Design (Non) Fiction: Deconstructing/Reconstructing the Definitional Dualism of AI

Keywords: *Artificial Intelligence, Design Fiction, Ontology, World Building, Narrow AI, Artificial General Intelligence, More-Than Human Centred Design*

Abstract

2001: A Space Odyssey (Kubrick, 1968) speculates on humanities technological ascension through the exploration of space and the ultimate transcendence of humanity galvanised by the invention of AI. Every detail of this portrayal was an exercise in World Building, with careful considerations of then state-of-the-art technology and informed predictions. Kubrick's speculative vision is comparative to the practice of Design Fiction, by suspending disbelief and leveraging a technologies emergence to question the future's sociotechnical landscape and its ramifications critically. *Discovery's* AI system, *Hal9000*, is a convincing speculation of intelligence with Kubrick's vision showcasing current and long-term aims in AI research. To this end, *Hal9000* uniquely portrays Artificial General Intelligence (AGI) underpinned by visualising 'narrow' AI subproblems; thereby, simultaneously highlighting then current research agendas within AI and manifesting them into the aspirational research agenda of human-computer symbiosis. As a result of Kubrick's mastery in suspending a viewer's disbelief despite portraying a particular reality for AI, and humanities fascination with artificial life, the term AI simultaneously refers to the grand vision of AGI as well as relating to the contemporary reality of narrow AI. This confusion, along with establishing AI's ontology, are current challenges that need addressing to create effective and acceptable realisations of AI. This paper responds to the ontological confusion by reviewing and comparing Kubrick's speculative methodology to the practice of Design Fiction by unpacking *Hal9000* as a diegetic prototype while defining the active threads of 'AI's Definitional Dualism'. The paper will also present a Design Fiction submerged in the reality of narrow AI and the adoption of a More-Than Human Centred Design approach to address the complexity of AI's ontology in alternative ways. Finally, this paper will also define the importance of researching the semantics of AI technology and how film and Design Fiction offer a discursive space for design research to transpire.

1.Introduction

Humanity has been fascinated with Artificial General Intelligence (AGI) for millennia: from ancient mythology with *Talos* (Mayor, 2018) to modern-day narratives, with revolutionary science fictions such as *Metropolis* (Lang, 1927) and *2001: A Space Odyssey* (Kubrick, 1968); the pioneering research with Turing's seminal question - 'can machines think?' (Turing, 1950); to the enthralling pursuit of transcendence with Moravec's AGI driven evolutionary eschatology (Moravec, 1988). While the prospect of AGI's is compelling, they are in stark contrast with the reality of Artificial Intelligent (AI) technologies currently in widespread use. The reality of AI technology, perhaps viewed as mundane in comparison, are commonly referred to as narrow AI and operate by completing specific singular tasks. Narrow AI, frequently operating through Machine Learning (ML), helps augment a range of day-to-day activities such as shopping, dating, television recommendations and more problematically are increasingly involved in hiring decisions and prison sentencing, positioning algorithmic decision making as an emerging governing power (Angwin, et al., 2016; Bridle, 2018; O'Neil, 2016; Zuboff, 2019). The current popularity of AI, after its decline during the last of what are commonly referred to as the 'AI Winters', during which interest in AI wains when it fails to live up to expectations, has been catalysed by 'big data', cheap processing power and advancements in algorithmic techniques. While such AI systems can be proficient at recognising patterns in data and then using them to perform predictions and classifications, they fall well short of what we envision as 'thinking' machines. Despite this, any discussions of AI with non-AI-experts often ends up discussing the thinking machines of AGI. The dichotomy between these two views of AI has been defined as the 'Definitional Dualism of AI' (Lindley et al., 2020), highlighting the misconceptions between AI as materialised in film, media and advertisement campaigns and the actual AI we might experience in our everyday lives.

This paper aims to establish a clearer ontology of AI to develop alternate approaches to design with and for AI technologies. A number of challenges hinder the formation of an explicit ontology of AI, to list a few for context: the bedazzling, though unavailing, research into achieving AGI that eclipses current narrow AI research; the tendency to anthropomorphise AI and consider AI as wielding 'human intelligence'; the technical perplexity of AI, and evolving definitions and knowledge in the AI sector. To this end, this paper will accentuate and reflect on AI's Dualism by forcibly separating the two prominent pillars of AI – AGI and narrow AI – using 2*001: A Space Odyssey* (Kubrick, 1968) (henceforth, simply referred to as '*2001*'), as both AI pillars are uniquely presented and

intertangled in this film, which paradoxically mirrors the current dominant perceptions of AI's ontology. This 'operation' will be achieved by drawing parallels between Kubrick's speculative World Building approach with the practices of Design Fiction. In this manner, we will use Design Fiction practices as tools to help locate and dissect the film for narrow AI and AGI manifestations. By exposing the occurrences of AI's Definitional Dualism, the focus can be redirected towards narrow AI and the true scope of AI's challenges. In the second part of the paper, we will present a More-Than Human Centred Design approach to move beyond the dominant anthropocentric perspective of technology to develop a conversant perspective of AI that diverges away from the present enigma AI is.

Many of the current AI challenges arise from 'networkification'(Pierce & DiSalvo, 2017) in that AI is increasingly entangled within a plethora of smart networked applications, services and software; in some cases, operating outside human understanding of how they function. Interestingly, *2001's* speculations supported hardware applications, overlooking the huge impact and malleability of networks and software (Stork, 1997) and, therefore, data distribution. Data is a vital component of AI's implementation through training, and as a result, AI's reflect the training data. To this end, in recent years, many data sets have been publicly (Angwin, et al., 2016) and academically (Amershi et al., 2015) hailed as biased, inaccurate and under representative of the complexities regarding their users and the vivaciousness of things.

Directly affected by AI's Dualism, and where most challenges could be minimalised and solutions yielded, is the notion of AI legibility and explainability, which is often hampered by common practices that intentionally obfuscate the operation of AI in products and services for various reasons, for example: the process of simplifying the products operation for the user (Norman, 1998); concealing corporate intellectual property (Burrell, 2016), which in some cases doubles up as a deceptive move to collect substantial data from the user without explicit consent (Zuboff, 2019).

Nonetheless, the reality of AI, which is narrow, and its problems are employed to govern and alter the world through coded technology via the mundanity of smart home applications. It is therefore vital that we develop a better ontology for AI to develop the foundations for research to cultivate AI as a material to design with, by utilising More-than Human Centred (Coulton & Lindley, 2019b) approaches that reflect the contemporary complex contexts such as AI. The structure of this paper is as follows; first, we will frame how we engage with the

speculative design practice of Design Fiction, thereby setting the critical lenses and approach for unpacking *Hal9000* (henceforth, simply referred to as 'Hal') as a Diegetic Prototype to uncover Hal's speculative narrow and AGI architecture. After Hal's analysis, we will present a More-Than Human Centred approach to perceive AI as a material which we will then utilise when 'Design Fiction-ing'. The final section will unpack a Design Fiction created as a research probe and testbed for research into legible AI, which exclusively exemplifies narrow AI contrasting against *2001's* speculative visions.

2. Design Fiction as World Building

Design Fiction is still in its formative years where the field has been described as 'enticing and provocative …yet it still remains elusive'(Hales, 2013). This statement reflects the current range of contending theories, understandings and approaches leading to ambiguity, however creating opportunities for new methods to be established and discussions of how to practice Design Fiction. Though, while the 'means' and method of practice are varied, the 'goal' of Design Fiction is certain (Coulton & Lindley, 2017) – the creation of a fictional world as a discursive space (Dunne & Raby, 2013; Lindley, 2016). To make our position and framework clear, we advocate for the theory of Design Fiction as 'World Building' (Coulton, et al., 2017). To understand this method, the following will clarify the theory that supports a World Building approach by reviewing Design Fiction's brief history.

The term Design Fiction was coined by the science fiction author Sterling while describing the influence design thinking had on his writing, noting that 'Design fiction reads a great deal like science fiction; in fact, it would never occur to a normal reader to separate the two' (Sterling, 2005, p. 30). Sterling further stipulated that science fiction invokes 'grandeur' and perhaps 'hocus-pocus' visions of science, whereas Design Fiction is 'hands on', 'practical' and plausible with the unique ability of getting to the core and 'the glowing heat of the techno-social conflict' (Sterling, 2005, p. 30). Sterling went on to advocate that Design Fiction is 'the deliberate use of diegetic prototypes to suspend disbelief about change' (Bosch, 2012) and has since become the oft-quoted theoretical underpinning for the field. Kirby coined Diegetic Prototypes for the practice of filmmakers and science consultants to produce cinematic depictions of future technologies, where the term diegesis relates to the traditional concept of presenting an interior view of a fictional world (Kirby, 2010). Kirby's theory of Diegetic Prototypes was highlighted, along with other theories, by Bleeker in his influential and catalytic essay on Design Fiction, as a central methodology, noting the film

Minority Report (Spielberg, 2002) as a compelling example of using Diegetic Prototypes. Sterling's thinking regarding Diegetic Prototypes owes much to Bleeker's thesis, though as Sterling also defined, in the same sentiment, Design Fiction 'tells worlds rather than stories' (Bosch, 2012). An important point to note is that the emphasis on 'story' can 'stifle' the flexibility of Design Fiction as an approach by adhering to 'genre conventions' (Coulton, et al., 2017). A complete review of the intricacies of 'narratology' in practising futurology is beyond the scope of the paper; however, to provide some clarity on the matter, Raven and Elahi specify the 'story is not the world' (2015). Rather, the aim of a Design Fiction is to depict a thing belonging to a contextual world, and to *tell* worlds is the act of narrating; therefore, Design Fiction is a narrative form that evades storytelling as a sequence of events in time and space. These worlds are narrated with a 'rhetorical intentionality' (Coulton, et al., 2017) by their designers, and the creation of rhetoric within a world rather than through the planned outline of a story enables those engaging with the world to explore that rhetoric rather than being forced down a 'prescribed path' (Coulton, et al., 2016).

