed: 1 December 2022)

Hertzmann, A. (2018) 'Can computers create art?' *Arts* 7(2). Available at: https://doi.org/10.3390/arts7020018

Hertzmann, A. (2020) 'Computers do not make art, people do', *Communications of the ACM*, 63(5), pp. 45-48.

Hilpinen, R. (1992) 'Artifacts and Works of Art', *Theoria*, 58(1), pp. 58–82. Available at: doi:10.1111/j.1755-2567.1992.tb01155.x

Hitt, T. (2019) 'Meet the Mormon college student behind the viral A.I. game that took Dungeons & Dragons online" *The Daily Beast.* 9 December. Available at: https://www.thedailybeast.com/meet-the-mormon-college-student-behind-viral-artificial-intelligence-game-ai-dungeon (Accessed: 11 December 2022).

Ho, J., Jain, A. and Abbeel, P. (2020) 'Denoising diffusion probabilistic models', *Advances in Neural Information Processing Systems*, 33, pp. 6840-6851.

Holliday, J. (2016) 'Emotional intimacy in literature' *The British Journal of Aesthetics,* 58(1), pp. 1-16. Available at: https://doi.org/10.1093/aesthj/ayx033

Howell, N. R. (2007) 'Uniqueness in context', *American Journal of Theology & Philosophy,* 28(3), pp. 364–77. Available at: http://www.jstor.org/stable/27944418

Hugging Face (no date) *Stable Diffusion 2.1 Demo*. Available at: https://huggingface.co/spaces/stabilityai/stable-diffusion (Accessed: 18 December 2022).

Hume, D. (1882) 'Of the Standard of Taste', in: T. H. Green (ed.) *Essays Moral, Political, and Literary*, *Vol. 1.* London: Longmans, Green, and co.

Hutchinson, E. (2019) 'The One Show may just have welcomed its creepiest guest ever', *Digital Spy*, 12 June. Available at: https://www.digitalspy.com/tv/reality-tv/a27963213/the-one-show-creepiest-guest-ever-robot-artist/ (Accessed: 12 December 2022).

IBM Cloud Education. (2020a) *Neural Networks* Available at: https://www.ibm.com/cloud/learn/neural-networks (Accessed: November 13, 2022).

IBM Cloud Education. (2020b) *What is Machine Learning?* Available at: https://www.ibm.com/cloud/learn/machine-learning#toc-what-is-ma-qhM6PX35 (Accessed: November 13, 2022).

Internet World Stats (2019) *Internet Growth Statistics.* Available at: https://www.internetworldstats.com/emarketing.htm (Accessed: 29 June 2019)

Jacob, P. (2019) 'Intentionality', In E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy.* Available at: https://plato.stanford.edu/archives/win2019/entries/intentionality/ (Accessed: 16 December 2022).

Jaiswal, A., Ramesh Babu, A., Zaki Zadeh, M., Banerjee, D., and Makedon, F. (2021) 'A survey on contrastive self-supervised learning', *Technologies,* 9(2). Available at: https://arxiv.org/abs/2011.00362

Jang, J. and Open AI (2022) 'DALL·E 2 research preview update' *OpenAI.* Available at: https://openai.com/blog/dall-e-2-update/ (Accessed: 1 December 2022).

Jaynes, T. L. (2020) 'Legal personhood for artificial intelligence: citizenship as the exception to the rule', *AI & Society,* 35, pp. 343–354. Available at: https://doi.org/10.1007/s00146-019-00897-9

Jones, J. (2018). 'A portrait created by AI just sold for $432,000. But is it really art?' *Guardian,* 26 October. Available at: https://www.theguardian.com/artanddesign/shortcuts/2018/oct/26/call-that-art-can-a-computer-be-a-painter (Accessed: 1 November 2022)

Jones, J. (2022) 'Incoherent, creepy and gorgeous: we asked six leading artists to make work using AI – and here are the results' *Guardian,* 1 December 2022. Available at: https://www.theguardian.com/artanddesign/2022/dec/01/six-leading-british-artists-making-art-with-ai (Accessed: 15 December 2022)

Jones, K. and Bonafilia, D. (2017) 'Gangogh: Creating art with GANs'. *Towards Data Science*. Available at: https://towardsdatascience.com/gangogh-creating-art-with-gans-8d087d8f74a1 (Accessed: 11 December 2022).

Jordanous, A. (2016) 'Four PPPPerspectives on computational creativity in theory and in practice', *Connection Science*, 28(2), pp. 194-216.

Jumalon, G. (2022) ‘TL;DR — Someone entered an art competition with an AI-generated piece and won the first prize. Yeah that's pretty fucking shitty.’ [Twitter] 30 August. Available at: https://twitter.com/GenelJumalon/status/1564651635602853889 (Accessed: 12 December 2022).

Kant, I. (2000) *Critique of the Power of Judgment*. Translated from German by Paul Guyer and Eric Matthews, Cambridge: Cambridge University Press.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T. (2019) ‘Analyzing and improving the image quality of StyleGAN’, *arXiv*. Available at: https://arxiv.org/abs/1912.04958 (Accessed: 1 May 2020).

Kemp, G. (2021) ‘Collingwood’s aesthetics’, in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy.* Available at: https://plato.stanford.edu/archives/win2021/entries/collingwood-aesthetics (Accessed: 10 December 2022).

Kennick, W. E. (1958) ‘Does traditional aesthetics rest on a mistake?’ *Mind,* 67(267), pp. 317-334. Available at: https://www.jstor.org/stable/2251530

Kiejzer, F. (1998) ‘Some armchair worries about wheeled behaviour’, in R. Pfeifer, B. Blumberg, J-A. Meyer and S. W. Wilson (eds.) *From Animals to Animats 5,* pp.13-21, Cambridge, MA: MIT Press. Available at: https://www.rug.nl/staff/f.a.keijzer/keijzer1998armchairworriespreprint.pdf (Accessed: 1 July 2019).

Kim, J. (2011) *Philosophy of Mind,* 3rd edn, Boulder, CO: Westview Press.

Knight, W. (2022) ‘When AI makes art, humans supply the creative spark’, *Wired,* 13 Jul. Available at: https://www.wired.com/story/when-ai-makes-art/ (Accessed: 12 December 2022).

Koza, J. R. (1994) *Genetic Programming: On the Programming of Computers by Means of Natural Selection,* Cambridge, MA: MIT Press.

Kronfeldner, M. E. (2009) ‘Creativity naturalized’, *The Philosophical Quarterly*, 59(237), pp. 577–592. Available at: https://doi.org/10.1111/j.1467-9213.2009.637.x

Kurzweil, R. (2002) ‘Locked in his Chinese room: Response to John Searle’, in J. W. Richards (ed.) *Are We Spiritual Machines? Ray Kurzweil vs. the Critics of Strong AI,* pp. 128–171.

