

AI as a Material for Design



Franziska Pilling

**This dissertation is submitted for the degree of Doctor of
Philosophy**

September 2023

Design

For Matthew

“The future was, very literally, in their own hands.”

2001: A Space Odyssey, Arthur C Clarke 1968.

Declaration

I hereby declare that this thesis titled “AI as a Material for Design” represents my research and work done during my PhD in Design at Lancaster University. The concepts and ideas resulting in my work are stated here are my own words, and where I include the ideas of others I have cited and referenced the original sources accordingly. This body of work has not been submitted in support of an application for another degree at this or any other institution. Many of the ideas in this thesis were the product of discussions with my supervisor Professor Paul Coulton. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented, fabricated, or falsified any ideas/data/fact/source in this submission or the course of research. I understand that any violation of the above will be cause for disciplinary action by the University or other related sources.

Excerpts of this thesis have been published in the following manuscripts:

Lindley, J., Akmal, H. A., Pilling, F., & Coulton, P. (2020a). Researching AI Legibility through Design. CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–13. <http://doi.acm.org/10.1145/3313831.3376792>

Lindley, J., Akmal, H. A., Pilling, F., & Coulton, P. (2020b). Signs of the Time: Making AI Legible. Proceedings of Design Research Society Conference 2020. DRS 2020, Australia. <https://doi.org/10.21606/drs.2020.237>

Lindley, J., Green, D. P., McGarry, G., Pilling, F., Coulton, P., & Crabtree, A. (2023). Towards a master narrative for trust in autonomous systems: Trust as a distributed concern. *Journal of Responsible Technology*, 13.

Pilling, F., Akmal, H. A., & Coulton, P. (2020). Researching and Designing Uncanny AI to Legible AI. International Transdisciplinary Conference.

Pilling, F., Akmal, H. A., Gradinar, A., Lindley, J., & Coulton, P. (2020). Legible AI by Design: Design Research to Frame, Design, Empirically Test and Evaluate AI Iconography. *Common Good Framing Design through Pluralism and Social Values: Design as Common Good*, 2442–2459.

Pilling, F., Akmal, H. A., Gradinar, A., Lindley, J., & Coulton, P. (2021). Using Game Engines to Design Digital Workshops for AI Legibility. 14th International Conference of the European Academy of Design, Safe Harbours for Design Research, 394–403.

Pilling, F., Akmal, H. A., Lindley, J., & Coulton, P. (2022). Making a Smart City Legible. In S. Carta (Ed.), *Machine Learning and the City* (pp. 453–465). John Wiley & Sons, Ltd.

Pilling, F., Akmal, H. A., Lindley, J., Gradinar, A., & Coulton, P. (2022). Making AI Infused Products and Services more Legible. *Leonardo*, 1–11.

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. (2019). Forget the Singularity, its mundane artificial intelligence that should be our immediate concern. *The Design Journal*, 22(sup1), 1135–1146. <https://doi.org/10.1080/14606925.2019.1594979>

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/10.21606/drs.2020.218>

Pilling, F., & Coulton, P. (2021). *Carpentered Diegetic Things: Alternative Design Ideologies for AI Material Relations. The Ecological Turn. Design, Architecture and Aesthetics beyond 'Anthropocene'. The Ecological Turn, Bologna, Italy.*

Pilling, F., Lindley, J., Akmal, H. A., & Coulton, P. (2021). Design (Non) Fiction: Deconstructing/Reconstructing The Definitional Dualism of AI. *International Journal of Film and Media Arts*, 6(1), 6–32.

Pilling, F., Stead, M., & Gradinar, A. (2022). *The Prometheus Terminal: Worlding Games for the Adoption of Sustainable Datafication and Cybersecurity practices. Cumulus Detroit 2022: Design for Adaptation. Cumulus, Detroit.*

Stead, M., Pilling, F., Gradinar, A., & Forrester, I. (2022). *SUSTAINABLE X SECURE EDGE: Design Guidelines for Future Data-Driven Edge-IoT Devices and Services. PETRAS Research Centre of Excellence.*

Abstract

From Netflix recommendations to Amazon Echos sitting proudly on kitchen countertops, artificial intelligence (AI) has been inserted into the mundane settings of our everyday lives. These ‘smart’ devices and services have given rise to the collection of data and processing within everyday objects, accumulating new challenges, particularly in legibility, agency, and negotiability of interactions. The emerging field of Human Data Interaction (HDI) recognises that these challenges go on to influence security, privacy, and accessibility protocols, while also encompassing socio-technical implications. Furthermore, these objects challenge designers’ traditional conventions of neutral interactions, which only work as instructed. However, these smart objects go beyond typical human-object relationships functioning in new and unexpected ways, creeping in function, and existing within independent and interdependent assemblages of human and non-human actants—demanding alternative considerations and design practice.

This thesis aims to question the traditional practice of considering and designing for AI technology by arguing for a post-anthropocentric perspective of things with agency, by adopting the philosophical approach of Object Orientated Ontology with design research. This research ultimately presents and builds (a currently) unorthodox design approach of Human-AI Kinship that contests the design orthodoxies of human-centred design. Conclusively, this research seeks to bring into being AI as a material for design and justify through the case study of AI legibility.

A More than Human Centered Design approach is established through a transdisciplinary and iterative Research through Design methodology, resulting in the design of AI iconography that attempts to communicate and signify AI’s ontology to human users. This thesis is concluded by testing the legibility of the icons themselves and discussing philosophy as an asset for design research.

Acknowledgements

Matthew, I could not have done it without you –thank you. Thank you to Professor Paul Coulton for being generous with your time, help, and advice.

Contents

Abstract.....	6
Acknowledgements.....	7
List of Figures.....	12
Chapter One Introduction.....	15
1.1 Arrival’s Logograms (spoilers lie ahead).....	16
1.2 Research Problem & Significance.....	18
1.3 Research Questions.....	20
1.4 PETRAS IoT Hub and Research Track.....	20
1.5 Why Philosophy.....	21
1.6 Why Fiction.....	21
1.7 Structural Outline.....	22
Chapter Two Seeing AI.....	25
2.1 A Brief AI History – The Evolution of Two Categories of Intelligence; Artificial General Intelligence / Machine Intelligence.....	26
2.2 Towards Responsible AI Through Legibility: Guidance and Frameworks.....	27
2.3 Introducing an AI History.....	29
2.3.1 A Brief Interlude in Human Intelligence with an AI agenda.....	30
2.3.2 Pre-20 th Century; Mechanical Imitation, Hoaxes and Arranging Knowledge.....	31
2.3.3 The 20 th Century.....	33
2.3.4 The ‘Good Old-Fashioned AI’ Days.....	35
2.3.5 The Modern Approach to AI.....	41
2.3.6 The Rise and Troubles with Big Data.....	47
2.4 The Definitional Dualism of AI; A Confused Ontology.....	49
2.4.1 Evolving Definitions.....	49
2.4.2 AI Hype Cycle and The Rebranding of AI.....	51
2.4.3 Science-Fiction and Anthropomorphising AI technology.....	52
2.4.3.1 Examining Hal’s Definitional Dualism.....	53
2.4.3.2 HAL – “Thank you for an enjoyable game” – AI game playing.....	55
2.4.3.3 HAL will see you now – AI vision.....	56
2.4.3.4 HAL More-Than just A Chatbot – Natural Language Processing.....	58
2.4.3.5 Anthropomorphising AI.....	59
2.4.3.6 Metaphorical Anthropomorphisation.....	62
2.4.3.7 Believable Perceptions.....	63
2.4.4 Magic and Metaphors: Is It a Kind of Magic?.....	64
2.4.5 Alien Technology; Creating Their Own Representation of The World.....	66
2.5 Conclusion.....	68
Chapter Three Groundworks.....	70
(Understanding AI).....	70
3.1 Introduction.....	71
3.2 Transdisciplinary Research a Postmodern Turn: Promiscuous Monsters on the Prowl.....	71
3.3 Design as Wicked problems, The advantage of Transdisciplinary Design Research.....	73
3.4 Crafting Hinterlands through Method Assemblages.....	76
3.5 Forging A Transdisciplinary Hinterland.....	78
3.6 The Hinterland of AI as a Material for Design: a thesis pattern.....	78
3.7 Transdisciplinary Assimilation; Adapting Philosophy.....	79
3.8 A Philosophical Intermission - Adapting Philosophy; a case study of Jean Baudrillard and The Matrix Trilogy.....	80

3.9 The Philosophical Imaginary; Philosophical Tools beyond the Written Word	81
3.10 Metamorphosis; adapting philosophy	82
3.11 In Summary: The Matrix Trilogy as philosophy	86
Chapter Four Methodologies.....	89
(Understanding AI)	89
4.1 Introduction.....	90
4.2 Design Research: revelling in ambiguity or just a nomadic practice	90
4.3 Design Research; defining Design.....	94
4.3.1 Design as a creative and iterative process.....	96
4.3.2 Design as an act of creative problem solving	98
4.4 Defining Research; Research is Design	100
4.4.1 Kinds of Design Research.....	101
4.5 Research Through Design	102
4.5.1 What to expect from Research through Design	106
4.5.2 Annotated portfolios	108
4.5.3 Research through Design = Practiced-Based Research.....	109
4.6 Conclusion and Going Forth	111
Chapter Five More Than Human-Centred Design: Shifting Perspectives through Philosophy.....	113
(Being AI).....	113
5.1 Introduction.....	114
5.2 Human-Centered Design: A Concise Background.....	114
5.3 AI by Human-Centered Design: Shifting Viewpoints.....	117
5.4 Towards Human-Centered AI.....	118
5.5 Human-Centered Design Research in Artificial Intelligence	120
5.5.1 Human-Centered Computing for Human-Centered AI.....	120
5.5.2 Interaction Design for Human-Centered AI	121
5.5.3 Simplicity by Design	123
5.5.4 Persuasive Design for Human-Centered AI.....	124
5.5.4.1 Background on Persuasion and the Art of Rhetoric	124
5.5.4.2 Persuasive Strategies	126
5.6 A Brief Ethical and Closing Note	128
Part Two More-Than Human-Centred Design.....	130
5.7 Shifting perspectives to A More-Than Human-Centred AI.....	131
5.8 Posthumanism as presented here: A Speculative Realist Tint	133
5.8.1 Speculative Realism	134
5.9 Phenomenology: A Short Historical and Theoretical Synopsis.....	136
5.9.1 A Subjective Appearance and Reality of Thing	137
5.9.2 Heidegger's Phenomenology	139
5.10 Beyond Human Experience	140
5.11 Object-Orientated Ontology.....	141
5.11.1 To be Object-Orientated.....	141
5.11.2 The meaning of Object.....	143
5.11.3 Object Ontology: Levels of Objects	145
5.11.4 Unit Operations.....	147
5.11.5 Vicarious Causation & Relations	148
5.11.6 Quantum Causation: Virtual Particles Mediating Agency of AI.....	150
5.12 An ideological interlude: The Case of Materialism and Immaterialism	151
5.12.1 Vibrant Objects.....	153
5.13 Concluding on a More-Than Human-Centred Design for AI.....	155

Part Three Human-AI Kinship	157
5.14 Introduction	158
5.15 A Short Introduction to Postphenomenology	159
5.16 Human-Technology Relations	163
5.17 Background relations: Notes on Engagement	166
5.18 The Evolution of Hermeneutic Relations to Digital Hermeneutic Relations	167
5.19 Machine Hermeneutics	170
5.20 Concluding on OOO and Postphenomenology: Namely Cultivating Object Empathy for Human and AI Kinship (despite Thing-Transcendentality)	173
5.21 Conclusion	175
Chapter Six Design Fiction: Adapting Philosophy for Design	177
(Being AI)	177
6.1 Introducing the Carpentry of Things	178
6.2 Constellations with a side of Onto-Cartography	180
6.3 Constellations for A Horizonless Perspective	182
6.4 Alien Phenomenology	185
6.5 Speculation and Design Fiction	187
6.6 Design Fiction: An Overview	187
6.7 Design Fiction as World Building	188
6.8 A Philosophical Interlude: Philosophical metamorphosis through Worlding Constellations	191
6.9 Framing Futures	193
6.10 Rendering Emerging Technologies as Mundane	197
6.11 Design Fiction for a new material palette	201
6.12 Carpentered Diegetic Things	202
6.13 The More Than Human Centred Design approach to AI as a Material for Design	205
6.14 Conclusions	206
Chapter Seven Designing for AI Legibility	208
(Designing for Human-AI kinship)	208
7.1 Introduction	209
7.2 Explainability, Interpretability and Transparency	209
7.3 Interpretability or Explainability?	212
7.3 Mechanisms for users	213
7.4 Limitations of Transparency: Seeing without Knowing	216
7.5 AI Legibility	218
7.6 Guidelines for Legible Human-AI Interactions	219
7.7 Ways through the Communication Challenge	221
7.8 Designing for Legibility: A Case for Icons	223
7.9 Background for Designing the Semiotics of AI	224
7.10 Defining the Interpretant: AI Attributes, Dimensions and Properties (AI's Ontology)	228
7.11 Reinstating A Philosophical Perspective: Aesthetics Is the Root of All Philosophy	231
7.12 The Icons Design and Refinement Process	233
7.13 Conclusion	235
Chapter Eight AI Legibility Workshops and Iterative Icon Development	237
(Designing for Human – AI Kinship)	237
8.1 Introduction	238
Part One	240
8.2 Designing and Building Workshops for Intuitive Testing	241
8.3 Workshop Exercises	243
8.4 The Analyser	246

8.5 First Icon Iteration Results Overview: Making Connections.....	248
8.6 What's in My AI: Scenarios.....	250
8.6.1 Training Data	251
8.6.2 Learning Scopes.....	252
8.6.3 Processing Location.....	252
8.6.4 Data Provenance	253
8.7 Draw Your Own: Co-Designing Icons and Introducing the Second Iteration of Icons	253
8.7.2 Participants' Re-designs.....	254
8.7.3 New Categories.....	256
8.8 Finalising the Second Iteration	259
Part Two.....	262
8.9 Introduction.....	263
8.10 The Workshops: Second Iteration	263
8.11 Second Iteration Scenarios: What's in Spotify's AI.....	265
8.11.1 Training Data	265
8.11.2 Learning Scope, Processing Location, Training Data Origin & AI-Assisted Decisions	266
8.12 Designing a User Priority Arrangement: What Matters?	268
8.13 Version 2: Draw Your Own.....	270
8.13.1 Redesigns	271
8.13.2 New Categories: Data and Common Good Designs	274
8.13.3 Social Good Designs	276
8.14 New Categories: What is Intrinsic Labour.....	277
8.14.1 Work Replacement and Value Gained.....	278
8.14.2 Human-in-the-Loop & Human-out-of-the-Loop.....	278
8.14.3 Climate Change and AI & Cost of using AI.....	279
8.15 To Note: Bringing the Human back into the Equation.....	281
8.16 Part One & Two Conclusion.....	281
Part Three: Machine Learning in the City	283
8.16 Introduction.....	284
8.17 AI for Lancaster.....	284
8.18 Rights and Wrongs: AI and Surveillance.....	284
8.19 Designing a Certification Body	286
8.20 In the Wild.....	288
8.21 Conclusion: The Truth, The Whole Truth, and a little bit more	293
Chapter Nine Conclusion.....	295
9.1 Introduction.....	296
9.2 Research Questions.....	296
9.2.1 RQ1: AI as A Material for Design & RQ2: AI Ontographs	297
9.2.2 RQ3: The Insights of a More Than Human-Centred approach	299
9.2.3 RQ4: Practical Designs for AI	300
9.3 Contributions.....	301
9.3.1 A More Than Human Design Approach	301
9.3.2 Philosophy and Design.....	303
9.3.3 AI Legibility.....	304
9.3.4 Workshopping during a pandemic.....	305
9.3.5 A Transdisciplinary Hinterland	305
9.4 Limitations	306
9.5 Going Forward: Future Research.....	307
9.5.1 Ontological Design Research.....	307

9.5.2 Legible Diagrams	308
9.6 Summary	309
Bibliography	310

List of Figures

Figure 1: A basic logogram with inky tendrils (Morrison, ND).	17
Figure 2: The barrier between the aliens and humans is also the site where the aliens share the technology (00:39:10) (Villeneuve, 2016).	17
Figure 3: The red lines and pie-sliced sections highlights how the software finds patterns within the logograms (Morrison, ND).	18
Figure 4 Visually displays the ontography of this thesis and highlights the parts of its unique assemblage.	23
Figure 5: Talos as seen in the film <i>Jason and the Argonauts</i> (00:41:57) (Chaffey, 1963).	31
Figure 6: Interior of Vaucanson's Automatic Duck (Homn, 1738).	32
Figure 7: Mechanical Turk with chess player hidden underneath (Racknitz, ND).	33
Figure 8: The robot Maria from the film <i>Metropolis</i> (00:43:10) (Lang, 1927).	34
Figure 9: Robby the robot from <i>Forbidden Planet</i> (00:13:00) (Wilcox, 1956).	36
Figure 10: Cog and Kismet robots with anthropomorphic features such as eyes and human form (MIT Museum, ND).	40
Figure 11: ASIMO's hand is a highly functional compact multi-fingered hand, which has a tactile sensor and a force sensor imbedded on the palm and in each finger (Honda, ND).	43
Figure 11: Figure 12: Playing chess with HAL was through voice interaction. (01:06:06) (Kubrick, 1968).	55
Figure 13: 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 (01:07:45) (Kubrick, 1968).	56
Figure 14: The black veneering around the focus of the lip's signals to the audience that this is Hal's visual perspective as a single and circular lens (01:27:17) (Kubrick, 1968).	57
Figure 15: The Voigt-Kampff Test uses a machine to focus in and look at the suspected Replicant's eyes (00:05:27) (Scott, 1982).	60
Figure 16: The experience of looking through the Terminators 'eyes' (01:00:47) (Cameron, 1984).	61
Figure 17: An appropriation of Hodge's (1995) teratogenesis of disciplines (Akmal, 2020).	73
Figure 18: Visualising inter, multi, cross, inter and transdisciplinary approaches (Jensenius, 2012).	75
Figure 19: The mirror liquid raising (00:31:40) (Wachowski & Wachowski, 1999).	85
Figure 20: Apoc's screen showing a tunnel down through the different hyperreal worlds (00:32:12) (Wachowski & Wachowski, 1999).	86
Figure 21: The Design Family Tree with CAD residing at the top with craft at the tree's roots. An appropriation of Walker's diagram (1989).	96
Figure 22: Design as a process (Cooper and Press, 1995).	97
Figure 23: The Design square by Hatchuel et al. (2004) explores the problem-solving process of design moving between spaces of concept (C) and knowledge (K).	99
Figure 24: By seeing research as a subset of design, Faste and Faste (2012) propose a view that design embodies research with practice embodying all.	100
Figure 25: Cyclic relation between kinds of design research according to Frankel and Racine (2010).	104
Figure 26: Cyclic diagram showing the process of action research, emphasising the approach is not linear but rather iterative. Adaption from Carroll & Kellogg (1989).	106
Figure 27: Research through Design is hands on in the process of creating knowledge through design. Faste & Faste (2012).	110
Figure 28: Illustrates the interrelation of human-AI system-context. Artificial Intelligence exists only within this relationship and not only in the AI system or the interactions. Auernhammer (2020).	121
Figure 29: Modes of Rhetoric according to Aristotle, appropriated from Coulton (2015).	125
Figure 30: Rhetorical mediums Coulton (2015).	126
Figure 31: Since objects cannot exist without qualities and vice versa, there are only four possible combinations, indicated by the four lines between the circles above. Appropriated for Harman (2018).	146
Figure 32: Kuhn's ontograph framework is a graphical notation for representing types of relations in controlled natural languages where simplification is required such as technical documentation (Bogost, 2012).	147

Figure 33: The Necker Cube is an optical illusion with no visual cues to its orientation, so it can be interpreted to have either the lower-left or the upper-right square as its front side.	162
Figure 34: Ben Fry’s Deconstructulator highlighting the sprite pieces and colour palette currently in memory during gameplay. Taken from Bogost’s Alien Phenomenology or What it’s like to Be a Thing (Bogost, 2012).	179
Figure 35: An example of the many possible Alexa constellations noting some of the possible independent perspectives and interdependent relationships.	181
Figure 36: Constellations count as a small world reconfiguration as they are drawn up to map the assemblage of particular interest for design research, with the designer knowing that the points of interest have a big world impact beyond the constellation.	185
Figure 37: This diagram aids in communicating how both world building and diegetic prototypes help synthesise one another (Coulton et al., 2018).	190
Figure 38: Artefacts at different scales create a richer and more detailed fictional world (Coulton et al., 2018).	191
Figure 39: This diagram shows the trajectory of different types of futures, including wildcard futures. Diagram appropriated from Voro (2003).	194
Figure 40: This futures ‘cone’ has been adapted and integrates Gonzatto et al’s. (2013) research, whose hermeneutic model represents the ‘interpreted present’ as an interplay between past, future, reality, and fiction.	196
Figure 41: Multiple artefacts construct the world at different entry points. Appropriated diagram from Coulton & Lindley (2017).	198
Figure 42: The Near Future Laboratory’s Ikea catalogue looks just like a real Ikea catalogue but with a glimpse of the future with the addition of gardening drones (2015).	199
Figure 43: World Fairs were built to be temporary insights into the future (Comstock, 1964).	200
Figure 44: Even though this looks like a fully functional piece of technology it is not. Looks can be deceiving and that is precisely what makes great diegetic prototypes (Wilson, N.D.).	201
Figure 45: Akin to the Ikea catalogue this diegetic prototype uses familiar cues and visualisations of a typical Amazon advertisement, with the Frankenstein app part of the app range anyone could speculatively get.	204
Figure 46: Explaining individual predictions. An AI model predicts that a patient has the flu, and LIME highlights the symptoms in the patient’s history that led to the prediction. Sneeze and headache are portrayed as contributing to the flu prediction which aids the doctor to make an informed decision about whether to trust the model’s prediction (Ribeiro et al. 2016).	211
Figure 47: Label created by the Data Nutrition project showing a breakdown of the data used for the New York City tax bills with an alert count, use cases and iconography badges (The Data Nutrition Project, n.d.).	222
Figure 48: A range of typical AI iconographies.	224
Figure 49: The Peircean Triad for the iconic save icon.	225
Figure 50: Examples of indexical, symbolic, and iconic signs.	226
Figure 51: Three different style variations Pictorial, Textual and Abstract.	227
Figure 52: Version 1 of the AI icons.	233
Figure 53: Icons applied speculatively to AI-infused products Amazon Alexa and Spotify.	235
Figure 54: Participants during a face-to-face workshop using the physical cards as seen on the right.	241
Figure 55: The Making Connections GUI. Participants dragged and dropped the cards into the textual positions they thought matched.	244
Figure 56: The What’s in My AI GUI. Here participants read the scenario and clicked on to the icons they felt were in operation. Selected icons greyed out to show they were selected.	245
Figure 57: The Draw Your Own GUI. Here participants used the tools found on the left-hand side of the GUI and drew their icons in the diamond shaped icon template.	246
Figure 58: The What’s an AI’s Intrinsic Labour GUI. This was the most basic GUI designed as participants simply typed their thoughts into the box and clicked finished once they had completed an entry.	246
Figure 59: The Analyser from the matching exercise. The correct matches are box bounded in purple. The magnified section shows extra tabs for the following exercises, while underneath, one can see the tally of matches per icon.	247
Figure 60: It is fair that the icons could be exchanged for the other textual description, and the symbology would still work.	249
Figure 61: This icon design could also fall into the definitional dualism category because there is a brain drawn in this icon.	254
Figure 62: A participant’s Cloud Processing design.	255
Figure 63: The participant explained that this icon showed processing was happening at three different places internally, at the edge and externally.	255

Figure 64: An example of the participant’s biometric design; this icon signifies face scanning. The participant also connected their design and mimicked the developed symbology with the rest of the icons in that grouping.	256
Figure 65: The icon on the left and in the middle are the participants’ designs. On the right is the icon that has been designed as a response to purely signify AI is present.	257
Figure 66: On the left-hand side is the participant’s design, which is inspired both contextually and symbolically the icon on the right, which is the final design for Trained Using User Data.	257
Figure 67: Shows different ways of communicating the application of ‘classification’.	257
Figure 68: The first two columns were design ideas and suggestions, although, as explained these ideas could easily be confused with other contexts: therefore, the icons on the right, which are just the first letters of the different outcomes were used.	258
Figure 69: Version 1 and 2 of static AI.	259
Figure 70: A comparison between icons from versions 1 and 2, noting the minor adjustments to the iconic, indexical, and symbolic elements. Participants continually interpreted the X as closed and unattainable rather than the black filled circle in version 1.	260
Figure 71: Version 2 of the icons with definitions.	261
Figure 72: A design idea for the Behavioural Training Data icon of a human hand interacting with a smart object, which fits in with the semiotic design of the other icons in the training data group.	266
Figure 73: This is a developmental icon for the External Dynamic application. The icon has been developed by moving the learning scope outside the diamond shape to indicate that learning is taking place from a different location.	267
Figure 74: Icons positioned in a hierarchical order (detailed in the following passage).	269
Figure 75: A screenshot of the exercise What Matters? The columns only have 20 spaces even though there are 21 icons, meaning that participants could not place all in one section—they had to make a choice.	270
Figure 76: The participant’s designs as detailed.	271
Figure 77: The X could also be mistaken for X marking the spot.	272
Figure 78: AI to AI learning re-designs.	272
Figure 79: One of the participant’s designs for Training Data Auditable.	272
Figure 80: Participant’s Generative re-designed icons as described.	273
Figure 81: Participants’ Classification icon redesigns as described.	274
Figure 82: Appliances have a rating using an alphabetical scale. Here the participant has given a low rating of F to an AI -infused device.	274
Figure 83: The participant’s icon as described.	275
Figure 84: The participant’s icons as described.	275
Figure 85: A simple but effective design. This icon would have the same issues as the AI-assisted Decision icons of being translated into different languages.	276
Figure 86: The participant drew a planet with a blue ‘ribbon’ around it to emphasise global good.	276
Figure 87: This icon uses the icon User Training Data cupped by two hands to signify care is taken with the data and used for good purposes.	277
Figure 88: Human In The Loop idea came from a participant’s idea of Turing’s red flag.	277
Figure 89: A first iteration design of Human Out of the Loop, which is the opposite of Human in the Loop icon with the hand outside of the icon diamond to signify humans had no part of the computation.	279
Figure 90: Map of Lancaster Market Square detailing the camera and microphone positionings.	286
Figure 91: The IOAIL class Mars act as a traffic light system for quickly communicating AI legibility.	287
Figure 92:Based on the modular framework an online report was generated of the AI security systems. Different facets of the system as detailed using the icons were assessed giving the system a mark of IOAIL 2.	287
Figure 93:A Design Fiction mock-up of a news article about Roomba’s receiving a low IOAIL mark highlighting the impact such a classification mark system would have on adoption of technology.	288
Figure 94:A series of informational signs were designed and placed around the square signifying AI was being used in the Market Square.	289
Figure 95:As well as the informational signs, signs were created of the AI ontographs and placed near or next to where the AI security was installed.	290

Chapter One Introduction

1.1 Arrival's Logograms (spoilers lie ahead)

The film *Arrival* (Villeneuve, 2016), based on the short story *Story of Your Life* by Ted Chiang (1998), focuses on linguist Louise Banks (Amy Adams) and physicist Ian Donnelly (Jeremy Renner) trying to communicate with mysterious aliens, called Heptapods. The Heptapods appear around the world simultaneously in pairs, to give the human race technology so that in return, in 3000 years, humans will help save their civilisation. The technology is logograms— a timeless language that unlocks the ability to see the future. The moral of the narrative is that communication, as a technology, is key to thriving, not only as a nation but together as a larger community worldwide. This sentiment bears some resemblance to this research as it is concerned with creating a system to communicate AI's functions and operations, where many have also described AI as alien technology (Bogost, 2012; Lindley & Coulton, 2020; Weld & Bansal, 2019a). This description, however, is describing AI's strange and unfamiliar characteristics rather than to signify from another world; though it does convey that the way to communicate AI's being must also be emblematic of its alien existence.

Speaking of *Arrival's* linguistics— “[w]e wanted to create a language that is aesthetically interesting,” says production designer Patrice Vermette. "But it needed to be alien to our civilization, alien to our technology, alien to everything our mind knows” (Vermette quoted in Rhodes, 2016). Vermette knew the alien language would appear in circles— the screenwriter, Eric Heisserer, stated as much in the script. This is because the Heptapods regard time as non-linear, and the language needed to reflect that; it needed to be more than human. However, consultations with graphic designers and linguists kept leading to fictional alphabets that Vermette says “hewed too closely to familiar systems like hieroglyphics, or code” (Ibid): their attempts were too human.

Vermette's artistic wife, Martine Bertrand, took the lead in visualising the written language, creating around 100 swirly circular symbols for the alien language (Figure 1). Vermette and his team then assigned meaning to the inky tendrils that emanate from each ring, developing a dictionary of 100 symbols. A logogram can express a simple thought or a complex sentence; the distinction lies in

the complexity of the shape. A logogram's weight carries meaning, too: a thicker swirl of ink can indicate a sense of urgency; a thinner one suggests a quieter tone; a small hook implies a question.



Figure 1: A basic logogram with inky tendrils (Morrison, ND).

Throughout the film, two Heptapods draw circular designs on a giant transparent wall (that acts as an atmospheric barrier separating the aliens and humans) to communicate back and forth with Louise and Ian, unbeknownst to them that the logograms are the technology (Figure 2).



Figure 2: The barrier between the aliens and humans is also the site where the aliens share the technology (00:39:10) (Villeneuve, 2016).

The production team brought in the founder of Mathematica coding software Stephen Wolfram and his son, Christopher Wolfram, to analyse the language the way Louise and Ian eventually do on screen. The Wolframs cut the logograms into pie-shaped sections of 12 and, through their software, found that specific patterns repeat, marking intent in the logograms (Figure 3).

Figure has been removed due to copyrights restrictions

Figure 3: The red lines and pie-sliced sections highlights how the software finds patterns within the logograms (Morrison, ND).

The reason for the brief venture into the film *Arrival* (Villeneuve, 2016) is that part of this research attempts to create a symbolic system to communicate AI functions through a More Than Human-Centred approach and, in the process, develop an (alien) interpretation of AI as a material for design.

1.2 Research Problem & Significance

AI is becoming increasingly ubiquitous. Enabling service providers to monitor in significant detail users' behaviour through data (often without explicit consent (Zuboff, 2019)) and subsequently turn this data into decisions and predictions, which are increasingly cited as potentially producing harmful results (Angwin, et al., 2016). AI technology is implemented into a wide range of everyday applications from social media, shopping, and media recommendations and is increasingly making decisions about whether we are eligible for a loan, health insurance and potentially if we are worth interviewing for a job. This proliferation of AI brings many design challenges regarding bias, transparency, fairness, accountability, trust, etc. It has been proposed that these challenges can be

addressed by considering user agency, negotiability and legibility as defined by Human Data Interaction (HDI). These concepts are independent and interdependent, and it can be argued that by providing solutions towards legibility, we can also address other considerations such as fairness and accountability. This design research perceives the challenge of legibility as a case study for investigating AI as a material for design, while illustrating how design-led research can deliver practical solutions towards legible AI.