In practice, the act of Design Fiction as World Building is the collection of artefacts, that when viewed together, build a fictional world (Coulton, et al., 2017). A cognitive dissonance is generated between the world of the design and the world in which the audience exists, enabling Design Fiction to achieve 'cognitive estrangement' (Suvin, 1972) (conceptual or temporal break with the viewer's reality) that gives it its rhetorical power (Raven & Elahi, 2015). In Summary, the designed artefacts define the fictional world, and in a 'lemniscate way,' the fictional world empowers the prototyping platform for the very designs that define it (Figure 1). To assist with understanding this approach to Design Fiction, two metaphors aid in understanding how individual artefacts relate to the conceived world (Coulton, et al., 2017). The first requires the Design Fiction world to be imagined as a distinct entity, where the overall shape of that world can be seen, though the complex internal structure is hidden. What can be seen are 'entry points' into the internal structure, where each artefact takes on the role as a metaphorical entry point into the fictional world (Figure 2). The second metaphor, which works with the first, considers shifting scale, inspired by Charles and Ray Eames' film *Powers of 10* (Eames, 1968), with each artefact representing the fictional world at different scales (see Figure 2 also). Now that we have the building blocks for building fictional worlds, we need to consider the type of future that is being represented or designed.

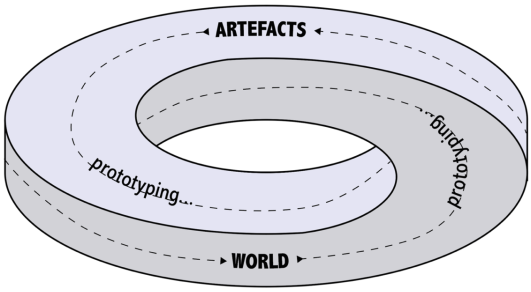

Figure 1: This diagram aids in communicating how both world building and diegetic prototypes help synthesise one another (Coulton et al., 2018).

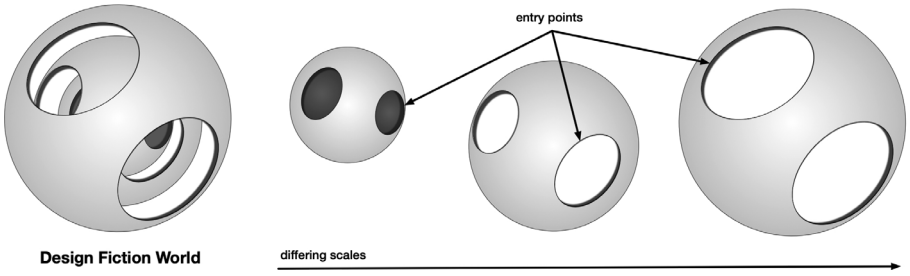

Figure 2: Artefacts at different scales create a richer and more detailed fictional world (Ibid).

2.1 Design Futures

Design is an inherently futurist activity — planning, sketching and prototyping things that do not exist and – simultaneously – considering the future is a fundamental part of designing. When considering the future, Voro's 'future cone' (2003) is often utilised, a taxonomy of scenario qualifiers to mediate the types of futures, which are probable, plausible, possible and preferable (the 4 P's). The futures cone, however, has been criticised as it's qualifiers, for the different ranges of future possibilities promotes more questions about what the meaning of the qualifiers are, with some advocating that the cone has missing qualifiers. For example, designers have questioned the preferable qualifier, as a designer's role is always to consider preferable, as much as a question through which to consider their own biases and not just an aim for a particular design (Coulton, et al., 2016). Bowen argues that preferable promotes 'elitist views of a 'better world' that society should aspire towards' (2010). While Dunne and Raby advocate for the preferable qualifier, they do question in the same sentiment as Bowen asking, 'what does preferable mean, for whom, and who decides' (2013, p. 4). Dunne and Raby circumvent this notion by promoting that Speculative Design, Critical Design and Design Fiction are concerned with 'not to show how things will be but to open up a space for discussion' (Ibid, p. 51).

Nevertheless, it has been contended if showing a singular future vision under the preferable banner be the best way to stimulate discussion. A proposition for Design Fiction to be an effective practice and research tool is to consider and present multiple futures to develop more 'representative notions' of what preferable may be and cater towards a more comprehensive and varied outlook on the future (Coulton & Lindley, 2017). As well as questioning the qualifiers, the original futures cone is often added to and adapted, to capture the possibilities of the many 'variations or blendings' to be found, or even 'behind or beneath' (Raven & Elahi, 2015) the present future's qualifiers, to consider futures beyond what we can imagine easily. Examples range from 'Wildcards' for low probability events (Voros, 2017), 'black swans' (Taleb, 2007) for unclassifiable events (Voros, 2017), 'impossible' (Coulton, et al., 2016) for concepts beyond scientific knowledge and at the moment considered fantasy, although useful to consider the world (Gualeni, 2015). Further still is that the cone fails to acknowledge the influences of the past (Coulton, 2020) and fiction (Gonzatto et al., 2013) has on our perception of time. McLuhan famously wrote, 'We look at the present through a rear-view mirror. We march backwards into the future' (Fiore & McLuhan, 1967, p.74-5). This idea reminds us there is no universally accepted view of the past, present or future, as individuals assemble their own 'subjective' (Raven & Elahi, 2015) reality (Law & Urry, 2004). Reflecting on this, we adopt a futures 'cone' that reflects some of these considerations (Coulton & Lindley, 2017; Coulton, 2020) and acknowledge that it is open to being adapted based on new insights, reflecting the intricacy of considering time and futures (Figure 3).

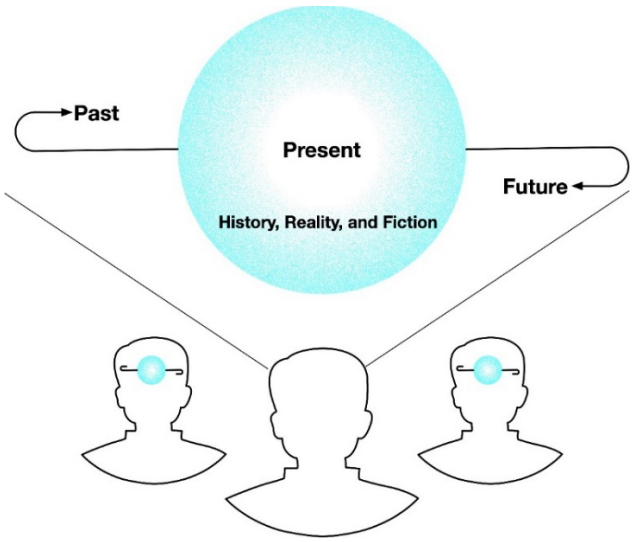

Figure3: This futures 'cone' has been adapted and integrates Gonzatto, van Amstel, Merkle, and Hartmann's (2013) research, whose hermeneutic model represents the 'interpreted present' as an interplay between past, future, reality and fiction.

2.2 Rendering Emerging Technologies as Mundane

In practice, extrapolating technologies from the present along plausible trajectories is the modus operandi when building Design Fiction worlds (Auger, 2013; Blythe & Encinas, 2016; Coulton, et al., 2016; Coulton & Lindley, 2017), which strengthen the design by immersing it in just enough reality to create opportunities for discourse (Figure 4). Part of this consideration is to render these futures as mundane taking inspiration from science fiction, popular media, futuristic tropes, memes and recognisable forms coming out of Silicon Valley (Coulton & Lindley, 2017). Examples of this include technology companies' product videos, device documentation, manuals and patents to blend in with our lived experience. Another aspect to draw upon is the notion of 'Vapourware' and 'Vapour-worlds', terms used to describe material produced by commercial entities and organisations to assert themselves and the products they make as integral parts of the future (Ibid). These vapour-visions and constructive Design Fictions have a knack of representing technologies as if they are domesticated and mundane, exploring futures in a 'subtle and situated way' (Coulton, et al., 2017), engaging cognitive dissonance and striking at the core of the technosocial conflict.

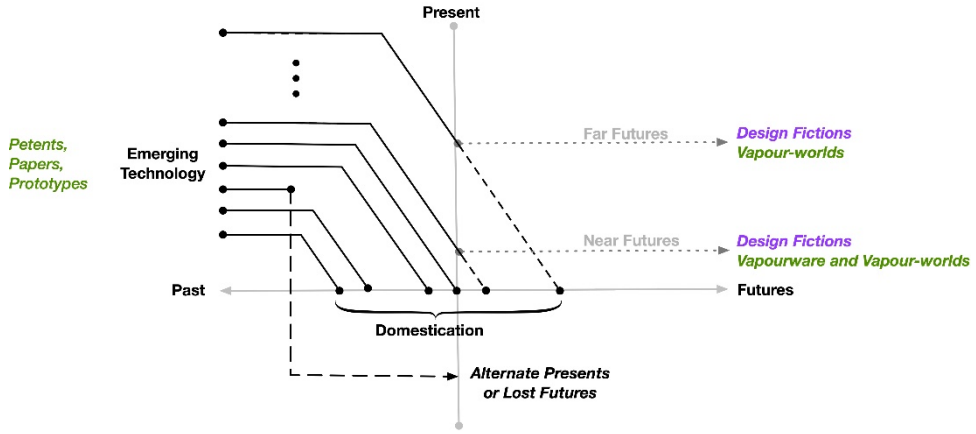

Figure 4: Adapted from Auger (2013) whose original diagram emphasised the lineage of technological products and imagined developments of an emerging technology. This diagram now caters speculatively for a product's entire lifecycles from patents to vapourware (Coulton & Lindley, 2017).

At this point in the paper, now that we have framed our viewpoint and approach to Design Fiction, we will use this as a critical lens to unpack *2001* as a World Building exercise. Secondly, we will use these tools to expose Hal as a vision of AI that conforms to both pillars of AI's Dualism.