Lawson-Tancred, J. (2022) 'Will A.I. usher in the end of human artists? Fear not, some say*', Artnet,* 14 September. *Available at:* https://news.artnet.com/art-world/ai-generated-art-debate-2175570 (Accessed: 14 November 2022).

Leemurz (2021) *Ai-Da,* Available at*:* https://commons.wikimedia.org/wiki/File:Ai-Da.jpg (Accessed: 18 December 2022).

Leviathan, Y. and Matias, Y. (2018) 'Google Duplex: An AI system for accomplishing real-world tasks over the phone', *Google AI Blog,* 8 May. Available at: https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html (Accessed: 29 March 2019).

Levin, J. (2023) 'Functionalism', in E. N. Zalta & U. Nodelman (eds.) *The Stanford Encyclopedia of Philosophy,* Available at: https://plato.stanford.edu/archives/sum2023/entries/functionalism/ (Accessed: 1 October 2023).

Levinson, J. (1979) 'Defining art historically', *British Journal of Aesthetics,* 19(3), pp. 232-250. Available at: https://doi.org/10.1093/bjaesthetics/19.3.232

Levinson, J. (1989) 'Refining art historically', *The Journal of Aesthetics and Art Criticism*, 47(1), pp. 21-33. Available at: https://doi.org/10.2307/431990

Levinson, J. (2002) 'Hume's standard of taste: The real problem', *The Journal of Aesthetics and Art Criticism*, 60(3), pp. 227-238.

Levinson, J. (2003) 'Elster on artistic creativity', in B. Gaut and P. Livingston (eds.) *The Creation of Art: New Essays in Philosophical Aesthetics,* pp. 235-256.

LG (2016) *'Low Battery Anxiety' Grips 9 out of Ten People*. Available at: https://www.lg.com/us/PDF/press-release/LG_Mobile_Low_Battery_Anxiety_Press_Release_FINAL_05_19_2016.pdf (Accessed: 20 August 2019)

Liao, S. and Gendler, T. (2020) 'Imagination', In E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*. Available at: https://plato.stanford.edu/archives/sum2020/entries/imagination/ (Accessed: 16 December 2022).

Liu, Y., Jun, E., Li, Q. and Heer, J. (2019) 'Latent space cartography: Visual analysis of vector space embeddings', *Computer Graphics Forum*, 38(3), pp. 67-78. Available at: https://par.nsf.gov/servlets/purl/10172008 (Accessed: 16 December 2022).

Livingston, P. (2009). 'Poincaré's "Delicate Sieve": On creativity in the arts', in M. Krausz, D. Dutton, and K. Bardsley (eds.) *The Idea of Creativity*, Leiden: Brill, pp. 129-146.

Li, X., Liu, S., Kautz, J. and Yan, M-H (2018) 'Learning linear transformations for fast arbitrary style transfer', *arXiv.* Available at: https://arxiv.org/pdf/1808.04537.pdf (Accessed: 1 July 2020).

Loftus, E. F., and Pickrell, J. E. (1995) 'The formation of false memories', *Psychiatric Annals,* 25(12), pp. 720–725.

Luke, S., Hamahashi, S. and Kitano, H. (1999) '"Genetic" programming', *Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation (GECCO'99),* 2, pp. 1098–1105.

Luty, J. (2017) 'Do Animals Make Art or the Evolutionary Continuity of Species: A Case for Uniqueness', *AVANT*, 8(1), pp. 107-116.

Maes, P. (1995) 'Artificial life meets entertainment: Life like autonomous agent' *Communications of the ACM* 38(11), pp.108-114.

Mag Uidhir, C. (2013) *Art and Art-Attempts,* Oxford: Oxford University Press.

Mandelbaum, M. (1965) 'Family-resemblances and generalizations concerning the arts', *American Philosophical Quarterly,* 2, pp. 219-228.

Marr, D. (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information,* San Francisco: W. H. Freeman and Company.

Ma, S. (2018) 'Microsoft expands presence of AI platform XiaoIce', *China Daily,* 28 July. Available at: http://www.chinadaily.com.cn/a/201807/28/WS5b5baf5ea31031a351e90b14.html (Accessed: 22 March 2019).

Matravers, D. (2000) 'The institutional theory: A protean creature', *British Journal of Aesthetics*, 40, pp. 242–250.

Matthias, A. (2004) 'The responsibility gap: Ascribing responsibility for the actions of learning automata', *Ethics and Information Technology*, 6(3), pp. 175-183.

McCulloch, W. and Pitts, W. (1943) ‘A logical calculus of the ideas immanent in nervous activity’, *Bulletin of Mathematical Biophysics*, 5, pp. 115-133.

McGregor, S., Wiggins, G., and Purver, M. (2014) ‘Computational creativity: a philosophical approach and an approach to philosophy’, *5th International Conference on Computational Creativity (ICCC), Ljubljana, Slovenia*. 10-12 June. Available at: http://computationalcreativity.net/iccc2014/wp-content/uploads/2014/06//12.2_ McGregor.pdf

Metz, R. (2019) ‘These people do not exist. Why websites are churning out fake images of people (and cats)’, *CNN.* Available at: https://edition.cnn.com/2019/02/28/tech/ai-fake-faces/index.html (Accessed: 1 July 2020).

Metz, R. (2022a) ‘AI made these stunning images. Here’s why experts are worried’ *CNN Business*, 2 August. Available at: https://edition.cnn.com/2022/06/30/tech/openai-google-realistic-images-bias/index.html (Accessed: 17 December 2022).

Metz, R. (2022b) ‘AI won an art contest, and artists are furious’ *CNN Business,* 3 September. Available at: https://edition.cnn.com/2022/09/03/tech/ai-art-fair-winner-controversy/index.html (Accessed: 11 December 2022).

Milkowski, M. (no date) ‘Computational theory of mind’, *Internet Encyclopedia of Philosophy,* ISSN 2161-0002. Available at: https://iep.utm.edu/computational-theory-of-mind/ (Accessed: 30 August 2023).

Miller, A. I. (2019) *The Artist in the Machine: The World of AI-Powered Creativity*. Cambridge, MA: MIT Press.

Miller, K. D. (2015) ‘Will You Ever Be Able to Upload Your Brain?’, *The New York Times*, 10 October. Available at: https://www.nytimes.com/2015/10/11/opinion/sunday/will-you-ever-be-able-to-upload-your-brain.html (Accessed: 29 March 2019).

Mingers, J. (1991) ‘The cognitive theories of Maturana and Varela’, *Systems Practice*, 4, pp. 319-338. Available at: https://www.researchgate.net/publication/226739585_The_cognitive_theories_of_Maturana_and_Varela (Accessed: 1 July 2019).

Miranda, E. R. (ed.) (2022) *Quantum Computer Music,* New York, NY: Springer.