In an attempt to combat harm, we have seen a proliferation of frameworks, principles, and guidance documents for AI. Of particular note regarding legibility are the themes of transparency and explainability, which are considered principal challenges impacting AI implementation (Fjeld et al., 2020). Whilst design thinking is cited in many frameworks as a means of potentially addressing AI concerns, design, in these instances, can merely be viewed as the outlining of problems rather than providing practical responses, which may be seen as the false promise of design thinking (Kolko, 2018). In reality, it perhaps reflects the need to articulate better how designers can provide approaches which traverse the current gap between abstract principles and specific implementation. To address this issue and taking inspiration from lived experience, this thesis presents research that practically addresses AI legibility through an iterative Research through Design (RtD) and More Than Human Centered Design (MTHCD) enquiry into AI iconography. The AI icons are a visual AI lexicon that communicates and diffuses the complexity of AI functions and raises user awareness of how AI operates within the products and services they use. Essentially the visual AI icons divulge an AI's ontology through a series of graphics that have been semiotically designed to represent the various functions and operations of an AI. The prospect of communicating AI's ontology, to establish a mechanism for AI legibility, was conceived by speculating on AI's being through a MTHCD investigation of AI. In total, through an iterative design, twenty-one icons were designed that articulates different functioning attributes of AI, such as an AI's learning scope, the type of training data required for functionality, the processing location of the training data, the provenance of the training data and the overall productivity of an AI, for example providing a recommendation or a generative output. Furthermore, this thesis will provide the theoretical underpinnings that led to the

design artefacts and detail the process of iterating the AI icons via a series of workshops using bespoke tools.

1.3 Research Questions

The core research question explored through the research in this thesis is:

RQ1. How can we craft an approach that explores how the materiality of AI manifests itself in design practice, using lenses derived from object-oriented ontology and postphenomenology?

Using the explorative and iterative RtD methodology (Frayling, 1993; Gaver, 2012) the following research questions emerged through the research process, which are answered throughout this thesis as insights and approaches.

RQ2. How can we design philosophical probes to explore design challenges such that they produce practical outcomes that explore the materiality of AI, such as an AI lexicon that contributes to AI legibility?

RQ3. Can the adoption of a More-than Human-Centered Design approach aid in the creation of alternative perspectives of the materiality of AI that challenge the dominance of science fiction renderings?

RQ4. How do we apply the consideration of AI as a material so that it produces practical solutions for living with AI?

Each of these four questions are addressed within the forthcoming chapters.

1.4 PETRAS IoT Hub and Research Track

This research was enabled by support of The PETRAS IoT Hub Project.¹ The funding was part of the IoT UK government-funded programme seeking to advance the UK's leadership in IoT technology. PETRAS has presented many findings spanning different tracks relating to ethics, trust, acceptability, adoption, security, and reliability. The track this research focusses on is the future adoption and acceptability of so-called smart devices and smart homes which would utilise IoT technologies infused with AI technology. These devices are typically designed for ease of use, with

¹ For more information, see: <http://www.petrashub.org/>.

their complex underlying procedures (intentionally) obfuscated, while explaining particular outcomes is hampered by their inherent ambiguity. This lack of legibility leads to misconceptions about how AI works. Through design research, this thesis addresses the challenge of AI legibility, conceiving it as a material for design, by designing AI iconography as an accessible way to communicate and better understand the role AI and data increasingly play in our everyday interactions.

1.5 Why Philosophy

While RQ2 specifically mentions philosophy it important to note this research is not *on* philosophy but instead, uses philosophical discourse and thinking to influence the practice of design. This research aims to develop a More-Than Human perspective of AI through a philosophical lens for incubating an alternative comprehension and design approach for designing with and for AI technology. The critical point here is the development of a perspective that perceives AI as an object within itself beyond the boundaries of human involvement.

In other words: this research endeavours to create a discussion about AI technology that is not based on biased design tenets that confirm to human-centredness. Instead, this research attempts to cultivate ‘object empathy’ through an Object-Orientated Ontology perspective. However, the human is reintroduced later in the philosophical discussion regarding human relations with technology and measures designed to have better relations with technology through a postphenomenological reasoning. Throughout the thesis, the reader will be able to notice alternative states of AI: Chapter Two, *Seeing AI*, Chapter Three and Four, *Understanding AI*, Chapters Five and Six, *Being AI* and Chapters Seven and Eight, *Designing for Human-AI Kinship*.

1.6 Why Fiction

In Chapter Three, *Groundworks*, an argument is made for the adaption of philosophy by describing *The Matrix*'s (Wachowski & Wachowski, 1999) adaption of Jean Baudrillard's concept of hyperreality. This research also draws on science fiction and technoscience in contemporary films, as cinema and television are great litmus tests for social unease and yearnings, from alien invasions that paralleled Cold War paranoias of the '50s to contemporary anxieties of creating artificial intelligence making humans redundant. The film researcher Aylish Wood writes:

The cultural products of any given period both expose and explore the concerns of that period, and whilst they are certainly fictionalised and packaged to fit the conventions of different kinds of genres, these products nonetheless touch on very real questions...[many such questions are] about what it means to be human in the late twentieth century (Wood, 2002).

For viewers, science fiction is a critical source of images and portrayals of science and technology. Films greatly influence our perspectives of technology and have aided in confusing what AI technology means. Although, as will be seen in Chapter Two, *Seeing AI*, the creation of sentient life was the foremost intention in AI research, which in turn has been reflected in many films.

1.7 Structural Outline

This thesis comprises nine chapters (Figure 4), Chapter One introduces the research questions, and an argument for the value of this research has been made. Chapter Two is a literature review of AI technology, recounting the perception of AI through a brief history of the technology coupled with inspecting AI as it is portrayed in popular media and films. This chapter features the ‘Seeing AI’ phase of this thesis. Chapter Three presents the method assemblage and the transdisciplinary structure of this thesis. Chapter Four presents the overarching RtD methodology of this thesis, explaining design research and how findings and further questions have emerged as part of the iterative process. Both Chapters Three and Four outlines the theoretical framework for ‘Understanding AI’.

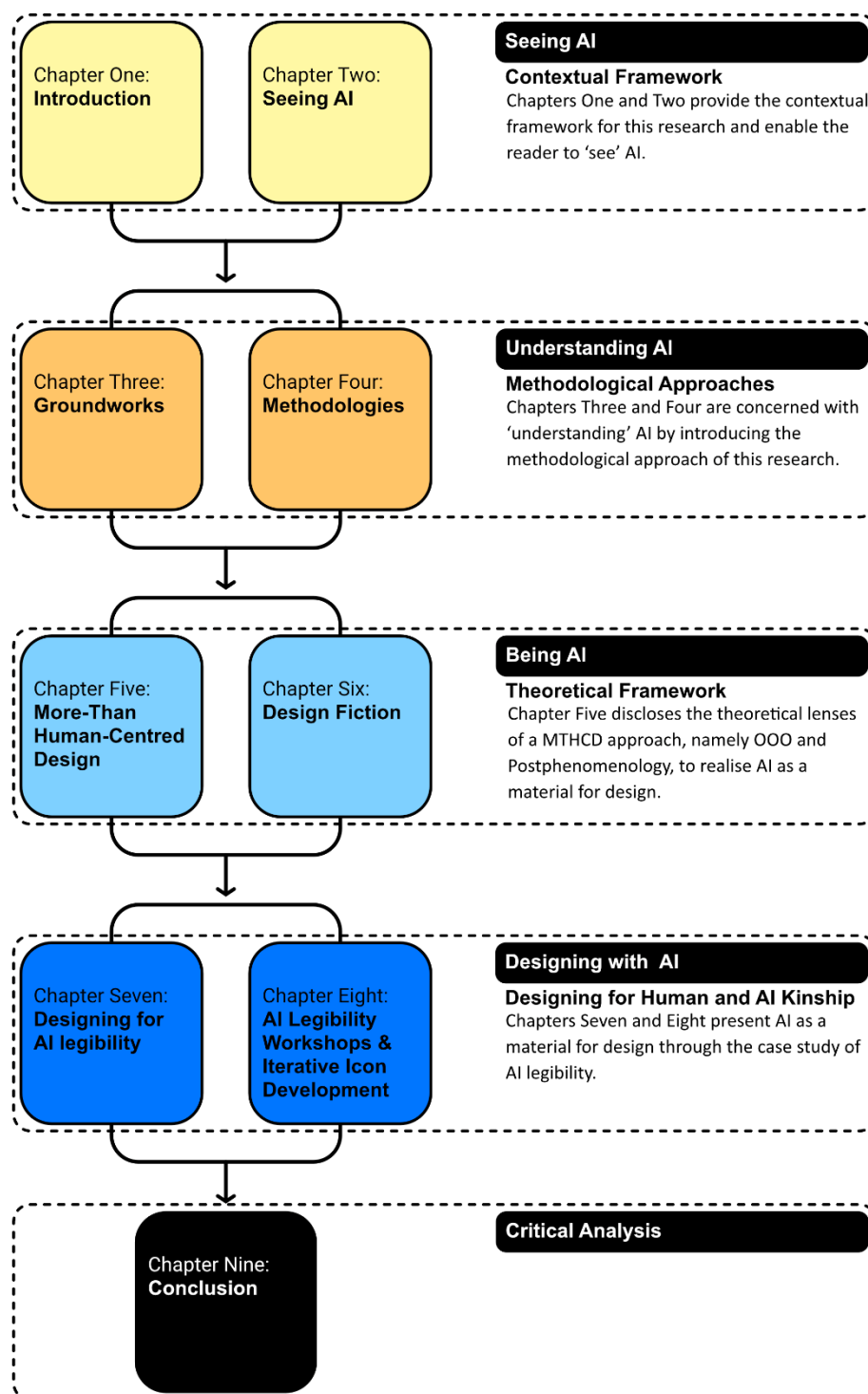


Figure 4: Visually displays the ontography of this thesis and highlights the parts of its unique assemblage.

Chapter Five introduces the concept of 'Being AI' in three parts; part one presents an overview of Human-Centered Design (HCD) and how this approach is reflected in current design thinking for AI, precisely through the concepts of simplicity by design, interaction design and persuasion by design. After reviewing accountabilities found in HCD, justifying a MTHCD approach, the second part of the chapter presents an outline of Object- Oriented Ontology to be metamorphosed through adaptation with speculative design in Chapter Six. Part three reintroduces the human user

back into the approach, developing a Human-AI Kinship through applying a post-phenomenological lens as, ultimately, the design artefacts are designed for human use. Chapter Six, as already stipulated, is concerned with practising philosophy through design via the approach of Design Fiction.

Chapter Seven presents ‘Designing for Human-AI Kinship’ by comprehensively viewing the concept of legibility and explains the difference between explainability, interpretability, and transparency. This chapter also introduces the system of icons and the rationale behind their design. Chapter Eight is in three parts. Part One presents the design and development of the online workshop tool and the testing of the first iteration of icons with the development of the second iteration. Part Two showcases the second iteration of the workshop for testing the second iteration of icons and analyses the results. Part Three presents a short-term project of the icon’s deployment into the wild.

Finally, Chapter Nine concludes this thesis by describing how this research has answered the questions set by this research and demonstrates how this thesis has contributed to new knowledge.

Chapter Two Seeing AI

2.1 A Brief AI History – The Evolution of Two Categories of Intelligence; Artificial General Intelligence / Machine Intelligence

Technological determinism would have us believe that any given problem can be solved by applying computation and, by extension AI. Nevertheless, AI functions and operations are opaque and illegible: taking place inside machines, data-gathering services, and governing systems, hidden by the complexity of systems' architecture and code. The illegibility of these systems also stems from Human-Centred Design's (HCD) axiom of simplicity for interaction with technology (Norman, 1998), whereby many underlying AI operations are obfuscated from the user experience, escalating technological illiteracy. For experts, an understanding of AI is also in constant flux, with definitions of AI often contested due to evolving theories, findings and perceptions (Elish & Boyd, 2018a; Hawley, 2019). Even the language we use in the context of AI is misleading (Elish & Hwang, 2016), habitually evoking anthropomorphised renderings.

This ontological review will consider AI beyond a mathematical understanding. Although it will not be a More-Than Human perspective yet, it will, however, focus on the blended and confused perceptions of AI stemming from the amalgamation of science and science fiction representations. This chapter will also highlight the socio-technical challenges flourished by particular encoded agendas and how design research can facilitate solutions towards offsetting the ambivalence of AI technology (Feenberg 2002, pp. 91–113). To articulate the challenges and the ambiguity surrounding AI, the following section will summarise current frameworks and guidance relating to responsible AI, situating this research's focus for designing possible solutions towards legible AI. Subsequently, a brief account of AI's history will unpack this multifaceted field's technical practices and positionings. The second part of this chapter will be a closer inspection of the challenges that confuse users' insights of AI through popular cultural representations and perceptions of AI, developing into rationales towards possible ways design solutions can counteract AI's pitfalls.

This review's practical and initial grounding definition is "Artificial Intelligence can be broadly understood as a characteristic or set of capabilities exhibited by a computer that *resembles* intelligent behaviour" (Elish and Hwang 2016, p. 8, emphasis added). M.C Elish and Tim Hwang further elaborate that "[d]efining what constitutes intelligence is a central, though unresolved,

dimension of this definition” (Ibid). The undetermined nature of intelligence associated with AI will be demonstrated in this review, as well as highlighting the effects of an entangled perception of AI emerging from fictional AI representations, therefore building the case for adopting a non-human perspective for Human-Computer-Interaction (HCI) design for AI.

2.2 Towards Responsible AI Through Legibility: Guidance and Frameworks

The perplexity of AI’s computational nature infused into black boxes and the supplementary obfuscation introduced for corporate secrecy (Burrell, 2016) is an obvious challenge for those working in the domain of AI, such as designers, researchers, theorists, and policymakers. This situation has resulted in initiatives, companies, researchers, and governing bodies developing frameworks and guidance documents promoting the advancement and use of responsible AI (Algorithm Watch 2019; European Commission 2021; Pichai 2018; Microsoft ND; Amazon ND). While the spectrum of proposed interpretations for responsible AI is broad, there is a shared rationale between them, with a recent analysis by Fjeld et al. (2020) distilling the main challenges into a singular guidance for responsible AI, with the key themes being:

Privacy – relating to how data is collected, stored, managed, and used, with much of the governance and safeguards enshrined in human rights laws such as GDPR (European Union, 2016).²

Accountability – considers who is accountable in the age of autonomy, with the necessary mediation and analysis of liability and legal responsibility, which include impact assessments and verifiability of AI functionality and appeal processes.

Safety and security – beyond science fiction imaginaries, there is a real danger of harm when something goes wrong with AI, either in the remit of harming its environment or someone physically or digitally.

Fairness – this relates to if AI-infused systems are creating decisions that go on to detrimentally impact or privilege particular populations. This is due to systems trained on biased and unrepresentative data and the classification design for decisions.

² GDPR stands for the General Data Protection Regulation and is a Regulation in EU law on data protection and privacy in the EU and the European Economic Area.

Agency – is concerned with giving users the capacity to act within these AI systems, such as control, inform and correct gathered data, and having opt-in or opt-out options.

Transparency, Legibility & Explainability – these terms are used almost interchangeably; however, they describe subtly very different things in the context of AI (Lindley & Coulton, 2020). Transparency concerns how open the data and algorithms are to outsider auditing to be verified or challenged. This openness is suggested for the whole AI system's design and implementation. Explainability relates to making AI systems and their decisions understandable to experts. The legibility tenant of responsible AI will be the focus of this research and will be analysed in more detail in Chapter Seven. Though, as a brief synopsis, legibility is concerned with how we can make AI systems and their decisions understandable and readable to non-AI experts (Lindley et al., 2020; Pilling et al., 2020; Pilling et al., 2022; Pilling et al., 2022; Pilling et al. 2020).

Human Centeredness – pertains to designing and developing AI systems that are easily operated by their users and serve humanity's best interests and values regarding inclusivity, social norms, and cultural beliefs.

In relation, the well-established pragmatic and empirical approach of Human-Centred Design is known to make sense of the world, consider every human the design has an impact on, and focus on the users' needs and requirements (Giacomin, 2014). Though, this approach often results in designs obscuring the complexity of technology for simplicity and ease of use by disappearing them into the background for seamless interaction, causing a dissociation between user and technology that can lead to varying degrees of harm, particularly in the new age of IoT and AI technologies with their intangible disposition. More-Than-Human Centred Design is a counter-response to this, which is a topic that will be covered in more depth in Chapter Five; however, to summarise, it is an approach that considers the independent and interdependent perspective of *every* thing (human and non-human) in a framed ecology of interaction (Coulton & Lindley, 2019; F. Pilling & Coulton, 2021).

Design thinking is cited in many frameworks as a means of potentially addressing AI concerns. However, it is merely outlining problems rather than providing practical responses. On this note, this may be seen as the false promise of design thinking (Kolko, 2018), though in reality, it perhaps reflects the need to articulate better how designers can provide approaches which traverse the

current gap between abstract principles and specific implementation. In response to this, this research will focus on designing legible interactions with AI technology using a More-Than-Human-Centred Design approach.

2.3 Introducing an AI History

The following AI timeline will not be a complete account of AI but rather a sketch showcasing a research and social division between two categories and embodiments of intelligence in AI. These are: the quest for AI technology to exhibit –human general intelligence– and the creation of man-made artificial beings known as Artificial General Intelligence (AGI) influenced by science fiction renderings; and secondly, Elish and Hwang’s concept of – machine intelligence – framing “the capacities and limitations of what intelligence, of all degrees, may look like embodied in a machine” (2016, p.12). In other words, the mundane reality of narrow AI, a class of technology with features of AI and automation effectively defined as “a device or system that accomplishes (partially or fully) a function that was previously, or conceivably could be, carried out (partially or fully) by a human operator” (Parasuraman, Sheridan, and Wickens 2000, p.287).

In the current era, the socio-technical landscape has become more infused with AI, driven by the growth in the availability of large data sets, significant progress in cheap computational power, and developments in data science. These advancements have permitted powerful algorithm-based technologies and methods, dubbed AI but implemented through Machine Learning (ML), Deep Learning and Neural Networks, which have become increasingly ubiquitous in our daily activities by empowering smart thermostats, streaming services, and AI assistants, such as *Alexa*. Domingos states that “[m]achine learning is not magic; it cannot get something from nothing”, going on to explain what technology does do, is get more from less (2012, p. 81). Pedro Domingos likens the process of ML to farming; first, you prepare the seeds and nutrients and let nature do the work: “learners combine knowledge with data to grow programs” (Ibid). Incidentally, data is a vital component of any AI discussion, as it is the primary driver for the current resurgence of AI research with the promise to present insights on par with human intelligence through the “purportedly neutral collection” and analysis of Big Data through AI (Elish and Boyd 2018a, p. 58). These Big Data practices have been

revealed to encode and magnify social values (O’Neil 2016; d’Alessandro, O’Neil, and LaGatta 2017); thus, both AI and data should be considered “social-technical concepts” (Elish & Boyd, 2018a), which will also be probed in the following AI history, presenting advocacy for designing legible interactions with AI.

2.3.1 A Brief Interlude in Human Intelligence with an AI agenda

Notoriously, AI is compared to human intelligence with the aim of emulating it in some form. As a general concept, intelligence has been described as the mental ability for reasoning, problem-solving, and learning through cognitive functions such as perception, attention, memory, language, or planning (Colom et al., 2010). However, the field of human intelligence is often disputed with no standard definition of what exactly constitutes ‘intelligence’.

Alan Turing’s influential paper *Computing Machinery and Intelligence* (1950) begins with a question that continues to dominate the technological discourse of AI: “Can machines think?” (Turing 1950, p. 433). Turing proposed to answer this question through an “imitation game” with the objective of a machine imitating the behaviour of a human player by providing answers “that would naturally be given by a man” (Ibid, p.459), noting that:

Intelligent behaviour presumably consists in a departure from the completely disciplined behaviour involved in computation, but a rather slight one, which does not give rise to random behaviour, or to pointless repetitive loops (Ibid).

Regarding the current “ontological system operation” of AI, the nature of machines is correlated to the nature of human experience (Bogost 2012, p. 13). In his game’s concept, Turing attempted to avoid “the normal use of words” (Turing 1950, p. 433) to define ‘machine’ and ‘think’, “drawing a fairly sharp line between the physical and the intellectual capacities of man” (Ibid, p.434) in the optimism that “machines will eventually compete with men in all purely intellectual fields” (p.460). In the six decades since Turing’s question, the operation or the thought of machines has been entangled with humanistic conditions— with human intelligence as a goal, or at least an impersonation. Consequently, the construction, programming, and improvement of machines are a global industry worth billions, where there is little room to understand the machine as a thing in itself.

2.3.2 Pre-20th Century; Mechanical Imitation, Hoaxes and Arranging Knowledge

Humanity has been fascinated with artificial beings for millennia. Talos, first mentioned in 700 B.C., was a giant bronze humanoid automaton that circled the island shores of Europa three times daily to protect it from invaders (Figure 4). The myth of Talos offers one of the earliest conceptions of a robot (Mayor, 2018), challenging what it means to be human and questioning the humanistic values embedded in technology. The ancient Greeks were fascinated by the manipulation of natural life and metaphysical enquiry, with their myths prolifically studying bio-techne (bios = life, techne = crafted by science) and mortality. In this period, the earliest known organisation of intelligence existed in Babylonia, a library with a collection of 30,000 clay textual tablets inscribed in various languages.



Figure 5: Talos as seen in the film *Jason and the Argonauts* (00:41:57) (Chaffey, 1963).

Several hundred years later, the Greek philosopher Aristotle (384 B.C. – 322 B.C.) formalised ‘logic’ revolving around the notion of deduction (sullogismos), whereby conflicting premises of an argument reaches a conclusion through deductive reasoning. The core algorithmic techniques used in AI for inductive and probabilistic reasoning can be traced back to Aristotle’s theory of Syllogism. To this end, several philosophers consider AI research and development a philosophical inquest (Bringsjord and Govindarajulu 2018).

Ismail al-Jazari (1136 CE – 1206 CE) was a scholar, artist, mechanical engineer, and known as the ‘father of robotics’. However, he is most famous for writing *The Book of Knowledge of Ingenious Mechanical Devices* (1974), which describes the design and construction methods of 50

mechanical devices. These included a range of humanoid automata servants. One automaton's purpose was to wash a user's hands using a technical system that utilises the flush mechanisms in modern toilets.

Jumping ahead, during the 17th and 18th Century the art and popularity of mechanical statues swept over Europe, reaching its zenith with a mechanical duck by the inventor Jacques de Vaucanson, which could flap its wings, eat, digest, and excrete grain through elaborate tubing in its stomach (Figure 5). Vaucanson was deliberately vague regarding the construction of the duck, however, stating that the creation of the duck was not intended as a perfect imitation regarding the nourishment and blood processes; rather, it was for the simulation of more prominent features such as intake and excretion. It has been recognised in the AI field that Vaucanson's aim of imitation should be considered when we come to the simulation of human thought processes instead of striving for replication (McCorduck 2004, p 17).



Figure 6:Interior of Vaucanson's Automatic Duck (Homn, 1738).

In 1769, the Hungarian inventor Wolfgang von Kempelen built and toured a humanoid automaton that played chess, the famous *Mechanical Turk*. The *Turk* was an elaborate mechanical illusion, with a human chess master concealed inside the automaton's base, who would operate the automaton's features to move the chess pieces (Figure 6). Today, Amazon pays homage to the *Mechanical Turk* by naming one of their online services *MTurk*, a crowdsourcing website for businesses hiring humans to perform 'Human Intelligence Tasks'. These tasks are beyond the scope of what computers can currently do, such as identifying and labelling specific content that goes on to be data used for training AI programs. In a similar trick of illusion, the operation of human data labellers

is obfuscated from the process, with AI algorithms performing the final output or service; consequently, users are led to believe that current AI systems are more sophisticatedly intelligent.



Figure 7: Mechanical Turk with chess player hidden underneath (Racknitz, ND).

During the 19th Century, artificially intelligent beings flourished in literary narratives, such as Mary Shelly's *Frankenstein, Or the Modern Prometheus* (1818), questioning the relationship between science and nature with the creation of sentient beings (Hammond, 2004). Coincidentally this period also marked the start of Empirical Psychology.

In 1822 Charles Babbage and Ada Lovelace started, although never finished, the *Analytical Engine*, a mechanical general-purpose computer. Lovelace is often considered the first computer programmer, as she was the first to recognise that machines have applications beyond pure calculation and could extend towards algorithmic problem-solving.

Towards the end of the century, Samuel Butler wrote *Erewhon*, a narrative focused on a speculation similar to the Singularity theory, which combined the rapid onset of the Industrial Revolution (1760 -1840) with Darwin's theory of Evolution towards machine consciousness, eventually superseding humanity.³

2.3.3 The 20th Century

³ The Singularity, according to Ray Kurzweil's theory, is a future epoch, where technological progression will be swift, bearing insurmountable impact – transforming humanity irreversibly (Kurzweil 2013).

In 1920, Karel Capek invented the word ‘robot’ for his play *Rossum’s Universal Robots*, where robots serve human beings and are deemed more consistent than humans by their inventor. Eventually, the robots rise up, threatening the human race to extinction. Capek’s play is the inception of the classic AI narrative, explored years later by Fritz Lang’s seminal masterpiece *Metropolis*, which brought anthropomorphised and artificially intelligent beings to the silver screen in 1927 (Figure 7).



Figure 8: The robot Maria from the film *Metropolis* (00:43:10) (Lang, 1927).

Only ten years later, in 1937, Turing conceived the basic principle for modern computers in an abstract concept known as the *Universal Turing Machine* (UTM). The UTM, in theory, would read and execute coded instructions inscribed on its tape, essentially the 'stored program' concept. Around the same period during World War II (1939-1945), Isaac Asimov published the ‘Three Laws of Robotics’ in the short story *Runaround* (1942). The laws were created as a framework to avoid writing stories about robots who would senselessly turn on their creator, therefore circumventing the penning of another Faustian punishment story (Asimov, 1964). The laws, as Asimov noted, were obvious safety mechanisms to be attributed to all tools and humans, though pragmatically observing that at times “[t]he safety may not be perfect (what is?)” (Ibid, p. 17). Interestingly, the infamous AI

researcher Marvin Minsky accredits the story *Runaround* for starting his lifelong contemplation on how minds might work (Markoff 1992, para 17).

After the war, Vannevar Bush, the director of the American *Office of Scientific Research and Development*, called for a “new relationship between the thinking man and the sum of our knowledge” (Bush 1945, para 1), and for the development of technology that would promote “the application of science to the needs and desires of man” (Ibid, section 8, para 10). Bush’s idea was a speculative vision in a time of information overload stunting the growth of new knowledge. The solution was a *Memex* machine, which would tag information with ‘trail codes’ and retrieve information through association or ‘information curating’(ibid). *Life Magazine* (1945) published Bush’s essay under the title “Machines will start to think”, characterising a long tradition of hyperbole in the media, misinforming the true capabilities of technology.

2.3.4 The ‘Good Old-Fashioned AI’ Days

A few years after Turing wrote his influential work on *Computing Machinery and Intelligence* (1950), in 1955, the cognitive scientist John McCarthy coined the term ‘Artificial Intelligence’, in the proposal for the influential Dartmouth conference, proposing that “every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it” (McCarthy et al. 1955a, para 1). However, Allen Newell and Herbert Simon referred to the field as *Complex Information Processing* rather than AI. They strove to develop the *Logic Theory Machine*, researching the suitability for a computer to demonstrate problem-solving in non-numeric domains by heuristically searching using humanoid heuristics (Simon, 1998). Despite the same connotations of mimicking and being influenced by humanoid cognitive processes, given how AI technology operates and how it is used, the moniker ‘Complex Information Processing’ would have arguably led to fewer misunderstandings about the field (Lindley and Coulton, 2020). Although, McCarthy used the term AI to conceivably create a buzz and distinguish the field from Norbert Wiener’s field known as Cybernetics (derived from Greek to mean ‘the art of steering’), which focused on controlling the flow of information with feedback loops in biological, mechanical,

cognitive, and social systems (1948). An interesting fact is that Machine Learning technology originated from Cybernetics and was later adopted by the AI research field, overshadowing its origins.

Only a year later, in 1956, the historical and culturally significant film *Forbidden Planet* was released. The film was noteworthy due to the visual effects and the plot regarding the materialisation of monsters from a human's psyche (Wilcox). It was the first visual account of a robot that resembled a humanoid form rather than a tin can while also displaying a distinct personality (Figure 8), defining the film as investigating the complexities of the psyche in varying entities.



Figure 9: Robby the robot from *Forbidden Planet* (00:13:00) (Wilcox, 1956).

In 1964, Joseph Weizenbaum published an early Natural Language Processing (NLP) computer program, or what is now referred to as a chat-bot, called *Eliza*. *Eliza* would conduct 'therapy sessions' for the user by interacting and responding to the user by imitating a therapist's typical response and questioning tactics. Essentially, the program picked out the user's constituent parts of speech from their inputs and then fed them back to the user by rephrasing the input in a manner sustaining the conversation. Weizenbaum's computerised therapist, with its pre-programmed responses mimicking human dialogue, is an example of 'Classic AI' or 'Good Old-Fashioned AI' (GOFAI) (Haugeland, 1985). The Classic AI approach attempted to fabricate an AI system to copy and mimic human intelligence, whereby one could argue, by definition, undermines what intelligence is if a machine is simply copying. Classic AI tried 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 (Warwick, 2012). The research was successful in concluding that the human brain

was capable of reasoning and deciding on an answer, forming an early depiction of what intelligence is comprised of, and establishing the basis for the IF (condition) THEN (conclusion) statement within computing (Ibid).

A few years later, in 1968, the visionary film *2001: A Space Odyssey* was released (Kubrick). Stanley Kubrick's catalogue of cinematic works is renowned for exploring humanity's complicated relationship with technology. The film presents a near and future world, speculating on technological ascension that permits the exploration of the far corners of space, the creation of AGI, and the transcendence of humanity— from the tool-wielding Dawn of Man to the ultimate birth of Star Child, driven by the invention of AGI. Discovery's AI system, *Hal9000*, is a convincing speculation of intelligence, with Kubrick's vision showcasing current and long-term aims in AI research, such as AI chess playing, AI vision, AI translation and language, and ultimately intelligence, both machine automated and a human level of intelligence.

Minsky served as an AI consultant for *2001: A Space Odyssey*. At the beginning of the 1960s, he speculated that he would witness machines surpassing humans in general intelligence in his lifetime. Although, by the end of the 60s, it was becoming evident that this speculation would not transpire. To this effect, Minsky co-authored *Perceptrons* (1969) with Seymour Papert, which pointed to key problems with the promise of neural networks and a need to understand how computers compute. *Perceptrons* has been identified for redirecting funding away from AI research, bringing forth the dawn of AI's first winter. On this very note, of examining the expectations of AI technology, Hubert Dreyfus devoted his career to determining whether a difference existed between man and machine. He was valiant in calling out current trends and myths within AI research, mainly through his book; *What Computers Still Can't Do* (1972), which was criticised until it became evident that the speculative promise of computation was not going to be realised. According to Dreyfus, there is an essential difference between human beings and computers:

[t]he human world, then, is prestructured in terms of human purposes and concerns in such a way that what counts as an object or is significant about an object already is a function of, or embodies, that concern. This cannot be matched by a computer, for a computer can only deal with already determinate objects[...] (Dreyfus, 1972, p. 173)

The last straw for the UK AI research program came a year later when Sir James Lighthill wrote a report for the British Government. The report concluded, “in no part of the field have the discoveries made so far produced the major impact that was then promised” (Lighthill, 1973, p. 9), which resulted in freezing the UK’s funding for AI research.