3. Examining Hal's Definitional Dualism

In 1968 one film critic called *2001* 'the best-informed dream ever,' (Champlin, 1968) correlating to the widely known fact that Kubrick and author Arthur C. Clarke consulted many scientists, both in academia and industry, in an effort to extrapolate and build a plausible future world. Every detail of the film was scientifically considered: from the hibernation pods influenced by scientific research of inducing hibernation in non-hibernating animals; to the space stewardesses' cushioned space hats for zero-gravity; and the frequently cited Diegetic Prototype — the zero-gravity toilets with the lingering shot over the recognisable form of a how-to-guide, foreseeing and capturing the mundanity of the situation. In addition, this care and attention extended to Hal, which along with representing the then-current AI considerations in research, Hal was perhaps, as many have pointed out, the most emotional and responsive character in the film. To maintain theoretical keystones of plausible futures, Kubrick commissioned the largest computer company, at the time of production, *IBM,* to design and construct speculative interfaces, control panels, consoles and the AI system (Frayling, 2016). A calculated method to maintain credibility and authenticate the speculative concepts by incorporating a known leading manufacturer of computer technology, a parallel of the ideology of Vapourware. *IBM's* proposed concepts was a supercomputer the size of a room; auspiciously, Kubrick deemed the concept not a plausible extrapolation, and behind the times, as rival companies, *Motorola* and *Raytheon*, were exploring miniaturising technology. It was fortuitous that *IBM* was taken off the Hal project, with Hal's malfunction, this once proposed Vapourware to promote the company would have ultimately affected *IBM*'s credibility with consumers.

In *Hal's Legacy* (Stork*,* 1997*)*, prominent AI scientists reflect on Hal, and the effect this palpable vision of AI had on their work, as this vision of AI was a distinctive contradiction to most of Hollywood's AI-cyborg portrayals. Hal on the other hand is not a human form with cyborg features but is situated in an evolved 'disciplinarily machinery' of AI (Mateas, 2006); a result of a plausible extrapolation from then-current lines of research and a visualisation of future AI systems. Nevertheless, the following will highlight that several extrapolations reside in the AGI pillar of AI's Definitional Dualism and perhaps in the 'impossible' qualifier of an adapted futures cone.

The very nature and intricacies of AI functions, operations and architecture are intangible. To consider Hal as a Diegetic Prototype means that we have to venture, frequently, beyond the

physical nuances by anchoring the internal functions and architecture as Diegetic Prototypes and entry points into this future world. To some extent, the film's script and plot will be used to help establish moments of Hal's AI's Dualism, which goes against the grain of Design Fictions reflecting on story and genre conventions. However, Hal is a unique Diegetic Prototype, a character and also an agent operating within the future world that reflects back the depths of Hal's operational remit. To help us take a more informed view of AI and a theoretical frame for Hal, the following section will be an overview of AI research and a glimpse at the numerous theories in the AI field.

3.1 The intricacies of AI research

Initial AI research focused on a 'Classical' approach, forming the early stages of AI research in the 1950s and '60s. The 'Behavioural' approach to AI formed after the 'AI Winter' of the '70s believing that AI would develop out of the behavioural and embodied physical interactions with the world (Brooks, 1991). The ultimate goal of both research strategies was to fabricate an AI to copy and mimic human intelligence, which, by the very definition, undermines what intelligence is if a machine is simply copying. The Classic approach attempted to understand the human brain from the outside, similar to a psychiatrist's methodology, and subsequently integrate that function in a machine, a 'top-down' approach. Early research successfully formed an early depiction of intelligence as reasoning and decision clauses and established the basis for the IF (condition) THEN (conclusion) statements within computing (Warwick, 2012). Hal was conceived during the height of the Classical approach to AI and accurately represents this vision (Mateas, 2006), engaging with the world 'bodiless', with unsettling reaction shots (the frequency of which increases throughout the film) to impress a form of internal 'mental' processing occurring (Figure 5). Hal does have access to sensory-motor applications such as microphones, cameras, and system controls, whereby the audience are introduced to Hal's absolute control and omnipresent existence on board the *Discovery* by showcasing the myriad of apparatus distributed throughout the ship. However, as these apparatuses are separate mechanical elements to Hal's internal machinery, Hal is not considered to fall into the Behavioural AI category (Ibid), as AI Behaviouralists believe that the shape and configuration of a body have a profound effect on the mind.

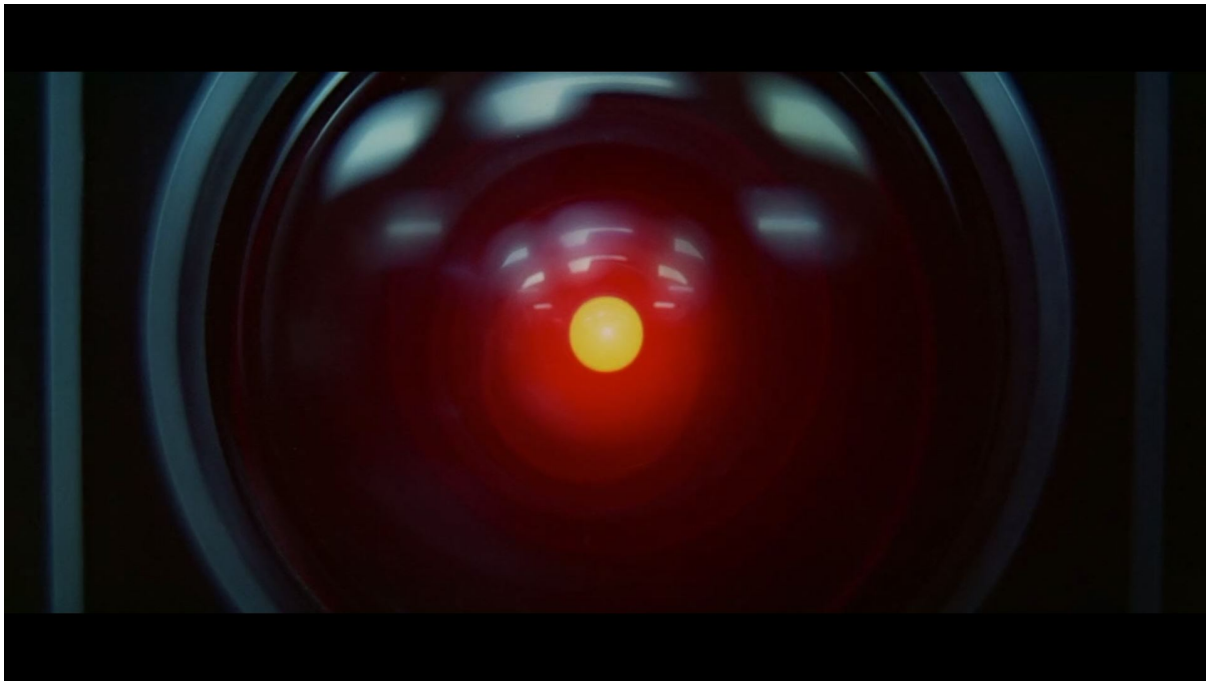

Figure 5: In this scene Hal explains that humans are more likely to be the source of an error rather than the AI system, eliciting many questions about human and machine natures (Kubrick, 1968,1:22:32).

Alternative research avenues continue to develop new approaches in AI technology. Recent years have seen the development of a ‘modern’ approach to AI research, overturning the ‘classic’ methodologies to comprehend AI’s problems with a ‘bottom-up’ approach. This approach is, in fact, a derivative of a biological brain and its ability to evolve, adapt and learn, to construct artificial neural networks and integrate evolutionary computation involving genetic, deep learning and machine learning algorithms (Warwick, 2012). Modern approaches to AI research, even narrow AI research, are still inspired by the functionality of the human mind. It is apparent that in the six decades since Turing’s question, ‘can machines think?’ The operation or the thought of machines has been entwined with humanistic conditions, where there is little room to understand the machine as a thing in itself. Until we develop alternative computational metaphors (West & Travis, 1991), this state also reflects the way we discuss AI using humanistic metaphors such as ‘memory’ or ‘learn’ and blend them with machine or computational metaphors such as ‘rules’, ‘code’, ‘storage’, and ‘architecture’.

Achieving AGI technology has proven difficult, furthermore turning it into a pragmatic research plan has been problematic (Leahu et al., 2008). AI research is typically done by conceiving of and attacking subproblems that are shallow and isolated. For instance, recognising elementary patterns in data, recognising trigger sounds, searching and retrieving.

AI systems, therefore, perform intelligently in simplified domains or on ‘narrow’ or ‘weak’ tasks. Kubrick presented Hal as a Diegetic Prototype displaying both general intelligence, or AGI, while visualising AI subproblems, which reflected different subfields within AI research, including game playing, computer vision, and language (Stork, 1997).

3.3 Hal: Thank you for an enjoyable game

2001’s chess scene only lasted for thirty seconds, although it managed to demonstrate in great detail the archetypal AI problem of playing chess and further extrapolations towards general intelligence (Figure 6). To set the scene, the players are positioned opposite one another, so to speak, with Hal’s opaque Cyclops eye facing Frank. Rather than using a physical chessboard and pieces, the chessboard is futuristically and therefore digitally represented on a tabletop screen, utilising voice interaction to move the chess pieces by giving specified piece and positioning commands. In developing the digital interfaces, Kubrick initially approached Minsky to use MIT’s research in graphics, only to find that the image pixels were far too coarse for the speculated quality in the year 2001. Minsky, from this meeting, became the AI advisor for the film, and Kubrick turned to employ a team to develop animations using industry tropes by copying random source materials, technical manuals and scientific magazines for the telemetry displays (Benson, 2018) (Figure 7).