Montero, B. G. (2022) 'Behaviourism', *Philosophy of Mind: A Very Short Introduction*, Very Short Introductions, Oxford: Oxford Academic.

Mordvintsev, A. Olah, C. and Tyka, M. (2015) 'Inceptionism: Going deeper into neural networks', *Google Research,* Available at: https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html (Accessed: 1 December 2022).

Moura, F. T. (2017) 'AIVA technology: An extraordinary music AI start-up', *LiveInnovation.org,* 17 October. Available at: https://liveinnovation.org/aiva-technology-extraordinary-ai-music-start-up/ (Accessed: 10 December 2022).

Müller, V. C. (2012), 'Autonomous cognitive systems in real-world environments: Less control, more flexibility and better interaction', *Cognitive Computation,* 4(3), pp. 212-15. Available at: https://doi.org/10.1007/s12559-012-9129-4

Murphy, B. (2022) 'No, the Lensa AI app technically isn't stealing artists' work – but it will majorly shake up the art world' *The Conversation,* 14 December. Available at: https://theconversation.com/no-the-lensa-ai-app-technically-isnt-stealing-artists-work-but-it-will-majorly-shake-up-the-art-world-196480 (Accessed: 17 December 2022)

Negrotti, M. (2019) 'Hubert Dreyfus, the artificial and the perspective of a doubled philosophy', *AI & Society,* 34, pp. 195–201. Available at: https://doi.org/10.1007/s00146-018-0800-5

Nemire, B. (2018) 'New AI style transfer algorithm allows users to create millions of artistic combinations', *Nvidia Developer News,* 17 August. Available at: https://news.developer.nvidia.com/new-ai-style-transfer-algorithm-allows-users-to-create-millions-of-artistic-combinations/ (Accessed: 1 September 2020).

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I. and Chen, M. (2022) 'GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models', *arXiv.* Available at: https://doi.org/10.48550/arXiv.2112.10741

Noordewier, M. K., Breugelmans, S. M. (2013) 'On the valence of surprise.' *Cognition and Emotion*, 27(7), pp. 1326-1334. Available at: doi:10.1080/02699931.2013.777660.

Novitz, D. (1999) 'Creativity and constraint', *Australasian Journal of Philosophy*, 77 (1), pp. 67-82. Available at: https://doi.org/10.1080/00048409912348811

NVIDIA Tech Demo Team (2020) 'CVPR 2020: Synthesizing High Resolution Images with GANs' Nvidia Developer. Available at: https://developer.nvidia.com/techdemos/video/dcv20 (Accessed: 1 July 2020).

O'Connor, R. (2022) 'How DALL-E 2 actually works', *AssemblyAI,* 19 April. Available at: https://www.assemblyai.com/blog/how-dall-e-2-actually-works/ (Accessed: 1 December 2022).

Ofcom (no date) *Adults' Media Use and Attitudes*. Available at: https://www.ofcom.org.uk/research-and-data/media-literacy-research/adults/adults-media-use-and-attitudes. (Accessed:29 March 2019).

Oleson, A. (2015) 'China Has Its Own Anti-Vaxxers. Blame the Internet.' *Foreign Policy,* 16 March 2015. Available at: https://foreignpolicy.com/2015/03/16/china-has-its-own-anti-vaxxers-blame-the-internet/ (Accessed: 29 March 2019).

OpenAI, (2022a) *DALL·E 2.* Available at: https://openai.com/dall-e-2/ (Accessed: 16 December 2022).

OpenAI (2022b) *DALL·E Now Available Without Waitlist.* Available at: https://openai.com/blog/dall-e-now-available-without-waitlist/ (Accessed: 29 November 2022).

Oppy, G. and Dowe, D. 'The Turing Test', in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*. Available at: https://plato.stanford.edu/archives/win2021/entries/turing-test/ (Accessed: 1 December 2022).

Pattee, H. H. (1985) 'Universal Principles of Measurement and Language Functions in Evolving Systems', in J. Casti and A. Karlqvist (eds.) *Complexity, Language, and Life: Mathematical Approaches.* Berlin: Springer-Verlag, pp. 168–281.

Peterson, M. (2019) 'The value alignment problem: a geometric approach', *Ethics and Information Technology*, 21(1), pp. 19-28.

Pentina, I., Hancock, T. and Xie, T. (2022) 'Exploring relationship development with social chatbots: A mixed-method study of replika.' *Computers in Human Behavior*, 140. Available at: https://doi.org/10.1016/j.chb.2022.107600

Place, U. T. (1956) 'Is consciousness a brain process?', *British Journal of Psychology*, 47, pp. 44-50.

Polger, T. W. (no date) 'Functionalism', *Internet Encyclopedia of Philosophy,* ISSN 2161-0002. Available at: https://iep.utm.edu/functism/ (Accessed: 30 August 2023).

Pritchard, H. (2017) 'A BMW Driver Followed His Sat Nav Straight into a River.' *Wales Online,* 6 February 6. Available at: https://www.walesonline.co.uk/news/wales-news/bmw-driver-followed-sat-nav-12564727 (Accessed: 29 March 2019).

Preston, B. (2020) 'Artifact', in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*. Available at: https://plato.stanford.edu/archives/fall2020/entries/artifact (Accessed: 1 October 2022).

Putnam, H. (1963) 'Brains and behavior', in R. J. Butler (ed.) *Analytical Philosophy, Second Series*, Oxford: Basil Blackwell, pp. 211-235.

Putnam, H. (1964) 'Robots: machines or artificially created life?' *Journal of Philosophy*. 61(21), pp. 668–691. Available at: doi:10.2307/2023045

Putnam, H. (1967) 'Psychological predicates', in W. H. Captain and D. D. Merrill (eds.) *Art, mind and religion,* Pittsburgh: University of Pittsburgh Press, pp. 37-48.

Putnam, H. (1973) 'Meaning and reference', *The Journal of Philosophy* 70(9), pp. 699–711.

Putnam, H. (1975) 'The meaning of 'meaning'', in *Language, Mind, and Knowledge: Minnesota Studies in the Philosophy of Science,* 7, pp. 131–193. Available at: https://doi.org/10.1515/9783110871241.110

Quick, T., Dautenhahn, K., Nehaniv, C. L. and Roberts, G. (1999) 'On bots and bacteria: ontology independent embodiment', *Proceedings of the Fifth European Conference on Artificial Life*. Heidelberg: Springer. Available at: https://link.springer.com/chapter/10.1007/3-540-48304-7_45 (Accessed: 1 July 2019).

Raamkumar, A. S. and Yang, Y. (2022) 'Empathetic conversational Systems: A review of current advances, gaps, and opportunities', *IEEE Transactions on Affective Computing*. Available at: https://doi.org/10.48550/arXiv.2206.05017

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S.A. and Botvinick, M. (2018) 'Machine theory of mind'. *Proceedings of the International conference on machine learning*, *PMLR*, 80, pp. 4218-4227. Available at: http://proceedings.mlr.press/v80/rabinowitz18a.html (Accessed: 15 December 2022).

Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. (2018) 'Improving language understanding by generative pre-training'. Available at: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf (Accessed: 12 December 2022).

Radford, A., Wu. J., Amodei, D., Amodei, D., Clark, J., Brundage, M. and Sutskever, I. (2019) 'Better language models and their implications' *OpenAI*. Available at: https://openai.com/blog/better-language-models/ (Accessed: 10 December 2022).

Raftery, B. (2017) 'The surreal comedy bot that's turning AI into LOL' *Wired,* 23 October. Available at: https://www.wired.com/story/botnik-ai-comedy-app/ (Accessed: 11 December, 2022).

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. and Chen, M. (2022) 'Hierarchical text-conditional image generation with clip latents', *arXiv.* Available at: https://doi.org/10.48550/arXiv.2204.06125.

Ramesh, A., Pavlov, M., Goh, G. and Gray, S. (2021) 'DALL·E: Creating images from text' *OpenAI.* Available at: https://openai.com/blog/dall-e/ (Accessed: 1 December 2022).

Rayner, A. (2016) 'Can Google's Deep Dream become an art machine?', Guardian, 28 March. Available at: https://www.theguardian.com/artanddesign/2016/mar/28/google-deep-dream-art (Accessed: 17 December 2022).

Realdreams (2022) Available at: https://www.realdreams.io/ (Accessed: 16 December 2022)

Rea, N. (2019) 'Sotheby's is entering the AI art fray, selling a surreal artwork by one of the movement's pioneers this spring', *ArtNet News.* Available at: https://news.artnet.com/art-world/sothebys-artificial-intelligence-1460332 (Accessed: 1 May 2020).

Rescorla, M. (2020) 'The computational theory of mind', in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy,* Available at:

https://plato.stanford.edu/archives/fall2020/entries/computational-mind/ (Accessed: 30 August 2023).

Riegler, A. (2002) 'When is a cognitive system embodied?' *Cognitive Systems Research,* 3(3), pp.339-348. Available at: https://www.univie.ac.at/constructivism/people/riegler/pub/index.cgi?showabstract=24 (Accessed: 1 July 2019).

Rey, G. (1986) 'What's really going on in Searle's "Chinese room" ', *Philosophical Studies*, 50, pp. 169–85.

Richter, D. (no date) 'G. E. M. Anscombe (1919—2001)', *Internet Encylopedia of Philosophy*. ISSN 2161-0002. Available at: https://iep.utm.edu/anscombe/ (Accessed: 10 December 2022).

Ridler, A. (2017) 'Misremembering and mistranslating: GANs in an art context', *31st Conference on Neural Information Processing Systems (NIPS 2017),* Long Beach, CA, 4-9 December. Available at: http://annaridler.com/gans-in-art (Accessed: 11 December 2022).

Ridley, A. (2002) 'Congratulations, it's a tragedy: Collingwood's remarks on genre', *British Journal of Aesthetics*, 42(1), pp. 52–63.

Rivron, V., Khan, M. I., Charneau, S., and Chrisment, I. (2016) 'Exploring smartphone application usage logs with declared sociological information', *Proceedings of the 2016 IEEE International Conferences on Big Data and Cloud Computing,* pp. 266–273. Available at: https://doi.org/10.1109/BDCloud-SocialCom-SustainCom.2016.49

Robinson, H. (2023) 'Dualism', in E. N. Zalta and U. Nodelman (eds.) *The Stanford Encyclopedia of Philosophy*. Available at: https://plato.stanford.edu/archives/spr2023/entries/dualism/ (Accessed: 30 August 2023).

Robinson, J. (2005) *Deeper than Reason: Emotion and its Role in Literature, Music, and Art,* Oxford: Oxford University Press. Available at: https://doi.org/10.1093/0199263655.003.0009

Rodger, P. (2019) 'World's first robot ARTIST can sketch people from sight - and it's even getting an exhibition', *Mirror,* 11 February. https://www.mirror.co.uk/tech/worlds-first-robot-artist-can-13983066 (Accessed: 1 July 2019)

Romero, A. (2022) 'DALL·E 2, Explained: The Promise and Limitations of a Revolutionary AI', *Towards Data Science*, 16 June. Available at: https://towardsdatascience.com/dall-e-2-explained-the-promise-and-limitations-of-a-revolutionary-ai-3faf691be220 (Accessed: 16 December 2022).

Roose, K. (2022) 'An A.I.-generated picture won an art prize. Artists aren't happy.' *New York Times*, 2 December. Available at: https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html (Accessed: 12 December 2022).

Rowlands, M (2010) *The New Science of the Mind: From Extended Mind to Embodied Phenomenology*, Cambridge, MA: MIT Press.

Rowlands, M., Lau, J., and Deutsch, M. (2020) 'Externalism about the mind', in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy,* Available at: https://plato.stanford.edu/archives/win2020/entries/content-externalism/ (Accessed: 30 August 2023).

Russell, S. (2015). *Value Alignment*. World Economic Forum Ideas Lab. Available at: https://www.youtube.com/watch?v=WvmeTaFc_Qw (Accessed: 9 May 2022).

Russell, S. (2016) 'Should we fear supersmart robots?' *Scientific American*, 314(6), pp. 58-59.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. London: Allen Lane.

Russell, S. (2021) 'Human-compatible artificial intelligence' in S. Muggleton, and N. Chater (eds.) *Human-Like Machine Intelligence.* Oxford: Oxford University Press, pp. 3-23.

Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. 3rd edn. Upper Saddle River: Prentice Hall.

Ryle, G. (2009). *The Concept of Mind: 60th Anniversary Edition.* London: Routledge.

Samuel, S. (2019) 'Artificial intelligence can now make art. Artists, don't panic.' *Vox,* 17 May. Available at: https://www.vox.com/2019/5/10/18529009/ai-art-marcus-du-sautoy-math-music-painting-literature (Accessed: 12 December 2022).

Sandred, O., Laurson, M. and Kuuskankare, M. (2009) 'Revisiting the Illiac Suite–a rule-based approach to stochastic processes', *Sonic Ideas/Ideas Sonicas*, 2, pp. 42-46.

Sapontzis, S. F. (1981) ‘A critique of personhood’, *Ethics*, 91(4), pp. 607-618.

Sarker, I.H. (2021) ‘Machine learning: Algorithms, real-world applications and research directions’, *SN Computer Science.* 2. Available at: https://doi.org/10.1007/s42979-021-00592-x

Sayre, K.M. (1986) ‘Intentionality and information processing: An alternative model for cognitive science’, *Behavioral and Brain Sciences*, 9(1), pp. 121-138. Available at: doi:10.1017/S0140525X00021750

Schlosser, M. (2019) ‘Agency’, in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy.* Available at: https://plato.stanford.edu/archives/win2019/entries/agency/ (Accessed: 10 November 2020).