The USA’s AI winter was already in full effect due to a report by the United States Government in 1966, which concluded that research into AI was futile based on low success rates (Dreyfus 1972). Before the report, AI research was heavily funded in an attempt to conceive a machine for Machine Translation (MT) that would operate a realistic set of translation rules using algorithms for simple recognition routines, ultimately replacing human translators. This goal gained additional traction with the onset of the Cold War, with the USA actively pursuing an advantage over the USSR (1947-1991).⁴ The hype for MT started as early as 1954, with a public event held by the computer hardware company International Business Machines Corporation (IBM), showcasing their successes stemming from their early research into machine translation. The *IBM 701* system automatically translated 60 sentences from (Romanized) Russian to English. This was enough for IBM to embellish the truth of the technologies’ capabilities in a press release calling the 701 computer a “versatile electronic brain” (IBM 1954, para 7). An active response from the Soviet government only fuelled the hype, corresponding with an increase in funding from the US government (Hutchins, 1996). Time showed that MT research followed the same trajectory as previous AI research, whereby the technology would not achieve the promises it once presented, greatly disappointing the government and the general public alike (Ibid). Funding was withdrawn from all fields relating to AI. Although progress into AI continued slowly under the guise of different research headings, such as ‘informatics’ and ‘pattern recognition’, the term artificial intelligence was avoided “for fear of being viewed as wild-eyed dreamers” (Markoff 2005, para 2).

Regardless of the AI winter, the late 70’s generated significant science fiction film franchises with AI characters having, or playing, central roles in the story arcs such as *Star Wars* (Lucas, 1977) and *Alien* (Scott, 1979). *Star Wars* was notable in its depiction of honourable AIs with good intentions

⁴ The Union of Soviet Socialist Republics (USSR) or the Soviet Union was a transcontinental country that spanned much of Eurasia.

towards humanity and for its anthropomorphic and zoomorphic representations of *C3PO* and *R2D2*.⁵ Nevertheless, the entertainment world and its audiences were still captivated by science fiction stories of murderous AI and sentient robots. Such as the uncanny depiction of humanoids revolting in *Westworld* (Crichton, 1973) and the AI that controlled the operation systems of a house, imprisoning the occupant for procreation in *Demon Seed* (Cammell, 1977).

The first AI winter thawed in the 1980s when the consumer market started implementing AI technology known as Expert Systems, with the field maturing into engineering and shifting the focus from ‘intelligence’ to ‘knowledge’. These AI systems were simplistic and less ambitious than the speculative AI systems of previous decades, with the commercial systems moving away from general intelligence to performing narrow and automated tasks through extremely specific rules. Such systems were implemented to perform various but explicit tasks; for example, the Digital Equipment Corporation (DEC) created a system for configuring compatible computer parts for sale. Developed in collaboration with Carnegie-Mellon University, this transition highlighted AI research shifting from academia to industry and, with it, the vastly different expectations of AI research solving real-world problems (Polit, 1984). Data was gathered by one or two human experts and encoded into rules composing the computer system as an attempt to emulate a human’s decision-making. However, these Expert Systems were criticised as being “brittle” (Forsythe 1993, p. 466) and further condemned by McCarthy for lacking common sense regarding their own limitations. Using the example of the Expert System *MYCIN* used to assist physicians, McCarthy explained that the system would proscribe a treatment of antibiotics for the *Vibrio cholerae* infection that, in due course, do as intended and eradicate the bacteria but would also kill the patient of Cholera before that (J. McCarthy, 1984).

As Expert Systems fell out of favour, new approaches for intelligence emerged from the reigning method of logic and reasoning, abstractly representing models of the world (Elish & Boyd, 2018a). The mid 90’s brought the development of the Humanoid Robotic department at MIT, a moment in time when the research field and the science fiction world unashamedly unite, with notable

⁵ Anthropomorphic means to ascribe human characteristics to nonhuman things or nonhuman things having a human form or human attributes. Zoomorphic means to attribute animal forms or animal characteristics to other animals or things other than an animal.

research projects such as *Kismet*, an expressive robotic ‘creature’ and *Cog*, a ‘human-like robot’ (Figure 9). *Kismet* was an experiment at developing an ‘expressive anthropomorphic’ robot, which engaged with human counterparts by processing visual and auditory input as triggers to motor outputs to ‘act’ out in real-time a response. From head and eye movements to vocalisation described as a ‘baby’s babble’.




Figure has been removed due to copyrights restrictions

Figure 10: Cog and Kismet robots with anthropomorphic features such as eyes and human form (MIT Museum, ND).

Cog was the brainchild of Rodney Brooks, director of the MIT AI Laboratory (1997-2007), who was inspired by Mark Johnson and George Lakoff’s theory that “we categorise as we do because we have the brains and bodies we have and because we interact in the world as we do” (Lakoff and Johnson 1999, p. 36). Brooks theorised that to develop a human level of intelligence, one would have to build a physical entity to interact with the world, which was the definitive belief of the ‘Behavioural’ approach to AI (Brooks, 1991). This research was heavily criticised by academics alike, including Minsky, due to the laboured effort and expense of building an anthropomorphic representation for embodiment within the world rather than simply simulating the conditions using software. It was no surprise that the research ended in 2003 with no success.

Despite no technological success of AGI or AI, the 00’s were overwhelmed with huge Hollywood blockbusters accentuating society’s prominence and fascination with AI. IMAX screens were illuminated with narratives of dystopian futures and simulated realities, with sentient machines capturing humanity and harvesting their bodies’ heat and electrical activity to maintain the energy grid in the film trilogy *The Matrix* (Wachowski & Wachowski, 1999, 2003a, 2003b), followed by the

‘perplexing’ logic (although, ironically performing its programming) of the Red Queen in *Resident Evil* (Anderson, 2002). The Red Queen is the main antagonist of the film, depicted as an AI security system which seals the entrance to an underground chemical weapons laboratory when a deadly virus is released, but kills the non-infected to reduce the statistical probability of the virus's release to the world. The 00's even saw the return of the media's favourite fearmongering speculation Skynet, in *Terminator 3: Rise of the Machines* (Mostow, 2003). Correlating to the media's sensationalism of AI, Ray Kurzweil published his book *The Singularity is Near* (Kurzweil 2005), resurfacing John von Neumann's theory of the Singularity to the fore (1958).

2.3.5 The Modern Approach to AI

A breakthrough in AI voice recognition happened in 2008, initially thought of as a simple problem but proved elusive until companies could store and compile vast amounts of data to build a statistical language model using ML. In contrast to the Classic approach, ML works with numerical data rather than symbolic data, performing statistical inference from large datasets using iterative optimisation, with some researchers preferring the term ‘statistical learning’ (Hawley, 2019). GOFAI programs ‘were notoriously brittle’, suffering the rigidity of staying within the conditions set, and in the event of missing or contradictory data, would result in ‘nonsensical’ outputs (Frankish and Ramsey 2014, p. 93). Whereas the numerical nature of ML allows for more “graceful degradation” for “imperfections in the data ... to lead to proportionally imperfect but often acceptable performance” (Ibid, p. 94). ML enabled Apple to launch the first version of the AI assistant *Siri* in 2011 for the iPhone 4S. This was a long-running technological ambition for Apple, which started in the 1980s when the company commissioned director George Lucas to create a concept film for a speculative idea known as the *Knowledge Navigator* (1987).⁶ The short film shows an iPad type of device, similar to the IBM Newspad in *2001: A Space Odyssey* (Kubrick, 1968), depicting a humanoid AI assistant on screen who voices out the day's schedule for a professor. This diegetic prototype also showcased the AI assistant's ability to retrieve knowledge, such as word files, akin to Bush's speculative *Memex*.⁷

⁶ See the film at <https://www.youtube.com/watch?v=umJsITGzXd0>

⁷ For reference: a diegetic prototype is an artefact, not limited to a specific materiality, that presents an interior view of a fictional world in status. The specificities will be further explored in Chapter 6 *Design Fiction*.

With the innovation of voice recognition, the era of Chat-bots commenced in 2014 with the bot *Eugene Goostman* reported as the first bot to pass the Turing Test by tricking 33% of a panel of judges that Eugene was a 13-year-old from Ukraine who did not speak English fluently. However, the media's excited reports were in the realms of science fiction, claiming the invention of true artificial intelligence (see *BBC News*, 2014). Many academics in the field claimed Eugene was simply a clever coded piece of software that managed to trick less than half the judges during a 5-minute conversation that should have lasted at least hours, if not days, to really test the capacities of an AI (Edgar, 2014).

2014 also saw the release of the conversational agent *Alexa* relayed through the Amazon Echo. *Alexa* functions by using Deep Learning technology for voice recognition and is currently being developed to fully operate a smart home, moving beyond the initial release functions of music playback and information retrieval. Deep Learning is a subset of ML that uses artificial neural networks designed to imitate how humans learn through neurons. These neural networks have layers of nodes with signals travelling between them, corresponding to assigned weighted inputs to produce an output.

The idea of the conversational agent has been a surprise billion-dollar opportunity for Amazon, and unexpectedly a technological, social actor brought into people's homes, highlighting the extent users will personify and interact with virtual agents. The AI program *Alexa* has been designed to encourage users to anthropomorphise it via strategically designed affordances that obscure and ease the pre-programmed cues of the AI into social life, such as specifying a name, gender and, to some extent, a personality. In turn, the interaction warrants human-like treatment, with researchers Purington et al. finding that a higher degree of personification results in more social interactions with the device and, therefore, an increased level of satisfaction using the product reported by the users (2017). In his book *Bodies in Technology*, the designer Don Ihde claims that 'alterity relations' occurs when interacting with technologies, taking on a "quasi-other" projection by enacting a presence within a device with which human users can interrelate with (Ihde, 2002, p. 81). To this end, the roboticist Cynthia Breazeal observes:

...we treat a computer not unlike the way they would treat each other [...]So, when you present our brain with things like these technologies that can over time mirror these abilities [such as voice interaction], our social brain just kicks in (Breazel quoted in Green 2017 para 5).

The opportunity and the placement of social agents into everyday life evoke the moving depiction of *Robot & Frank* (Schreier, 2012) about an ageing man with dementia and the budding personal relationship he has with his domestic robot. The diegetic prototype of the cinematic robot is suggestive of Honda's *ASIMO* humanoid robot, first introduced in 2000 and continues to be developed to be fully autonomous in a social environment (Figure 10).



Figure 11: ASIMO's hand is a highly functional compact multi-fingered hand, which has a tactile sensor and a force sensor imbedded on the palm and in each finger (Honda, ND).

Continuing the Chat-bot theme, Microsoft activated *Tay* in 2016, a chat-bot that took less than 24 hours to "go off the rails" (Price 2016, para 1). Deployed on Twitter and designed to engage in playful conversations and gradually learn from dialogue with other users. However, *Tay* was corrupted by learning from tainted data sets curated from conversing tweets and, from these, learnt to tweet racist and Neo-Nazi slurs back out into the world. *Tay* was a very public example of how AI can be corrupted by prejudiced data and optimising algorithms with racially discriminating patterns. Just like the moral of *Frankenstein*, insufficient consideration was given to its creation and its impact on the world and vice-versa (Dove & Fayard, 2020).

In the same year, the field of AI game playing had enormous success with Google's AI program *AlphaGo*, which succeeded in beating the Go champion Lee Sedol (2016). As previously

noted, game playing has had a long history within the field of AI as a way to demonstrate learning in the form of game strategies, demonstrated by the *Mechanical Turk* and *IBM'S Deep Blue* chess AI defeating the 1996 world chess champion, Garry Kasparov. *AlphaGo* was taught to play Go using a deep neural network, after which reinforcement learning was used by playing against another *AlphaGo* AI, thus learning by tracking moves and 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 and plan better. *Alpha-Go'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 more straightforward game than chess which has more rules, the space of possibility is ultimately much more extensive, making it more difficult for a computer to learn. Therefore, this was a high achievement in automated intelligence, which for the general public, was easy to perceive as a sign of a machine performing better than human intelligence. Though *AlphaGo*, like *IBM's Deep Blue*, calculated moves in a brute-force manner, prompting Kasparov to observe that "quantity had become quality" (Murray S 1997, p 86).

Moving on from imitating strategic thinking to replicating human characteristics, the term '*Deepfake*' was coined in 2017 by a Reddit user of the same name, who created an online space for sharing pornographic content that used open-sourced face-swapping technology. The term *Deepfake* has now expanded to count for all synthetic media creations using AI technology, such as voice generation and *StyleGANs* that create images of non-existing people, and the original application of face-swapping. Data manipulation, both digital and analogue, has had a long history. Benjamin Franklin (1706-1790) wrote and curated a fake newspaper to encourage a peace treaty between America and Britain by tapping into and influencing the views and opinions of the British public. Franklin penned an article with gruesome detail regarding a fake discovery of a Native American Indian bag containing scalps of British prisoners. The article, as intended, was picked up by British newspapers and reprinted, as Franklin paid specific attention to detail and skilfully designed his hoax paper. However, there were signs that the original article was fake, such as the wrong typeface and

detail inconsistencies, although at this point, sensationalism took the article forward, much like the media's embellishment of AI's functions. In this way, Franklin's operation is comparable to that of deepfakes, from the data manipulation to the digital footprints of fabrication (nonetheless, this continues to improve with advancing technology), through to the moral and ethical dilemma of creating fake media artefacts and is further complicated when artefacts are created posthumously.

Continuing with the simulation of human characteristics, Google announced their new AI assistant *Google Duplex* in 2018, which was built to book appointments over the phone, with a seemingly uncanny imitation of human-sounding speech. This imitation is achieved through recurrent neural networks to work with unsegmented and uncorrelated data and machine learning. By using training sets, the neural network trains itself by guessing the answer and adjusting to get closer to the solution. Google used its entire collection of conversation data and its Automatic Speech Recognition (ASR) technology to provide data points for the ML. Additionally, to these interweaving technologies, Google used sound and text synthesises and text-to-speech engines to initiate and control the intonation of the AI assistant based on the conversation. To further develop a natural human-sounding response, Google added filters in the AI assistant's response, such as 'hmm' and 'umm'. The pursuit of AI technology to perfectly imitate human speech is a clear-cut attempt to mislead a user into believing they are speaking to a human, which ultimately raises ethical questions regarding the deception and increasing lack of AI legibility.

Consequently, after a backlash, Google has now implemented a disclaimer response voiced by the AI assistant before a user interacts with the assistant. However, what is perceived as a step forward in human-computer interaction, the application still requires a human to monitor and step into the call if the AI gets confused. An ironic handicap and reversal of jobs in the age of automation.

Following the tradition of rooting intelligence in language impersonation, Microsoft introduced its '*Turing Natural Language Generation*' (T-NLG) in 2020. At the time of release, the T-NLG was the most extensive language model— with 17 million parameters, which is part of the model derived from learnt historical data. T-NLG utilises open-source deep learning technology and, designed by Google, a 'Transformer', a type of attention mechanism that connects the encoder and decoder of a model for better generative results (Vaswani et al., 2017). This means the model can

generate words and finish sentences by responding as fluently as a human would. The transformer technology has been implemented in Google’s NLP with a 1.6 trillion parameter model, making it the largest language model to date. However, bigger does not always equate to better, as the large models amplify encoded biases and increase the risk of perpetuating hate speech, abusive language, stereotypes, and other dehumanizing languages towards specific and minority groups. These risks were highlighted by Google’s leading AI researcher Timnit Gebru in a co-authored paper (Bender et al., 2021); as a result, she was publicly forced to leave her role at Google in 2020.

As a final note regarding AI’s history James Vincent, a senior reporter at Verge, remarks that “AI is killing the old web, and the new web struggles to be born” (Vincent, 2023, para 1). What Vincent refers to is that generative AIs, like Chat GPT and Midjourney, are changing the economy and standards of the web by generating lower quality content as generative AIs are predisposed to create plausible rather than accurate content. For background: the web was first visioned in 1989 by the British computer scientist Sir Tim Berners-Lee, and by 1990 the first web page was served on the open internet.⁸ In 1993, the underlying code was made available royalty-free, and by 1994, websites became available for general public use. The World Wide Web became a huge conglomeration of information created, curated, and fact-checked by humans; however, now, with the ability to scrape information from the open web and refine it into machine-generated content, the content becomes cheap and scalable, however less reliable. Essentially, the web is now flooded with AI junk and fake news, and the imminent future will be a “battle over *information* –over who makes it, how you access it, and who gets paid” (Ibid, para 22).

With a thorough history of AI and the pursuit of machine intelligence, the following section will focus on data.

⁸ The Internet started in the 1960s as a way for government researchers to share information. However, membership was limited to certain academic and research organisations. In response to this, other networks were created to provide information sharing. In 1983 A new communications protocol was established called Transfer Control Protocol/Internetnetwork Protocol, allowing computers on different networks to “talk” to each other, thus the birth of the Internet.

2.3.6 The Rise and Troubles with Big Data

Etymologically the word ‘data’ is derived from the Latin word *dare* to mean *give*, describing the process wherein a phenomenon gives something of an element over to be recorded and measured. However, as Rob Kitchin points out, those elements are taken through observations, computations, and record taking, leading to the Latin translation of *capta* (2014, p. 2). The term data, akin to AI, is the product of a falsely given eponymous, highlighting the false perception a given term can conjure. Further still, the term Big Data is a neologism, as the practice of collecting data for statistics can be traced back as early as the Han Dynasty (206 BC–220 AD) and the Roman Empire (27 BCE-14 CE) for measuring and managing population, commodities, and viable soldiers. The techniques of Big Data now mean big business (Elish & Boyd, 2018a; Zuboff, 2019). Modern practices in Big Data date back as early as the 1990s, with the principles behind the concept flourishing in business discourse via a Gartner report, defining Big Data as volume, velocity and variety: the “3Vs” (Laney, 2001), thus turning into a new paradigm in the business sector by 2010 (Manyika et al., 2011). As a result, technological companies emerged, offering services to clients with an opportunity to ride this new digital wave and ‘get smart’ by storing and managing data through cloud and software packages. As swift as the embrace of Big Data was, so too were the raising concerns of the purpose, value and impacts of bias generated in its use, which Geoffery Bowker speculated on saying, “[r]aw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care” (2005 p. 200).

An interesting change in the public response to Big Data can be gleaned from the shifting attitudes presented in the Obama White House reports especially commissioned to investigate Big Data (Elish & Boyd, 2018a). Exhilarated about the economic and social potentials of Big Data, a panel of commissioned experts in 2014 wrote an optimistic report concerned with seizing the opportunities afforded by Big Data while upholding the values of protecting privacy, ensuring fairness, and preventing discrimination (Podesta et al., 2014). Only two years later, a second report painted a very different and pessimistic picture regarding data discrimination and the vast amount of personal data being collected and sold (Munoz et al., 2016).

On this note, the term Surveillance Capitalism coined by Shoshana Zuboff is the use of data beyond product or service improvement, with the extra data collected declared by tech companies as “behavioural surplus”, which is fed into machine intelligence and fabricated into “prediction products that anticipate what you will do now, soon and later” (2019, p. 8). Zuboff elaborates further that the prevailing predictive behavioural data is derived from an intervention in the state of interaction by nudging and coaxing the user’s behaviour. The goal of Surveillance Capitalism is no longer the automation of information flows about users but the automation of users (2019). Zuboff’s forensic analysis on the use of Big Data is cautious to observe that Surveillance Capitalism is a choice of how to wield technology, reminding us of Kranzberg’s famous first law “[t]echnology is neither good nor bad; nor is it neutral” (Kranzberg 1986, p. 547). The extent of the Surveillance Capitalism networks is hidden from view as a design choice, capitalising on the fact that the majority of users are unaware data is gathered from each interaction and button pressed, thus galvanising for a design solution to promote agency and negotiability when using AI and IoT technology (Mortier et al., 2014).

This brief AI and data history has demonstrated how humanity has been impelled to collect, store, and disseminate what is known and manifest knowledge into computational thought with the original goal of creating AGI and sentient beings. In a departure, however, AI research has in recent years excelled at evolving ‘machine intelligence’ performing, in light of AGI, ‘narrow’ and ‘weak’ computational tasks due to Big Data, computational power, and ‘networkification’ (Pierce & DiSalvo, 2017). As demonstrated, the term AI has had a long and diverse history, with AGI as the prevailing perception when people think of or hear the term AI, which is only perpetuated by the vast amount of science fiction portraying sentient killing robots. This paradox of misinterpretation between these two deviating, though entangled concepts of AI has been defined as the ‘Definitional Dualism of AI’ (Lindley et al., 2020; F. Pilling, Lindley, et al., 2021). The following section will further detail the factors that complicate and obscures an understanding of AI with the familiar and interlinking thread of “AI[’s] misinformation epidemic” (O. Schwartz, 2018) acting as a mutual catalyst and simultaneously setting up a design space for readdressing this.

2.4 The Definitional Dualism of AI; A Confused Ontology

The following sections will analyse the challenging factors that confuse AI's perception. These challenges are 1. the various evolving definitions of AI; 2. AI technology is considered the standard measure of applying technology to a situation, and therefore going unquestioned and re-classified as not being AI technology: a symptom of the 'AI Hype cycle' (O. Schwartz, 2018); 3. The false dichotomy of AI's qualities confused with science fiction renderings of sentient beings, extending to the anthropomorphising of AI and considering AI applications as transpiring through magic; 4. Creators of AI have argued that they do not understand how their AI systems reach a decision or understand the techniques used to build the programs in the first place, describing AI as a type of "alien technology" (Rahimi quoted in Hutson 2018, para 1). For instance, engineers have developed deep learning systems that 'work' by automatically detecting the faces of cats or dogs—without necessarily knowing why they work or being able to see the logic behind a system's decision (Ananny & Crawford, 2018). This situation can cause serious complications, such as when Google's Photo app unexpectedly tagged Black people as "gorillas" (Dougherty, 2015) or when Nikon's camera perceived Asian people were blinking.

2.4.1 Evolving Definitions

As showcased, the term 'Artificial Intelligence' conjures a manifold of meanings, which reflects the numerous definitions being developed to describe AI research, with each newly created definition attempting to establish a robust and singular term to understand the technology and its goals. However, new definitions bring specialised nomenclatures to distinguish research (Hawley, 2019), further adding to the confusion. Nevertheless, in defining AI, human intelligence is the comparative constituent. The first formalised definition can be found in the Dartmouth Summer Proposal –

For the present purpose the artificial intelligence problem is taken to be that of making a machine behave in ways that would be called intelligent if a human were so behaving (J. McCarthy et al., 1955a).

Furthermore, Minsky defined AI as “the science of making machines do things that would require intelligence if done by men” (1968, p. 5). Both these definitions were formed at the height of AI research when predictions about greater-than-human capabilities of AI dominated. Even though these predictions have somewhat subsided in AI research, these greater-than-human capabilities still tinge perceptions of AI and, to some extent, contemporary definitions of AI, using human intelligence as a milestone. Computer scientists Stuart Russell and Peter Norvig have argued that the history of AI, far from discerning any particular definition of intelligence (Elish & Boyd, 2018a), have been concerned with four interrelated but distinct goals: “systems that think like humans, systems that act like humans, systems that think rationally, systems that act rationally” (Russell and Norvig 1995, p. 5).

The UK parliament in 2018, observing the lack of a conclusive definition of AI, adopted the definition used in the 2017 Industrial Strategy White Paper:

Technologies with the ability to perform tasks that would otherwise require human intelligence, such as visual perception, speech recognition, and language translation (HM Government 2017, p. 37).

Moving forward, the European Union published in 2021 its AI proposal for regulating its use and transparency. The act interestingly attempted to provide “a single future-proof definition of AI” (European Commission 2021, p. 3):

[An] ‘artificial intelligence system’ (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I [machine learning, symbolic approaches, and statistics] and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with (Ibid, p. 18).

The European Union’s definition still uses a human level of intelligence as a comparison, though it avoids using the word intelligence, opting for ‘human-defined objectives’ instead. In relation, Turing noted that “the idea behind digital computers may be explained by saying that these machines are intended to carry out any operations which could be done by a human computer”, going on to say that

“[t]he human computer is supposed to be following fixed rules; he has no authority to deviate from them in any detail” (Turing, 1950, p. 436).

Human computers were an occupational title for mathematicians, meaning ‘one that computes’ and performed long and complex calculations before the invention of digital computers. A famous human computer was Katherine Johnson, who calculated orbital mechanics for Nasa’s first crewed space flight, which is now done, as Turing predicted, through digital computers. The shift to digital is straightforward; humans do not process as much information or as quickly. Nevertheless, humans can predict through a logical deduction of facts and, diverging from digital computers, use tacit knowledge of the world that can deviate from strict calculations and often be more representative of the spontaneity of reality.

2.4.2 AI Hype Cycle and The Rebranding of AI

AI definitions, as evidenced habitually, include a comparison to human intelligence, and with that, there is an unbridled optimism for technology that performs a *form* of intelligence. The current enthusiasm for AI intelligence is to improve users’ interaction with products and services, such as predictions for entertainment or distinguishing supposedly ideal candidates for an employer to recruit. Ironically, machine intelligence is considered intelligent if successful at a proscribed task, though soon after success, a machine will frequently be declassified as non-intelligent or simply the process of crunching data.

With data in mind, a form of rebranding occurred with the surge of criticism aimed at Big Data because of the suspicions raised from the surveillance and recording of personal data. Consequently, many Big Data companies rebranded themselves as AI companies to avoid criticism, promoting instead the technical methods for the analysis of data while concealing that the acquisition of data was the same (Boyd, 2016; Levy, 2016).

Established data analysis techniques are being rebranded as AI, more specifically ML, taking advantage of the current AI hype, with many of these processes not regarded explicitly by many as AI methods. Nevertheless, due to their illusive implementation as core AI methods, these techniques are brought to the forefront of the hype (Hawley, 2019). This is attributable to the fact that the trademark

‘statistics’ gleams less of a buzz from the general public, with the term ‘AI’ inviting both good and bad attention from the media (Ibid); as the saying goes ‘all publicity is good publicity’. Conversely, due to the plethora of ambiguities associated with the term AI, many researchers avoid using the term, preferring to use specific ML algorithms such as Deep Learning, Neural Networks. Though to non-AI experts, the reverse is true, with the terms such as AI and Deep Learning used interchangeably due to a lack of knowledge which is hindered by information overload and the false reporting of information (Elish & Hwang, 2016).

As well as the confusion gleamed with rebranding and the difficulties of keeping abreast of changes, it is well known in the AI research field that the media notoriously cultivates a flawed interpretation of AI, using science fiction as the inception of these narratives to generate clickbait with fantasies of “artificial brains” (Bello Del 2018, para 1). Thus, producing unrealistic AI technology expectations for the general public (O. Schwartz, 2018).

The manufactured hype additionally exacerbates confusion to ramp up sales of products and services using AI technology, with the promise of innovations outstripping current capabilities known as ‘vapourware’ (Coulton & Lindley, 2017) and the selling of AI-snake oil (Elish & Boyd, 2018a). On a separate note, further enabling the “hype-driven ecosystem” of AI is the strained implementation of AI systems to stay relevant in the era of smart devices, which ultimately circumvents the discourse about the appropriate and practical design of such technologies resulting in the assembly and facilitation of poorly produced AI models and unsound practices (Elish and Boyd 2018, p. 74).

2.4.3 Science-Fiction and Anthropomorphising AI technology

In 1895 Herbert George Wells published the science fiction novel *The Time Machine* (1895), while simultaneously, the Lumiere brothers held the first public screenings for their new cinematography machine in France. With this in mind, James Bell asks, “[i]s it just a coincidence that the cinema and science fiction as we know it were both born in the same year?” (2014, p. 6) Science fiction has the proficiency for igniting our imaginations, reflecting our fears and wonders of the what the future could be. Moreover, science fiction renderings are often taken to be factual leading to many misconceptions about technology, impacting adoption and use— as the Thomas theorem reflects “[i]f

[people] define situations as real, they are real in their consequences” (Thomas & Thomas, 1928, p.571-572)

As evidenced earlier, a highly proscribed science fiction subgenre is AI, with narratives of artificial sentient beings, cyborg and robotic forms, and overzealous machines enslaving humanity. As there are many seminal titles to choose from within this subgenre, this section will concentrate on select titles that influence general users’ perception of AI by falling into the category of AI’s definitional dualism and highlighting the consequences of this confusion.

2.4.3.1 Examining Hal’s Definitional Dualism

An unsurpassed example of AI characterisation in science fiction and one that complicates AI’s ontology is *2001: A Space Odyssey* (Kubrick, 1968) (henceforth simply referred to as ‘2001’). A cinematic experience that speculates on humanity’s technological ascension through the exploration of space, and the ultimate transcendence of humanity, galvanised by the invention of AI bringing about the Singularity. One film critic called *2001* “the best-informed dream ever” (Champlin, 1968), correlating to the widely known fact that Kubrick and the author Arthur C. Clarke consulted many scientists, both in academia and industry, to extrapolate and build a plausible future world.

The central ‘character’ in *2001* is the ship’s AI system, *HAL9000* (henceforth, simply referred to as ‘HAL’), where many film critics have pointed out that despite being a machine, HAL is the most emotional and responsive character in the film. HAL is a unique example of AI being depicted in fiction due to displaying AGI underpinned by visualising ‘narrow’ AI subproblems. HAL simultaneously embodies and illuminates then-current (at the film’s release date) research agendas, present-day advances within AI, and, further still, developments that are yet to be achieved. HAL manifests these AI functions into the aspirational and hypothetical research agenda of human-computer symbiosis and the enthralling pursuit of transcendence. The collective movement known as Singularitarianism attempts to propel the creation of AGI for the goal of transcendence to digital immortality (Geraci 2010), catalysed by Moore’s law.⁹ Though it may be the context of science

⁹ The perception that the number of transistors on a microchip doubles every two years and the cost of halves. Therefore, Moore’s Law states the speed and capacity of computers increase every two years and cost less (Moore, 1965). Though, it has been criticised that Moore’s Law is an observation not a law (McCorduck, 2004).

fiction, notable scientists such as Hans Moravec and Kurzweil have put forward speculations and, at one time, worked towards an AGI-driven evolutionary eschatology (Kurzweil, 2013; Moravec, 1988).

Every detail of *2001* 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. Special care and attention extended to HAL, with Kubrick commissioning the most prominent 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 computer technology manufacturer, conjuring the ideology of vapourware (Coulton & Lindley, 2017).¹⁰ IBM's proposed concept 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 the book *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. On the other hand, HAL is not a human form with cyborg features but is situated in an evolved 'disciplinarily machinery' of AI (Mateas, 2006), resulting from a plausible extrapolation from then-current lines of research and visualisation of future AI systems.

Kubrick presented HAL as a diegetic prototype displaying AGI while visualising AI subproblems, which emulated different subfields within AI research, including game playing, computer vision, and language (Stork, 1997). The following sections will examine HAL's definitional dualism by unpacking HAL as a diegetic prototype to expose HAL's speculative narrow and AGI

¹⁰ The term vapourware stands for technological artefacts that are imagined to create a buzz about the future and the company's image although are never intended to be produced. This notion will be explored in more detail in Chapter Six.

architecture. The nature and intricacies of AI functions, operations and architecture are intangible. To consider HAL as a diegetic prototype means that we have to venture beyond the physical nuances by anchoring the internal functions and architecture as diegetic prototypes through the film's script, mise-en-scène and plot.¹¹

2.4.3.2 HAL – “Thank you for an enjoyable game” – AI game playing

2001's chess scene only lasted for thirty seconds; however, it demonstrated in great detail the archetypal AI problem of playing chess and further extrapolations towards general intelligence (Figure 11). To set the scene, the players are positioned opposite one another, so to speak, with HAL's opaque cyclops eye facing Frank (Gary Lockwood). Rather than using a physical chessboard and pieces, the chessboard is futuristically and digitally represented on a tabletop screen, utilising voice interaction to move the chess pieces.