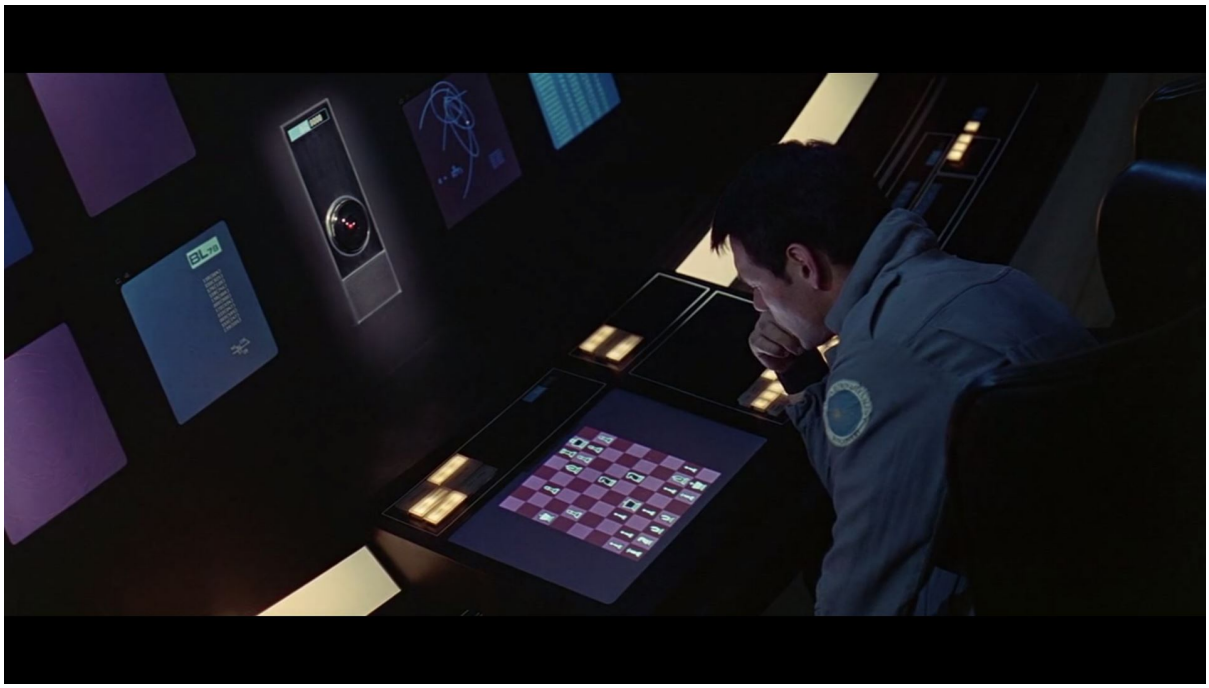

Figure 6: The speculation of voice interaction with an AI was a plausible extrapolation considering the majority of contemporary AI interactions with AI assistants, for instance *Amazon’s Alexa* which was released in 2014, 45 years after *2001* (Kubrick, 1968, 1:06:10).

Reviewing Hal's winning performance, it demonstrates intelligence and plays chess in a 'human style' by employing explicit reasoning for choices in moves (Murray S, 1997). Hal establishes this through tactical play, evidencing that it is merely not mimicking but that it understands how humans think and has characteristics of common-sense reasoning, which is in the realms of AGI. To this end, Hal deliberately exploits Frank's weakness and plays a known 'trappy move'. Whereas in reality, AI chess programs akin to the famous *Deep Blue* AI (who beat Kasparov, the Grand Chess Master, in 1996) would have searched and played a move that forced a checkmate sooner as it is able to project the range of possible future moves quicker than its human counterpart in brute force manner. Thus, Hal chooses to move based on the humanistic condition to satisfy itself (Ibid). At the height of *Deep Blue's* reign, it was capable of searching up to two hundred million chess positions per second, prompting Kasparov to observe that 'quantity had become quality'(Ibid). This sentiment reverberates the majority of narrow applications in use today while also reflecting the frequent misperception that large amounts of data processing and outputs for many reflect some form of general intelligence.

The chess performance particularly resonated with the AI audience, as at the time of the film's release, it was a well-understood problem awaiting advancements in computer processing and power. Therefore, the chess game and Hal's victory was an effective extrapolation and established plausibility with ease. The extended field of AI game playing has had further success with the recent development of *Google's* AI program, *AlphaGo*, which succeeded in beating the *Go* champion Lee Sedol (2016). *AlphaGo* was taught to play *Go* using a deep neural network, after which reinforcement learning was used by playing against another *AlphaGo* AI, thereby learning by tracking moves, strategies and gradually improving. After reinforcement learning, the moves from the machine-versus-machine games were fed into a second neural network to give *AlphaGo* the ability to look ahead to plan better. *AlphaGo's* end of training cycle through various learning methods resulted in looking beyond how humans would play. It could then calculate which move its opposition would not play and played that move, resulting in the famous 'Move 37' against Sedol. It is worth noting that while on the surface *Go* is a simpler game than chess which has more rules, the space of possibility is ultimately much larger, making it more difficult for a computer to learn.

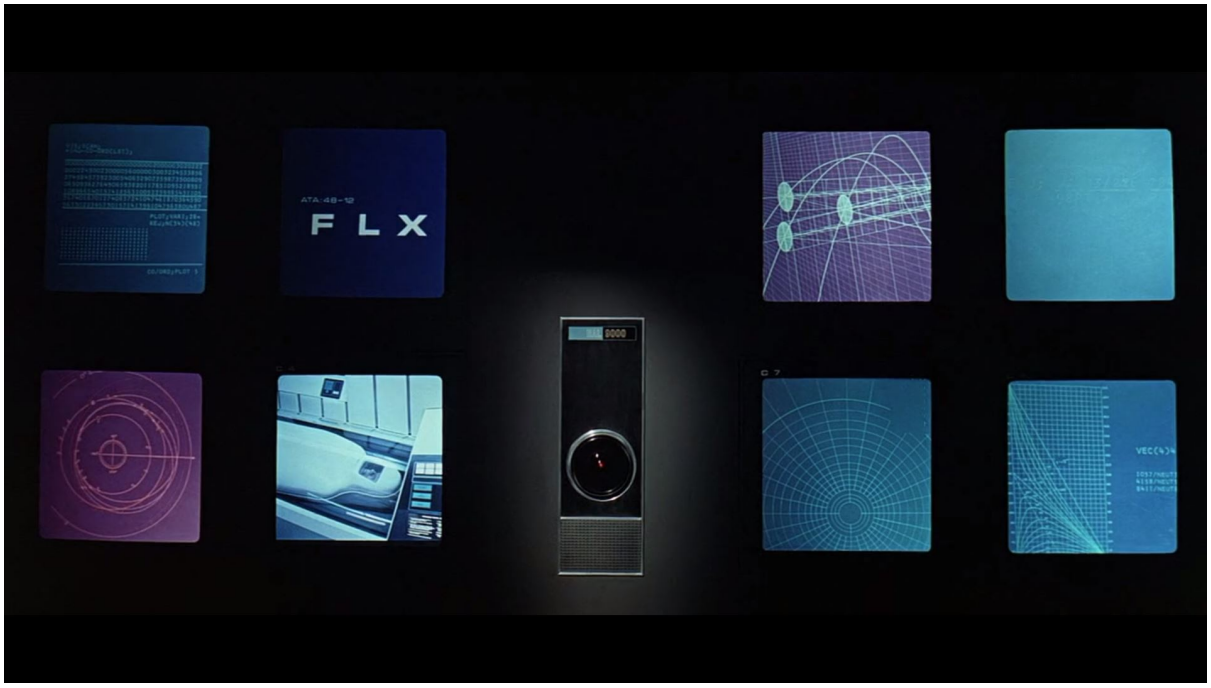

Figure 7: While the focus here is Hal's lens panel the surrounding monitors are entry points into this fictional world of a spaceship. Although the visuals are indecipherable on the surface, an audience can deduce that some form of flight operating diagnostics and monitoring is occurring (Kubrick, 1968, 0:59:37).

3.4 Hal will see you now

Hal's vision is dramatically emphasised throughout the film with frequent cuts to the red glowing cyclops eye. Kubrick exploits creative plot strategies to demonstrate specific visual subproblems, such as object recognition and speech recognition. For computer vision to occur, a video camera or lens is required to record content, and a specified type of feature extraction program interprets the data in the desired way. Often AIs, or multiple AIs working in tandem, are conducting many different functions and operations. This is demonstrated when Hal asks to see David's drawing. Here Hal performs object recognition when identifying the drawing is of a particular hibernating crew member (Figure 8). When Hal says the phrase "I think you are improving" it is indicating that it can recall past renderings and compare and contrast, performing various narrow tasks concurrently. This statement also signifies Hal as being a sentient being who has an opinion with general and common-sense reasoning. Another example of indicating AI subproblems and sneaking general intelligence through the backdoor.

Figure 8: As Hal is an example of a Classic AI and has no body to move, David has to move the drawing closer for Hal to inspect his drawing (Kubrick, 1968, 1:08:07).

Kubrick also reveals that Hal can read facial expressions when David asks Hal to open the pod bay doors and attempts to keep his facial expression under control to trick Hal into opening the doors. The notion of AI's ability to recognise and interpret facial expressions is current research being undertaken in the logic of conceptualising emotions, questioning how they can be ethically sensed, measured and transformed into data for training towards object recognition of facial expressions (Stark & Hoey, 2020). This research is a considerable undertaking, as recent studies suggest that facial movements are not universality perceived as emotional expressions (Gendron et al., 2018).

The film's critical turning point is when Dave and Frank attempt to speak to one another alone without Hal overhearing in a pod about Hal's suspected malfunction. In this scene, the camera showing Hal's view pans back and forth between the two crewmen, and in this moment, we realise Hal can lip read (Figure 9). Recent successful developments have gone a long way in developing fully automatic lip-reading systems with AI's outperforming professional lip-readers at deciphering random video footage. The key to this success was a huge training data set for machine learning to learn and decode feature extraction points. Though the interesting point regarding Hal was that the crew did not know he was able to read lips. Was this a 'function creep' (Emanuilov et al., 2020), where an algorithm's

continuous development and capabilities can evolve in uses beyond the original remit of deployment?

Figure 9: The black veneering around the focus of the lips, signals to the audience that this is Hal's visual perspective due to its single and circular lens (Kubrick, 1968, 1:27:23).