Schneider, C (2015) ‘The surreal artwork of Artificial Intelligence’, *Mental Floss.* Available at: https://www.mentalfloss.com/article/65458/surreal-artwork-artificial-intelligence (Accessed: 1 May 2020)

Schneider, S (no date) ‘Identity theory’, *Internet Encyclopedia of Philosophy,* ISSN 2161-0002. Available at: https://iep.utm.edu/identity/ (Accessed: 30 August 2023).

Sclafani, R. J. (1971) ‘‘Art’, Wittgenstein, and open-textured concepts’, *The Journal of Aesthetics and Art Criticism,* 29(3), pp. 333-341. Available at: https://doi.org/10.2307/428976

Searle, J. R. (1980) ‘Minds, brains, and programs’, *Behavioral and Brain Sciences,* 3(3), pp. 417–424. Available at: doi:10.1017/S0140525X00005756

Searle, J. R. (1982) ‘Proper names and intentionality’, *Pacific Philosophical Quarterly,* 63(3), pp. 205-225. Available at: https://doi.org/10.1111/j.1468-0114.1982.tb00100.x

Searle, J. (1990) ‘Is the brain’s mind a computer program?’, *Scientific American,* 262, pp. 26-31.

Searle J. R. (1999) ‘The Chinese Room’, in R.A. Wilson and F. Keil (eds.), *The MIT Encyclopedia of the Cognitive Sciences,* Cambridge, MA: MIT Press.

Setiya, K. (2022) ‘Intention’, In E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy.* Available at: https://plato.stanford.edu/archives/win2022/entries/intention/ (Accessed: 15 December 2022).

Shambler, T. (2019) 'Meet the robot taking the art world by storm' *Esquire Middle East.* https://www.esquireme.com/meet-the-robot-taking-the-art-world-by-storm (Accessed: 1 July 2019).

Shane, J. (2020) *AI Weirdness.* Available at: https://aiweirdness.com (Accessed: 12 June 2020)

Shapiro, L. and Spaulding, S. 'Embodied cognition', in E. N. Zalta (ed) *The Stanford Encyclopedia of Philosophy.* Available at: https://plato.stanford.edu/archives/win2021/entries/embodied-cognition/ (Accessed: 18 December 2022).

Shen, S. T (2015) 'The digital generation: Comparing and contrasting smartphone use in the digital age', *Journal of Internet Technology,* 16(1), pp. 121-127. Available at: https://doi.org/10.6138/JIT.2014.16.1.20140509

Siddiqui, J. R. (2022) 'Diffusion Models Made Easy', *Towards Data Science.* Available at: https://towardsdatascience.com/diffusion-models-made-easy-8414298ce4da (Accessed: 1 December 2022).

Simon, J. (no date a) *Ganbreeder.* Available at: https://ganbreeder.app/category/random (Accessed: 29 March 2019).

Simon, J. (no date b) 'Ganbreeder.' *Github.* Available at: https://github.com/joel-simon/ganbreeder (Accessed: 29 March 2019.

Simon, J. (no date c) 'Ganbreeder.app.' *JoelSimon.net.* Available at: https://www.joelsimon.net/ganbreeder.html (Accessed: 28 March 2019).

Simonite, T. (2018) 'When it comes to gorillas, Google Photos remains blind', *Wired*, 1 January. Available at: https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/ (Accessed: 28 March 2019).

Simonton, D. K. (1999) 'Creativity as blind variation and selective retention: Is the creative process Darwinian?', *Psychological Inquiry*, pp. 309-328.

Skirry, J. (2006) 'Descartes: the mind-body distinction' *Internet Encyclopedia of Philosophy,* ISSN 2161-0002. Available at: https://iep.utm.edu/rene-descartes-mind-body-distinction-dualism/ (Accessed: 30 August 2023).

Smart, J. J. C. (1959) 'Sensations and brain processes,' *Philosophical Review*, 68, pp. 141-156.

Smart, J. J. C. (2022) 'The mind/brain identity theory', in E. N. Zalta and U. Nodelman (eds.) *The Stanford Encyclopedia of Philosophy,* Available at: https://plato.stanford.edu/archives/win2022/entries/mind-identity/ (Accessed: 30 August 2023).

Smith, M. N. (2018) 'Political creativity: a skeptical view', in B. Gaut and M. Kieran (ed.) *Creativity and Philosophy.* New York, NY: Routledge

Smuts, A. (2005) 'Are video games art?' *Contemporary Aesthetics*, 3. Available at: https://digitalcommons.risd.edu/liberalarts_contempaesthetics/vol3/iss1/6/ (Accessed: 16 December 2022).

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. and Ganguli, S. (2015) 'Deep unsupervised learning using nonequilibrium thermodynamics', *Proceedings of the 32nd International Conference on Machine Learning,* pp. 2256-2265.

Solaiman, I., Jack Clark and Miles Brundage. (2019) 'GPT-2: 1.5B Release' *Open AI, 5 November.* Available at: https://openai.com/blog/gpt-2-1-5b-release/ (Accessed: 11 November 2019).

Sotheby's (2019) *Memories of Passerby I.* Available at: http://www.sothebys.com/en/auctions/ecatalogue/2019/contemporary-art-day-auction-l19021/lot.109.html (Accessed: 1 November 2022).

Sparrow, R. (2007) 'Killer robots', *Journal of Applied Philosophy*, 24(1), pp. 62-77.

Sprevak, M. (2007) 'Chinese rooms and program portability', *British Journal for the Philosophy of Science*, 58(4), pp. 755–776.

Sprevak, M. (2009) 'Extended cognition and functionalism', *Journal of Philosophy*, 106(9), pp. 503–527.

Stability AI (2022) *Stable diffusion public release*. Available at https://stability.ai/blog/stable-diffusion-public-release (Accessed: 1 December 2022).

Stanciu, M. M. (2015) 'Embodied creativity: a critical analysis of an underdeveloped subject', *Procedia-Social and Behavioral Sciences*, 187, pp. 312-317.

Stecker R. (1997). *Artworks: Definition Meaning Value,* University Park PA: Pennsylvania State University Press.

Steinert, S. (2017). 'Art: Brought to you by creative machines', *Philosophy & Technology*, 30(3), pp. 267-284.

Sternberg, R. J., and Davidson, J.E. (1995) *The Nature of Insight*, Cambridge, MA: MIT Press.