Figure 11: Figure 12: Playing chess with HAL was through voice interaction. (01:06:06) (Kubrick, 1968).

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 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

¹¹ Mise-en-scène refers to everything that appears before the camera. It also accounts for their arrangement in terms of composition, props, sets, actors, costumes, and lighting.

reality, AI chess programs akin to *Deep Blue* AI 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. Thus, HAL chooses to move based on the humanistic condition to satisfy itself (Ibid).

2.4.3.3 HAL will see you now – AI vision

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, conduct 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 12). When HAL says the phrase "I think you are improving" (01:08:22), it indicates that it can recall past renderings and compare and contrast, performing various narrow tasks concurrently. This statement also signifies HAL as a sentient being with an opinion with general and common-sense reasoning. Another example of indicating AI subproblems and sneaking general intelligence through the backdoor.



Figure 13: 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 (01:07:45) (Kubrick, 1968).

Kubrick also reveals that HAL can read facial expressions when David (Keir Dullea) 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 and Hoey, 2020). This research is a considerable undertaking, as recent studies suggest that facial movements are not universally perceived as emotional expressions (Gendron et al. 2018).

The film's critical turning point is when the crewmen 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 at this moment, we realise HAL can lip-read (Figure 13). 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 ML to learn and decode feature extraction points. The interesting point regarding HAL was that the crew did not know he could read lips. The question is, 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 14: The black veneering around the focus of the lip's signals to the audience that this is Hal's visual perspective as a single and circular lens (01:27:17) (Kubrick, 1968).

2.4.3.4 HAL More-Than just A Chatbot – Natural Language Processing

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; being used to decode spoken words (or lip reading); understanding 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 that HAL's language abilities transcend these. As technically, language is an amalgamate of subproblems, and current NLPs 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). Michael 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).

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. 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 the prevailing rhetoric and

scientific narratives stipulate AI is a future technology, in reality, it is here now, and so are its challenges.

2.4.3.5 Anthropomorphising AI

Pope Francis focused his monthly prayer intention during the pandemic on AI technology's safe, ethical, and beneficial development. As well as devoting his prayer to reducing inequality and for “progress to always “serve humankind,” and respecting human dignity”, Pope Francis’s final thoughts had an uncanny resonance about “taking care of Creation”: “Let us pray that the progress of robotics and artificial intelligence may always serve humankind... we could say, may it “be human.”(NA, 2020, para 1 & 11). The question of what it means to be human has concerned humanity for Millenia. Attentive to discover an answer was the 17th-century philosopher René Descartes, who founded upon his famous dictum “I think, therefore I am”, the proposal of the Cartesian Dualism, also known as the mind/body dichotomy. Descartes was primarily concerned with distinguishing us from animals; nevertheless, the materialisation of AGI can also be synonymous with his theoretical concepts. Inspired, the director Ridley Scott named his protagonist in the dystopian film *Blade Runner* (1982) Deckard (Harrison Ford) after Descartes.

Rick Deckard’s job as a Blade Runner was to hunt down and ‘retire’ any Replicants found, as the law forbade any existing on Earth, due to a bloody mutiny off-world initiated by a Nexus-6 model combat team vilified by their designated life span of four years. Replicants are bioengineered androids composed entirely of organic material; the only way to differentiate them from humans was by testing emotional responses and timings, using the film’s emblematic Voight-Kampff Test, and provoking a physiological response through a series of empathy-inciting questions, which they are considered to lack. Rather than focusing on the feature of androids exhibiting intelligence, the film explores beyond this purview towards emotional intelligence. Early in the film Deckard and his boss Bryant (Emmet Walsh), consider the failure of the Voight-Kampff tests on the Nexus-6 Replicants, entertaining the possibility that a replicant could be empathetic enough to be considered a human being. Andrew Norris writing on the film’s philosophy observes that the test at this point is not the search for the essential characteristics of a human being but for a mark of contingency, with another distinction

between the two must be found (2013). In a sense, it is a reclassification of what it means to be human, with the irony being that the Replicants demonstrate time and again a vibrance of life and empathy with suffering creatures (Ibid).

In contrast, another anthropomorphised science fiction rendering of AI is the Terminator (Arnold Schwarzenegger), who feels nothing; it is a killing machine and nothing more (Cameron, 1984). Another distinct feature of the Terminator cyborg models is that they are wearing a flesh suit, often torn, revealing their mechanical makeup. The Replicants, by contrast, cannot discard their skin; it is their living flesh and “a feature of their own experience, and not just the experience of those around them” (Norris, 2013, p. 22). Writing further about the Replicant’s own experience, Norris observes that this is symbolised using the Voigt-Kampff Test, which entails looking into the eyes of the Replicant and exposing what is unconsciously expressed in the organ through which they see and experience the world (Ibid) (Figure 14).



Figure 15: The Voigt-Kampff Test uses a machine to focus in and look at the suspected Replicant’s eyes (00:05:27) (Scott, 1982).

Once more, the Terminator's eyes also play an essential role in the film's communication of its existence, though in polarity to the Replicants, the audience gets to see what the cyborg sees, revealing to the audience that the Terminator is not human; it is a machine with the programmed intelligence to kill (Ibid) (Figure 15).



Figure 16: The experience of looking through the Terminators 'eyes' (01:00:47) (Cameron, 1984).

Both cinematic masterpieces, *Blade Runner* (Scott, 1982) and *The Terminator* (Cameron, 1984), and many other narratives of anthropomorphised AI, are thought experiments concerning different scopes of what artificial life and its creation means. However, the reality is that AGI that would conceive artificial life remains in the realms of fiction, with the reality of narrow AI yet to compute seamlessly, let alone develop a consciousness. Not to mention the famous uncanny valley phenomena demonstrated by humanoid robots (Mori, 1970) employed through clever design implementations resulting in human imitation, such as the Google *Duplex*, which confuse perception and encourage believability of artificial intelligence.¹²

Descartes speculated, before Turing, the need for a test to discern whether something was human or machine because of imitation. The test he detailed was in two parts; the first, almost akin to Turing's, is an examination of language and communication. Though, Descartes stipulates a competence of communication beyond corresponding and notifying "a change in its organs" from touch or damage, as he deems "men of the lowest grade of intellect can do" this (Descartes 2008, p.

¹² The uncanny valley hypothesis predicts that an entity which appears almost human-like will risk eliciting revulsion and eerie feelings in viewers.

44). The second part of the test is finding fault in the machine's execution of an action. Descartes does not particularise what the action may be but explains that machines may get to a point where they might "execute many things with equal or perhaps greater perfection than any of us", though there will be a point of failure; exposing the fact that the machines do not act on knowledge and reason, rather through the "particular arrangement" and "diversity of organs sufficient to enable it to act in all the occurrences of life" (Ibid, p. 45).¹³ Both Turing's and Descartes's tests can be used to expose narrow AI crafted to imitate human responses, and furthermore, they could also be used to test hypothetically creations of AGI. However, a test specifically devised to test if a machine has a conscious has been devised called the AI Consciousness Test (ACT), which comprises a set of specially designed questions to gauge what is going on 'inside' the machine akin to the Voigt-Kampff Test. At this point in time, it is just a thought experiment. However, the developers believe the test could facilitate "consciousness engineering" (Schneider and Turner 2017, para 16). Perhaps this type of engineering is just sensationalism, but the test's development highlights the roots of AI's definitional dualism again and, inevitably, the confused perception users have of AI and, *conceivably*, its future.

2.4.3.6 Metaphorical Anthropomorphisation

In 1912 the philosopher Julien Offray de La Mettrie published the famous *Man a Machine*, where he compared the human body to a "watch", one that is constructed by nature with "such skill and ingenuity" that is far beyond the skill set of Vaucanson used to make his mechanical duck (Mettrie, 1912, pp. 140–141). On that note, due to the obscurity of a simple and unconforming description for the nature of computers, Marakas et al. contend that this has resulted in the use of metaphors used to communicate and consider computers, embodying the reversal of La Mettrie's treatise: *Machine a Man*, using "the most familiar foundations to build upon: ourselves" (Marakas, et al., 2000, p 722).

¹³ The organs, Descartes focuses attention on, can be considered technological sensors, a concept the films *Blade Runner* and *The Terminator* capitalises on.

We, as humans, have a spirited tendency to anthropomorphise and give human characteristics to non-human things (Stebbins, 1993). While still an evolving theory, the psychology of anthropomorphism is considered an attempt to make sense of the surrounding world by using the same neurological mechanisms and tactics humans use to decipher other humans (Urquiza-Haas & Kotrschal, 2015). Anthropomorphism intuitively extends to the ‘metaphorical personification’ to understand the complexities of computing technologies (Marakas, et al., 2000; West and Travis, 1991). For instance, computers ‘read’, ‘write’, and ‘catch viruses’, and AI programs ‘learn’ and ‘train’. Marakas et al. further observe; that technology and intelligent machines introduce a new vocabulary from new findings and innovation, where overtime they become intertwined with the manner to describe *us*, a process they call “technomorphism” (Marakas, et al., 2000, p. 722). Marakas et al. argue that despite the vast amount of knowledge, “we may never acquire enough understanding to describe the totality of the computer in terms of more familiar objects” aside from ourselves (Ibid, p. 737). Stemming from this challenge is creating a different perspective for AI, a beyond human perspective.

2.4.3.7 Believable Perceptions

The anthropomorphic personification embedded in metaphors not only aims to capture the physical perspective of *what computers do*. A socially constructed perspective correspondingly occurs to understand *what computers are*, which too is greatly influenced by science fiction renderings and the successive media distorted narratives. In this regard, the physical and the social perspective are mutually associated and perpetuating one another (Ibid).

Unpacking the social perspective, Marakas et al. see two distinctive perspectives we adopt to conceptualise and perceive intelligent machines. On the one hand, those who perceive the role of the computer as an extension and as a tool for magnifying the mind and body into the realms of achievement (Zuboff, 1998) with awareness and distinction that machines are created, programmed, instructed and alterable by humans (Marakas, et al., 2000). This view is inclusive towards an understanding that machine intelligence is programmed to learn autonomously through code to perform without supervision. The opposing perspective views a computer with human-like attributes,

creating the impression of a rational actor with autonomy (Ibid), such as computer inhibiting personalities that are psychologically real to users (Moon & Nass, 1996).

Various Human-AI interaction research ventures enrich the latter perspective; for instance, the designing of live imitation strategies for positive interaction between humans and virtual agents, which influences part of the brain that triggers anthropomorphism (Numata, et al. 2020). Furthermore, the development of algorithms that enable learning-by-imitation of human social behaviour to enable robots to interpret and respond to cues given for effective interactions (Doering, et al.2019). In the paper, *The Art of Designing Socially Intelligent Agents: Science Fiction and The Human in The Loop*, Kerstin Dautenhahn provides a general overview of what socially intelligent agents (SIAs) are and a framework for their effective design that permits a “cognitive fit” between agent-human interaction (1998). Dautenhahn argues that the theories for intelligence are formidably debated and cannot be defined objectively. Instead, intelligence should be viewed as constructed and attributed by humans through interaction rather than a phenomenon inside something. In such a circumstance, the author advises that aspects of human social psychology should be considered and implemented into social agents, such as “storytelling, empathy, embodiment and historical and ecological grounding for a *believable* and cognitively well-balanced design” (Ibid, p 573 (italicised for emphasis)). Using a range of examples from cyber-pets (Tamagotchi (Maita, 1996)) to Pixar’s short animation *Luxo Jr* (Lasseter, 1986), Dautenhahn observed that human’s nature is inclined to judge any artefact according to its believability, through crucial attributes of interactivity, natural expressiveness and imitation of behaviour.¹⁴ SIAs take advantage of the human tendency to anthropomorphise through the use of HCD principles identifying the need to implement a shorthand to communicate the factuality of the interaction with a non-human thing.

2.4.4 Magic and Metaphors: Is It a Kind of Magic?

When magic and technology are considered together, more often than not, Arthur C Clarke’s third law, “[a]ny sufficiently advanced technology is indistinguishable from magic” (1976, p. 39), is often quoted and with it an impulse for using magic as a metaphor when describing the unknown

¹⁴ Pixar’s short Luxo Jr can be watched at <https://www.youtube.com/watch?v=FI0T00j7WFE> .

modus operandi of a technological operation. Clarke's quote is repeatedly taken out of context; rather, his laws were meant to express his aspiration for humanity's technological endeavours rather than contribute and encourage the obscuring of technologies' proper remit.

From the perspective of the user, when technology is said to 'work like magic', a recognisable idiom, we understand this as a way to communicate seamless functionality, whereby the overall experience is fulfilling, and the means of the effect is overshadowed, becoming irrelevant to the less discriminating (Elish and Boyd, 2018a). On this note, the anthropologist Alfred Gell writes, "[p]roduction 'by magic' is production minus the disadvantageous side-effects, such as struggle, effort, etc." (1988, p. 9) From this we can say the description of technology working – like magic – is a common expression in the marketing of technology, especially AI (Elish & Boyd, 2018a). In his paper *Venerating the Black Box*, William Stahl conducted a systematic search into the media's role in shaping public perceptions and responses about technology, finding that overt magical and occult language was used to describe technology (1995). This tactic was specially deployed in the 80s, a pinnacle moment in computing history with the widespread launch of personal computers from Apple, Commodore, and Atari. Stahl reflects, "[w]hen a technology is a black box it becomes magical", a strategy for essentially obscuring the operational remit of technology and fabricating an alternative reality for users; that technology in all its power should not be feared: but harnessed for our own needs (Ibid, p. 252). The description of 'magic' also reinforces a sense of how technology works is unfathomable, a condition that especially happens with AI (Bridle, 2018; Elish & Boyd, 2018a; Selbst, 2017). In essence, to "evoke magic", Elish and Danah Boyd observe, is to "minimize attention to the methods and resources required to carry out a particular effect" (Elish and Boyd 2018, p. 63).

The common misperception of Clarke's earlier quote echoes the statement in Leigh Brackett's short story *The Sorcerer of Rhiannon* — "Witchcraft to the ignorant, ... simple science to the learned" (Brackett 1942, p. 39). However, the illiteracy of AI technology is not at fault with the users. A combination of opacity and complexity makes AI processes illegible with coding a specialised skill

(Burrell, 2016). As James Bridle states and initiates this research's design challenge, "you should be able to understand technological systems without having to learn code at all" (Bridle 2018, p. 4).

2.4.5 Alien Technology; Creating Their Own Representation of The World

Ali Rahimi, a leading research scientist at Google, boldly stated that "machine learning has become alchemy", arguing that even though alchemy 'worked' the foundations of alchemy were formed upon unverifiable and, for modern times, dubious theories (Rahimi quoted in Elish and Boyd 2018b, para 1). The ancient Hermetic art of alchemy was the practice of transmuting a lesser material into a greater material; in comparison, machine learning models are insufficiently understood and used to make life-altering judgments on individuals through opaque transmutations of data.

Domingos identifies that the development of "successful machine learning applications requires a substantial amount of 'black art'", giving a sense that mastery is not easily found in textbooks but through hearsay and experience (Domingos 2012, p. 78). He reminds us that the goal of learning predictive models is to use them as "guides to action", whereby correlation and prediction do not mean causation (Ibid, p. 87). That is to say, ML algorithms that are soundly designed generalise beyond the training set for more accurate predictions on what – may – happen, though many pitfalls easily occur in development; such as, overfitting where the data is not sufficient to ultimately determine the correct data label classifier, leading to generalisations with errors of bias (learning the wrong thing) or variance (learning random things irrespective of the signal); and the notorious 'curse of dimensionality', where generalising correctly becomes exponentially harder as the number of features (measurable characteristics of the data) (dimensionality), or columns, of the examples, grow leading to the model overfitting (pp. 81-83). The black art, Domingos surmises, is found in features design, not the intuition of adding in more features; a trial and error of taking raw data and constructing features with the frontier of ML to automate this process more and more (p. 84).

Nevertheless, feature engineering is difficult because it is domain-specific while learners are general purpose, though once through training, ML systems can perform exceedingly well in their

explicit domain, giving the impression of generality. That is to say, a system from DeepMind learned to play Atari games outperforming the human benchmark (Volodymyr et al., 2013) can inevitably lose all its learnt abilities when the operating environment changes somewhat, such as the pixels in the frame slightly moving (Stockton, 2017).

ML is applied to various ‘problems’ for which encoding an explicit logic of decision-making does not work, and the act of coding is a two-sided operation and communication, where the human codes for the machine to learn (Burrell, 2016). Simply put, the algorithm evolves beyond human intelligibility and understanding to work out problems in an albeit sensitive way conceivably too vulnerable for many real-world applications, concealed within the matrix of the machine’s logic. Automation of these systems thus become ‘Human Out-of-the-Loop’ systems that learn to perform tasks in a literal way, although ultimately incorrect by human standards; we can only adjudicate the results. Frank Lantz’s 2017 *Universal Paperclips* shines a comical light on AI’s literal rationality. In this game, the user plays the role of an AI programmed to produce paperclips. They first click on a box to create a single paperclip at a time, followed by options to sell paperclips to finance machines that automatically produce huge quantities of paperclips without human intervention. The game ends when the AI succeeds in converting the entire universe into paperclips: destroying the world (Rogers, 2017).

It is no wonder Rahimi anguishes that “[m]any of us feel like we’re operating on an alien technology” (Rahimi quoted in Hutson 2018, para 2). A situation facilitated by the black box nature of AI, the apprehension of code and data, and, as Facebook AI researcher Dhruv Batra remarks, “we do not understand what they are basing their decisions on” (Abhishek et al., 2017; Rutkin, 2016). AI processes attempt to determine indeterminacy by bringing the indeterminacy of the world and lived experience into computation, whereby users believe AI decisions are an accurate and unquestionable representation of the world (Bridle, 2018; Fazi, 2018). Bridle notes that throughout the history of computation, we have been conditioned to believe and depend on computers rendering the world clearer and more efficiently than our own perceptions; that they reduce complexity and decipher

better solutions to the world's problems while expanding our agency (Bridle 2018, pp. 26-27). However, computation is a concentration of power into the narrow domains of those who control, model, and operate these systems “[b]y conflating approximation with simulation, the high priests of computational thinking replace the world with flawed models of itself” (Ibid, p. 27). Then, it is necessary to observe AI technology for what it is – as Alien technology – a thing unlike and beyond human intelligence. From this positioning, a new metaphor can be developed to understand and view AI technology from an alternative non-human perspective and establish an approach to defuse the complexity and illegibility of interacting with AI technology.

2.5 Conclusion

This chapter has provided an ontological review of AI technology through an AI history and a review of popular culture renderings of AI, shaping public perceptions of AI technology. The main thread of the review concerned the type of intelligence exhibited by these AI visions or research enterprises, categorised as attempting to display or act out human or machine intelligence. Towards the end of the chapter, AI's definitional dualism was explored by reviewing the films *2001* (Kubrick 1968), *Blade runner* (Scott 1982) and *Terminator* (Cameron 1984), finding that science fiction representations greatly influence both expert and public's perception of AI. Thereafter AI's definitional dualism was further investigated by opening the scope of how AI is perceived as alien technology, as magic, or as a reflection of humanity.

These various ramifications hinder the legibility of AI technology, and as such, there are currently no supportive and standardised ways of communicating the ‘shapeless and faceless, everywhere and nowhere’ (Pierce & DiSalvo, 2017) constructs of AI. Rarely can we say for sure why an AI has reached a particular decision, even with the aid of expert knowledge. Nevertheless, often all that users have to work with is metaphors that confuse the reality of AI technology, shrouding the real threats of governing and data-gathering technology. It is essential to limit the indeterminacy of information and pave the way for general AI literacy, and make human-AI interaction design accountable (Pilling, et al., 2022). Advocation for ‘interactive explanation systems’ (Weld & Bansal,

2019) is in high demand, as evidenced by the diverse authored frameworks and guidelines for future AI implementation. The next step is designing possible solutions and going beyond written guidance.

This research pursues the design challenge of materialising a method for legible AI by creating an accessible and uniformly constructed AI lexicon not only to demystify AI and the effect this mystification may have on users (confusion, uncertainty and/or erroneous use) but also to make it possible for the general user to understand and assimilate future AI developments while remaining aware and, where needed, critical of intentional or unintentional obfuscation of AI processes.

In the following chapters, this research turns to explain the research methodology. What would usually be one chapter, this research has presented its methodology across two chapters. This is because Chapter Three presents the method assemblage structure employed to contend and understand the ‘messy’ reality (Law, 2004) of AI. This chapter also showcases an argument for metamorphosing philosophy. Chapter Four will be the conclusive part of the methodology section for this research, further explaining the method assemblage and the iterative nature of this research in which much of the research was conceived through thinking through design.

Chapter Three Groundworks

(Understanding AI)

3.1 Introduction

The previous chapter has laid out an ontological perspective of AI by mapping out current understandings and thinking within the field. As this research is concerned with establishing AI as a material for design via a transdisciplinary scheme, it necessitates the development of a methodological approach that validates crossing disciplinary boundaries such as philosophy and design, transcending to a type of ‘design as philosophy’. Further still, this research has not been a linear journey rather, it has been generative to adapt to the complexity and messiness of reality (Law, 2004).

This chapter lays the methodological groundwork by establishing the transdisciplinary nature of this research. The title of this chapter, ‘Groundwork’, has been chosen due to the term referring to work done to prepare sub-surfaces to start construction work, a preparatory stage that makes or breaks the final finished building, or research in this case. Furthermore, the ‘method assemblage’ is presented as an underlying methodological model, with the function to accommodate, assemble and silo additional theoretical elements into the research and design practice for generating knowledge. Therefore, the structure of this thesis echoes the method assemblage ensemble and the manifold of elements that constitute this research, which will be explained in this chapter.

In this regard and as previously noted, there will be two chapters on methodologies; Chapter Three *Groundworks* and Chapter Four *Methodologies*. The Methodologies chapter presents the overarching design approach of Research through Design as an iterative and generative method to conduct design research. This research is, therefore, transdisciplinary and affords the integration of supplementary disciplinary theories, namely philosophical thinking, for a More-Than Human perspective of AI. Consequently, this thesis does not follow the traditional formulation of a doctoral thesis. This point will be addressed throughout this chapter: first introducing a transdisciplinary enterprise’s inherent features, then detailing the method assemblage model of this research.

3.2 Transdisciplinary Research a Postmodern Turn: Promiscuous Monsters on the Prowl

The academic Bob Hodge presents an argument for the “postmodern turn”, a “revolution” in social science research he calls the “new humanities” (1995). Hodge remarks, at the time of the paper’s publication, on a new trend of PhD students conducting transdisciplinary theses and research that inadvertently ran the risk of being rejected by not proscribing to the criteria applied (then and to some extent now) for conducting research (p.35-7). In addition to advocating for a change in the marking guidelines for PhD research, Hodge encourages that post-modern humanities should be transdisciplinary by endorsing the “breeding of monsters” (p.36-7). In other words, “radical” approaches to knowledge production ushering in “a set of monsters waiting to come into the light” from the darkness of the unknown and yet to be discovered (p.39) (Figure 16).

Expanding on the philosopher Michel Foucault’s original thinking, Hodge describes “the ideal image of a disciplinary organisation of knowledge” (p.36-7) as a set of ellipses of light surrounded by darkness in which monsters live, breed and “prowl” (Foucault, 1976. p.224). Hodge and Foucault refer to monsters as knowledge and the potency to incorporate supplementary disciplines beyond interdisciplinary ones to solve problems. These monsters are, however, yet-to-be-realised pockets of knowledge that are impure, promiscuous, messy, and fertile, rather than the typical elicit response of *fear* the term ‘monsters’ usually conjures. These monsters challenge and contest disciplinary boundaries and hide in the research hinterland (Law, 2004)– beyond the interdisciplinary – less travelled, which could be considered ‘transdisciplinary hinterlands’.

Returning to the diagram, Hodge highlights an intense focus on discipline knowledge and expertise at the centre of each discipline ellipse. The unexpected can be found in the boundaries in-between surrounding the disciplines, perceived as the space for interdisciplinarity research to occur in a configuration that reinforces knowledge. Consecutively, a transdisciplinary formation can be illustrated by folding the ellipse set onto itself, thus creating an opportunistic advantage to “see what disciplines are necessarily super-imposed in the common space of [the] problem”, whereby a new centre is formed by “chaotically overlapping with outgrowths of other disciplines” (p.37). The viewpoint and perspective gained from venturing into a transdisciplinary hinterland facilitate a greater

overview and dexterity in tackling a problem with new knowledge. The following section will expand on the advantage of a transdisciplinary approach to design research and tackling ‘wicked problems’.



Figure 17: An appropriation of Hodge’s (1995) teratogenesis of disciplines (Akmal, 2020).

3.3 Design as Wicked problems, The advantage of Transdisciplinary Design Research

The well-known term “wicked problems” was coined by design theorist Horst Rittel in the 1960s. The theory was later developed in the seminal publication *Dilemmas in a General Theory of Planning*, co-authored with fellow design theorist Melvin Webber, to draw attention to the complexities and challenges within social and urban planning (1973), so-called – wicked – due to their “malignant” complexity and resistance to a solution (p.160). Unlike the “tame” or “benign” problems of science, as framed by Rittel and Webber, who reflected that within these circumstances, unlike wicked problems, “the mission is clear. It is clear, in turn, whether or not the problems have been solved” (Ibid). In recent years problems such as climate change and sustainability, amongst many others, have been labelled as wicked problems due to the lack of clarity in aims, solutions and their subjectivity to real-world constraints, hindering risk-free attempts to find a solution.

A succinct summarisation of wicked problems can be gleaned from the first publication of Rittel’s idea in a guest editorial.

[P]roblems which are ill-formulated, where the information is confusing, where there are many clients and decision makers with conflicting values, and where the ramifications in the whole system are thoroughly confusing (Churchman, 1967, B141).

This description of wicked problems resonates with the confrontations that designers experience with every new challenge, with Richard Buchanan pinpointing the fundamental issue lying behind design

practice – “the relationship between *determinacy and indeterminacy* in design thinking” with “no definitive conditions or limits to design problems” (1992, p.15-6). This particular nature of design, Buchanan theorises, is the reason why design problems are wicked. A consequence of design having no predetermined subject matter and design’s universal reach and scope towards tackling *any* problem via the intrinsic quality of “establishing a principle of relevance” from other disciplines without reducing design’s to another (p.15-8).

With reference to Buchanan (1992), McDermott et al. identify one of design’s pertinent characteristics is the ability to work in cross-sectional teams in an unrestricted approach, where collaborations can adapt and take different forms of integration betwixt other disciplines (2014). Examining the benefits and challenges of transdisciplinary collaboration in a university setting and its part in pedagogy, McDermott et al. describe a series of collaborative projects realised in their teaching curriculum. The authors explain that in the early stages of a collaborative project, students were required to drop their disciplinary affiliations to conduct heterogeneous research and integrate auxiliary “expertise and variation in thinking to handle ... complex challenges” (Ibid). From experience, the authors summarise that multidisciplinary is *just* the coming together of several disciplines to tackle a problem. In contrast, transdisciplinary is the “deeper integration” of disciplines required for enigmatic issues, with the cohort of disciplines actively shaping together the designed output (Ibid). Corresponding to these definitions is the research compiled by Alexander Refsum Jensenius (2009, 2012), whose own investigation into the relationships between intra, cross, multi, inter, and trans-disciplinarity formations were shaped by the academic Marilyn Stember’s (1991) implore to advance the field of social sciences through interdisciplinary enterprise; encouraging an evolution from the standards set solely by scientific research. Jensenius offers a concise summarisation of the different levels of disciplinarity definitions (numbered list as follows with additional examples given), and further visually disseminates and translates these into a diagram, based initially on E.F Ziegler’s Interdisciplinary model (1990), and in the process, evolving the diagram to include a transdisciplinary constitution (Figure 17).



Figure 18: Visualising inter, multi, cross, inter and transdisciplinary approaches (Jensenius, 2012).

As visually presented in Figure 17:

1. **Intradisciplinary**: working within a single discipline.
2. **Multidisciplinary**: a collaboration between different disciplines, each drawing on their disciplinary knowledge.

An example would be hospital teams with members with specific roles and duties in patient care, such as a Psychiatrist, Psychologist, Occupational Therapist, Surgeon, and Nurse. The academic Hugh Petrie specialising as a philosopher of education notes that multidisciplinary projects are short-lived and that there is “seldom any long-term change in the ways in which disciplinary participants ... view their own work” (1992, p. 303).

3. **Crossdisciplinary**: viewing one discipline from the perspective of another. Stember offers the example of a physics professor describing the physics of music (1991).
4. **Interdisciplinary**: integrating knowledge and methods from different disciplines, using a synthesis of approaches.

Nasa describes that their research centre provides unique interdisciplinary scientific expertise and capabilities that advance human understanding of the galaxy by configuring teams of members from divergent disciplines.

5. **Transdisciplinary**: creating a unity of intellectual frameworks beyond the disciplinary perspectives.

Petrie writes that transdisciplinary—

exemplifies one of the historically important driving forces in the area of interdisciplinarity, namely, the idea of the desirability of the integration of knowledge into some meaningful whole (1992, p. 304).

The emphasis, Petrie continues, is “the grand synthesis of knowledge”, referencing Marxism structuralism and feminist theory as examples (p.305). Another illustration could be the Institute for Social Futures at Lancaster University, which is committed to establishing frameworks that integrate a wide range of disciplinary expertise for the construction of social futures lenses for the consideration and prominence of *social futures* in futures research.

Referring back to this research subject matter: AI technology can be justified as a wicked problem when considering its adoption is widespread, obscured by its success, and subsequent lack of knowledge grounded in the reality of how it works, which despite a lack of working knowledge is being used for socially consequential classifications “valoris[ing] some point of view and silenc[ing] another” (Bowker & Star, 1999 p.5). By cultivating a More-Than Human Centred approach for designing with AI, thus considering the technology from an alternative perspective than human, this research aims to venture into the transdisciplinary realm via the assembly of various approaches and disciplines into a ‘meaningful whole’. The construction of method assemblages will realise the congregation of disciplines and theories.

3.4 Crafting Hinterlands through Method Assemblages

In *After Method*, the sociologist John Law (2004) presents an argument for alternative approaches in research methods for the social sciences. This line of enquiry was conceived from a need to move beyond research methods that work on the assumption that the world could be understood “*as a set of fairly specific, determinate, and more or less identifiable processes*” (Ibid, p.5). Instead, Law pursues methods that comprehend that “reality is ephemeral... [and] the world is complex and messy” (p.2), as methods “not only describe but help to *produce* the reality that they understand” (p.5). In a synonymous sentiment, Daniel C Edelson reports that the “relationship between design and ... research is changing” due to the evolving complexity of ‘design’ required for

the increasingly multifarious challenges of the world, presenting challenges for traditional and fixed research methodologies (2002, p. 106).

Law advocates for the crafting of ‘method assemblages’ to grapple with the world’s ‘messy’ and ‘slippery’ intricacies. To aid in unpacking and define what method assemblage means, Law turns to the philosopher Jacques Derrida to understand the concept as a verb, as well as a noun (2004, p.42):

...the word sheaf seems to mark more appropriately that the assemblage to be proposed as the complex structure of a weaving, an interlacing which permits the different threads and different lines of meaning – to go off again in different directions, just as it is always ready to tie to with others. (Derrida 1982, p.3)

The concept of method assemblages is the process of “assembling”, “bundling”, and an aptitude for “recursive self-assembling” (Law, 2004, p. 42). Clarifying the theory further, Law quotes the philosopher Gilles Deleuze and collaborator Claire Parnet, who explains that when considering the multiplicity of elements together, the critical part is what occurs in-between these elements (Deleuze & Parnet, 1987, p. viii), evoking Hodges and Foucault’s channelling monsters into the light. Law continues to observe that elements which compose method assemblages are not fixed in shape, thus empowering a flexible and entangled existence impelled to grow organically by not being pre-fixed by any restricting guidelines (Law,2004, p.42). Though the construction of assembles is not random, with Law proscribing, there is a choice regarding which “realities it might be best to bring into being”, quoting the philosopher Isabelle Stengers –“the question: can I incorporate this ‘thing’ into my research” (1997, p.83) remains at the core of designing a method assemblage for any research.