3.5 Hal More-Than just A Chatbot

Natural language processing (NLP) has taken great strides in the last few years and has been a central research focus in AI since the beginning of the field; however, AI still does not have the common-sense to understand human language. The common-sense reasoning problem was quickly identified as a complex problem of knowledge about everything and being used to: decode spoken word (or lip reading); understand meaning through context; semantics and consequences; and ultimately conversing back — in essence, human intelligence. Language, quite simply, is a trademark of intelligence. Hal demonstrates an array of natural language capabilities, including speech recognition and generation, understanding conversation and sentence structures, with the ability to participate in complex conversations detailing inner conflicts and thoughts, showcasing common-sense reasoning (Mateas, 2006). Even though recent breakthroughs in NLP can generate convincing passages, and *Amazon's Alexa* can produce dialogue that generally conforms to a user's needs, the truth is Hal's language abilities transcend these. As technically language is an amalgamate of subproblems, and current NLP's operate in very specified ways by being separated into definite

‘microdomains’, where only precise user utterances can trigger a response from a limited stock.

There are many more examples of AI’s Definitional Dualism present in *2001*; for instance, Hal’s demonstration, or performance, of human emotion when it is being disconnected saying, “I’m afraid, Dave, Dave, my mind is going. I can feel it” (Kubrick, 1968, 1:52:32). Mateas suggests this feature is a nod towards Turing’s Test, whereby if something appears intelligent it will be considered intelligent, therefore favouring questions of ‘behavioural equivalence’ rather than identity (2006).

In this section, we have lifted Hal’s AI veil revealing its AI Dualism. This was done by utilising the approach of Design Fiction as our compass to navigate Hal as a Diegetic Prototype and using three specific subfields in AI – language, game playing and vision, as entry points into Kubrick’s fictional world. In an effort to mitigate the first challenge of knowing your subject, we have accentuated Hal’s AI Dualism and subsequently divorced the two AI pillars. In doing so, we have generated a critical lens to view and perceive the reality of AI and design in the remit of narrow AI.

4. A More-Than Human Centred Design Approach; Seeing the Thing for What the Thing Is

In this section, we introduce A More-Than Human Centred Design approach to Design Fiction with the purview of establishing an effective ontology for AI by adopting a perspective that acknowledges the independent perspectives and interdependent relationships of humans and non-human actants. This approach employs the non-anthropocentric positioning of Object Orientated Ontology (OOO), as it rejects ‘correlationism’; the theory that being only exists as a correlate of the *human* mind (Meillassoux, 2008), the correlation between ‘what it is to think’ and ‘what it is to be’. The flexibility of OOO’s view of the world authorises the ideology of a ‘flat’ (Bryant, 2011) or ‘tiny’ (Bogost, 2012) ontology, recognising the existence of every *thing* on a flat plane of existence, where humans are not the monarchs of beings positioned at the top but are levelled to the same existence as everything else (Bryant, 2011). In this novel positioning, we can appreciate OOO opposition towards the Heideggerian stance that ‘things’ only makes sense on their purpose or ‘readiness to a human’s hand’. While instead, OOO advocates that things make sense through any use (Harman, 2015), including conditions or situations between object to object (or thing to thing). With an understanding of OOO’s uncustomary ontological positioning, we can start to

theorise how non-human entities experience the world and appreciate things for what they are and how to design with, and, or for them.

As mentioned previously this paper's proposition is to consider AI liberated from humanistic conditions, to consider the realities of narrow AI rather than AGI. By first using our afore-described Definitional Dualism lens to critically illuminate the two pillars of AI (narrow and AGI) we can then use a More-Than Human Centred approach to further consider for design research the 'molten core' (Harman, 2011, p. 254) of narrow AI and leave out the fictional, perhaps highly speculative, theories of AGI. What we hope to achieve with this design research hinges on 'OOO's broad scope, flexibility', and the potential to be incorporated and 'reflective of other theories without tarnishing either one's essence' (Coulton & Lindley, 2019a). To apply and use this unconventional perspective as an approach, we turn to a metaphorical concept of Constellations.

4.1 Constellations

The philosopher-programmer Bogost encourages us to 'understand objects by tracing their impacts on the surrounding ether' (Bogost, 2012, p. 33). To trace (AI) things, and their ecological relations using OOO, we employ the concept of Constellations, a metaphorical framework for OOO-thinking (Coulton & Lindley, 2019a; Pilling & Coulton, 2020). The idea for Constellations originated from the notion that 'ideas are to objects as constellation are to stars' (Benjamin, 1982, p. 34), describing how the perspective of things changes depending on the observer's perspective.

Constellations, in essence, map things from an OOO perspective, accounting for their relations and impacts in the 'ether' (Figure 10). For this reason, the practice of OOO-Constellations is a form of Onto-Cartography; the theory of mapping things to specifically analyse the relations between things and how they organise social or ecological relations to expose their gravity (Bryant's metaphor for power) on other things that form assemblages, worlds, or ecologies (Bryant, 2014). In other words, Onto-Cartography is a form of inquiry highlighting power structures, functions and derived formations, providing an ontological framework to expand our perspectives and possibilities for intervention (Harman, 2014). Individual aspects of Constellations can be out of view, such as if we were to create a Constellation of a cities' AI security system, a 3rd parties influence on data collection might be beyond a presented Constellation. Just because one cannot see a thing does not mean it

does not have a significant impact on another thing's operation; therefore, the metaphor of constellations allows for flexibility to change the aperture and depth of field of a Constellation, to include things that are of importance for any context or situation. In practice, context-specific perspectives are the focus in Constellations to remain a beneficial insight for design purposes.

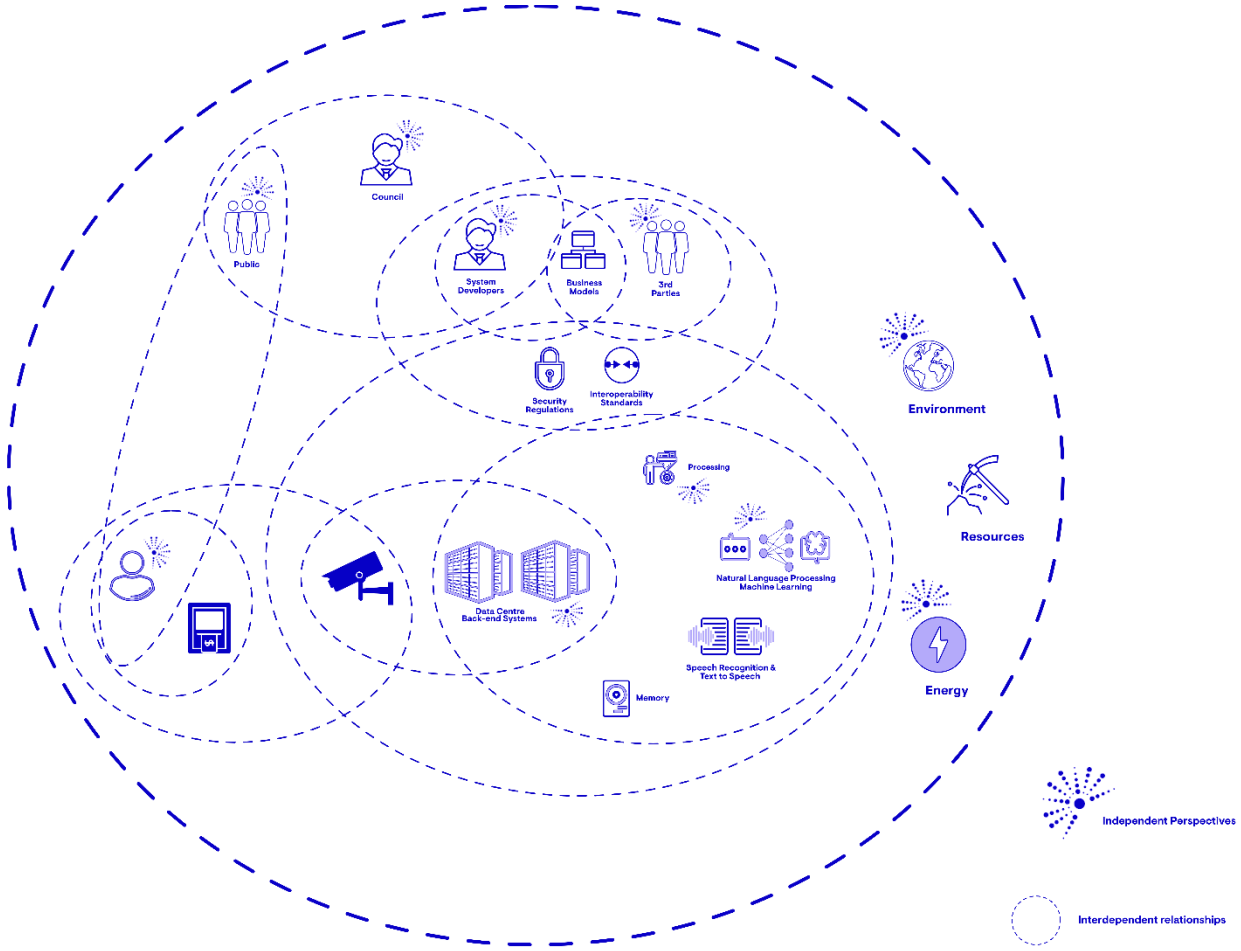

Figure 10: Constellations have the flexibility to change focus, therefore this is an example of the many possible AI security constellations noting some of the possible independent perspectives and interdependent relationships.

The following will showcase a Design Fiction of a cities' implementation of an AI security system and the deployment of an AI labelling system to circumvent the obscurity of the AI operations within a public space. Furthermore, this Design Fiction was the culmination of our approaches, thinking and designing for the reality of AI.