Stock, M. (2019) 'Ai-Da, the humanoid robot artist, gears up for first solo exhibition', *Reuters,* 5 June. Available at: https://www.reuters.com/article/uk-tech-robot-artist-idUKKCN1T61Z1 (Accessed: 1 July 2019)

Stokes, D. (2008) 'A metaphysics of creativity', in K. Stock and K. Thomson-Jones (eds.) *New Waves in Aesthetics,* Basingstoke: Palgrave Macmillan, pp. 105–24.

Stokes, D. (2011) 'Minimally creative thought', *Metaphilosophy,* 42(5), pp. 658-681.

Stokes, D., and Bird, J. (2008) 'Evolutionary robotics and creative constraints', in B. Hardy-Vallée and N. Payette (eds.) *Beyond the Brain: Embodied, Situated, and Distributed Cognition,* pp. 227-245. Newcastle: Cambridge Scholars Publishing.

Stokes, P. D. (2005) *Creativity from Constraints: The Psychology of Breakthrough*, New York: Springer.

Stokes, P. D. (2008) 'Creativity from constraints: What can we learn from Motherwell? from Modrian? from Klee?' *The Journal of Creative Behavior,* 42, pp. 223-236. Available at: https://doi.org/10.1002/j.2162-6057.2008.tb01297.x

Stravinsky, I. (1997) 'Poetics of music', in F. Barron, A. Montuori, and A. Barron (eds.) *Creators on Creating: Awakening and Cultivating the Imaginative Mind.* NY: Putnam, pp. 189-194.

Subagdja, B. and Tan, A-H. (2019) 'Beyond autonomy: The self and life of social agents', *International Conference on Autonomous Agents and Multiagent Systems (AAMAS) Proceedings*. Available at: http://www.ifaamas.org/Proceedings/aamas2019/pdfs/p1654.pdf (Accessed: 1 November 2019).

Sutton, B. (2019) 'An artwork created by AI sold for £40,000 at Sotheby's, failing to generate the fervor that propelled another AI work to sell for 40 times its estimate last year.' *Artnet,* 6 March. Available at: https://www.artsy.net/article/artsy-editorial-artwork-created-ai-sold-40-

000-sothebys-failing-generate-fervor-propelled-ai-work-sell-40-times-estimate-year (Accessed: 11 December 2022).

Symes, C. (1999) 'Writing by numbers: OuLiPo and the creativity of constraints', *Mosaic: a Journal for the Interdisciplinary Study of Literature,* 32(3), pp. 87-107.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D. and Goodfellow, I. (2014) 'Intriguing Properties of Neural Networks', *arXiv*. Available from: https://arxiv.org/abs/1312.6199 (Accessed: 9 May 2022).

Tang, G., Wu, K., Wu, Y., Liao, H., Guo, D. and Wang, Y. (2020) 'Quantifying Low-Battery Anxiety of Mobile Users and Its Impacts on Video Watching Behavior', *arXiv*, Available at: https://doi.org/10.48550/arXiv.2004.07662

Tate (no date) 'Dorothea Tanning, *Eine Kleine Nachtmusik.* 1943', *Tate*. Available at: https://www.tate.org.uk/art/artworks/tanning-eine-kleine-nachtmusik-t07346 (Accessed: 30 April 30 2021).

Tate (2020) *Photorealism.* Available at: https://www.tate.org.uk/art/art-terms/p/photorealism (Accessed: 1 July 2020).

Tavinor, G. (2011) 'Video games as mass art', *Contemporary Aesthetics*, 9. Available at: http://hdl.handle.net/2027/spo.7523862.0009.009 (Accessed: 16 December 2022).

Taylor, K. and Silver, L. (2019) 'smartphone ownership is growing rapidly around the world, but not always equally', *Pew Research Center*. Available at: https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/ (Accessed: 29 July 2019).

Tech Demo Team, Nvidia (2020) 'Synthesizing high resolution images with GANs', SIGGRAPH 2020, 17-28 August, online. Available at: https://developer.nvidia.com/techdemos/video/dcv20 (Accessed: 1 September 2020).The Alan Turing Institute (no date) *Data Science and AI Glossary.* Available at: https://www.turing.ac.uk/news/data-science-and-ai-glossary (Accessed: 1 November 2022)

The Art Newspaper (2018) 'Aican the AI artist: putting the art in to artificial intelligence' *The Art Newspaper,* 25 December. Available at: https://www.theartnewspaper.com/2018/12/25/aican-the-ai-artist-putting-the-art-in-to-artificial-intelligence (Accessed: 1 October 2022).

The Barber Institute of Fine Arts (no date) *Edgar Degas (1834-1917): Dancer Ready to Dance, the Right Foot Forward*. Available at: https://barber.org.uk/edgar-degas-1834-1917-4/ (Accessed: 1 October 2022)

Theis, L., Oord, A.V. D. and Bethge, M. (2015) 'A note on the evaluation of generative models'. *arXiv.* Available at: https://doi.org/10.48550/arXiv.1511.01844

The One Show (2019) BBC One, 12 June, 19:00. Available at: https://www.bbc.co.uk/programmes/m0005xp7 (Accessed: 30 June 2019)

Tilghman, B. R. (1973) 'Wittgenstein, games, and art', *The Journal of Aesthetics and Art Criticism*, 31(4), pp. 517-524. Available at: https://doi.org/10.2307/429325

Tiu, E. (2020) 'Understanding latent space in machine learning' *Towards Data Science,* 4 February. Available at: https://towardsdatascience.com/understanding-latent-space-in-machine-learning-de5a7c687d8d (Accessed: 16 December 2022).

Tollefsen, D. P. (2006) 'From extended mind to collective mind', *Cognitive Systems Research,* 7(2–3) pp. 140–150.

Totschnig, W. (2020) 'Fully autonomous AI.' *Science and Engineering Ethics*, 26(5), pp. 2473-2485.

Tran, L., Alter, R. and Flattum-Riemers, T. (2019) 'Anti-vaxx propaganda is flooding the internet. will tech companies act?' *The Guardian*, 5 March 2019. Available at: https://www.theguardian.com/commentisfree/2019/mar/05/anti-vaxx-propaganda-internet-tech (Accessed: 29 July 2019).

Trazzi, M. (2020) 'more frustrating than compelling. i tried super hard to find a joke but all sentences sound boringly human' [Twitter] 19 April. Available at: https://twitter.com/MichaelTrazzi/status/1251995692421316608 (Accessed: 1 June 2020)

Tresset, P. (2011) *New Work.* Available at*:* https://patricktresset.com/new/project/tender-pixel-2011/ (Accessed: 11 December 2022).

Tromp, H. (2010) *A Real Van Gogh: How the Art World Struggles with Truth.* Amsterdam: Amsterdam University Press.

Trujillo, G. M. (2022) 'AI art is art.' *Aesthetics for Birds*, 2 November. Available at: https://aestheticsforbirds.com/2022/11/02/ai-art-is-art/ (Accessed: 10 December 2022).