Throughout his book, Law describes the praxis of method assemblage using accounts of his fieldwork and determines as a research practice, it “works in and ‘knows’ multiplicity, indefiniteness and flux” (Ibid, p.14); a method for detecting and amplifying reality (p.116). As previously mentioned, Buchanan discussed the indeterminacy in design thinking with no conditions and limitations set while embarking upon a design problem. In correlation, Law’s method assemblage works and thrives in these indeterminate states to formulate knowledge, multifariously operating with an intrinsic and adaptable characteristic in the messy reality of wicked problems and the world they encumber, perfectly suited for design research.

The second enactment of method assemblages is that they craft their own hinterlands and can also grow out of hinterlands. A more detailed account of a ‘hinterland of methods’ is that they “*enact* realities ...[a]nd those realities then enact the conditions of possibility of further research” (p.38). Realising Law’s application of Deleuze and Parnet’s description indicates that the space or possibilities forged between elements are paramount to crafting and developing research hinterlands. In other words, akin to Hodge’s observation, the space between methods and disciplines is bountiful in new knowledge opportunities.

3.5 Forging A Transdisciplinary Hinterland

The structure of this thesis reflects the specific construction of its method assemblage and, consequently, the formation of the research’s unique hinterland. The divergence from the “traditional” pattern of a thesis (introduction, literature review, methods, results, discussion, conclusion) is common practice, notes the academic Brain Paltridge (2002), whilst researching how published advice on ‘how to write a PhD’ varies from actual practice. Paltridge reviews alternative thesis patterns and identifies these as “complex” (p.131), observing that the investigation of more than one topic reforms the structure from a traditional thesis and identifies Hodge’s theory of the ‘postmodern turn’ for a transformation in how theses may be theorised, researched, and written. The divergent themes of interest for this research, such as philosophy and Human-Computer Interaction (HCI), have their own established forms of presentation, with Paltridge quoting both Tony Dudley-Evans (1993) and Paul Thompson’s (1999) findings that there is considerable variation in expectation and values of the academic discipline in which a thesis is produced and assessed. As this is a design research observation, the topics of concern using Law’s theory of method assemblage will be integrated into one and pivoted towards design as the resonating and perennial approach to this research.

3.6 The Hinterland of AI as a Material for Design: a thesis pattern

In the forthcoming chapters, the overall method assemblage for this research will be introduced, echoing the reality of the research journey taken when supporting approaches were assimilated for the problem at hand. Some chapters will be dedicated to one topic, such as philosophy, with some chapters composed of various supplementary and supporting approaches weaved together,

spawning secondary or offshoot assemblages. The components of assemblages used will be reviewed and signposted throughout the thesis when appropriate. The ‘in-between’, as kindled by elements of an assemblage, is the resulting research and generation of knowledge that composes the exclusive transdisciplinary hinterland of AI as a material for design.

As noted, devising a transdisciplinary assemblage requires merging them into a meaningful whole by adapting disciplines to work together effectively. However, specific disciplines like philosophy are notoriously distinct, with figureheads often ostracising attempts to tamper with theories, believing that any adaption will destabilise a presented theory (Constable, 2009; Le Doeuff, 1989). The final part of this chapter will present a methodology for adapting philosophy facilitating the incorporation of philosophical thinking into a design research assemblage.

3.7 Transdisciplinary Assimilation; Adapting Philosophy

In their paper, *Research Through Design and Transdisciplinarity*, Findeli et al. observe that the number of disciplinary perspectives on a single phenomenon can be at risk of blurring the picture and “consequently render it difficult to grasp and operationalise in design terms” (2008 p.79). The authors go on to state that every discipline carries and is driven by philosophical sensitivities and a specific *Weltanschauung* (worldview) that influences the way “it beholds the world” (Ibid). To counterbalance the intrinsic values of disciplines when assimilating and assembling them, Findeli et al. recommend that it is:

...essential to be epistemologically awake ... to draw the right conclusions as to the consequences of our choice on the orientation and limits of the research, and on its expected and necessary relevance for design (p.79).

Findeli et al.’s assessment of transdisciplinary methods is congruent with Petrie’s (1992) thinking, in which assembled disciplines should be “integrated into one whole” through a single disciplinary perspective and into a “common problematic” (Findeli et al., 2008, p. 80-1). In his book *Alien Phenomenology, or What it’s like to be a thing*, Ian Bogost uses the philosophy of Object Orientated Ontology (OOO) to develop a phenomenological approach of viewing objects as actors in their own right, creating an unconventional perspective for the way things act as they do. As such, the

philosophical insight of OOO was specifically chosen as it correlated to the problem this research endeavours to answer in developing an alternative approach to perceive and consider AI differently, as currently epitomised when designing for AI technology (Lindley & Coulton, 2020). With the problem already detailed, the next section is concerned with integrating philosophy into the disciplinary perspective of design by investigating how one would adapt philosophy to incorporate it with another field and generate new knowledge.

3.8 A Philosophical Intermission - Adapting Philosophy; a case study of Jean Baudrillard and The Matrix Trilogy

A benchmark for assimilating philosophy and influencing the freedom to adapt philosophical sources is the work of Catherine Constable. Working in the emerging interdisciplinary field of Film-Philosophy, Constable develops a methodology for inter-relating philosophy and film that goes beyond the concept of philosophy on film, offering instead philosophy of film via the function of imagery within philosophical discourse (2009, p.149). Her research intricately investigated ways in which *The Matrix Trilogy* (Wachowski & Wachowski, 1999, 2003a, 2003b) adapts Jean Baudrillard's *Simulacra and Simulation* (1994) and creates its own unique philosophical position (Constable, 2006, 2009).

Baudrillard's *Simulacra and Simulation* (1994) is at the helm regarding philosophical source material for *The Matrix trilogy*, in which Constable using her formulated methodology, demonstrates that the trilogy both emulates and further adapts Baudrillard's concept of the 'hyperreal'. For context, Baudrillard's concern is with the role of the *image* in contemporary society and argues that the current postmodern condition is in a crisis between what is real and what is fiction. Such simulations have escalated to the point of composing how we understand reality, where ultimately, the consequential 'simulacra' is indistinguishable from the real, which is to say, an image so realistic that it is taken as reality rather than as representative – known as the 'hyperreal'. With this seamless amalgamation, such as Disney World, there is no clear distinction between where one ends and the other begins.

Famously, *Simulacra and Simulation* was the required pre-reading for cast members set by the Wachowski sisters, with a version of the book notably used as the hiding place for Neo's (Keanu Reeves) virtual reality contraband in *The Matrix* (1999). Constable draws attention to the production

design of the on-screen version of Baudrillard's book, which optimises an "eminent literary classic" leather bounded and gilded, a distinction from the slim-line paperback available to purchase, thus acting as a "visual...prefigure" of "the trilogy's take up and transformation of Baudrillard's key concepts and arguments" (Constable, 2009, p. 126). Conversely, Baudrillard has publicly condemned the trilogy for misrepresenting his philosophical work (Constable, 2009), and as a result, many treat the film negatively as a misinterpretation and a distortion of the source text; however, many identify the trilogy as a positive contribution and a "beginner's guide to philosophy" positioning the films as "useful examples that make the theories or text accessible" (Constable, 2009, p. 1; Irwin, 2002).

3.9 The Philosophical Imaginary; Philosophical Tools beyond the Written Word

Constable critically questions "what is the philosophical project of the *films* themselves" (Ibid) and endeavours to answer by initiating an inquiry into how philosophical thought is reconstituted within the communicative proclivities of moving images and figuration (Fisher, 2013).¹⁵ The *groundwork* approaches Constable engages in developing her methodology for viewing the trilogy is the philosophical work of Michéle Le Doeuff, who questioning the boundaries of philosophy, wrote about the role of imagery within philosophical language and demonstrated that philosophical and filmic texts are "profoundly linked through their reliance on symbolic figuration" (Constable, 2009, p. 150).

Constable highlights (2006) that Le Doeuff's main argument in *The Philosophical Imaginary* (1989) is that philosophical discourse is defined as "the rational, the concept, the argued, the logical, the abstract" and is overtly contrasted with the "myth, fable, [and] the poetic" of literary discourses which constitute "the domain of the image" (Le Doeuff, 1989, p. 1). This division between the philosophical and the poetic has endured since Plato expelled poets from his ideal republic because their work interferes with the search for the truth— "the *raison d'être* of philosophers" (Constable, 2006, p. 234). The historian Jean-pierre Vernant points out that the threat of images for Plato was

¹⁵ In this context, figuration is meant as the act or an instance of representation in figures, objects, and shapes.

because they “create a semblance, an appearance, through a colourful glitter of words and rhythms that produce an effect of fascination and a vertigo of the mind” (1992, p. 177). However, Le Doeuff observes that Plato himself falls foul of using imagery in his own work and cannot separate philosophy from the poetic as he draws upon mythic elements from Greek poetic heritage (1989, p. 5). Subsequently, Le Doeuff determines that in western philosophy, the image falls into two divergent roles despite the “common failure of recognition” and denial from the philosophical “enterprise” (Ibid, p.7). The image is “seen as a distraction, [and] an embellishment that should be expunged from truly philosophical discourse”, and the image acting “as an illustration, translating complex ideas into an accessible form for the less able reader” (p. 6). The crux of Le Doeuff’s argument for Constable is that despite philosophy attempting to separate itself from the image, there is ironically – an abundance of imagery in philosophical texts – which importantly serves as the means through which philosophical concepts are created and expressed (2006, p. 235). Constable summarises that Le Doeuff’s method reconceptualises the relationship between film and philosophy because imagery can draw out the conceptual implications of philosophical texts and “sustain or destabilize the concomitant philosophical system” (Ibid, p.237). This research further posits that design embodies the same characteristics as film and imagery through the development of tangible artefacts that disseminate meaning through their considered construction and curation and can themselves be tools that enact philosophical discourse (Akmal, 2021; Lindley et al., 2018; F. Pilling & Coulton, 2020). In particular reference to images and figuration, the artefacts designed as part of this research are a collection of graphical symbols designed to communicate AI’s working parameters and operational ontology.

3.10 Metamorphosis; adapting philosophy

An additional and critical element of Constable’s methodology is the work of Kamilla Elliot and her thesis on the filmic adaption of literature as a form of metamorphosis (2003), enabling one to trace changes and transformations between the source and the adaptation. The methodology is also supplemented with Christian Metz’s thesis on cine-semiotics (1982), specifically the interrelation of

metaphor and metonymy, and how visual symbols embodied in specific objects within the *mise-én-scène* assemble set up symbolic narratives (Constable, 2009, pp. 41–68).

According to Constable, the evolution of a philosophical text is the point at which a film, or thing, goes beyond imitation to the inception of novel philosophical ideation via the process of - ‘adaption as metamorphosis’. Adopting Elliot’s position, Constable enlightens that “the figure of adaptation as metamorphosis ...occurs when a series of changes are seen to create and sustain a new whole” (2009, p. 152). The idea behind Elliot’s concept of metamorphosis is established through the presentation of the White Queen’s metamorphosis from Lewis Carroll’s *Alice Through The Looking Glass* (1961, p. 174), in which a sheep knits a shawl that the Queen has already worn and will go on to wear, thus setting up a

...cyclical model of inter-relations between adaptation and original, a transformation that ensures that each return to the original is a moment in which it is viewed afresh (Constable, 2009, p. 64).

On this note, the Matrix Trilogy takes and signifies through filmic imagery an original concept from Baudrillard’s philosophy and adapts it further, changing the way we view the original philosophical concept and, in turn, alters our appreciation of the trilogy as a philosophical concept. However, with the argument against images mobilised by Plato, Constable draws on Elliot’s idea for the cognition of words and images, with Elliot highlighting (2003, pp. 221–222) that they are “both objects of perception” and that “perception is indivisible from comprehension, [as] written words and visual images are said to engage the same parts of the brain but in reverse order” (Constable, 2009, pp. 48–49). Elliot’s model of a “looking glass” depicts the inverse processes of the cognition of words and images “if a verbal metaphor raises mental imaging, then *conversely and inversely*, a pictorial metaphor raises mental verbalising” (Elliott, 2003, p. 221, authors emphasis). With Elliot’s analysis of the relationship between words and images emulating a looking glass, Constable further singles out Elliot’s metaphor of film as a:

multi-faceted prism as filmic figures run multifariously and complexly through the multiple channels of filmic signification (acting, costumes, props, sets, music, sound, dialogue, cinematography, editing and more), creating figurative resonances every bit as dense as (one

can even argue more dense than) literary figuration because of the many and varied sign systems film engages (Elliott, 2003, pp. 232–233, authors emphasis).

With an approach for identifying the transformative power of imagery, Constable draws out complex visual, verbal, and aural expressions and observes the film's 'precession of simulacra' towards hyperreality. The adaption as metamorphosis is the construction of the trilogies' distinctive philosophical position –the presentation of a series of differentiated hyperreal spaces– rather than a singular hyperreal. The trilogy observes a tryptic constitution, with the 'differentiated hyperreal spaces' presented in the film, distinguished through distinctive colour palettes "to delineate the three main hyperreal worlds: the green of the matrix, the blue/browns of Zion and the oranges/reds of machine city" (Constable, 2009, p. 91). This differentiation within the hyperreal "draws attention to Baudrillard's reliance on binary opposition... and the erasure of the two in the construction of the one, singular hyperreal" (Ibid), theorising "the hyperreal as a *single* 'universe of simulation' "(Baudrillard, 1994, p. 125) (Constable, 2009, p. 145). Consequently, *The Matrix Trilogy* has taken the source of a single hyperreal and, through a metamorphosis, constructed a narrative of three to consider the existence of many hyperspaces and a proposal of film as philosophy.

Constable's philosophical analysis of the trilogy also focuses on the adaptive and transitional functions of mirrors, screens, and code. Mirrors are significant motifs in *Simulacra and Simulation* that denote the double/binary and are often referenced in analysing a range of topics from cinema to cloning (Baudrillard, 1994). One of the significant transitions is Neo's from the Matrix by the assimilation of the liquid mirror transpiring through "[t]he shimmering viscous substance travel[ing] up Neo's arm and over his body" (p.80) (Figure 18).

Figure has been removed due to copyrights restrictions

Figure 19: The mirror liquid raising (00:31:40) (Wachowski & Wachowski, 1999).

In the referenced scene, Apoc (Julian Arahanga) attempts to track and lock on to Neo's body in the incubation towers of the power plant that feeds the machine city, using a signal initiated in the Matrix and displayed on a digitised scope (p.80). As Apoc shouts, "lock, I've got him", the monitor goes from depicting the spiral tunnel of the 'other reality' to snapping together a series of opposing vertical lines, indicating the trace a success. The silver mucilaginous liquid travelling up Neo's arm in the Matrix then consumes and spirals down Neo's throat, echoing the digital scopes illustration of the other hyperreality (Ibid). "[A] moment of metamorphosis, transforming Neo into a simulacrum", which Constable observes engages with key elements of Baudrillard's figure of the mirror as a Mobius strip that folds over on itself as "[t]he silvering of Neo's throat undoes the— opposition between inside and outside, retaining the conceptual implication of the metaphor" (2009, p. 80).¹⁶ Constable further interprets the transition to the incubation vats setting up the possibility for differentiation of the hyperreal, noting that "the moment of transition on Apoc's screen suggests that the relation between the matrix and the vats is that of different levels in a computer game", therefore the "presentation of a series of hyperreal worlds" (p.81) (Figure 19).

¹⁶ Kellner informs us the twisted Mobius strip represents the twisting of meaning in our society. He goes on to say that "[m]eaning is distorted by excess information and by the blurring of the distinction between reality and simulation. . . . understanding the Mobius strip is key to understanding Baudrillard's work and ideas (Kellner, 1994, p. 85).



Figure 20: Apoc's screen showing a tunnel down through the different hyperreal worlds (00:32:12) (Wachowski & Wachowski, 1999).

As well as mirrors, the analysis presents an in-depth concept of code within these hyperreal worlds. Famously, a way to read the Matrix by way of luminous green lines of digits spawning and travelling vertically on monitors and as an evolution of Neo's strengths as the 'One' after his death and resurrection code can be viewed directly in the Matrix through a point-of-view-shot. In the second film, code is embodied as characters serving as computer programs "capable of [positive] change" (p.150), most notably the Oracle aiding humanity in her ability to predict and guide the future.

Constable observes the positive framing of new technologies ushers a comparison to the prominent postmodern position of Donna Haraway (2009, p. 151), who uses the figure of the cyborg to trace the philosophical potential of new technologies serving as an essential guide for the design and development of future technologies. Haraway observes that:

modern medicine is ... full of cyborgs, of couplings between organism and machine,' producing a figure of a postmodern self whose openness to forms of intimate inter-relationship with the machinic provides the means for continual change (1995, p. 150).

These positive technological narratives, Haraway continues, are "not just literary deconstruction but liminal transformation" (Ibid p. 177). Constable elaborates further that these enlightened and philosophical narratives "have the capacity fundamentally to affect our future experience of technology"(2009, p. 151).

3.11 In Summary: The Matrix Trilogy as philosophy

The trilogy offers, as well as the principal narrative of humanities' downfall due to the invention of AI and subsequently the human rebellion against the tyranny of machines, a "reconceptualization of the hyperreal as a space of progression and potential" thus substituting the nihilistic trajectory of Baudrillard's narrative (Constable, 2006, p. 249).¹⁷ Consequently, there are many complex threads to Constable's argument for *The Matrix Trilogy* contributing to postmodern philosophy. Researching the intersection of popular culture and philosophy, Wartenberg accredits the role of many films is the illustration of philosophy, observing that "many fiction films embody philosophical ideas ... by providing vivid examples that make it clear what the stakes are in an otherwise quite abstract philosophical debate" (2007, p. 8). That being so, "the language of embodiment presents film [and design] as a materialisation of philosophical abstraction"(Constable, 2009, p. 159). Accordingly, the transdisciplinary approach this research will adopt when utilising philosophical theory, such as Object Orientated Ontology, will imbue the fact that philosophical "texts cannot be held to a simple, single interpretation; they always mean more than their author/s know" (p.157). On this note, Bogost argues against the need for writing as the only way to scholarly productivity, especially when 'doing philosophy', and offers two points of contention for his thinking (Bogost, 2012, pp. 88–92). The first is the argument against the ideology for the scholarship to be considered – real – it should be written, not to be "*read* but merely to *have been written*" (Ibid, p. 88). Bogost uses the example of when scientists conduct experiments in the tangible realm, the results and the practical applications are only accountable when they are written up. While the process of peer review and transparency is a good reason for written scholarship, it is often done under the guise of "academic mumblespeak" (Morris quoted in Bogost, 2010) or incomprehensible jargon. The second point Bogost observes is that writing is dangerous for philosophy as, ontologically speaking, "writing is *only one form* of being" (Bogost, 2012, p. 90), which theoretically will be probed further in the coming chapters.

In summary, we do not relate to the world through the written word alone but through confrontations with things (M. B. Crawford, 2009, p. 199). Ironically, this is a written thesis;

¹⁷ A straightforward description of nihilism is the rejection of religious and moral principles, believing that nothing in the world has a real existence.

nevertheless, it is an annotation of things made through design practice that perform philosophy. On this very note, one of the research aims is to take an abstract notion of textual philosophy and craft a tangible probe that, on some levels, *does* philosophy for viewing AI technology through design.

The aim of showcasing Constable's approach was to provide evidence that philosophical concepts can be adopted, adapted, amplified, and cast into other forms, media, and representations beyond the original text, without forfeiting validity or authority. This position was evaluated by briefly outlining how *The Matrix Trilogy* presents a series of hyperreal worlds through mirrors and code that addresses a key question from *Simulacra and Simulation* (1994) – the possibility of radical and positive change within a pre-programmed system (p.150). To this end, as analysed in this chapter, the notion of adaptation shall be a vital element entwined into the method assemblage to establish a transdisciplinary approach of philosophical thought and observation for design. A different concept of metamorphosis will be considered in the forthcoming chapters when viewing the phenomenology of things using Bogost's approach to OOO. However, as noted in this chapter, Constable's use of 'adaption as metamorphosis' by which philosophical concepts are adapted to create new manifestations or externalisations will also be appropriated and weaved into the assemblage to be engaged through design practice.

Chapter Four Methodologies

(Understanding AI)

4.1 Introduction

As established earlier in Chapter Three *Groundworks*, the methodological approach of this research will be developed over two chapters. The first methods chapter has conveyed the transdisciplinary nature of this research's hinterland by presenting Law's theory on method assemblage and how this would be translated for design research. The following methods chapter will present the overarching design approach of Research through Design (RtD) as an iterative, adaptive, and generative method to conduct design research, which responds creatively and flexibly to the research required when exploring AI. Furthermore, the design approaches and theoretical underpinnings presented here are to be considered as elements that form the design research component of the method assemblage for generating knowledge. This will begin by replicating many design theorists' approach, tackling what design research is, and revisiting and reflecting on a fundamental understanding of 'design' and 'research' (Buchanan, 2001; Faste & Faste, 2012; Frayling, 1993; Friedman, 2000). Faste and Faste observe, while attempting to demystify design research, that the process of going back to the staples, "design researchers will be able to situate their work in the larger research landscape, and explain their activities more clearly to others" (2012). Therefore, this chapter aims to 'situate' itself in the research landscape by outlining the way this research perceives design research and the methods undertaken. This chapter is in two parts; the first part outlines what design research is, which in turn sets up the foundations for the second part, detailing how this research uses and undertakes RtD as a methodology.

4.2 Design Research: revelling in ambiguity or just a nomadic practice

In their paper *The Complex Field of Research* (2010), Lois Frankel and Martin Racine provide a historical overview of design research, starting with the 'Design Science' movement of the 1960s (Hubka & Eder, 1996). The authors call attention towards a 1962 conference on design methods, which ignited the drive for design to be a "valid scientific research subject" (Frankel & Racine, 2010). Subsequently, this was quickly disseminated by leading design thinkers such as Herbert Simon (1969) and Buckminster Fuller, who advocated this period in design as a 'design science revolution' (1971). These attempts to 'scientise' design were influential throughout the 1970s,

driving the belief that research in design should be founded on scientific objectivity. This principle continues to reverberate in practice today, particularly in fields such as HCI, due to its shared roots with a positivist engineering tradition (Dourish, 2004; Rauterberg, 2003; Zimmerman et al., 2010). However, a postmodernist philosophy challenged the positivist formula, encouraging a more tolerant and pluralistic approach (Swann, 2002). The design science approach, with its “sequential structured” methods, proved “inadequate... for understanding complex design problems” (Frankel & Racine, 2010). The postmodernist design approach, celebrated by those such as Bruce Archer and Nigel Cross, ushered in the intuitive process of design, building upon design’s innate ability to tackle *wicked problems* (Buchanan, 1992; Rittel & Webber, 1973) by responding to “the complexity, uncertainty, instability, uniqueness, and value conflicts” (Schön, 1983, p. 14). Cal Swann, in his paper *Action Research and the Practice of Design* (2002), credits the integration of design with social science approaches, which brought forth a wealth of alternative ways to validate research, information, and knowledge realised through these approaches, having more of an “affinity with design processes than the science/engineering model” (p.50). In addition, Cross famously coined the term “designerly way of knowing” (2006) to describe an epistemology that has its own “appropriate intellectual culture” that does not disregard other cultures, where he emphasises has “much stronger histories of enquiry, scholarship and research than we have in design” (Ibid, p. 100). Further still, Cross accentuates that “we have to treat design as a mysterious, ineffable art... that design has its own distinct intellectual culture; its own designerly “things to know, ways of knowing them, and ways of finding out about them”(Archer et al., 1979)” (Cross, 1999, p. 7).

Notwithstanding design research’s flourishing in the early part of this century with diverse methods, ideologies and the knack of designs interdisciplinary ethos (Cooper et al., 2018), Buchanan echoes an “uncertainty” in design research and its positioning in the larger context of other disciplines writing;

Despite a growing body of research and published results, there is uncertainty about the value of design research, the nature of design research, the institutional framework within which such research should be supported and evaluated, and who should conduct it. (2001, p. 3)

Here, Buchanan highlights a trend in thinking as to what design research – *is* – prompting many design researchers to “disambiguate[e] the domain” (Lindley & Coulton, 2020). However, William Gaver cautions that observations should be “wary of impulses towards convergence and standardisation”, which would most likely “stifle the unique character of design research” (Green & Lindley, 2021,p.2) and mirror a design science paradigm. Emphasising the ambiguity of design research and positioning it as an asset in design is the academic Miguel Ángel Herrera Batista, who remarks that “design was born with an interdisciplinary tradition” (2021, p. 5). Thus reflecting on the value of contributions made to design from other fields and the application of “adding the design looks ‘from within’” to other disciplines (Batista, 2021, p. 5).

Trygve Faste and Haakon Faste explain that the term ‘Design Research’ has “become part of the common vernacular in the field of design” (2012), remarking that the term is increasingly used to label and describe a myriad of approaches, perspectives, philosophies, and methods that have merged while doing design research. Design researchers, David Green and Joseph Lindley, describe Design Research as “powerful, promising and increasingly popular, but also ambiguous, broad and contested” (2021, p.1), quoting Gaver that as a field, it is in a state of being “pre-paradigmatic” (2012, p.5). This state could be due to Buchanan’s earlier observation that design has no ‘predetermined subject matter’, which catalyses design’s universal scope of tackling any problem. In the same sentiment, the design historian Victor Margolin affirms, “[b]ecause the subject of design research, then, is not only products but also the human response to them, the research techniques for design must of necessity be diverse” (2000, pp. 1–2). Design reflects the world it designs for where “[a]fter all, the everyday world itself is inherently ambiguous: most things in it have multiple possible meanings” (Gaver et al., 2003, p.233).

However, this diversity, as Green and Lindley point out, has the knock-on effect of invalidating the field with “contemporary scholars grappling with means to define archetypes, typologies and taxonomies for design research” with “inward discussions relating to the field’s still-maturing epistemologies, methods, and conventions” (2021, p.2). Margolin identifies the way to work – even revel in the ambiguity –observing that:

[w]hen issues, rather than methods, are central research concerns, then it is possible to acknowledge different modes of research and give them value in terms of their contributions to a particular question or set of questions (2000, pp. 2-3).

This concept Gaver calls “the many worlds of design”, where research and practice in design cannot be described as a single and independent world; instead, it is multiple and generative – the “proliferation of new realities” that co-exist together with its burgeoning of methods and techniques (2012, pp. 941– 943). Gaver explains that the design process fails to emulate the convergence process in science, where the discipline is fulfilled by building on accepted results. Instead, design (and the arts) “are cumulative in the way a conversation is, elaborating on what has gone before, but seldom aiming for or finding resolution” (Gaver, 2012, p. 942). This ‘heterogenetic’ nature of design, Green and Lindley reflect, is invoked by the “panoply of methods” design researchers exercise and is what makes design research “so powerful” (2021, p. 2). However, it makes the field liable to the pre-paradigmatic mould and the blowback of questions about the nature of design research due to its unrestrained, unpredictable, and ambiguous nature (ibid). The academic Ron Wakkery, however, would find strength in the ambiguous nature and correlate it to ‘Nomadism’ as described by Deleuze and Felix Guattari (Deleuze & Guattari, 1987). Wakkery writes:

Nomadism... refigures design from a single territorial discipline to a multiplicity of concurrent, allied, non-allied, collaborative, competitive, contradictory, or aligned practices of design marked by who gathers around a particular something to design. There is a plurality of gatherings that traverse across a landscape, territorializing and deterritorializing as they go, following the somethings they design for wherever that may lead, often crossing paths to contest or form allegiances with other nomadic practices (Wakkery, 2021, p. 53).

Consequently, the methodology used in this research has been developed to work in a transdisciplinary approach, well-suited to design’s nature, as Jacobs describes design as the “scavenging” methods across disciplinary traditions (Jacobs quoted in Green & Lindley, 2021).

This sentiment and approach for conducting design research reflects the diversity and complexity of the incited problem of AI legibility. Embracing the ambiguity of design research as a strength comes with the flexibility to adapt, borrow, and utilise various methodologies, theories, and

ideologies from a range of fields siloed into a method assemblage, as detailed previously. To this end, a careful balance shall be struck between kindling and framing the context of design research to avoid being “overly rigid”, echoing that this is just one of many ways to conduct research (Margolin, 2000) in its nomadic and ambiguous nature. The following sections of this chapter will define what design research is through a rudimentary comprehension of design and research and their merger.

4.3 Design Research; defining Design

Design research comprises two distinctive words – design and research – which Green and Lindley emphasise forms an ‘open compound’ word that habitually inherits and relinquishes meanings and associations (2021, p.2). Consequently, a consideration of what meanings are and are not inherited in the context of this research shall follow, starting with the term design.

The word design in the English language articulates a multitude of connotations from its use as a verb, for instance, as an activity ‘to design’ (lower case), as a noun such as a design, a “purpose,” “plan,” “intention,” “goal,” “malicious intent,” “plot,” “form,” or “fundamental structure” (Flusser & Cullars, 1995), and as an adjective, as in something is ‘designer’. Or fulfilled by design (Friedman, 2000, p. 9; Glanville, 1999, p. 88; Lawson, 2005, p. 3; Walker, 2018, p. 1). Ken Friedman (2000, p. 5-9) traces the act of design and its entwined existence with the evolution of humanity to the present day. From design’s dawn over half a million years ago with *homo habilis* crafting tools to more specialised tools such as spears and “information tools” some 20,000 years ago. Friedman observes the design of these later information tools as the undertaking of humans externalising the “representation of knowledge” by carving onto bones or antlers, thus forging tools that “reshape the way we think” (Ibid,7). The early inauguration of information technology. Today, architectural theorists Beatriz Colomina and Mark Wigley observe that “[t]he average day involves the experience of thousands of layers of design that reach deep into the ground and outer space but also deep into our bodies and brains” (2016, p. 9) showcasing how design occupies multiple disciplines, spaces and manifestations. Notwithstanding design’s ability to form tangible outcomes and solutions to a problem, design also offers the opportunity for theory development, with design procedures about

specific domains in a descriptive and prescriptive manner for the formulation of frameworks, guidelines, and methodologies for research (Edelson, 2002, pp. 112–116).

The design academics Rachel Cooper and Mike Press suggest that design is an umbrella and a generic term for many specialised disciplines, such as graphic design, industrial design, and fashion design etc. (1995, p. 26). To investigate the interrelationships between these diverse design disciplines, Cooper and Press present a historical linear model and visual metaphor of design using Walker's 'The design family tree' diagram (Figure 20) (p.23). The historical mapping of design's disciplines is presented using the trees' botanical growth and structure to map time. As an example, based at the tree's roots is the practice of 'drawing' presented as an act of craft tradition, with 'computer-aided design' (CAD) residing at the uttermost reach of the trees canopy, illustrating a derivative progression in technological sophistication (based on technology in 1989 when the diagram was created). The tree also maps an association of design disciplines with art and science. The placement is decided upon by a tenant of sensibilities associated with specific disciplines, for instance, fashion, a discipline depicted as more in line with art and disciplines such as structural engineering, veering over to the science branch. However, one could argue that due to the interdisciplinary and transdisciplinary nature of design, flexibility should be presented in the design tree with the ability to move disciplines to either art or science branch. For example, fashion design could be closer to science, depending on the interests of the project at hand.