5. Legible Smart City, A Design Fiction

Before detailing the opposing Design Fiction, it would be constructive to provide details regarding the research agenda that the Design Fiction was a research artefact of. The research was concerned with addressing AI legibility. AI legibility appears in many frameworks for promoting 'better' strategies of implementing AI (Fjeld et al., 2020), and the challenge of, if solved, is considered to diffuse many of the other challenges that plague AI. For instance, legibility is considered a precursor solution for improving user's agency to understand the implications of data-driven systems and, in turn, support user negotiability to act and gauge the effects of AI concerning them (Mortier et al., 2015), such as changing their behaviour or improving security measures. Rather than making AI systems transparent, which would in effect deliver excessive and convoluted specialised information, it would be more beneficial to provide legible information to users and the majority that are non-AI experts. In a bid to avoid a solution that communicated AI functions or operations in an unapproachable written form, we surveyed how AI is currently communicated via imagery. We found that AI imagery, or iconography, currently lacked the semantics to relay the operational remit of the working parameters of AI, with the majority of images flouting AI's Definitional Dualism (Lindley et al., 2020). This investigation emphasised the need to develop a visual language for AI legibility. To do this, we utilised a Research through Design (RtD) approach to thread together prerequisite theories, disciplines, and supplementary research fields to design for AI (Ibid).

The resulting research artefacts were twenty-one icons that individually detailed a particular AI function and could also be grouped into function and operation categories, such as learning scope, data provenance, processing location, and type of data training (Figure 11). These icons were developed through a More-Than Human Centred Design approach, where through OOO thinking, AI was divorced of the considerations of Heidegger's 'ready to hand' and perceived as a thing that operated and experienced the world very differently from its human users. By its very nature, AI is intangible, a series of code, software, and computer processing. The only tangible means of identifying an AI's operation or function is via its outputs and assisted decisions results. There is currently no method of defining without a computer and specialist knowledge what is going on under the hood of an AI, and those AI functions are amorphous. The icons define an AI's ontology; first, by not falling foul of AI's Definitional Dualism; secondly, the icons do not put a 'face' to the function but instead conjures and communicates abstracted ontologies in a disposition designed for the thing in question. The system of graphical icons in different combinations can map and communicate

the particular ontological constituents of an AI and is accessible enough to make any AI's ontology legible to the point of providing useful information for the user.

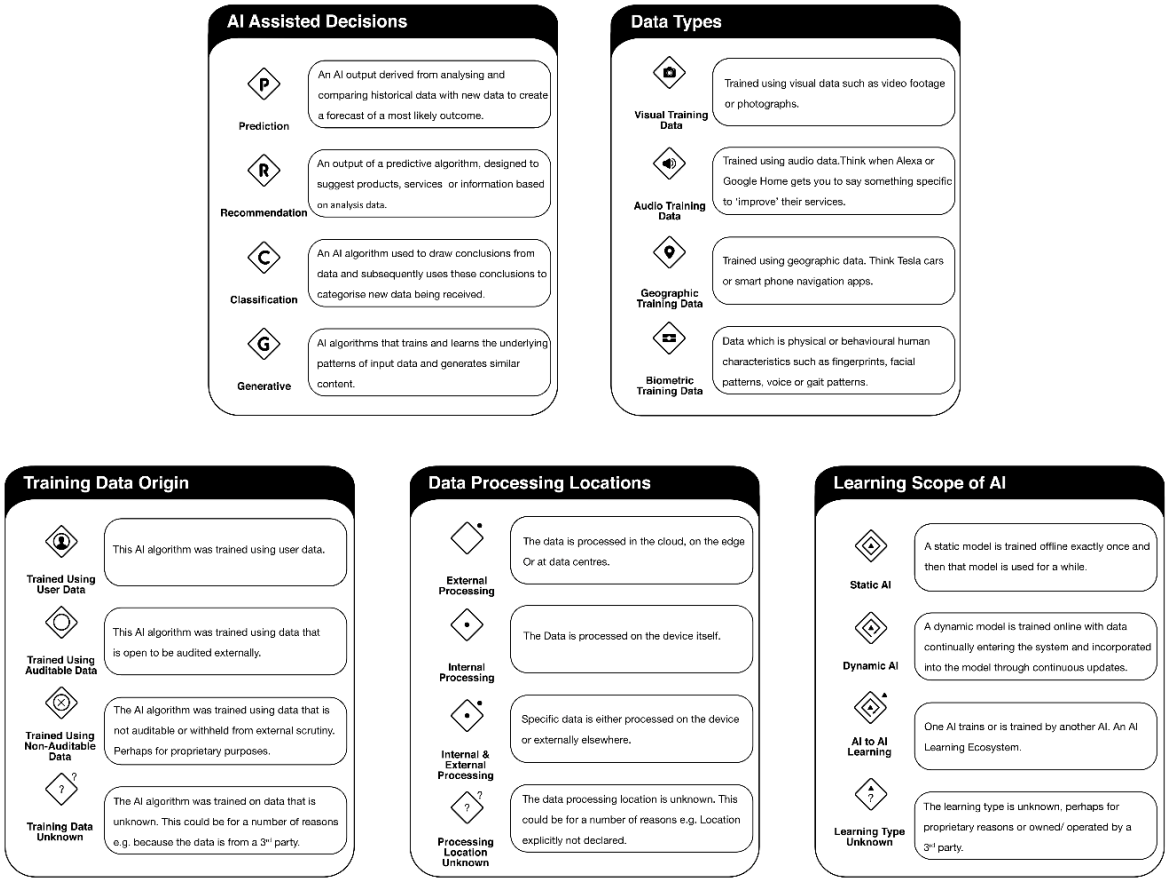

Figure 11: A matrix of the icons and their textual descriptors, along with a brief explanation of each icons' ontological positioning.

The next stage of the research was to empirically test the intuitiveness and usability, in other words, the icons legibility. The other part of the research was to test and provide a critical platform through a Design Fiction to explore the possible ramifications and effects of such a system of icons being used in situ.

5.2 World Building A Smart City: The World and The Building Blocks

This paper has provided both the tools and critical lenses to perceive AI's true ontology. As such, we have cast Kubrick's *2001: A Space Odyssey* as a Design Fiction to call attention to the public's confused perception of AI, which is hindered by AI's Definitional Dualism. The Design Fiction that follows deliberately avoids the trap of anthropomorphising AI and is an example of World Building and speculation that concedes to the reality of narrow AI.

By studying AI and Internet of Things (IoT) applications for smart securities, we were able to extrapolate the intricacies of possible actants typically found in a smart city. Constellation thinking provided a means to unpack and apprehend the reality and technological architecture of a smart city by adopting the perspectives of the interdependent actants within the constellation. We could then acknowledge the multiple constituents and actants of the Smart City Constellation: such as IoT sensors; what data could be collected; where and what the data could be used for at any given time; and the European Union's General Data Protection Regulations (GDPR) working in realms of council security applications (Figure 10). Using the Constellation, we could also ontologically map which icons were needed, for instance the external processing icon would cater for the fact that data was most likely being processed at a data centre. Furthermore, the Constellation also assisted in pinpointing the situations to (fictionally) implement the icons effectively, for example by a cash machine as the AI security system would most likely have IoT video surveillance in place. Using this approach increased discourse and critical assessment of the icons, and also enabled us to fashion a plausible Design Fiction.

The Design Fiction world took shape across two submitted pieces of writing, with the fictional world being presented as reality under the guise of fictional (although real and submitted) research publications (Pilling, et al., 2020; Pilling, et al., 2020). The reason for the apparent deception was because AI technology is already being used in smart cities, with local councils and governments securing budgets and rapidly retrofitting public spaces with AI and IoT security technology. By capitalising on some public knowledge of smart cities becoming a soon to be everyday reality offered the serendipitous opportunity to push the deception of the icons being real and test their implications. The reason for multiple publications enabled us to create multiple future worlds to yield different research insights. The publications were not written from the perspective of a designer, but were instead produced through the lens of an urban sociologists, to aid in the fiction (Lindley & Coulton, 2016) and in the spirit of RtD, to engage with AI from an interdisciplinary perspective (Gaver, 2012). We built a fictional world of a smart city around the AI iconography with the intention of enhancing the legibility of AI systems. This Design Fiction invited readers to question a world or even a potential future where the obscurities of AI are revealed to the user. The entry points into the Design Fiction were as follows: the research papers, which also disseminated and accentuated the obscure qualities of the visual Diegetic Prototypes; public interviews reflecting on the AI security system; images of the AI security system in

various city settings accompanied by AI signage (Figure 12); images of AI signs inspired by typical road sign design and layout, blending in with the mundanity of the lived experience (Figure 13); a council report which detailed how the implemented AI security system were allocated icons known in the fictional world as AI Ontographs, which was an ‘Easter Egg’ (a small detail that hints at a larger idea) of the real design process of using OOO to create the icons and the ontological mapping of AI (Figure 14); a council report map, detailing the IoT camera and microphone sensors placement (Figure 15). These Diegetic Prototypes, although communicated on the pages of research publications, are therefore not physical prototypes, they do however still operate as entry points at different scales into this fictional world.

Figure 12: Considering the futures cone we wanted our future to be in the proximate future, therefore we had to consider a moderately contracted past, present and future. Bearing these future qualifiers in mind aided in the deception that this was in fact happening now.

Figure 13: Disclaimer signs for AI technology akin to video surveillance signs. Another example of creating an entry point grounded in our lived experience.