Turing, A. M. (1936) 'On computable numbers, with an application to the Entscheidungsproblem', *Journal of Mathematics*, 58(345-363), pp. 230-265.

Turing, A. (1950) 'Computing machinery and intelligence', *Mind,* 59(236), pp. 433–60.

Unger, M. J. (2009) *Picasso and the Painting that Shocked the World*. London: Simon & Schuster.

University of Exeter School of Psychology (2009) 'The psychology of scams: Provoking and committing errors of judgement', *Office of Fair Trading*. Available at: https://ore.exeter.ac.uk/repository/bitstream/handle/10871/20958/OfficeOfFairTrading&hx0025;202009.pdf?sequence=1&isAllowed=y (Accessed: 1 July 2020).

Valstar, M. F., Colton, S. and Pantic, M. (2008) 'Emotionally aware automated portrait painting demonstration', *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition,* Amsterdam, Netherlands, 17-19 September. Available at: https://doi.org/10.1109/FG13733.2008

Van Arman, P. (2018) 'The first sparks of artificial creativity', *Cloudpainter.* Available at: https://www.cloudpainter.com/ai-art-blog/2018/1/27/the-first-sparks-of-artificial-creativity (Accessed: 1 November 2022).

Vincent, J. (2018) 'How three French students used borrowed code to put the first AI portrait in Christie's' *The Verge,* 23 October. Available at: https://www.theverge.com/2018/10/23/18013190/ai-art-portrait-auction-christies-belamy-obvious-robbie-barrat-gans (Accessed: 1 June 2020).

Voidlogic (2018) 'You've made a very successful horror image generator, call the developers of Resident Evil and Silent Hill', [Twitter] 27 March. Available at: https://twitter.com/voidlogic/status/979050000876523521 (Accessed: 1 June 2020).

Waddington, Simon (2018) 'That's errr...a bit terrifying...' [Twitter] 27 March. Available at: https://twitter.com/SimonWad/status/978747515263705089 (Accessed: 1 June 2020).

Walczak, S. and Cerpa, N. (2003) 'Artificial neural networks', in R. Meyers (ed.) *Encyclopedia of Physical Science and Technology.* 3rd edn. Cambridge, MA: Academic Press Inc.

Walton, K. L. (1970) 'Categories of art', The Philosophical Review, 79(3), pp. 334-367. Available at: https://doi.org/10.2307/2183933

Walton, K. L. (1984) 'Transparent Pictures: On the Nature of Photographic Realism', *Critical Inquiry,* 11(2), pp. 246-277. Available at: www.jstor.org/stable/1343394

Walrand, J. and Varaiya, P. P. (2000) 'Control of networks: mathematical background', in *High-Performance Communication Networks*. Cambridge, MA: Morgan Kaufmann.

Wang, P. (no date, a) *This Cat Does Not Exist.* Available at: https://thiscatdoesnotexist.com (Accessed: June 12, 2020).

Wang, P. (no date, b) *This Horse Does Not Exist.* Available at: https://thishorsedoesnotexist.com (Accessed: June 12, 2020).

Wang, P. (no date, c) *This Person Does Not Exist.* Available at: https://thispersondoesnotexist.com (Accessed: June 12, 2020).

Waskan, J. (no date) 'Connectionism', *Internet Encyclopedia of Philosophy,* ISSN 2161-0002. Available at: https://iep.utm.edu/connectionism-cognition/ (Accessed: 30 August 2023).

Webster, R., Rabin, J., Simon, L. and Jurie, F. (2019) 'Detecting overfitting of deep generative networks via latent recovery', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 11273-11282. Available at: https://doi.org/10.48550/arXiv.1901.03396

Weidinger, L. (2018) 'Can artificial intelligence replicate human mental processes?' *ResearchGate,* 2018. Available at: https://www.researchgate.net/publication/329092847_Can_artificial_intelligence_replicate_human_mental_processes (Accessed: 29 July 2019).

Weiner, K. (2018) 'Can AI create true art? And if it can, what are the implications for the future of creativity?', *Scientific American,* 12 November. Available at: https://blogs.scientificamerican.com/observations/can-ai-create-true-art/ (Accessed: 12 December 2022).

Weitz, M. (1956) 'The Role of Theory in Aesthetics', *Journal of Aesthetics and Art Criticism,* 15, pp. 27–35.

West, J. (2016) 'Microsoft's disastrous Tay experiment shows the hidden dangers of AI', *Quartz,* 2 April. Available at: https://qz.com/653084/microsofts-disastrous-tay-experiment-shows-the-hidden-dangers-of-ai/ (Accessed: 12 November 2019).

Wheeler, M. (2005) *Reconstructing the Cognitive World: The Next Step*, Cambridge, MA: MIT Press.

Wheeler, M. (2010) 'In defense of extended functionalism', in R. Menary (ed.), *The Extended Mind*, Cambridge, MA: MIT Press, pp. 245-270.

Wheeler, M. (2018) 'Talking about more than heads', in B. Gaut and M. Kieran (eds.) *Creativity and Philosophy*, New York, NY: Routledge.

Wikimedia Commons (2018) *Edmond de Belamy,* Available at: https://commons.wikimedia.org/wiki/File:Edmond_de_Belamy.png (Accessed: 18 December 2022).

Williams, J., Fiore, S.M. and Jentsch, F. (2022) 'Supporting artificial social intelligence with theory of mind', *Frontiers in Artificial Intelligence*, 5(750763). Available at: doi:10.3389/frai.2022.750763

Wilson, G. and Shpall, S. (2012) 'Action', In E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy.* Available at: https://plato.stanford.edu/archives/win2016/entries/action/ (Accessed: 16 December 2022).

Wiltsher, N. (2018) 'Feeling, emotion and imagination: in defence of Collingwood's expression theory of art', *British Journal for the History of Philosophy*, 26(4), pp. 759-781.

Wimsatt, W. K. and Beardsley M. C. (1946) 'The Intentional Fallacy', *Sewanee Review,* LIV, pp. 468–488.

Wittgenstein, L. (2009). *Philosophical Investigations,* 4th edn. Translated from German by G. E. M. Anscombe, P. M. S. Hacker and J. Schulte. Sussex: Wiley-Blackwell.

Wollheim, R. (1974) 'On an alleged inconsistency in Collingwood's aesthetic', *On Art and the Mind.* Cambridge, MA: Harvard University Press.

Wreen, M. (2002) ‘Forgery’ *Canadian Journal of Philosophy,* 32(2), pp. 143-166. Available at: doi:10.1080/00455091.2002.10716515.

Wreen, M. (2014) ‘Beardsley’s Aesthetics’, In E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy.* Available at: https://plato.stanford.edu/archives/win2014/entries/beardsley-aesthetics/ (Accessed: 15 December 2022).