Figure has been removed due to
copyrights restrictions

Figure 21: The Design Family Tree with CAD residing at the top with craft at the tree's roots. An appropriation of Walker's diagram (1989).

The definitions of design that concern this research are design as an act of problem-solving built upon design as a creative process. Notwithstanding Friedman's steadfast synopsis that there is "no common and well understood definition for design" (2000, p.9), Cooper and Press endeavour to define design in lieu with a myriad of definitions presenting the design family tree as a notion of 'design as a family of professions' (1995, p.25-7). The authors proceed with presenting design as a form; of art, an industry in its own right with consultancy applications, a process of planning and a creative act for achieving goals, and finally, the act of problem-solving (1995, pp. 7–47).

4.3.1 Design as a creative and iterative process

As a verb, design is a "dynamic process", according to Friedman, and explains this delineates "the ontological status of design as a subject of philosophical inquiry" realised through the role of a designer as a thinker and planner with skills of moving from thought to action (2000, pp. 9–10). This quality relates to a primary feature of design as an act of research and creation and habitually pursues to 'draw things together' (Binder et al., 2012).

Inspired by the Science and Technology Studies (STS) philosopher Bruno Latour's Object-oriented politics, in that "objects are always assemblies" of complexity, Binder et al. propose a view of design as unpicking this complexity by "accessing, aligning, and navigating among the "constituents" of the object of design"(p. 26). These constituents are the socio-material things Latour advocates and challenges designers to make public in their work. The authors go on to advocate that the object of design's public communication can be achieved via "creative design practice" and "imagination" inspired by the work of Donald Schön and the role of a 'reflective practitioner' for which "knowing and doing are inseparable" and "learning-by-doing" is seminal to the process (p. 24). Expanding on how to devise a creative process and influenced by John Dewey's philosophical works of 'experiences', Binder et al. promote that a creative process is a form of human experience for which design typifies the "inseparability of doing and experience" (p.25). Following Dewey's observation, Binder et al. explain that "all creative activities show a pattern of controlled inquiry framing situations, searching, experimenting, and experiencing" and conclude that observing is a vital

characteristic and that the process is “open-ended” (Ibid). In other words, the creative process is not a predetermined linear journey, rather it is nomadic. Harmonious with this view, Edelson characterises the design process as a “complex” and “open-ended” creative process, which is achieved through iterative cycles of design with an emphasis on the implementation of data and research into the subject matter to coherently inform both the design and the method to respond to the challenge in hand (Edelson, 2002, pp. 106–108)

From another point of view, outlining design as a creative process, Bryan Lawson examines how designers think, noting a binary division between rational, logical, and convergent thinking and intuitive, imaginative, and divergent thinking (2005, p. 142). Lawson contends that combining these two calibres of thinking is a crucial design skill and suggests the creative process of design is by no means a purely analytical task but also an imaginative insight into the interpretation and solution of a problem towards a goal. Lawson proposes five phases to make sense of the creative process: ‘first insight’, ‘preparation’, ‘incubation’, ‘illumination’, and ‘verification’ (Lawson, 2005, p. 148).

Inspired, Cooper and Press adapt these phases to emphasise how designers iteratively think through problems and describe this as an “internal creative process” of the designer resulting in a project output (1995, pp. 36–37). This process starts with defining and understanding the problem, developing ideas, and testing the designs. The authors accentuate that the process is rarely linear. With new information and data, designers return to earlier stages, underscoring that an iterative and creative process is indispensable in a design journey (Ibid) (Figure 21). On reflection, Swann observes that the design process can only be “effective” if it is in a “constant process of revisiting the problem, re-analysing it and synthesising revised solutions” (Swann, 2002, p. 53)



Figure 22: Design as a process (Cooper and Press, 1995).

As noted previously, this design process is one of many interpretations of how to practice design, conduct research, and explanations of design processes (Frankel & Racine, 2010; Friedman,

2000; Koskinen et al., 2011; Rhea, 2003; Sanders, 2008), consequently highlighting design's fluid and inclusive nature to cultivate cross, multi, inter and transdisciplinarity formations due to the expansive and wicked nature of problems design confronts.

4.3.2 Design as an act of creative problem solving

The design theorist Benjamin Bratton quips that “the job of Design in the 21st century is to undo (much of) the Design of the 20th” (2016). For this reason, the celebrated design researcher Donald Norman discerns three common reasons why design fails. Firstly, aesthetics over function; secondly, designers design for themselves, failing to understand the context of the use and function of the products; and thirdly, clients have the final say, often changing the original design pitch (1988). However, a seminal point made by Norman is that “most design is not done by designers, it is done by engineers, programmers and managers” who do not have training or aptitude for design (p. 156).

On the other hand, one can say that the array of design challenges is vast. Friedman presents the types of challenges in a taxonomy of design knowledge, emphasising core domains and disciplines of inquiry while also highlighting the implications a designer is faced with, should consider, and utilise (2000). These domains include skills for learning and leading, the human world, the artefact, and the environment, which collectively emphasise the multifaceted extent of knowledge vital to exercise design (p.11). Therefore, it is often considered that design is the act of problem solving (Edelson, 2002; Jones, 1980; Simon, 1969). Speaking from an industrial design perspective, Cooper and Press note, “[b]ecause the products of design fulfil a specified function, then design is an activity concerned, at *least in part*, with problem solving” (Cooper & Press, 1995, p. 16 (emphasis added)).

Similarly, the design academic Armand Hatchuel propositions that “there is no doubt that problem solving is part of a design process, yet it is not the whole process” (2002, p. 10). Hatchuel promotes that the design process – also – seizes the opportunity of “expandable rationality”, whereby “unexpected expansions of the initial concept” generate new and unconsidered problems since design does not conform to a contained format of logic found in maths or science to solve problems (Ibid, p.5). For this reason, Hatchuel (2002) believes that Herbert A. Simon's thinking on problem solving is

too restrictive due to his attempts to credit the process of creative problem solving to a formula, which reasonably was influenced by his work in AI and cognitive psychology (Simon, 1969).

Hatchuel's theory cultivates a design process that includes and goes beyond problem solving. In other words, a process open to arrive at novel and sometimes unexpected solutions, and in part, the welcoming of encountering unforeseen problems due to expansion for a more thorough exploration for a solution – mirroring the complexity of problems design undertakes. In a nutshell, this process rejects Simon's 'Bounded rationality' that promoted a "short list of actions instead of rich spaces of possibilities" (p.9). Furthermore, for Hatchuel and fellow academic Benoît Weil, this insight yielded the perception that upholding a more rigorous and precise design process limits creativity (2003). In response, the authors introduce a cyclic model for the design process known as C-K theory, which sees the reproductive effect of information shared between two spaces of concept and knowledge (Figure 22).



Figure 23: The Design square by Hatchuel et al. (2004) explores the problem-solving process of design moving between spaces of concept (C) and knowledge (K).

In a generative manner, new concepts expand the c-space, thus triggering the expansion of new knowledge. In a complementarity manner, new knowledge incites new concepts, which equals the

search to acquire new knowledge. In this sense, the design process becomes a form of research through ‘creation’ and the opportunity to scout for new knowledge and unanticipated hinterlands due to an aptitude to expand and continue to define and understand the problem space in greater detail.

4.4 Defining Research; Research is Design

The following section defines the ‘research’ component of design research. On a contradictory note to the previous line of thought – design as a form of research, Faste and Faste offer an alternative view which they visually present (Figure 23) and justify “that design research is not a “kind” of research, but rather that research is always a “kind” of design” (2012). “Practice”, the authors continue, “is the super set” with “research ... a subset of design practice at large, and that design research is simply the set of such methods not conventionally considered to be research” (Ibid).



Figure 24: By seeing research as a subset of design, Faste and Faste (2012) propose a view that design embodies research with practice embodying all.

Faste and Faste go on to explain that while typical research, such as science, narrows its focus towards specific solutions to well-defined problems, design research, as touched upon, expands and broadens the problem domain and, with it alternative solutions for wicked problems (Buchanan, 1992). In conclusion, “[d]esign research is really about the design of design” (Faste & Faste, 2012); that is to say, the act of design itself is a type of research, and research is a type of design with creative practice encapsulating the process. The importance of design as a practice-based discipline will be discussed at

length later in this chapter. Though as a prefix commentary, the creative practice introduced by the design process has been an essential application to the success of inter- and cross disciplinary research endeavours (such as those concerned with HCI) for overcoming limitations by expanding and reiterating the process of research, problem framing and knowledge generation (Frankel & Racine, 2010; Gaver, 2012).

4.4.1 Kinds of Design Research

Until this point, this chapter has discussed several facets of design research. It is drawing a fine line between defining how this research interprets design research and describing its pre-paradigmatic condition, which has been interpreted here as a strength by revelling in the ambiguity and expanding a more significant number of possibilities and knowledge avenues.

On the other hand, some qualities are known that disambiguate the domain of design research. A common misunderstanding in the design community Buchanan (2001) observes, is that the act of research can be reduced to a single activity. On that note, there are three forms of research identified by Buchanan and others (Frankel & Racine, 2010; Friedman, 2000): basic research, applied research, and clinical research. Friedman provides a short and simplified version of the following explanation: “research is a way of asking questions ... The different forms and levels of research ask questions in different [(basic, applied, and clinical)] ways” (2000, p.18).

Basic research concerns an empirical examination of fundamental and general principles that lead to developing theories with far-reaching implications (Buchanan, 2001; Frankel & Racine, 2010; Friedman, 2000). Friedman explains that “truly general principles” can apply beyond the field they orientated (2000, p.18). They are sometimes “the *first* principles – which govern and explain phenomena” (Buchanan, 2001, pp. 18–19). To help situate ‘basic research’ in a design context, Buchanan associates this type of research with design theory, a critical process he emphasises provides a foundation for all other activities in design (Ibid, p. 19). Applied research adapts the findings of basic research into a classification of problems (Friedman, 2000, p. 18) and develops “reasoning” (Buchanan, 2001, p. 18) and “several hypotheses” (Frankel & Racine, 2010), resulting in new knowledge that is effective and targeted at the problem through systematic inquiry (Buchanan,

2001, p. 18; Frankel & Racine, 2010; Friedman, 2000, p. 19). Buchanan highlights that clinical research is directed towards an individual case focusing on the problem for action, which involves the application of both 'basic' and 'applied research' findings (2001, p.17). Frankel and Racine provide an example of a designer tasked by a specific company with the design of a particular walking aid that would require research explicit, though wide-ranging (including basic and applied research findings) to that project, such as users, materials, environments, and competitive products (2010). The authors go on to mention that both Buchanan (2001, p.18) and Friedman (2000, p.18) identify a common trait in case studies that this research generates insight into problems beyond the problem in hand, which can be filtered back as the subject of 'basic' and, or, 'applied research' (Frankel & Racine, 2010). Subsequently, design research is the balance and engagement of basic, applied, and clinical research, for which reason Friedman distinguishes that "[t]he designer is a synthesist who helps solve problems" (2000, p.18).

4.5 Research Through Design

After defining a rudimentary understanding and background of design research, this part of the chapter now focuses on the design research methodology of Research Through Design (RtD), the lynchpin method for this thesis.

The origins of RtD are habitually traced back to Sir Christopher Frayling's renowned paper *Research in Art and Design* (1993), which was primarily grounded on previous work by Herbert Read about education through art and an enduring conversation with Bruce Archer and other members of the design research studio at the Royal College of Art. Frayling articulates the existence of a conceptual tension that "research practice" has been historically entwined with the scientific method and therefore championing "words not deeds" (p. 1) and suggests that, in reality, this signature is not representative of research practice in art and design. In summary, Frayling offers three modes in which research could be executed in both art and design practice, conceived by comparing the roles and activities of a researcher in the domains of science, art, and design. Since the publication of Frayling's paper, many design researchers inspired have gone on to accentuate these research modes for design research, with many authors such as Findeli et al. (2008), Zimmerman et al. (2010), Faste

and Faste (2012) and Gaver (2012) who have focused predominantly on RtD. A brief synopsis of Frayling's research modes in the context of design is as follows:

- **Research for Design** is research linked to the designer's practice and carried out in developing every design project. In his book *Design Research*, Peter Downton calls this research "research to enable design" (2003, p. 17) and is therefore known to all designers as it is part of their daily practice, with the designed artefact the final output for this mode of research.
- **Research into Design** sometimes referred to as research *about* design, is a research approach correlating to what Margolin calls *Design Studies*, often occurring in academia for the research concerned with developing a greater knowledge of design as a discipline. Buchanan also calls this mode of research "design inquiry" and perceives it as "an explanation in the experience of designers and those who use products", noting two research subject themes as "the discipline of designing" and "the creativity of the designer" (Buchanan, 2007, p. 58). Findeli et al. (2008) unpack the research categories further and perceive it as a probe into design with relation to its history, objects, processes, actors, meanings, and social impact.
- **Research through Design** is an approach where the researcher develops through practice "prototypes, products, and models to codify their own understanding of a particular situation and to provide a concrete framing of the problem" (Koskinen et al., 2011, p. 5). In this process, "transferrable knowledge" (Durrant et al., 2017, p. 3) is generated through the practice of design –with– as Frankel and Racine stress, "the emphasis is on the research objective of creating design knowledge, not the project solution" (2010).

In his paper *Research into, by and for design* (2008), Friedman acknowledges that Frayling's probe is a "worthy effort" (p.156); however, he criticises the fact that these categorisations of design should not be taken as a factual representation of design practice. To some extent, Frayling already pre-empted Friedman's sentiment noting "that research for art ... and design needs a great deal of further research", striking an invitation for debate and further research to clarify the design research phenomenon, which has unquestionably commenced in the years following the publication of the

paper (Frayling, 1993, p. 5). To give an example, building on Frayling’s research modes and another of Friedman’s queries of “how does new knowledge move from research into practice” (2000, p. 1), Frankel and Racine present a map illustrating the flow of knowledge between research *for* design, research *through* design and research *into* (about) design, and how each of these modes of research informs one another and subsequently corresponding to basic, applied, and clinical research (Figure 24) (2010). One could say that the map illustrates Friedman’s response to his own question, observing that:

The important issue is that a field must grow large enough and rich enough to shape results and circulate them. As this happens, the disciplinary basis of the larger field also grows richer. This leads to a virtuous cycle of basic results that flow up toward applied research and to clinical applications. At every stage, knowledge, experience and questions move in both directions... Practice tends to embody knowledge. Research tends to articulate knowledge (2000, p. 23).



Figure 25: Cyclic relation between kinds of design research according to Frankel and Racine (2010).

Frankel's and Racine's *Map of Design Research Categories* plots, unsurprisingly, RtD as an applied research approach – the systematic inquiry through design practice of adapting basic research for the development of new hypotheses – by executing action-reflection methods.

The routine association of RtD and Action Research can be traced back to Frayling's paper, for which he identified this as a method for reporting and communicating the results of research; consequently, the explicit "separate[ion of] research from the gathering of reference materials" (1993, p. 5). The psychologist Kurt Lewin, generally cited as the originator of Action Research and thereupon often utilised in the social sciences, observed, "[i]f you want truly to understand something, try to change it" (Lewin quoted in Tolman et al., 1996). Lewin here accentuated that the comprehension of something (research) and the improvement of something (design) coincide, an instinctive journey for the learner and designer, for which the prominent researchers *about* RtD, Pieter Jan Stappers and Elisa Giaccardi describe this statement as "close to the heart of designers" (2017).

Writing extensively on Action Research, Ortrun Zuber-Skerritt, describes the method in brief terms as a spiral of cycles of action and research consisting of four major moments – plan, act, observe and reflect (Figure 25) (Carroll & Kellogg, 1989) (Zuber-Skerritt, 2001). The reflecting stage may lead to identifying a new problem or problems creating a new cycle.




Figure has been removed due to
copyrights restrictions

Figure 26: Cyclic diagram showing the process of action research, emphasising the approach is not linear but rather iterative. Adaption from Carroll & Kellogg (1989).

Swann suggests that the cyclical process of Action Research bears a familiar resemblance to the design process, remarking on the accentuated notion of action combined with research manifesting as an “interplay of forces in the process of the activity, and this is precisely what designing is about” (2002, p. 56). On this note, Swann encourages the method of documentation, inciting the work of Schön’s *The Reflective Practitioner* (1983), which is an established methodology for Action Research highlighting the cross-fertilisation of methods and approaches adopted by social sciences and design. As previously mentioned, Schön formulates an epistemology for practice-based processes: the method by which practitioners reflect on their process during and after. For Schön, the key terms for inciting a reflective process is “in action” and reflection “on action” (1983).

The significance of this section was to show the roots of RtD as a brief retrospective, showcasing the affinity for design to adopt social science methods, therefore underlining how RtD has appropriated, innovated, and developed a unique approach for design research. The following sections will further elaborate on RtD as a methodology and stipulate how this research will utilise this approach.

4.5.1 What to expect from Research through Design

Returning to an earlier conversation regarding design’s pre-paradigmatic state manifests into RtD’s current situation. As noted by Gaver, there is “little agreement about the values for where we should design, the appropriate methods for doing so, standards for evaluation or agreed forms of output” (2012, p. 942). Zimmerman et al. advocate for a specific model and formalised approaches to RtD (Zimmerman et al., 2007, 2010). In short, calling for a RtD paradigm to be set, conceivably influenced by HCI’s positivist approach. Whereas Gaver, willing to be nomadic, warns:

such standards might lead to a form of self-policing that would be overly restrictive of a form of research that I value for its ability to continually and creatively challenge status quo thinking (2012, p. 937).

Gaver further emphasises, in his paper *What Should We Expect From Research Through Design*, “that attempts to establish disciplinary norms of process or outcome are political acts to be approached with

care” (Ibid, p. 945). Gaver speculates that RtD may develop through a discursiveness attitude and elaboration precipitated through subversion, suggestions of alternatives, and the establishment of “entirely new constructions” explored in a design community where “consensus can come to look like a constraint” (p. 946). This impression of RtD is a result of Gaver comparing two accounts from the Philosophy of Science, highlighting issues within these examples, and juxtaposing them with the nature of design as a research endeavour, ultimately exposing the “characteristics of theory” likely to be produced as a result of doing RtD (p. 939). The two examples are ‘Popperian falsifiability’ and Lakatos’ observations of ‘scientific research programmes’, chosen by Gaver to illustrate how “unsettled and controversial accounts of science are”, thus illuminating a diversity of philosophy also found in the sciences (p. 941).

As a brief overview, the philosopher Karl Popper proposed that for a hypothesis and, by extension, any derived theory to be considered scientific must be held accountable, tested, and falsified in principle. Popper noted that an endless number of confirmations could not prove a theory, which he observed often occurring in pseudo-sciences. Popper exclaimed that “[i]t is easy to obtain conformations or verifications, for nearly every theory – if we look for confirmations” (2002, p. 47).

Painting a very different picture of science is the philosopher Imre Lakatos’ account of scientific research programmes, which is characterised by an assemblage of a ‘hard-core’ theory surrounded by a ‘protective belt’ of additional theory, hypotheses and various approaches that attempt to answer any research programmes unanswered questions (Lakatos, 1977). These programmes are considered to evolve through adapting the protective belt rather than the hard-core theory, resulting in a “dynamic machine for generating new knowledge, new understandings and new discoveries” (Gaver, 2012, p. 940). Gaver observes that Lakatos’s theory for science is more familiar to a RtD approach than that of Popper’s, which “holds a potentially unflattering mirror up to the theories produced as a part of research through design” due to the account of ‘confirmation’ often practised in design and design theories often categorised as “vague” (Ibid). Nevertheless, observing this as a vantage, Gaver stipulates that design and RtD are generative *rather than* verifiable through falsification, which he stipulates if this was the case it would change the essence of design by practising it through “arranging tests to refute such statements”, whereas “theory should be allowed to

emerge from situated design practice” (p. 940-942). Here, Gaver specifies that RtD as an approach should focus on “theory-making ... as a way of capturing and communicating new learning to the research community and as a way of guiding design practice” by virtue of designed artefacts and annotated portfolios – collections of artefacts– that capture and communicate “the myriad of choices made by their designers” and the “implicit theories embodied” (p.943-944). With this in mind, Gaver moves the conversation past how scientific or unscientific RtD is for the design community and instead lends more weight towards how to derive theory, research, and knowledge from design practice.

4.5.2 Annotated portfolios

Continuing reviewing theory-making through RtD, Gaver highlights the problem that “theory underspecifies design”; that is, any given theory falls foul of encapsulating all the successful aspects of a single design (p. 940). Therefore, Gaver introduces annotated portfolios with his colleague John Bowers as a method to practically guide and conceptually develop theory through design (Ibid; Gaver & Bowers, 2012).

To grasp the underlying concept, Gaver introduces two metaphors for how design artefacts are paramount to design theory and its formation (p.944). The first is John Carroll and Wendy Kellogg’s concept of a design artefact exemplifying a ‘theory nexus’, where multiple elements and choices made by the designer are personified in a single thing, revealing both issues of importance and how to address those issues (1989). The second metaphor is philosophical in the sense that Gaver describes designs occupying a point in a design space. A collection of design artefacts, therefore, formulates a portfolio, inhabiting a point collectively in the design space, and individually each design forms a unique design space around itself. This notion could be described as taking on the materialisation and sensation of a design orrery, befitting Gaver’s observation that designs operate on their own path and in their own space but also converge with one another, whereby multiple examples can ‘tease’ out interrelated concerns and judgements from a particular configuration between examples, hence providing greater opportunities for the designer to make better design decisions and conduct design research (p. 944).

The takeaway ideology is that where “artifacts embody theory”, annotated portfolios “encode” theory by making “accessible” the thinking and decisions that encompass an artefact’s embodied theory and provide “dimensionality” to its design space (p. 944). In this way, Gaver transforms the relationship of theory and design by instead of presenting design examples as “mere illustrations” of theory, design theory is distinguished in its own right as annotation – a process of illuminating, amplifying and accentuating the “ultimate particulars” (Stolterman, 2008) or “truths of design” (Gaver, 2012, p.944). Gaver stresses that annotated portfolios are not the replacement for all other forms of design theory, such as theoretical writings, but provides a manner or device to unpick, analyse and communicate theory, research and ultimately knowledge, “[a]s artefacts are to theory, from this perspective design portfolios are to research programmes” (Ibid). With this overview of annotated portfolios, it is easy to trace this model back to Schön’s reflective practitioner (1983) and Action Research by annotating the designerly ways of knowing (Cross, 2006).

The concept of annotated portfolios is a strong argument for legitimising the research activity of RtD via the process of theory-making. It is, therefore, a robust method for this research, by which this thesis will be the embodiment of a design portfolio with annotated particulars of the designs and the charting of the design space, hence extracting the contribution of knowledge towards AI as a material for design.

4.5.3 Research through Design = Practiced-Based Research

As mentioned previously, design is rooted in practice – that is to say, hands-on – and RtD enthusiastically exemplifies this fact, as evidenced by the wide variety of design practitioners using RtD who make things for the objective of knowledge generation. Wherein the artefact codifies and unfurls a new understanding for the designer (Akmal, 2021; Basballe & Halskov, 2012; Gaver, 2012; Gaver & Bowers, 2012; Lindley et al., 2020; F. Pilling, Akmal, Gradinar, et al., 2020; F. Pilling, Akmal, Lindley, & Coulton, 2022; F. Pilling & Coulton, 2021; Zimmerman et al., 2010). In the paper, *Research Through Design: Twenty-First Century Makers and Materialities*, Durrant et al. (leading design academics who take part in researching *into* RtD) reflect on the process of practice-based inquiry, that they summarise generates transferable knowledge through the act of *making* (2017). The

authors use the RtD conferences (where the method of RtD is analysed and celebrated biannually) to reflect on and explore RtD as an approach. Their observations evoked the craft traditions of design, accentuated by recounting the anthropologists Tim Ingold's keynote speech from an earlier conference. Wherein Ingold physically demonstrated through the simple act of manipulating the lengths of strings in his hand that making is “*constitutive* of knowing and understanding” that ‘being’ of string (Ibid, p. 5, emphasis authors own), which Ingold inferred that “design is fundamentally processual and relational in a practice of ‘gathering’ and transforming materials” (Ibid).

To examine this aspect further and emphasise RtD as practice-based research, Faste and Faste present a taxonomy matrix of design research to “demystify” the different research modes via their attributes, consequently illuminating the practice-based emphasis of RtD (Figure 26) (Faste & Faste, 2012).




Figure has been removed due to
copyrights restrictions

Figure 27: Research through Design is hands on in the process of creating knowledge through design. Faste & Faste (2012). The matrix comprises four modes of design research they consider existing; the three ‘original’ modes of research Frayling presented – *for*, *into* and *through* design and the final mode called *Design of Research* – which, as the name suggests, is the process by which research activities are designed. Viewing the matrix, the horizontal line reflects research on the left and design on the right. The

authors explain that since research is a subset of design activity, as described earlier, the left half is a subset of the right. The vertical axis represents the ‘degree of the practitioner’s involvement’ incited as ‘hands on’ at the bottom and ‘hands off’ at the top. RtD is in the prime position of the bottom right corner. Subsequently, Faste and Faste offer an alternative name for RtD, "Embedded Design Research" due to the specific part of the method – whereby through the enactment of design – knowledge generation flourishes at the hands of the designer (Ibid). Analysing this, the authors explain that this mode of research is:

a combination of process and research culminate in an artifact as the embodiment of design research knowledge (e.g., an object, process, interaction, experience) ...the knowledge generated is contained in the cognitive processes and artifacts of the design activity performed (Ibid).

Accordingly, Zimmerman et al. summarise, “the artefact [constructed] is itself is a type of implicit, theoretical contribution” (2010, p. 314).

Reflecting on these concepts, Stappers et al. singles out that ‘design action’, or rather, the role of things being made (objects/ artefacts/ prototypes/ sketches/ frameworks) in practice-based research, forces the designer-researcher to go in and do it. Accordingly, “confront theory, confront the world, [and] ... evoke discussion and reflection” (2014, p.166-169). In reaching this conclusion, the authors compare design-inclusive methods with RtD and observe that the former is driven by theory and hypothesis testing, which is essentially a process dissociated from knowledge generation and therefore restricted, whereas the latter is “phenomenon-driven”, nomadic and “explorative in nature” (p.166-167). On this subject, Faste and Faste see an artefact as a result of an RtD investigation as “embody[ing] the answer to the research question” (2012, para, 22), which is how this research will perceive the creation of artefacts.

4.6 Conclusion and Going Forth

This chapter has established the methodological approach for this research, along with the previous chapter Groundworks, conveying the method assemblage in which RtD folds into as a catalytic force. The combination of the method assemblage and RtD is potent for generating

knowledge. Both methodologies work harmoniously together due to their resembling predispositions of flexibility in adapting and moulding approaches, styles, and ideologies. Subsequently, the amalgamation of methods facilitates a transdisciplinary approach for merging design practice, philosophical theories, and technical understandings of AI.

Before going into the theory behind RtD, this chapter outlined a comprehension and a brief history of design research, noting that there was little agreement in the field. With the presented contention in design, this research has been orchestrated with the conviction that there is merit in the ambiguity of what constitutes design research and practice, and with it, the ‘apparent’ lawlessness in using off-the-cuff approaches and theories within the field—enabling the scavenging and assembly of methods explicitly tailored to the problem in hand. Additionally, this chapter has also probed what it means to undertake research through design by cross-examining rudimentary definitions of both research and design, highlighting Faste and Faste’s observation of a reciprocal connection betwixt them both, due to design acting as a type of research, and research a type of design engulfed by creative practice and problem-solving. Consequently, this analysis precipitated in construction of a design research model paying homage to Gaver’s opinion of theory and knowledge emerging from design practice through annotation. Rather than the construction of artefacts to test and produce theory, archetypal of a science lineage. The next chapter will develop the More Than Human Centred Design approach before introducing elements to the method assemblage with the design practice of Design Fiction integrated as a method of adaptation for philosophy through metamorphosis.

Chapter Five More Than Human-Centred Design: Shifting Perspectives through Philosophy

(Being AI)

5.1 Introduction

This chapter provides the theoretical underpinnings of this research, as one of the contributions of this thesis presents an alternative design approach to the widely adopted human-centred approach and perspective in the design of AI services, products, governance, and implementations. The literature review provided an overview of our confused perception of AI's ontology and the impact this has, namely, users' perplexed understanding of AI technology and the ramifications of its use. The review also summarised the integral reason for this confused perception— attributed to synthesising artificial intelligence with human intelligence: primarily leading to not considering AI as a thing in its own right. By developing an alternative perception and approach for human-centred design, we can design differently for AI.

As a synopsis, the following will present a More-Than Human Centered Design (MTHCD) approach for designing with AI by adapting and metamorphosing the philosophical thesis of Object-Oriented Ontology (OOO) with design thinking. This chapter is the backbone of this research, presenting and assembling many interdisciplinary parts into a united whole, which has been segmented into three parts for ease of consumption for the reader. However, the reader should be aware of the size of this chapter, although large the contents of which have been condensed down to the fundamental concepts for this research. To begin, the chapter will provide a historical account of Human-Centered Design (HCD) and how the approach is reflected and embodied in current design thinking for AI. This part will give a platform to specify the counter-theoretical thinking of a MTHCD approach. The second part of the chapter will present an OOO rationale derived from both Graham Harman's and Bogost's conception of and adapt these theories into a MTHCD approach by metamorphosing it through a proposed model of speculative design. The third part of the chapter reintegrates the human user back into the approach to develop a Human-AI Kinship through the application of a post-phenomenological lens, as, ultimately, anything that will be designed will be for human consumption.

5.2 Human-Centered Design: A Concise Background

HCD is the design approach that centres on humans and their needs, behaviours, motivations, and emotions in the development of a design; essentially, the viewing of humans not as part of the system but as central in every aspect of the design. Numerous scholars have developed design approaches that centre on human values, for example, John Arnold (2016); Henry Dreyfuss (1955); Donald Norman (1988); Victor Papanek (1983), to name a pivotal few.

The cognitive psychologist and designer Donald Norman is often cited as a key instigator of HCD (1988; Norman & Draper, 1986). Norman started his research into HCD when invited to be part of a team to analyse the Three Mile Island (Pennsylvania) nuclear power plant accident in 1979 for causes and potential solutions. He remarked, “what we found was that all of the operators were very intelligent, and did the best they could in an environment that was horrifically designed” (Norman quoted in Long, 2021, para 2). During the inquiry, investigators found that warning lights and klaxons did go off as problems began with the reactor; however, designers noted that a red light could mean fourteen different things and not invariably fatal. Operator panels were also mapped incorrectly, “clustering bits of information in meaningless ways”, giving no sense of how the plant worked, with elevator failure lights next to the alerting panels for reactor leaks (Kuang, 2021, para 3). The pinnacle fault in the system's design was that the indicator light and the operational switch for the release valve of the cooling system were only wired to each other, with the light only relaying the flipping of the switch and not actioning the release of the value. Consequently, the plant was then retrofitted and designed to consider, anticipate, and accommodate human needs and the usership of the system. The design meant leaving no ambiguity in the operating system's navigation and suitable feedback to design conceptual models for users to form mental models of the system.