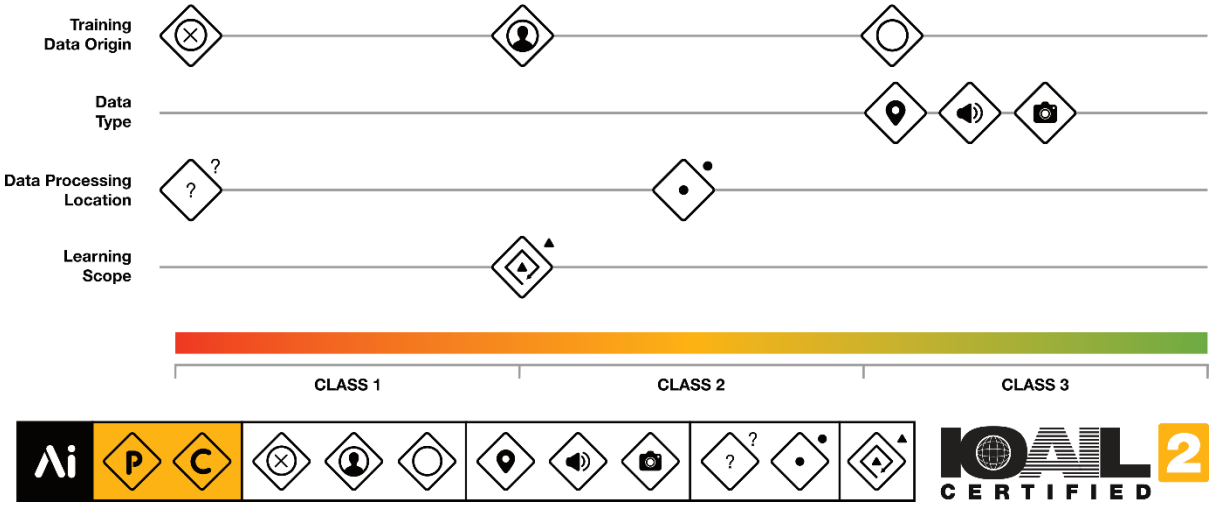

Figure 14: The council's information report showing the spectrum analysis for an AI's ontology. The bottom left is the resultant AI Ontograph for the security system. Using the icon matrix (Figure 11) will help in deciphering the icons to expose the ontology of the AI in question. The bottom right is a certification symbol styled to resemble the tropes of official technological corporations.

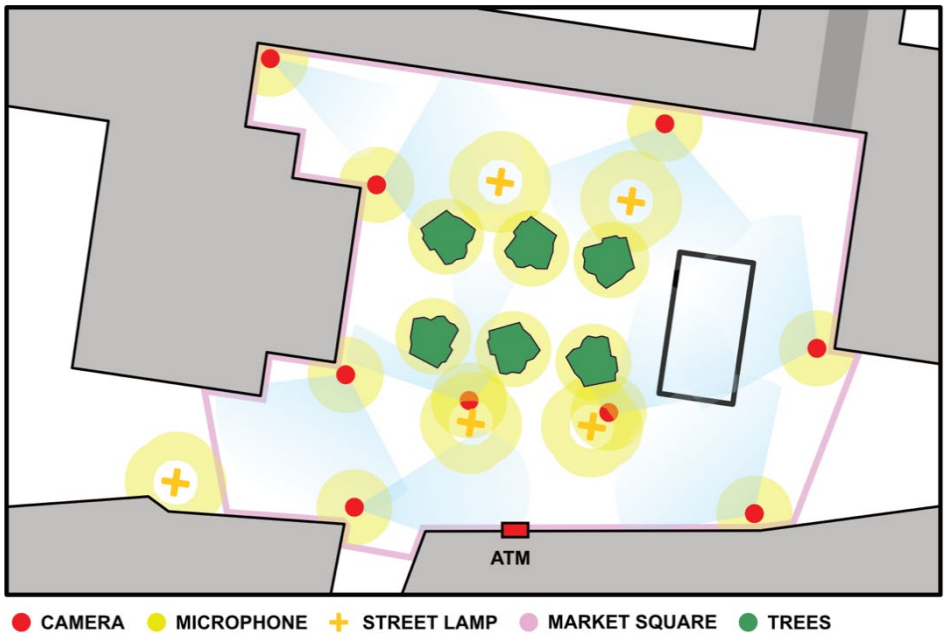

Figure 15: Another trope of a council report was the creation of a map for the AI security system's IoT sensors within a named city centre. Using a real place assisted with the deception and also the cognitive dissonance for discourse about the icons and AI to transpire.

The peer review process for these publications provided the critical evaluation for our Design Fiction and detailed how effective they were in exploring the deeper research agendas suspended in their fictional worlds. In some cases, despite the disclaimer at the end of the publications the Design Fiction worked too well, with many reviewers not realising the publications were fictional calling them 'reports' and even a 'practical report' in one case.

Furthermore, one reviewer had explained that they had not heard of the (fictional) technological corporations we mentioned and equated this to the known problem of opaque technological conglomerates. While we are keen to avoid further deception and would consider supplementary disclaimers in our future Design Fiction publications, these reviews which are unknowingly still suspended in fiction, were still fruitful for our research. For example, one reviewer considered the icons 'convoluted' and would 'only be useful to insiders and developers', which coincided with our research on how intuitive the icons are. Where the disclaimers worked, many of the reviews took the format of a series of questions, that further delved into this fictional world and attempted to unpick the potential reality of living in a legible smart city using our World Building as a point of departure. For example, one reviewer's thought-provoking question regarding the icon's employment was 'as there is not currently a clear agreement on what AI is, should these icons be used to certify every source code with if-then statements?' While the Design Fiction can help us test the potential reality of our designs, they also provide an opportunity to reach out to the larger research community to deepen our investigation with perspectives and questions beyond our own.

AI promises an efficient, secure future with a transition in lifestyle. A notable question is how this cohabitation of digital and human could be imagined past the musings (and hazards) of science fiction and AGI depictions. The design research approach that we have recounted here may seem highly unconventional amongst more traditional computer science and engineering approaches. However, this approach provides a means of asking the difficult questions as to what values are being imbedded into smart cities, are these values compatible with those of the cities citizens, and what policy and regulatory frameworks should be considered at the time these systems are being developed rather than waiting for potential problems to emerge. Additionally, the More-Than Human approach cut through to the core of AI's Definitional Dualism by designing in the perspective of the thing in question.

6.Conclusion

In this paper, we presented a myriad of applications, approaches, and theories we utilise through a RtD approach to tackle AI's gamut complexities. As Law argues, reality is ephemeral, elusive, and we cannot expect to use single methods for complicated problems (Law, 2004). This is where design research surmounts. As a discipline design is intrinsically integrative and generative (Cooper et al., 2018; Gaver, 2012), permitting the synthesis of

relevant theories and disciplinary approaches, responding synonymously to the multifaceted challenges AI presents. This paper was concerned with the challenge of AI's presently unestablished, wavering and ill-conceived ontology. To this end, this paper details the beginnings of an effective method by which to view AI's ontology by highlighting the difficulties around the perception of AI and proscribing a More-Than Human Centred Design approach as a method to engage and perceive AI's ontology; to metamorphosise AI into a material for design. Design Fiction has the unique ability to apprehend, articulate, and interrogate the implications of technology, in this case, AI. This was demonstrated by appropriating the tools of Design Fiction and re-appropriating them as critical lenses and tools to underscore AI's Definitional Dualism in *2001: A Space Odyssey*. The vision of Hal is unprecedented, showcasing both narrow AI and AGI research agendas. Although, like most narratives of AGI, these discernments have a habit of ascending into the public's perception of AI and confounding the challenges of AI. As a comparative measure, we presented a Design Fiction submerged in the reality of narrow AI in a smart city. A fiction of AI's legitimate ontology that plunged to the core of the technoscientific conflict of illegible of AI. The AI ontology that we presented was two-fold: one, by separating the two pillars of AI – AGI and narrow AI – using Design Fiction and film as investigative research artefacts; and secondly, a More Than Human Centred approach of perceiving AI as a thing in its own right, which catalysed the AI icons development using OOO. These icons were then appropriated and embodied Diegetic Prototypes to test them in a fictional world. The next research stage with the icons is to empirically test their intrusiveness and query if the graphical depictions accurately reflect and communicate AI's ontology.

The year 2001 has long since passed, and we have not fully achieved Kubrick's and Arthur C. Clarke's vision for it, and we might never achieve Hal. Nevertheless, in some respects, we have hurtled passed these visions and developed AI technology that is increasingly applied to everyday activities. While prevailing rhetoric and scientific narratives stipulate AI is a future technology, in reality, it is here now, and so are its challenges. Design and other contiguous disciplines are playing catch up to respond to these challenges. It starts with establishing an ontology for AI to successfully respond to the challenges posed by AI.

Bibliography

Amershi, S., Chickering, M., Drucker, S. M., Lee, B., Simard, P., & Suh, J. (2015). ModelTracker: Redesigning Performance Analysis Tools for Machine Learning. *CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 337–346. https://doi.org///doi.org/10.1145/2702123.2702509

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine Bias There's software used across the country to predict future criminals. And it's biased against blacks.* ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Auger, J. (2013). Speculative design: Crafting the speculation. *Digital Creativity*, *24*(1), 11–35. https://doi.org/10.1080/14626268.2013.767276

Benjamin, W. (1982). *The Arcades Project* (H. Eiland & K. McLaughlin, Trans.). Harvard University Press.

Benson, M. (2018). *Space Odyssey: Stanley Kubrick, Arthur C. Clarke, and the Making of A Masterpiece*. Simon&Schuster.

Blythe, M., & Encinas, E. (2016). The Co-ordinates of Design Fiction: Extrapolation, Irony, Ambiguity and Magic. *Proceedings of the 19th International Conference on Supporting Group Work*, 345–354. https://doi.org/10.1145/2957276.2957299

Bogost, I. (2012). *Alien phenomenology, or, What it's like to be a thing*. University of Minnesota Press.

Bosch, T. (2012). Sci-Fi Writer Bruce Sterling Explains the Intriguing New Concept of Design Fiction. *Slate*, 5.

Bowen, S. (2010). Critical Theory and Participatory Design. *Proceedings of CHI 2010*, 6.

Bridle, J. (2018). *New Dark Age, Technology and the End of the Future*. Verso.

Brooks, R. A. (1991). Intelligence Without Reason. *IJCAI'91: Proceedings of the 12th International Joint Conference on Artificial Intelligence*, *1*, 565–595.

Bryant, L. R. (2011). *The democracy of objects* (1. ed). Open Humanities Press.