Wright, A. (2022) ‘Stable diffusion 2 is here, but not everyone’s happy’. *How-to Geek, 25 November.* Available at: https://www.howtogeek.com/850931/stable-diffusion-2-is-here-but-not-everyones-happy/ (Accessed: 30 November 2022).

Wykowska, A., Chaminade, T. and Cheng, G. (2016) ‘Embodied artificial agents for understanding human social cognition’, *Philosophical Transactions of the Royal Society B,* 371(1698), pp. 1-9. Available at: https://doi.org/10.1098/rstb.2015.0375

XiaoIce (2018) *XiaoIce – full-duplex voice sense demo*, 14 September 2018. Available at: https://www.youtube.com/watch?v=z3jqIGT-kmg (Accessed: 29 March 2019)

Xu, Q., Huang, G., Yuan, Y., Guo, C., Sun, Y., Wu, F. and Weinberger, K. (2018) ‘An empirical study on evaluation metrics of generative adversarial networks’, *arXiv*. Available at: https://doi.org/10.48550/arXiv.1806.07755

Yudkowsky, E. (2008) ‘Artificial intelligence as a positive and negative factor in global risk’, in N. Bostrom, and M. Cirkovic (eds.) *Global Catastrophic Risks*. Oxford: Oxford University Press, pp. 308-345.

Yudkowsky, E. (2013) ‘Five theses, two lemmas, and a couple of strategic implications’, *Machine Intelligence Research Institute*. Available at: https://intelligence.org/2013/05/05/five-theses-two-lemmas-and-a-couple-of-strategic-implications/ (Accessed: 9 May 2022)

Yudkowsky, E. (2016). *The AI alignment problem: why it is hard, and where to start.* Symbolic Systems Distinguished Speaker Series. 5 May 2016, Stanford University. Available at: https://intelligence.org/stanford-talk/ (Accessed: 9 May 2022)

Zangwill, N. (1995) ‘Groundrules in the Philosophy of Art,’ *Philosophy,* 70: 533–544. Available at: https://www.jstor.org/stable/3751081

Zeki, S. and Ishizu, T. (2013) 'The "visual shock" of Francis Bacon: an essay in neuroesthetics', *Frontiers in Human Neuroscience,* 7(850). Available at: https://doi.org/10.3389/fnhum.2013.00850 (Accessed: 1 June 2020)

Zhou, L., Gao, J., Di, L. and Shum, H-Y. (2018) 'The design and implementation of XiaoIce, an empathetic social chatbot', *arXiv.* Available at: https://doi.org/10.48550/arXiv.1812.08989

Ziemke, T. (2001) 'Are robots embodied', *First international workshop on epigenetic robotics Modelling Cognitive Development in Robotic Systems*, 85, pp. 701-746. Available at: https://www.semanticscholar.org/paper/Are-Robots-Embodied-Ziemke/ce66e4006c9948b7ed080c239540dfd0746ff639 (Accessed: 1 July, 2019).

Ziemke, T. (2003) 'What's that thing called embodiment?' *Proceedings of the annual meeting of the cognitive science society*, 25(25). Available at: https://pdfs.semanticscholar.org/8eb4/b841ac4107602ab48b6196489900386e9fd9.pdf (Accessed: 1 July 2019).

Ziff, P. (1953) 'The task of defining a work of art', *The Philosophical Review*, 62(1), pp. 58-78.

Zucconi, A. (2018) 'Understanding the technology behind DeepFakes', *Alan Zucconi.* Available at: https://www.alanzucconi.com/2018/03/14/understanding-the-technology-behind-deepfakes/ (Accessed: 1 July 2020).

Appendix I - AI Art: Key Points in History

1957 - L.A. Hiller and L.M. Isaacson utilise random computer processes in composing the string quartet the *Illiac Suite* (Sandred, Laurson & Kuuskankare 2009)

1972-2010 - Harold Cohen develops various versions of Aaron, a computer programme that could draw (Cohen, 1995)

1990s - Boden writes extensively about AI creativity (and continues on into the 2010s) (Boden, 1994; 1995; 1998)

2011 - Patrick Tresset exhibits three "clumsy robots" which can draw faces it finds from scanning the room with a robotic arm (M Brown, 2011; Tresset, 2011).

2013 - The emotional Painting Fool is exhibited in Paris (Colton et al., 2014).

2014 - Generative Adversarial Networks (GANs) are invented by Goodfellow et al. (2014)

2015 - DeepDream is developed (Mordvintsev, Olah & Tyka, 2015).

2015 - Colton and his team integrate vision into the Painting fool (Colton & Halskov, 2015)

2016 - AlphaGo makes the famed 'move 37' (*AlphaGo,* 2017)

2016 - AIVA, an AI start-up aiming to develop 'emotional' soundtracks through AI is launched (Moura, 2017)

2017 - GANs for art start proliferating - for example, GANgogh, an adapted GAN, is applied to artworks (Jones & Bonafilia, 2017), Robbie Barrat modifies a GAN (Barrat, 2017) that will later be used by the art collective Obvious to produce the first AI artwork sold at auction (Flynn 2018), Anna Ridler presents and Neurips (Ridler, 2017).

2017 - Creative Adversarial Network (CAN) first developed (Elgammal et al., 2017).

2017 - Botnik, an 'AI-assisted humour application' (Raftery, 2017)

2018 - GPT (Generative Pre-Training), an impressive text-generating AI system (with several iterations since) is developed by researchers at OpenAI (Radford et al., 2018)

2018 - GAN-generated portrait *Edmond de Belamy* is sold at Christie's for $432,500 (Christie's, 2018a)

2018 - Cloudpainter wins robot art competition (Cloudpainter, no date).

2019 - In March Sotheby's sells their first AI piece, Mario Klingemann's *Memories of Passersby I,* for (the more modest sum of) £40,000 (Sutton, 2019)

2019 - In February OpenAI launches GPT-2 to much fanfare (Radford et al., 2019)

2019 - AI Dungeon, a text-based digital game which uses text generated by GPT-2 is launched (Hitt, 2019).

2020 - GPT-3 is launched (T Brown et al., 2020)

2021 - DALL-E, an impressive text-to-image generator is launched by OpenAI (Ramesh et al., 2021)

2022 - April - DALL-E 2 previews to some users (Jang & OpenAI, 2022)

2022 August - Stable Diffusion, a rival text-to-image AI system is made widely available to the public (Stability.ai, 2022)

2022 - September - AI art wins state fair, garnering global attention (Metz, 2022b)

An explosion of AI for artistic purposes occurred in 2017-2018 (when I started this topic for my MA, and proposed this PhD project). Notable artists who I have not listed here but have been influential in this field include: Memo Atken, Helena Sarin, Refik Anadol, Gene Kogan, and Sougwen Chung. For a comprehensive list of artists using AI, visit AIArtists.org.