Conceptual models are formulated as tools for understanding systems, which designers have strategically designed to disseminate into users' heads as mental models guiding their use of things (2011, pp. 35–40). Norman advises that conceptual models help us transform a system's complex reality into understandable, consistent, and intelligible images, from which users form mental models of the system (1983, p.13). The image, Norman continues, is what he calls the “system image”, an overall understanding and presentation of the system to users. If the system image is consistent with the conceptual model, then the users' mental model will also be compatible. Norman explains that for

this uniform configuration to happen, the conceptual model taught to the user must satisfy three characteristics: which are, learnability, functionality, and usability (1983, p. 12). Though, it is evident that unintentional and erroneous interactions with a system often happen as users attempt to compensate in their mental models due to a deficiency in any one of these aforesaid characteristics (Ibid, p.7-14).

During the commercial peak of personal computers in the '80s and just before the start of the Internet age in the '90s,¹⁸ Norman, along with co-author, Stephen Draper, published *User Centered System Design: New Perspectives on Human-Computer Interaction* (1986). The text represented a shift in designing technology using a HCD perspective rather than a techno-driven focus: an umbrella approach that also forefronts User-Centered Design (USD), concentrating on users' interests and needs. Since the publication, Norman coined the term 'user experience' while working at Apple in the '90s to "cover all aspects of the person's experience with the system including industrial design graphics, the interface, the physical interaction and the manual" (Norman quoted in Rutter, 2016, para 2). In the history of design thinking, Stefanie Di Russo sees this as a pivotal point where we see a design methodology manifesting as a philosophical mindset rather than simply a set of tools and frameworks, citing Rouse's text *Design for Success: A Human-Centered Approach to Designing Successful Products and Systems* (1991), that unpacks the ideology of HCD's mindset (Di Russo, 2012, para 16). Correlating with Norman's view, William Rouse observes the philosophy of HCD is to consider "the role of [the] human in complex systems" using three design objectives and orientations: first, the enhancement of human abilities; second, the assistance of overcoming human limitations and third, fostering user acceptance (Ibid, pp 4-5).

Now, in the AI age, HCD is increasingly becoming synonymous with the design of AI technology, placing humans at the centre of system design thinking. However, as previously noted in Chapter Two, the lines between AI (machine/technology) and human intelligence often blurs. The following section explores the shifting viewpoints in AI design, from the goal of creating a super

¹⁸ The Internet age refers to the period when the Internet became widely available to the public, changing its nature and opening global communication. The Internet was first restricted to government use in the 60s, and then due to limited resources used in academia before its commercialisation.

brain indistinguishable from human intelligence towards evolving systems that aims to augment and foster better user experiences.

5.3 AI by Human-Centered Design: Shifting Viewpoints

Terry Winograd (2006), who inspected the divergent characterisations of AI and HCI communities due to their opposing view of how humans and computers should interact, and their conceivment of knowledge and design, noted two distinct philosophic orientations underlying both disciplines— categorised as “rationalistic” and “design” orientations (Winograd & Flores, 1986). Explaining his source, Winograd cites John Markoff’s historical account *What the Dormouse Said: How the Sixties Counterculture Shaped the Personal Computer Industry* (2005),¹⁹ which describes conflicting views between pioneering AI researchers: McCarthy, whose rationalistic idea was to model people as cognitive machines and create artificial general intelligence or simply a “superbrain”; and Douglas Engelbart’s design idea of “augmentation”, philosophically opposing McCarthy’s approach by crucially not replacing the human in the loop and using computing to help augment human needs (Ubiquity staff, 2005, para 6).

Allen Newell and Herbert Simon’s Physical Symbol System Hypothesis is reasoned to be the most explicit expression of a rationalistic view. Theoretically, essential features of thoughts could be captured and expressed in a formal symbolic representation through well-defined algorithmic rules to create intelligent programs. This hypothetical approach has influenced a generation of AI and, to a degree, HCI researchers, as Newell’s theory contributed to establishing HCI as a discipline of cognitive engineering, which remains influential in the HCI community today (Card et al., 1983).

Moving on to the design approach, Winograd describes this as “harder to label”. However, noting that it has an affinity with those who call their approach “phenomenological, constructivist and ecological” (2006, p. 1257). Explaining that a critical difference in the design approach is the role of modelling, realised through an iterative process of prototyping, testing, and refinement, while also

¹⁹ John Markoff’s *What the Dormouse Said* (the title is taken from the lyrics of the Jefferson Airplane song “White Rabbit”, which itself is also a reference to Lewis Carroll’s *Alice’s Adventures in Wonderland*) is a historical account of the important period when the personal computer and the Internet as we know them came into being. In reference to the title’s inspiration Markoff also describes the newfound culture of sex, drugs, rock and roll manifesting at the same time as the computers, sometimes instigating computer development.

acknowledging the limitations of knowing and dealing with the complexities of the real human world (Ibid). In the development of AI, a design approach is seen as an interplay between adaptive mechanisms, applied examples for training cycles and world experience, which “leads over time to [(imitation of)] intelligent behaviours” (Ibid).

The characterisation of AI’s design approach complements the previous discussion on design research, its process, and its inherent nature to tackle real-world complexity (Rittel & Webber, 1973). However, while the rationalistic approach has declined somewhat, its embodiment and influence are found in the advancement of statistical languages, machine learning and neural networks forming the adaptive mechanisms of AI (Auernhammer, 2020; Winograd, 2006). Nevertheless, at this point in AI history, we see an abandonment of GOFAI approaches that were the result of pure science and technology – “push[ing] the computer metaphor on to all reality”, towards interdisciplinary design approaches resonating within HCI and HCD disciplines (Ibid, p. 1258).

5.4 Towards Human-Centered AI

In 2019, Stanford University founded the Institute for Human-Centered AI (HAI/HCAI) to advance and focus AI research for the creation and design of “AI applications that augment human capabilities” (Stanford, ND, para 2). Stanford’s focus on HCAI illustrates a movement in academia (Ramchurn et al., 2021; Shneiderman, 2020a; Stanford, 2020; S. J. H. Yang et al., 2021), industry (Lovejoy & Holbrook, 2017; Wortman Vaughan & Wallach, 2022), and government (European Commission, 2019) for the development of ethical and trustworthy AI applications that are human-centred.

Yang et al. (2021) theorise that HCAI can be interpreted from two perspectives; “AI under human control” (p.2), which sees a reliable, safe and trustful collaboration between human control and AI Automation for empowering human productivity (Shneiderman, 2020a), also known as human-AI partnerships (Ramchurn et al., 2021). The second perspective is “AI on the human condition” (S. J. H. Yang et al., 2021), which refers to the design of AI algorithms that are explainable and interpretable (Kaur et al., 2020; Wortman Vaughan & Wallach, 2022) with continuous adjustments to the algorithms that imbue human and social context, and the augmentation of human intelligence

(Stanford, 2020). These perspectives are firmly within the design approach rather than the rationalistic, directed towards the same goal of human empowerment and control of AI technology using a HCD mindset by tackling different technological and theoretical facets of AI.²⁰

With this in mind, Jan Auernhammer observes that the rationalistic and design approach addresses the question of ethics and the human impact of AI differently. To demonstrate, the author notes that the rationalistic approach focuses on developing aggregated and normative models that result in generalised principles and guidelines for ethical AI (Fjeld et al., 2020; Jobin et al., 2019), which do not represent real-world complexity (Auernhammer, 2020, p. 1317). In contrast, the design approach focuses on examining the messiness and complexity of the human situation through “enlightened trial and error” (Winograd, 2006, p. 1258) by prototyping and researching the emerging ethical dilemmas in interactions between AI systems and humans. Furthermore, Auernhammer argues that a ‘humanistic design approach’ is more suitable to examine the societal impact of AI for several reasons that parallel points made in Chapter Four, *Methodologies* when reviewing what design *is*. For instance, the design approach considers and addresses differences that may occur through diverse cultural and ethical perspectives; secondly, the approach can be focused and context-specific, whereas generalised guidelines fall short of being focused enough to guide in fixed circumstances, such as trust in autonomous driving; and thirdly, design is well-calibrated to problem-solving (2020).

Using methods such as Wizard of Oz prototyping, designers can examine in close detail user-experiences specifying their needs, behaviours and interactions in situ; for example, prototyping machine learning experiences for Explainable AI (XAI) (Browne, 2019).²¹ However, as Auernhammer clearly expresses, designers and design researchers need to consider various aspects of human implication and interaction beyond just paying attention to human and social factors. To this end, he outlines in his paper *Human-centred AI: The role of Human-centred Design Research in the development of AI*, different HCD approaches and the distinct standpoints they address when researching and designing for HCAI. The following section will unpack the HCD approaches

²⁰ Though methods and technical applications stemming from the rationalistic approach, as noted before, may be used and reformulated with a HCD purview.

²¹ The Wizard of Oz technique is a moderated research method in which a user interacts with an interface manned by a human who controls the system responses.

concerning this research regarding legible AI by viewing both the positive and significantly negative consequences of their implementations, thus establishing an argument for a More-Than Human perspective for AI design.

5.5 Human-Centered Design Research in Artificial Intelligence

Auernhammer comprehensibly outlines various HCD approaches in his paper, as each approach provides a distinctive perspective, implication, and value in researching and designing HCAI (2020, pp. 1318–1326).²² Of particular interest to this research is the theoretical thinking and methods of Persuasive Technologies, Interaction Design, and Human-Centered computing, as they are concerned with analysing human behaviour and interaction used to guide the design of HCAI systems. These approaches will be outlined in the following sections, as their underpinning theories and methods are consequential to the design of legible technology.²³

5.5.1 Human-Centered Computing for Human-Centered AI

The approach of Human-Centered Computing (HCC) is the analysis of ‘interspaces’ created by AI systems and how these impact (through augmenting/replacing/constraining aspects of) users’ lifestyles. To detail the elaborate nature of AI interactions, Auernhammer utilises Winograd’s concept of ‘interspace’, where interactions occur in space beyond the physical and two-dimensional, between a person and a machine. This space enables complex interactions between users, digital interfaces, various devices, corporations, and others. This concept can be further explained by the fact that cyberspace is termed a ‘space’, reflecting a profound metaphor of a kind that Lakoff and Johnson would say we “live by” (1980; Winograd, 1997). Thereon, Winograd describes an interspace as the product of designing new digital worlds and interactions (Ibid).

²² The rest are Human-Centered System (HCS), Social Design, Participatory Design, Inclusive Design, and Need-Design Response (NDR) with a detailed outline found in Auernhammer, 2020.

²³ Or currently not legible in practice.

Therefore, HCC, using a HCD mindset and integrating various views from computer science to psychology, is about understanding the interspace and the dynamic context in which human thought, behaviours and interactions occur (Brézillon, 2003). For an HCC framework to aid in designing HCAI, Ford et al. outline that the goal is to create “cognitive orthotics that can amplify and extend our cognitive abilities” (Ford et al., 2015, p. 7). A move away from the age-old notion of “*artificial* intelligence” and AI’s traditional Turing test ambitions of comparing AI to human performance towards “*amplified* intelligence” through the augmentation of human cognition (Ibid). The idea behind technological orthosis, the authors continue, is that technology is beneficial when it fits and how good that fit is, with two categories of “species fit and individual fit” realised through HCD (Ibid). For this reason, Auernhammer observes that intelligence is a consequence of a finely tuned through a designed combination of human-machine-context (Figure 27) (Auernhammer, 2020; Ford et al., 2015; Hoffman et al., 2001)



Figure 28: Illustrates the interrelation of human-AI system-context. Artificial Intelligence exists only within this relationship and not only in the AI system or the interactions. Auernhammer (2020).

5.5.2 Interaction Design for Human-Centered AI

It is widely known that the term Interaction Design was coined by Bill Moggridge and Bill Verplank in the mid-'80s (Moggridge et al., 2007; Verplank, 2009). While both pioneers had somewhat different approaches, the crux of Interaction Design is the design and the understanding of human interaction with machines by examining human behaviour, actions, and cognitive processes within these interactions (Norman, 1988). This concept is equivalent in AI research, whereby human

behaviour is observed during interaction with an AI system, often using the aforementioned Wizard of Oz prototyping.

Another example of this research method is driving simulators for testing autonomous driving interactions and human responses. To illustrate, Fu et al. investigated how drivers form mental models and their perception of the trustworthiness and reliability of the automated emergency braking system, which alerts the driver of approaching hazards and automatically brakes (2019). While these simulators do not entirely elicit real-world behaviours, the research found that if the driver observed that hazards were often missed being detected by the car's system, the driver remained vigilant in critical moments. Whereas, when the system had a perfect performance in detecting hazards, it led to driver complacency with negative responses. Intriguingly, complacency also occurred when the system gave false alarms of hazards – much like the boy who cried wolf – drivers would ignore these false alarms and, with it, their attentiveness to hazards (Ibid). It is possible that this situation transpired due to the drivers developing a mental model of prediction concerning the car's functionality with the anticipation that the alarm will always be false.

Revisiting an earlier mentioned concept, when modelling mental models, Norman outlines three essential properties of a mental model that should be considered when formulating the conceptualisation of the model (1983, p. 12). These are: Belief System, in that a user's mental model reflects their beliefs about the system, which is acquired through observation, inference or instruction; Observability, whereby aspects and states of a system can be observed; and Predictive Power, in which the purpose of a mental model is to enable the user to 'run' the models mentally, to understand and anticipate the behaviour of the system through procedural derivation. Superstitious behaviour also impacts interaction with a system. When superstitious behaviour ensues, it is through the formation of particular beliefs and behaviours, which are performed in a precise sequence of actions by the user in the hope that this interaction with the system will reduce or eliminate difficulty or error. Norman gives the example of the excessive pressing of the clear button on a calculator before the intended calculation, which results from users' prior difficulties or their own reasoned limitations and knowledge about the system stemming from an inconsistent conceptual model. Using Interaction Design, together with a HCD mindset, attempts to limit user error when interacting with an AI system

while identifying potentially harmful interactions to create valuable experiences with AI systems. However, as outlined in this chapter and Chapter Two, users' understanding and perception of AI systems are confused, leading to the formulation of incorrect mental models of AI systems, such as when users believe AI technology is alive and intelligent. Although superstitious behaviour is familiar with Amazon's Echo devices with *Alexa* always listening, although many users are still unaware of how much listening *Alexa* does.

5.5.3 Simplicity by Design

An early commentator on the proliferating presence of computers through “network-ification” (Pierce & DiSalvo, 2017, p. 1388) was Mark Weiser, who described the disposition of “Ubiquitous computing” as “[t]he most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it” (1991, p. 94).

Ubiquity materialises from IoT objects embedded with sensors, wireless receivers and transmitters, transcending the physical world into an “information system” (Chui et al., 2010, para 1) for users and data harvesting. The incorporation of AI for facilitating IoT operations extends and complicates user interactions. As a result, a core axiom of HCD is the *simplification* of interaction as proscribed by Norman, with the objective for technology to be “invisible, out of sight, out of mind” (1998, p. xii) so that “the emphasis is on the human activity the appliance is meant to serve” (Ibid, p. xi). In other words, designs that cut out excessive information.

Simplifying the complexities of computer operations is often realised through instilling ‘fictions’ and conceptual models into users’ interactions while concealing the underlying convoluted operations. Norman highlights that there are no ‘files’ or ‘folders’ inside computers; instead, the material is saved within the computer’s memory system through coded pointers tagging the file’s storage (2011, p. 35). For designers, creating conceptual models of AI presents a new challenge, which to some extent, is avoided, as a user does not typically interact directly with the AI system but rather supplies it with data points through interaction (Finn, 2017). This prevalent abstraction, and the introduction of conceptual fictions, circumvent a user’s recognition of an authentic conceptual model entirely.

The consequence of purposefully designing technology to disappear into the background seamlessly has made users unaware of its presence, with the detrimental effect of users (through no choice) being illiterate and unknowingly subjected to the conditions and ramifications of use. Furthermore, simplification often gives the software and product companies licence not to disclose the reality of technologies' functionality, using the fictional propellants of 'magic' and 'smart' to service human activities.

5.5.4 Persuasive Design for Human-Centered AI

An incarnation of interaction design, specifically persuasive design, can nudge and persuade users while interacting with AI systems to perform in ways that meet the service providers' goals, such as collecting detailed data from unsuspecting users. This design area capitalises on the fact that computers have a significant capacity to present information in various bespoke and interactive ways that can adjust as the situation evolves; people are not always influenced by the information itself but by how it is presented and when. A reoccurring point of this research is that AI technology is often embedded into services, applications and IoT products. So, while it may appear that a user is just interacting with an IoT product, AI technology, developed using HCD research, is obscured from users' knowledge, which in turn benefits from persuasive strategies.

To illustrate: one persuasive strategy is tailoring services and experiences, which is achieved through collecting users' data to train machine learning algorithms to predict users' likes and dislikes. The slightest suggestion of a curated experience will influence users to resubscribe to services or invest in other products of the same supplier (Eyal & Hoover, 2014). Alternatively, another HCD approach in developing AI systems is to provide knowledge about persuasive computers, thus allowing people to recognise and adopt them for their benefit (Auernhammer, 2020). For instance, fitness trackers track, predict and guide users towards better sports performance (Fogg, 1998).

5.5.4.1 Background on Persuasion and the Art of Rhetoric

For context, the study of persuasion has a long and varied history resulting in no single definition. Though often brought into the contemporary conversation as a fundamental point of view is Aristotle's concept of rhetoric and the art of persuasive speaking (Rapp, 2010), who defined

rhetoric as “the faculty of observing in any given case the available means of persuasion” (Aristotle, 1850, p.11). In this definition, William Fortenbaugh highlights that for Aristotle, rhetoric is not about actually affecting persuasion; instead, it is the capacity to consider each case and the possible means of persuasion (Fortenbaugh, 2007, p. 107).

The concept of rhetoric transpires across many domains and in various forms beyond speech. In philosophy, literature presents rhetoric to persuade and justify theoretical concepts. Design is also an argument for persuasion, personified through visual rhetoric found in graphic design and speculative visualisation (Kim & DiSalvo, 2010), and procedural rhetoric employed in game design (Bogost, 2007) and interactive design (Coulton, 2015).

As detailed in Aristotle’s *Art of Rhetoric*, *the composition of rhetoric* is problematic and complex (Fortenbaugh, 2007, p. 107). Though within the extensive study, Aristotle presents four elemental modes for establishing rhetoric: *Logos*, a sense of logic to articulate rational arguments forming syllogisms;²⁴ *Pathos*, appealing to emotion either through creating it or counteracting it, *Ethos*, concerned with credibility and authority; and *Kairos*, which pertains to the opportune moment, context and creating the right atmosphere (Figure 28).




Figure has been removed due to
copyrights restrictions

Figure 29: Modes of Rhetoric according to Aristotle, appropriated from Coulton (2015).

²⁴ On this note mentioned in Chapter Two the composition of syllogisms and logic forms the fundamental basis for algorithms.

Expanding on this concept further, Paul Coulton introduces the notion that procedural rhetoric can be applied to all interactive computer systems by exchanging concealed system logic for rule-based representations (2015, pp. 5–6) (Figure 29). These examples typify Buchanan’s famous observation of design as rhetoric, acknowledging both design’s ability to influence people and that design is not a neutral act (Buchanan, 1985).

Figure has been removed due to copyrights restrictions

Figure 30: Rhetorical mediums Coulton (2015).

5.5.4.2 Persuasive Strategies

Influenced by Aristotle and how persuasive technology could exercise the rhetoric modes, B.J. Fogg introduces the “functional triad” as a framework that illustrates the different persuasive roles computing technology can play (Fogg, 2003, 1998). These roles are: as a tool, in the way of an interactive product designed to change attitudes and behaviours, or both, by making the desired outcome easier to achieve; as media, that shapes attitudes and behaviour through simulated experiences; and as social actors, which persuade through a variety of social cues to elicit social responses from users. Designers can implement particular rhetoric strategies by knowing the role or the combination of roles a computer is playing and how to amplify the capitalisation of data points for rich data-driven analysis for service providers and third parties (Fogg, 2003, pp. 23–29; Singh et al., 2014; Zuboff, 2019). Incidentally, Fogg’s book, *Persuasive Technology*, aims to highlight persuasive strategies’ positive and ethically sound applications (2003). However, Fogg also questions, ‘can persuasion ever be ethical?’ In this question, the author acknowledges that with ease, persuasive strategies can be designed as unethical and insensitive to the values of its’ users and individuals who

are unknowingly using technology and their underlying impacts (2003, pp.211–235)—in short, using the insights obtained through HCD unethically.

To illustrate, tunnelling is a process that leads users through a predetermined sequence of actions that exposes them to information, activities, and opportunities of persuasion that the user would otherwise not have experienced. A positive intention of this method can be found in the process of installing software, with the user going through a linear and, therefore, an easy-to-navigate route through the process. A counter-example can be found during the mundane registration process for apps and websites to unlock content or services. In this tunnelling course, platforms gather user data through their inputs while making offers about premium services or other products. When users take a closer view of these other products on offer out of curiosity, the interaction is tracked and logged without consent and then used in targeted and predictive advertisements. This particular interaction exploits users' relative inexperience. A strategy often used in novel technology to distract users through the unfamiliarity or complexity of the interaction. To this end, Fogg warns that tunnelling strategies can border on deception and coercion, noting that often the scope of personal information collected is intimate. Ethically, users should always be aware of what is happening and how they can exit the tunnel at any time.

Another persuasive strategy in accord with the rhetoric element of Kairos pertains to suggesting behaviour at opportune or premeditated moments. Curating an opportune moment is difficult as timing involves many different contextual elements presented in a physical and social environment. To illustrate the difficulty of creating these moments, Fogg details a speculative example of how location-based persuasion may be facilitated using a McDonald's stuffed bear prototype fitted with GPS technology that could be given away as the free toy children's *Happy Meals* (Fogg, 2003, p. 43). Whenever the bear comes near the vicinity of a McDonald's, the bear sings how delicious McDonald's food is. Raising many ethical concerns relating to child manipulation, children cued by the bears' singing would nag their parent/s to take them to McDonald's. Fogg stipulates that the bear's singing may easily persuade a child's sudden urgency to go to McDonald's; however, the technology cannot cater for the state of the adult's mind (Ibid). Therefore, in many instances, technology and creative ideas may not be enough to compose the right moment. However, despite the

difficulty in persuasive timing, a configuration of this scenario has become a reality with the data collection of location history through mobile devices and the use of this data in targeted location-based advertisements (Associated Press, 2018). Articulating these operations in some way, as this research is concerned with, is a strong argument for design research.

5.6 A Brief Ethical and Closing Note

In response to the growing convergence of persuasive strategies afforded by computer technology in the '90s, Daniel Berdichevsky and Erik Neuenschwander developed a framework to reflect on designers' moral responsibilities and to minimise ethical abuse (1999). The authors' principles centre around 'The Golden Rule of Persuasion', cautioning that creators should not inflict something on a user they themselves would not consent to. Relating to the research focus of AI legibility, Berdichevsky and Neuenschwander outline the 'Disclosure Principle' advising that users should be made aware of persuasive mechanisms, motivations, methods and intended outcomes. The authors also note the 'Dual Privacy Principle' suggesting that collecting and relaying personal information to a third party must be scrutinised for privacy concerns.

Despite the careful deliberation for ethical frameworks, it is well-known that persuasive technology can have unintended consequences beyond reasonably predictable outcomes (Berdichevsky & Neuenschwander, 1999; Fogg, 2003, 1998; Verbeek, 2006). However, 'unintended' is far removed from being unable to identify the motivational strategies employed in AI technology that facilitates and benefits from persuasive mechanisms (Auernhammer, 2020). This situation is often contingent on the addition of machine learning once training evolves beyond human intelligibility, as mentioned previously, thus, resulting in research aimed explicitly at studying AI-infused persuasive strategies (Ndulue et al., 2022; Orji & Moffatt, 2016).

Each of the above HCD research approaches provides insights into their perspectives and developmental contribution towards ethical AI, explicitly examining the interaction and consequential user behaviour within these systems. These approaches have also been critically analysed because of their influence on the current illegibility of AI: explicitly, the design of simplified interactions, resulting in indiscernible user interactions and system operations, that employ fictionalised conceptual

models, imparting core tenets of HCD. The overview, however, should not be taken as a dismissal of HCD and HCAIs' benefits in the advancement of AI. Nonetheless, the summary does present an argument for an alternative point of view in design catering to the agency and vitality of things and our entanglement with them. The next part of this chapter will present the theoretical underpinning for developing a More-Than Human approach to AI.

Part Two More-Than Human-Centred Design

5.7 Shifting perspectives to A More-Than Human-Centred AI

Up until this point, this thesis has been forming a critical discussion for perceiving AI as a thing in itself (or many things in themselves), an argument for *dislodging* (Wakkery, 2021, p. 9) the dominant anthropocentric view of AI and justifying the metamorphosing of a philosophical lens for design. Like many design research endeavours (Akmal, 2021; Coulton & Lindley, 2019; Redström & Wiltse, 2018; Wakkery, 2021), this research uses philosophy and posthumanist thinking to mobilise a More-Than Human-Centred Design (MTHCD) perspective for AI.

In a diverging theory from a human-centred approach, Norman (who notably founded HCD) has also argued against an anthropocentric view in his article titled *Human-Centred Design Considered Harmful* (2005). Norman states that HCD has become a dominant theme in design, practised without thought or criticism leading to a blind commitment and attention to users' needs which, ironically, generates a lack of cohesion and increased complexity in designed artefacts. This condition, Norman observes, induces a misdirection of users' focus, especially when interacting with multiple and dynamic applications that execute overlapping tasks (Ibid, para 25), where users are only permitted to experience a fictionalised and representative interaction. Thus, hindering agency and negotiability in technology, ultimately designed to be human conscientious.

The computer scientist Ben Shneiderman considers that for humans to be truly at the centre of design thinking, design needs to evolve away from traditional methods and perspectives towards an approach for computers designed to be in themselves (2020a) while ensuring human control (2016, 2020b). Developing the concept of "Humans in the Group; AI in-the-loop" (an overturn of human-in-the-loop), Shneiderman looks to Lewis Mumford's thesis on the evolution of new technologies identifying a design pattern he called "The Obstacle of Animism" (2020a). The tendency of emerging technology to use animals and humans to guide design. Analysing many failed technological innovations, Mumford examines Leonardo DaVinci's unsuccessful attempt to create a flying machine by imitating the motion of birds' wings, asserting "[t]he most ineffective kind of machine is the *realistic* mechanical imitation of a man or other animal"; where "for thousands of years animism has stood in the way of ...development" (2010, pp. 32–33).

Dovetailing on the obstacle of animism, Wakkery draws on animal studies scholar Cary Wolfe's posthumanist definition to 'dislodge' the anthropocentric hold on our thinking and to recontextualize our relations with things for design (2021, p.9). Wolfe summarises a posthumanist view as:

It forces us to rethink our taken-for-granted modes of human experience, including the normal perceptual modes and affective states of *Homo sapiens* itself, by recontextualizing them in terms of the entire sensorium of other living beings and their own autopoietic ways of "bringing forth a world"—ways that are, since we ourselves are human animals, part of the evolutionary history and behavioural and psychological repertoire of the human itself. But it also insists that we attend to the specificity of the human—its ways of being in the world, its ways of knowing, observing, and describing—by acknowledging that it is fundamentally a prosthetic creature that has coevolved with various forms of technicity and materiality, forms that are radically "not-human" and yet have nevertheless made the human what it is (2010, p. xxv).

Wolfe specifies that a posthumanist view attempts to remodel humanism, which has been, as the fellow posthumanist philosopher Rosi Braidotti describes, the dominant model, measure, and structuring force of all things evolving from the antiquity period, propagated across the likes of political ideals, legal rights, and scholarship (2013). According to David Roden, a humanist philosophy is *anthropocentric* if it bestows humans a "superlative status". However, noting that human-nonhuman distinctions are not anthropocentric as they can provide beneficial descriptions, i.e. humans are the only animals that drive cars (2015, p. 11). Wolfe prompts a conscious foregrounding of humans and things into view, where both bring forth a world of their own. However, attesting that paradoxically the view will always be human, as we can never entirely escape the bias of human perspective. This point imparts the following discussion and structure of the remaining chapter.

This research aims to present a MTHCD approach by integrating philosophical theories that align humans and non-human things along a flat plane of being. However, the method should contend with and reflect on the human relationship with technology. Ultimately, the exercise here is to design and enable conscious human interaction with AI technology via a more considered approach and the illumination of the things in question. First, this intent will be realised by forming a design framework

for the speculation of non-human beings by employing a particular type of posthumanist phenomenology through the philosophy of Object Orientated Ontology (OOO), commissioning the works of Graham Harman, Ian Bogost and Levi Bryant. As a finishing touch, the theories taken from OOO will be crowned with insights from a materialism standpoint. It is atypical to mix OOO and materialism; nonetheless, this will be justified in the following text.

Thereafter, the research turns to the work of Don Ihde and Peter-Paul Verbeek, emerging from the field of post-phenomenology and the empirical investigations of the relations between human beings and technological artefacts and how technology shapes relations between human beings and the world. Part Three develops a 'Human-AI Kinship' by connecting OOO and post-phenomenology through what Yoni Van Den Eede calls "object empathy", suggesting "there is nothing wrong with humanism, as long as the humanism is inclusive" (Van Den Eede, 2022, p. 241) (cf. also Morton, 2017).

5.8 Posthumanism as presented here: A Speculative Realist Tint

Posthumanism is a highly contested term eliciting diverse meanings, though widely used among philosophers (Harman, 2018; Morton, 2017), theorists, futurists (Carrico, 2013), and designers (Forlano, 2017). For this reason, this will be a succinct overview for clarity on the term's usage in this research. Here, posthumanism is aligned with a Speculative Realist tint that will be expanded through an amalgamated post-phenomenology infusing Harman's and Ihde's theories together (Van Den Eede, 2022).

According to Roden, all forms of posthumanism criticise human-centred traditions of understanding life and reality; however, in distinct traditions. In his book *Posthuman Life*, Roden outlines the different flavours of posthumanism to isolate his credence of Speculative Posthumanism, which opposes human-centric thinking about distant future implications of modern technology (2015). The theory, at a glance, is not too dissimilar from the transhumanists model, in which followers believe the most critical application of the so-called "Nanotechnology, Biotechnology, Information Technology and Cognitive Science suite of technologies" will be to modify— the human – for unprecedented control over their nature (Geraci, 2010), for the ultimate goal of transcending to an

immortal existence (Roden, 2015, p. 13). For speculative posthumanists, special consideration is given to the *posthuman* (no longer humans) in the event of Vernor Vinge (1993) and Kurtzweil's (2013) prediction of the Singularity, what Roden calls the "wide descendants" of current humans (Roden, 2015, p. 22).

Conversely, Critical Posthumanists argue that western humanism is based on a dualist conception (Descartes) whose nature is transparent to itself and, as Veronika Hollinger puts it, "a scenario in which the human(ist) subject remains unmarked by its interactions with the object-world" (Hollinger, 2009, p. 273; Roden, 2015, pp. 23-31).²⁵ The Critical Posthumanist Braidotti (2013, p. 66) and fellow philosopher Claire Colebrook (2012) argue that liberal politics oriented towards the rights of humans are incapable of addressing issues such as climate change and the ecological depletion of the Anthropocene. Descartes' dualism is, however, eroding with the de-stabilization of 'the human' in our technoscientific era, championed by the likes of Haraway (1992, 1995, 2003, 2016). Thus, establishing a connection between biotechnical sciences and human and social sciences for levelling up non-human actors in a geopolitical and eco-philosophical manner (Braidotti, 2006).