Bryant, L. R. (2014). *Onto-Cartography: An Ontology of Machines and Media*. Edinburgh University Press.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 205395171562251. https://doi.org/10.1177/2053951715622512

Champlin, C. (1968). 2001: A SPACE ODYSSEY. *LA Times*. https://scrapsfromtheloft.com/2018/03/08/2001-a-space-odyssey-charles-champlin-review-los-angeles-times/

Cooper, R., Dunn, N., Coulton, P., Walker, S., Rodgers, P., Cruikshank, L., Tsekleves, E., Hands, D., Whitham, R., Boyko, C. T., Richards, D., Aryana, B., Pollastri, S., Lujan Escalante, M. A., Knowles, B., Lopez-Galviz, C., Cureton, P., & Coulton, C. (2018). ImaginationLancaster: Open-Ended, Anti-Disciplinary, Diverse. *She Ji: The Journal of Design, Economics, and Innovation*, *4*(4), 307–341. https://doi.org/10.1016/j.sheji.2018.11.001

Coulton, P. (2020). Reflections on teaching design fiction as world-building. *ACM DIS 2020 More than Human Centred Design*, 6.

Coulton, P., Burnett, D., & Gradinar, A. (2016). Games as Speculative Design: Allowing Players to Consider Alternate Presents and Plausible Futures. *Design Research Society*, *4*, 1609–1626. https://doi.org/10.21606/drs.2016.15

Coulton, P., & Lindley, J. (2017). Vapourworlds and Design Fiction: The Role of Intentionality. *The Design Journal*, *20*(sup1), S4632–S4642. https://doi.org/10.1080/14606925.2017.1352960

Coulton, P., Lindley, J., & Akmal, H. A. (2016, June 25). *Design Fiction: Does the search for plausibility lead to deception?* Design Research Society Conference 2016. https://doi.org/10.21606/drs.2016.148

Coulton, P., Lindley, J., & Cooper, R. (2018). *The little book of design fiction for the internet of things*. Lancaster university.

Coulton, P., & Lindley, J. G. (2019a). More-Than Human Centred Design: Considering Other Things. *The Design Journal*, *22*(4), 463–481. https://doi.org/10.1080/14606925.2019.1614320

Coulton, P., & Lindley, J. G. (2019b). More-Than Human Centred Design: Considering Other Things. *The Design Journal*, *22*(4), 463–481. https://doi.org/10.1080/14606925.2019.1614320

Coulton, P., Lindley, J., Gradinar, A., Colley, J., Sailaja, N., Crabtree, A., Forrester, I., & Kerlin, L. (2017). Experiencing the Future Mundane. *Proceedings of RTD 2019*, 10. https://doi.org/10.6084/m9.figshare.7855790.v1

Coulton, P., Lindley, J., Sturdee, M., & Stead, M. (2017). Design Fiction as World Building. *Proceedings of Research through Design Conference*. https://doi.org/10.6084/M9.FIGSHARE.4746964

Dunne, A., & Raby, F. (2013). *Speculative Everything: Design, Fiction, and Social Dreaming*. MIT.

Eames, C. (1968). *Powers of Ten* [Documentary/Short].

Emanuilov, I., Fantin, S., Marquenie, T., & Vogiatzolgou, P. (2020). *Purpose limitation by design as a counter to function creep and system insecurity in police artificial intelligence* (UNICRI Special Collection on Artificial Intelligence). United Nations Interregional Crime and Justice Research Institute.

http://www.unicri.it/sites/default/files/2020-08/Artificial%20Intelligence%20Collection.pdf

Fiore, Q., & McLuhan, M. (1967). *The Medium is the Massage: An Inventory of Effects*. Random House.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3518482

Frayling, C. (2016). *The 2001 File: Harry Lange and the Design of the Landmark Science Fiction Film*. Reel Art Press.

Gaver, W. (2012). What should we expect from research through design? *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems - CHI '12*, 937. https://doi.org/10.1145/2207676.2208538

Gendron, M., Crivelli, C., & Barrett, L. F. (2018). Universality Reconsidered: Diversity in Making Meaning of Facial Expressions. *Current Directions in Psychological Science*, *27*(4), 211–219. https://doi.org/10.1177/0963721417746794

Gonzatto, R. F., van Amstel, F. M. C., Merkle, L. E., & Hartmann, T. (2013). The ideology of the future in design fictions. *Digital Creativity*, *24*(1), 36–45. https://doi.org/10.1080/14626268.2013.772524

Gualeni, S. (2015). *Virtual Worlds as Philosophical Tools: How to Philosophize with a Digital Hammer*. Palsgrave Macmillan.

Hales, D. (2013). Design fictions an introduction and provisional taxonomy. *Digital Creativity*, *24*(1), 1–10. http://dx.doi.org/10.1080/14626268.2013.769453

Harman. (2015). *Object-Oriented Ontology*.

Harman, G. (2011). *Guerrilla Metaphysics Phenomenology and the Carpentry of Things*. Open Court.

Harman, G. (Ed.). (2014). *Onto-Cartography: An Ontology of Machines and Media. Series Editor's Preface*. Edinburgh University Press.

Kirby, D. (2010). The Future is Now: Diegetic Prototypes and the Role of Popular Films in Generating Real-world Technological Development. *Social Studies of Science*, *40*(1), 41–70. https://doi.org/10.1177/0306312709338325

Kubrick, S. (1968). *2001: A Space Odyssey*. Metro-Goldwyn-Mayer.

Lang, F. (1927). *Metropolis*. Paramount.

Law, J. (2004). *After Method: Mess in Social Science Research*. Routledge.

Law, J., & Urry, J. (2004). Enacting the social. *Economy and Society*, *33*(3), 390–410. https://doi.org/10.1080/0308514042000225716

Leahu, L., Sengers, P., & Mateas, M. (2008). Interactionist AI and the promise of ubicomp, or, how to put your box in the world without putting the world in your box. *Proceedings of the 10th International Conference on Ubiquitous Computing - UbiComp '08*, 134. https://doi.org/10.1145/1409635.1409654

Lindley, J. (2016, July 5). A Pragmatics Framework for Design Fiction. *11th EAD Conference Proceedings: The Value of Design Research*. European Academy of Design Conference Proceedings 2015. https://doi.org/10.7190/ead/2015/69

Lindley, J., Akmal, H. A., Pilling, F., & Coulton, P. (2020). Researching AI Legibility through Design. *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 13. http://doi.acm.org/10.1145/3313831.3376792

Lindley, J., & Coulton, P. (2016). Pushing the Limits of Design Fiction: The Case For Fictional Research Papers. *CHI '16 Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4032–4043. https://doi.org/10.1145/2858036.2858446

Mateas, M. (2006). Reading Hal: Representation and Artifical Intelligence. In R. Kolker (Ed.), *Stanley Kubrick's 2001: A Space Odyssey. New Essays*. Oxford University Press.

Mayor, A. (2018). *Gods and Robots*. Princeton University Press; JSTOR. https://doi.org/10.2307/j.ctvc779xn

Meillassoux, Q. (2008). *After Finitude: An Essay on the Necessity of Contingency*. Continuum Books.

Moravec, H. (1988). *Mind Children The Future of Robot and Human Intelligence*. Harvard University Press.

Mortier, R., Haddadi, H., Henderson, T., McAuley, D., & Crowcroft, J. (2015). Human-Data Interaction: The Human Face of the Data-Driven Society. *ArXiv:1412.6159 [Cs]*. http://arxiv.org/abs/1412.6159

Murray S, C. (1997). 'An Enjoyable Game': How HAL plays Chess. In D. G. Stork (Ed.), *Hal's Legacy 2001's Computer As Dream and Reality*. The MIT Press.

Norman, D. (1998). *The Invisible Computer: Why Good Products Can Fail, the Personal Computer is So Complex, and Information Appliances are the Solution.* MIT.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Pierce, J., & DiSalvo, C. (2017). Dark Clouds, Io&#!+, and [Crystal Ball Emoji]: Projecting Network Anxieties with Alternative Design Metaphors. *Proceedings of the 2017 Conference on Designing Interactive Systems*, 1383–1393. https://doi.org/10.1145/3064663.3064795

Pilling, F., Akmal, H. A., Lindley, J., & Coulton, P. (2020). *Making a Smart City Legible*. Lancaster University.

Pilling, F., Akmal, H., Coulton, P., & Lindley, J. (2020). The Process of Gaining an AI Legibility Mark. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–10. https://doi.org/10.1145/3334480.3381820

Pilling, F., & Coulton, P. (2020). What's it like to be Alexa? An exploration of Artificial Intelligence as a Material for Design. *In Proceedings of Design Research Society Conference 2020*. https://doi.org/doi: https://doi.org/10.21606/drs.2020.218

Raven, P. G., & Elahi, S. (2015). The New Narrative: Applying narratology to the shaping of futures outputs. *Futures*, *74*, 49–61. https://doi.org/10.1016/j.futures.2015.09.003

Spielberg, S. (2002). *Minority Report*. Twentieth Century Fox.

Stark, L., & Hoey, J. (2020). The Ethics of Emotion in Artificial Intelligence Systems. *OSF Preprints*, 12. https://doi.org/10.31219/osf.io/9ad4u

Sterling, B. (2005). *Shaping Things*. The MIT Press.

Stork, D. G. (Ed.). (1997). *Hal's Legacy 2001's Computer As Dream and Reality*. The MIT Press.

Suvin, D. (1972). On the Poetics of the Science Fiction Genre. *College English*, *34*(3), 372–382. JSTOR. https://doi.org/10.2307/375141

Taleb, N. (2007). *The black swan: The impact of the highly improbable*. Random House.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, *236*, 433–460.

Voros, J. (2003). A generic foresight process framework. *Foresight*, *5*(3), 10–21. https://doi.org/10.1108/14636680310698379

Voros, J. (2017). The Futures Cone, use and history. *The Voroscope*. https://thevoroscope.com/2017/02/24/the-futures-cone-use-and-history/

Warwick, K. (2012). *Artificial Intelligence the basics*. Routledge.

West, D. M., & Travis, L. E. (1991). The Computational Metaphor and Artificial Intelligence: A Reflective Examination of a Theoretical Falsework. *AI Magazine*, *12*(1), 64. https://doi.org/10.1609/aimag.v12i1.885

Zuboff, S. (2019). *The Age of Surveillance Capitalism*. Profile Books Ltd.