5.8.1 Speculative Realism

An alternative posthumanist argument (granted does have overlapping points with the latter) is Speculative Realism, which opposes the philosophical privileging of the human-world relationship, human subjectivity, and conceptual thinking found in Kantian (Kant, 1996, 1998) and post-Kantian transcendental philosophy (Hegel, 1977). Speculative Realists unanimously reject the philosophy of idealism found in Immanuel Kant's *Critique of Pure Reason* (1998). To reflect: Kant formed a modernised version of metaphysics by critically observing Plato's metaphysics. A simplified definition of metaphysics, or ontology, is the science of that which lies beyond physical things (Tampio, 2015, p.1).²⁶ In transforming this, Kant defines metaphysics as the conceptual scheme that makes possible (human) experience concerning arguments around consciousness, morality, space, and

²⁵ To many Rene Descartes' is considered to be the arch-humanist with the philosophy of Cartesian dualism in which a self-transparent subject (an idea of the mind, therefore contains no doubt) represents a mind-independent nature.

²⁶ In avoiding "historical and mystical baggage", Harman uses metaphysics and ontology as synonyms to avoid repetition and uses ontology to simply mean 'the study of being' (Harman, 2018, pp. 13–14)

time (1998); concepts which go beyond the scope of the research at this time, though substantiates the emphasis Kant placed on the human subject. The political researcher Nicholas Tampio succinctly summarises Kant's metaphysics for us as– “not describing features of nature in itself; rather, such a metaphysics describes and justifies a system of categories that make possible *human* investigations of nature” (2015, p. 5 (italicized for emphasis)).

The ideology of Speculative Realism moves away from this subjectivity and dislodges anthropocentrism and human exceptionalism towards the cosmic throngs of unhuman things (Roden, 2015, p. 31). A signature text for Speculative Realism is *After Finitude: An Essay on the Necessity of Contingency*, authored by Quentin Meillassoux, who refers to any philosophy that upholds the idea that “we only ever have access to the correlation between thinking and being, and never to either term considered apart from the other” as *correlationism* (2011, p. 5). Correlationism, that is, is the view that things can only exist in relation to human's thinking and never in isolation, where subjectivity and objectivity are already intertwined in human cognition. A thing can never be understood separately or in isolation (Zahavi, 2016, p. 293). Both Harman and Meillassoux have argued that reality must be thought of as independent from human subjectivity, a departure from Kant's correlationist circle, as summed up by Harman:

Inspired ultimately by Immanuel Kant, correlationists are devoted to the human-world correlate as the sole topic of philosophy, and this has become the unspoken central dogma of all continental and much analytic philosophy. Speculative realist thinkers oppose this credo (though not always for the same reasons) and defend a realist stance toward the world. But instead of endorsing a commonsensical, middle-aged realism of boring hands and billiard balls existing outside the mind, speculative realist philosophies are perplexed by the strangeness of the real: a strangeness undetectable by the instruments of common sense (2011, pp. vii-iii).

Yet, Kant warned us “never to venture with speculative reason beyond the boundaries of experience” (1998, p. xxiv). The philosopher Steven Shaviro contests, Speculative Realism urges us to do precisely that – “[p]lace Kant, we must think outside of our own thought; and we must positively conceive the existence of things outside our own conceptions of them” (2011, para 6).

Since the inception of Speculative Realism in 2007 at a Goldsmiths Workshop by distinguished philosophers Harman, Meillassoux, Ray Brassier and Iain Hamilton, their own philosophical interpretations have somewhat furcated and broken apart into various splinter groups bearing little resemblance to one another. Conceived from Harman's expression of Speculative Realism is OOO, a radicalization of phenomenology, which, again, is a theory that offers various formulations by philosophers. Though what unites these philosophers is the pursuit to reverse Kant's human-world duopoly and the anthropocentric bias of 'classic' phenomenology through the acknowledgement of a variety of different phenomenologies, where humans are not the measure of all things (Harman, 2018, p. 45). Theorems for the "pluralisation of perspectives", rather than the eradication of human perspectives, whilst recognising that a rock, the wind, a law, or a computer program also have phenomenologies, or ways of "apprehending the world" (Bryant, 2012, para 2). The following chapter will present conceptualisations emanating from OOO to formulate a philosophical model for a MTHCD approach to design for AI futures. To demonstrate: Harman advocates using Latour's actor-network theory that sees all things as actors no different from us, creating a status of consensus; as all actors (human and non-human) try to form links with other actors to become stronger (2018, pp. 57–58). Bryant contends in his book, *The Democracy of Objects*, that every object is an observer with a particular point of view on the world and purpose (2011). In *Alien Phenomenology*, Bogost proposes a type of phenomenology speculating how nonhuman things encounter the world (2012).

To appreciate different phenomenologies and their approaches, the following section will briefly overview the archetypal understanding of phenomenology through Edmund Husserl's (1859-1938) and Martin Heidegger's (1889-1976) conceptualizations. Both their work forms the foundational concept of Harman's OOO via a juxtaposed and synthesised scheme by the philosopher, in which 'hidden' relations, qualities, and causations of objects can be phenomenologically inquired through speculation.

5.9 Phenomenology: A Short Historical and Theoretical Synopsis

The field of phenomenology is vast; therefore, this text will not cover every aspect of the topic but rather a comprehension that contributes to the discussion here. The simplest definition of phenomenology is the study of phenomena. That is, the everyday *experience* of objects can serve as the point of phenomenological investigation. A phenomenologist is concerned not with the nature of the external world as a metaphysician would be but with our mode of access to it (Smith, 2016, p. 3). In other words, the experience of how it appears to us (Zahavi, 2019, p. 1). From different standpoints, Husserl, and Heidegger, were immersed in what the act of observing these things meant for us: the accessibility of rather than their hidden depths.²⁷ Inversely, when Rahimi described AI as “alien technology”, he was referring to the evolving and unknown nature of AI’s being (Hutson, 2018, para 2). The field of phenomenology, or a version of it through OOO, suggests seeing AI’s being as a phenomenon that can be understood through its experiences and how the world/things appear to it – speculatively that is.

5.9.1 A Subjective Appearance and Reality of Thing

Two figureheads of the phenomenological movement were Husserl and his considered successor Heidegger, who eagerly contradicted expectations of carrying Husserl’s theoretic torch by taking phenomenology in a new direction. Husserl’s method was to bracket all consideration of the outside world and concentrate on the phenomena that appeared to consciousness. On the other hand, Heidegger draws our attention to what lies behind all phenomena to provide a sense of reality beyond science’s knowledge (Harman, 2011b, p. 36).

Husserl’s views are the touchstone account presented as a form of Transcendental Idealism heeding Kant’s rhetoric (Zahavi, 2022). On this note, Kant introduced the notion of phenomena – things as they appear – and noumena –things as they are in themselves – which we never experience directly since we remain in the conditions of the human experience (Kant, 1998). Phenomenologists reject the Kantian noumena, studying only phenomena, everything humans can encounter, perceive,

²⁷ Although arguably, Heidegger did acknowledge that things do have withdrawn qualities, which will be explained in the coming chapter.

think about or use. Therefore, the elementary principle of phenomenology is to describe what *appears* to us rather than “speculate on hidden causal mechanisms” (Harman, 2018, p. 152).

For Husserl, the study of phenomena was to go “back to the things themselves!” (2001, p. 168). Paradoxically, this considers things only insofar as they appear via ‘phenomenological reduction’. Describing what is given and the different types of givenness – “purely as it is given” – attending to our experience of the thing *rather* than their qualities (Smith, 2016, p. 11). Reduction, an underpinning principle in phenomenology, diverged away from the empiricists’ view, promoted by the likes of David Hume (1711-1776), who saw little evidence for objects as unified things—instead perceiving an object as just a bundle of qualities. For Hume, there is no such thing as an ‘apple’ but only the palpable qualities of red, juicy, sweet, and hard (1878). A bundle of qualities which appears together as an apple, where the ‘object’, according to Hume, adds nothing to our perception (Harman, 2018, p. 76). However, Husserl completely inverted this empiricist way of looking at things by going back to the object itself, rather than its qualities, due to the shifting nature and discrepancies that can impact said qualities (such as light level, varying distance and angle to object) from one instance of perception to another (Ibid). In this regard, the viewer accepts the apple because of the many factors that appear when experiencing it (light, shape, colour, texture). Collectively these given qualities construct the phenomenon of experiencing the apple for what it is. In his book *The Quadruple Object*, Harman exposes Husserl’s tension between objects and their qualities by accentuating and categorising ‘Sensual-Objects’, which are the correlates of our own experience of the apple and that of the apple’s ‘Sensual-Qualities’ (Harman, 2011b, p. 26).

Another perception to briefly touch upon is Husserl’s concept of ‘intentionality’ as it is fundamental to the phenomenological enterprise (Harman, 2011b, pp. 21–22).²⁸ Dan Zahavi explains that we are always conscious of a thing in a particular way from the presentation, description, or perspective of (2019, p. 17). A thing can be intended in different ways through a ‘type of intentional experience’, as one is always conscious of a thing in a particular way (Ibid). For instance, a smartphone is – intended by a user – as a means of communication, a present received from a partner,

²⁸ It is important to note the concept here is summarised to an elementary understanding, highlighting the experiential subjectivity of things.

or a source of irritation (as the model can no longer update to the latest operating system). Subsequently, the observer's subjectivity is the prevailing account of the thing experienced for a Husserlian phenomenologist both in terms of phenomenological reduction and intentionality—therefore equating to *no* autonomous reality of the thing in question with no inherent causation (Harman, 2011b, p. 22).

5.9.2 Heidegger's Phenomenology

Heidegger rejects Husserl's phenomenology, which uses descriptive methods via the 'reduction for a neutral view' of things around us, for a 'hermeneutic phenomenology' and the "meaning of being" (Heidegger, 1996).²⁹ The meaning of 'being' here is the "intelligibility of entities" to make sense of (the being of) things, drawing our attention to what lies behind the phenomena (Smith, 2016, p. 26).

For Heidegger, phenomena reveal themselves as being through our practical action with them (1996, p. 52), executing a global dualism and constant reversal of being. That is to say, entities withdrawing into a "silent underground", a reality in which things operate without our noticing them (the hammer doing the job it is meant to), and entities likewise exposing themselves to the presence by becoming explicitly noticeable for any reason (when a hammer is broken) (Harman, 2011b, p. 39, 2018, pp. 152–153). In summary: the phenomenological description of the 'world', according to Heidegger, is to address that which is 'present-at-hand' or 'readiness-to-hand', the notion that the world is made up of things waiting for human use. For Heidegger, one's understanding of the being of AI is manifested through utilising it, apprehending fundamental relation to the world as practical rather than cognitive (Smith, ND, para 25).

The tool-analysis forms a core foundational tenant in OOO by Harman's De-anthropocentrised reading of Heidegger, marked by objects enacting their own reality as they withdraw into a subterranean background when we are silently relying on tools (ready-to-hand) and the broken hammer as a 'disruptive phenomenon', with previously withdrawn qualities now

²⁹ Phenomenological investigations are not neutral for Heidegger as they rest on prior implicit conceptions of being. If we did not have this "pre-ontological" understanding of being then the world would remain hidden from us (Heidegger, 1996, pp. 10–11).

foregrounded and present-at-hand. On this note, Heidegger is usually perceived as a thinker who reduces reality to its accessibility to human Dasein (human experience / being-there in the moment). Although Harman considers Heidegger a realist metaphysician, focused on ambiguous states of specific instants for detailed analysis of phenomena. Here, Harman presents another tension and coupling between what he calls *Real-Objects* (objects in their own right) with sensual qualities that are translated into acts of sensual presences to the user (Harman, 2011b, p. 50).

Nevertheless, under these circumstances, phenomenology only tells us something about the apparent nature of things as they appear or reveal themselves to us through the experience of observing or our activity with them. To consider what a thing *really* is involves going beyond the phenomenal and subjective apparentness to consider the noumena of things speculatively. However, phenomenologists ostracise what is called the “two-world doctrine”: the approach of creating a distinction between the world that presents to and can be understood by us and the world as it is in itself (Zahavi, 2019, p. 14). This latter point is precisely the aim of this research, via speculation to transcend and to design more conversantly on AI’s being while simultaneously contending with how they appear or reveal themselves to us through post-phenomenology.

5.10 Beyond Human Experience

Husserl’s phenomenology precluded objects outside of human experience. In contrast, Heidegger’s phenomenology considers the reality of things being withdrawn from human access, only revealing themselves in a specific way to us. These phenomenological propositions correlate with Chapter Two’s themes of human experience, related interactions, and AI’s confused anthropomorphised ontology— symptoms of western perspectives of the world and, thereupon, the design of things.

As documented, AI-infused systems are designed to be outwardly a binary process, whereby much of the operation of AI is happening beyond human interaction and experience. For instance, a user may be unaware of the different algorithmic functions and data operations occurring when clicking the button promoted by Netflix to generate a personalised catalogue of films to view. As previously cited, this state of affairs can be the consequence of designing for: simplicity, where the

experience of finding something to watch is made more accessible, established from users being characterised and preferences monitored from viewing data and collected in ‘likes’; for persuasion (both moral and immoral intentions) to continue engagement with the platform, with some users believing the application is magic for knowing their preferences.

Additionally, inexplicable film suggestions can be the ramification of implementing AI technology into these services despite the unknown particularities of AI, even by experts. These circumstances, not to mention the illegibility of the user’s interaction with AI technology, can all reasonably be the consequence of not speculating on the deeper existence of AI and its operational practice through OOO.

5.11 Object-Orientated Ontology

According to Harman, the concept of idealism is precarious by not accounting for reality since it is always radically different from our formulation (2018, pp. 3–17). The external world exists independently from our awareness, whereby Harman and Bogost ask us to approach reality indirectly, appreciating that things withdraw from our direct access (Bogost, 2012; Harman, 2018).³⁰ Objection to the thesis of OOO usually ensues; however, it highlights a current popular modus operandi with two routes: statements of truth and poetic gestures on the other. Instead, OOO works in the cognition of metaphor, speculation, and philosophy, to name a few, rather than knowledge (Harman, 2018). A maverick move when currently knowledge is the cure to every ailment and problem, though OOO attempts to speculate and detect the gap between knowledge and reality (Bogost, 2012; Harman, 2018). Here lies the purpose of forming a OOO lens for design, as it provides the freedom to speculate on the deeper existence of things while tracing our current understanding.

5.11.1 To be Object-Orientated

The etymology of OOO ‘borrows’ object-oriented from computer science; however, not directly motivated by the field, it does shine a light on the meaning of ‘object-oriented’ in the context of OOO (Ibid, pp. 11-12). Older computer languages function with all their parts integrated into a

³⁰ Approaching reality indirectly is achieved through metaphor and speculation, which will be discussed later in this chapter.

program of a unified whole. Object-oriented programs use independent programming objects that interact with other objects, with their internal programming information hidden or withdrawn. This innovation means that computer programs are no longer written from scratch or exist as entire programs, whereby individual programming objects can come together in various programmed combinations for new uses. Harman draws on the synonymous idea in OOO and Object-oriented programming that objects never make complete contact any more than they do with the human mind. On this note, the OOO fulcrum is that all things are equal in existence, wherein humans do not fill up “fifty percent of ontology” with objects considered autonomous things (Harman, 2018, p. 56).

Borrowing the term ‘flat ontology’ from Manuel Delanda (2002), Bryant grants that all objects are equal beings, rejecting that different types of objects require different ontologies (Bryant, 2011, pp. 112-114). OOO is a post-humanist realist ontology, though as Bryant stresses, this does not mean it is an *anti-human* ontology, but rather “an ontology where humans are no longer monarchs of being but are instead *among* beings, *entangled* in beings, and *implicated* in other beings” (Ibid, p.40, original emphasis). This ontological positioning is uncustomary in modern philosophy since Hume and Kant, whose ideas embody the correlationist conceit – whereby we cannot think or speak of the world without humans or humans without a world (Bogost, 2012, pp. 14–15; Harman, 2018, p. 56). Alternatively, Harman refers to it as a “shoddy dualism ...[of an] implausible taxonomy between human thought on one side and *everything else in the universe* on the other” (Harman, 2018, p. 56 original emphasis). Bogost illustrates the human-world correlation using Turing’s famous question, “Can machines think?” As previously noted in the literature review, from this point in history, science assumes the nature of the computer is related to the nature of human experience. However, as Bogost stresses that “like everything, the computer possesses its own unique existence worthy of reflection and awe, and it’s indeed capable of more than the purposes for which we animate it” (Bogost, 2012, p. 16). By promoting a flat ontology, Bryant suggests this can synthesize the human and the nonhuman into a common collective and deeper examination of objects (2011, pp. 26–33). In a flat ontology, the

laboured data buffering at an edge gateway holds just as much interest as the IoT devices the data emanates from.³¹

5.11.2 The meaning of Object

Harman accentuates that “[t]o be an object means to be itself, to enact the reality in the cosmos of which that object alone is capable” (2011b, p. 74).³² In much the same way, Bogost employs the term ‘unit’ and uses a black hole as a simile for a unit’s intricacy of being, which can be perceived on the one side of being as an “unfathomable density” (Bogost, 2012, p. 22). However, on the other side exists a withdrawn content of individual units that compose it – “an entire universe of stuff” (Ibid). Therefore, a unit or an object is a peculiar umbrella term for everything, not just physical or ‘real’ things.

Yet, Harman describes the theory and history of objects as *naively* understood in philosophy and science (2011b, p. 7). Dissatisfied, Harman considers each object as unified things which withdraw or reveal their features to us (p.10). The ‘darkness of objects’ has been historically overlooked through acts of ‘overmining’, ‘undermining’ and duoming, the basic forms of human knowledge.

Overmining theories are used for objects considered “too deep” to perceive, subsequently reduced to their impact on us or each other, denying anything beyond such impact (Harman, 2018, p. 49). An act played out in phenomenology through the idealistic view of objects, denying the existence of an external world, typified by correlationism and the human-world relation. Undermining involves breaking objects down into their constituent parts. This custom occurs when objects are “too shallow to be the truth” and are not measured as having the same reality as their comprising objects of atoms or quarks, reasoned to have detailed realities within them (Harman, 2018, p. 46).³³ In this view, a chair receives all its properties from components: screws, wood, paint, etc. Reasonable enough, as

³¹ Riffing off of Bogost’s illustration of a flat ontology (see, Bogost, 2012, p.17)

³² Often, Harman also employs the synonyms of object, thing, and entities periodically, which is also imitated in this thesis.

³³ Related in the concept of ‘Smallism’, coined by the philosopher Sam Coleman, is the idea that real elements in any situation are the tiniest component that can be broken down to (2009).

without these components, the larger object would not exist. However, this argument misses the point of emergence for Harman, in which new properties appear when smaller objects are joined.

Harman's strongest criticism is the act of duoming, a combination of overmining and undermining, that customarily happens in materialism. The act of reducing down to the ultimate components while also treating them as bundles of qualities, thus making it rare to find the reduction types in isolation (2011b, p. 13). Harman clarifies that we can never entirely be sure which objects exist, and the 'considerations' of objects are merely the figments of the permeated illusions of undermining, overmining or duoming practices that have persisted since Pre-Socrates (470 BC-399BC). In this regard, objects' true relative independence is dismissed (Harman, 2016, p. 9, 2018).

According to Harman, the core axioms of materialism emphasise that everything constantly changes, has fluid boundaries, is contingent, intra-acts, and is multiple rather than singular and immanent.³⁴ In opposition to materialism, he offers the principles of *immaterialism* that priorities objects change intermittently with stability the norm, possessing definite boundaries, are limited in contingency, are autonomous and have withdrawn essences and realities that can interact (2016, p. 9 - 16). Harman and Timothy Morton remind us that we should not imagine objects as singular 'small' things. Objects come together to form new 'bigger' objects, or in Harman's terminology, "composite objects" (Harman, 2016). Probably most things in the world are composite objects.

The case study Harman uses to illustrate OOO immaterialism is a historical account of the Dutch East India Company (1602 – 1790), a trading, exploration, and colonization company detailed by the author in its various stages of its existence over time. To summarise, the company had no set place of existence in the same way as an atom, or a quark does. Nor was it just one thing, but a form comprised of ships, shareholders, sailors, different operation sites etc. This example also illustrates the notion of unified objects with *emergence* (Ibid). Going further, the notion of the hyperobject, as

³⁴ Karan Barad's concept of intra-action deserves a greater overview than what this thesis can offer. However intra-action replaces 'interaction', which entails pre-established bodies participating in action with each other. Instead, Barad's idea refers to the "mutual constitution of entangled agencies" whereby distinct agencies emerge through inter-action rather than precede it (Barad, 2007, p. 33). According to Barad "Things" don't preexist; they are agentially enacted and become determinately bounded and propertied within phenomena" – for Barad, the basic units of reality (Ibid, p. 150). Haman opposes the theory of new materialism, who considers this as an overmining and unattainable theory. He notes that it is an account of not being able to explain change instead posing relations generating their terms out of nothing (2018, p. 53 & 153).

forwarded by Morton (2013) based on Harman's work: is a thing that is everywhere and nowhere at the same time and that, by definition, escapes the empirical gaze, such as— climate change.

Harman details that fictions are also objects; Sherlock Holmes, his sidekick Dr Watson, and the fictional flat of 221B Baker Street all have their own realities in OOO and deserve equal attention as they are very much part of the human experience and animal life more generally (Harman, 2018, pp. 33-34); as do abstract or conceptual structures such as jealousy, pandemics and political advocacy (Bogost, 2006, p. 5). In this respect, OOO is not discriminatory of what conventionally may be considered an object, i.e., algorithms, pencils, Groggu, space, racial tension, and the Titanic, are all considered objects which must be accounted for by ontology rather than reduced to nullities (Harman, 2011b, p. 5).

5.11.3 Object Ontology: Levels of Objects

Objects in OOO are redefined in terms of a relation-substance structure (Van Den Eede, 2022). Harman presents a quadruple structure that combines the key insights of Husserl and Heidegger's phenomenological investigations for enquiring about objects' ontology via rifts and tensions within them. The four-fold model presents four tensions: *Real-Objects*, *Real-Qualities*, *Sensual-Objects* and *Sensual-Qualities* (Figure 30).




Figure has been removed due to
copyrights restrictions

Figure 31: Since objects cannot exist without qualities and vice versa, there are only four possible combinations, indicated by the four lines between the circles above. Appropriated from Harman (2018).

To explain: on the one hand, the model presents the phenomenal realm for us, displaying a tension between intentional objects and their shifting qualities based on Husserl's insights. Conversely, it also accounts for the 'subterrain' level of things, with Harman defining *Real-Objects* (RO) as the 'side' of objects that withdraw from our experience (Heidegger's hammer), where through mental exercises for instance speculation, we can approach the *Real-Qualities* (RQ), that a phenomenon needs in order to be itself or in other words an object's *essence* (what makes an apple an 'apple'). *Sensual-Objects* (SO) and *Sensual-Qualities* (SQ) exist in relation to that of a real object, as a correlate in our minds (Husserl's phenomenology and how we experience the apple).³⁵ It is helpful to understand these permutations of an object's ontology as levels and caricatures we perceive of its reality which reveals and withdraws from the world. Or, as Harman describes as an object's essence – "the essence of that thing" the tension between RO-RQ and an object's *eidos* –the tension between the correlate of our minds/experience and the object's qualities that exist whether we are aware of them or not— SO-SQ (2018, p. 9, p. 159).

For Harman, ontography is the consideration of the various possible combination of these poles "map[ping] the basic landmarks and fault lines in the universe of objects" (Harman, 2011b, p. 125). Essentially creating miniature worlds or assemblages full of relationships, perspectives, and possibilities an object may or may not experience.

Bogost takes a slightly different approach to ontology, included here, as his theses provides an attentive way to disseminate and explore ontography in design practice. In contrast, Harman's philosophy enables us to appreciate the complexity of objects. As Bogost confesses, his methods, too, are laced with correlationism but do "sow a promising seed" for a philosophical model for the investigation of things (2012, p. 37). In his book *Alien Phenomenology*, Bogost details the informaticist Tobias Kuhn, who developed an ontograph framework that depicts and evaluates controlled natural languages (CNLs), grammatically and semantically simplified language used in

³⁵ The four-fold model by Harman has been described here at an elementary level, sufficient enough to demonstrate a deeper ontology and existence of objects beyond the usual comprehension. For a detailed understanding of this model see Harman 2011b, 2018.

situations like technical documentation, where reduced ambiguity is best practised (Ibid).³⁶ The legibility of the CNL's textual language statements is assessed by users comparing with graphical notations Kuhn calls 'ontographs'. Each ontograph depicts the CNL subject within a mini world, noting different things and their relations (Kuhn, 2009) (Figure 31).



Figure 32: Kuhn's ontograph framework is a graphical notation for representing types of relations in controlled natural languages where simplification is required such as technical documentation (Bogost, 2012).

Exploded views likewise have the spirit of drawing fault lines between objects held together as a 'system', so to speak, in which such systems can be understood as units themselves (Bogost, 2006, p. 5). Bogost explains that these views draw our attention to an object's "configurative nature", a world usually unseen, recording the presence of unit operations (Bogost, 2012, pp. 50–52).

5.11.4 Unit Operations

A simple explanation of 'unit operations' is that "*units operate*"—reacting and acting, meshing with one another *configuratively*; thus, "worthy of philosophical consideration" (Bogost, 2012, p. 27-28). Bryant develops his ontology using 'machines' to account for "any entity, material or immaterial, corporeal or incorporeal, that exists" (2014, p. 15). Despite the historical and problematic connotations associated with the term machine, for Bryant, it elicits the sense in which entities

³⁶ Aviation English is another tangible example, which is the de facto language of civil aviation used for aeronautical radiotelephony communication to plan and maintain global flight paths.

operate, function and their divergent roles. Bogost maintains that objects have their own sense-making – tracing the reality of one another –via the process of engaging with their worlds (Ibid). This principle touches upon Harman’s ‘darkness of objects’ and the different tensions that exist within objects. Incidentally, unit operations that become relevant to another unit differ. Therefore, to perform philosophical work on unit operations is the practice of speculation, which will be discussed in Chapter Six: *Design Fiction: Adapting Philosophy for Design* (Bogost, 2012, p. 30). An additional note to consider at this point, and moves the discussion on to relations, is that something is always something else, an expansion of infinite possible arrangements: a relation in another assembly, a part in another system, a gear to another mechanism (Ibid, p.26). Beings can expand. A phenomenon in AI is function creep, where programs deviate from original programming and considered purpose, and similarly, when data is used for another purpose (Koop, 2021).

5.11.5 Vicarious Causation & Relations

Brassier critically probed at the Goldsmiths workshop that “the really significant challenge is explaining [(object’s)] their relations” (Brassier et al., 2007). In response, as detailed previously, Harman separated objects into two categories, real and sensual and as the designer Simon Weir describes, whose difference between the two is determined by their interactions and relations (2020, p. 148). Harman explains that Real-Objects are autonomous from relations with other objects and their own qualities or properties. In light of this, Real-Objects are incapable of touching one another. However, they “touch without touching” through indirect contact known as ‘vicarious causation’; the rift between the real and sensual, in that Real-Objects never touch and never exert causal forces on each other directly, but rather only come into contact via Sensual-Objects (Harman, 2018, p. 150; Weir, 2020). As Harman illustrates, a “rock strikes the sensual version of another, in such a way that there are retroactive effects on the real” (2018, p. 163). Every object is an island – with “[t]heir reality consisting solely in being what they are” (Harman, 2011b, p. 73), whose essential characteristics are independent of the interactions in which they are involved. Thus, understanding cause and effect relations in Harman’s interpretation of OOO takes a *sensual* form involving indirect, partial, distorted, translated, and representational relationships between objects (Harman, 2011b, p. 120). Likewise,

Morton emphasises that causality is “a matter of how entities manifest themselves for other entities[...]Nuclear radiation-for the flower turns its leaves a strange shade of red” (2013, p. 39) and as C.J. Davies notes, “though it might cause cancer in a human being” (2019, p. 101).

The theory of Causation is a colloquial term for the occasionalist tradition that originates from Islamic and early European rhetoric and thinking that God was the source and mediator of all causation: permitting two objects to interact with each other and evoking the hidden causality in objects. “Fire might appear to burn cotton, but in reality only God burns it” (Harman, 2009b, 2018, p. 164). Pivoting back to Harman’s four-fold model, the reality of objects is never fully deployed in their relations. When fire burns cotton, it does so by making contact with the flammability of the cotton, not its smell or softness (properties accessible to humans and others). Fire does destroy and change these properties that lie outside of its grasp; however, it does so indirectly. “The being of the cotton withdraws from the flames, even if it is consumed and destroyed” (Harman, 2011b, p. 44). Objects cannot exhaust the reality of other objects when their natures collide, and not all properties are relevant to the interaction (Davies, 2019). However, the fact remains that their *natures* do, in actuality, collide, and fire does burn cotton.

The critical point hinges on object’s tendency to withdraw. For instance, Bryant writes:

...entities or substances withdraw from one another insofar as no entity encounters another entity in terms of how that entity itself is, but rather every entity reworks ‘data’ issuing from other entities in terms of the prehending substance’s own unique organization (2011, p. 136).

Albeit, Davies observes that it may be more appropriate to say that “objects are excessive: their reality is in excess of their qualities and relationships. They might always surprise us” (Davies, 2019, p. 100). On this basis, as Harman and Morton assert, some data about each object is left out of the interaction, with only a partial version of each object interacting with the other. The very being of an object’s reality is not accounted for through an interaction with another.

This final statement inspires the following section concerning the nature and the process of data or information transferring from one object to another, or the speculation of ‘object agency’, later defined in this thesis as *Vibrant Objects* for reasons that will become apparent. Harman would conceivably disagree with the customisation of his theory of causation and the hidden causality in

objects with that of agency, as he does not consider the ultimate role of objects is *doing* (Harman, 2018, p. 241).³⁷ However, both Bryant and Bogost would disagree with Harman’s position.

Underscoring the argument for object agency will proceed by introducing Weir’s argument for living and non-living occasionalism as he describes it as ‘mediating agency of non-living’ things— an interpretation of Harman’s reading on occasionalism where contact is made between Real-Objects through virtual particles—a theory situated in Quantum Physics.

5.11.6 Quantum Causation: Virtual Particles Mediating Agency of AI

Harman questions:

How does a real object [...] make some sort of contact, however oblique, with another real one? Only the answer to this question will give us a clear understanding of the manner in which influence is a *pure gift* from elsewhere, without recompense (Ibid, p.98, italicised for emphasis).

Attempting to answer this question, Simon Weir proposes that the “pure gift of influence” is virtual particles located in the empty space, which contemporary physicists call the quantum vacuum (2020, p. 150). The point is that object’s interactions occur on an atomic level with particles rearranging themselves as needed. Theories of quantum mechanics are reasonably, compared to other disciplines, in its early stages of exploration, and as Weir discusses, have its fair share of correlationist views (Ibid pp. 152–155).³⁸ However, when space is discussed in quantum physics, it does not infer to the habitual understanding of ‘empty’; instead, it is theoretically accepted as containing virtual particles, which in turn draw their energy from the vacuum of empty space, thus able to realise their roles as carriers of forces in non-living causal interactions.

Weir continues to explain that Real-Objects can leverage virtual particles to enact forces, and in reverse, the same virtual particles act as Sensual-Objects for Real-Objects in interactions. Weir articulates:

³⁷ Possibly making OOO an ironic choice for design research, with the exception of Bryant and Bogost.

³⁸ Weir provides an in-depth description of both quantum mechanics and his synthesis of quantum theories with Harman’s OOO. The overview presented here only details the highlights of the arguments and for a deeper understanding refer to his paper (Weir, 2020).

In their favour, virtual particles are available locally in all locations at all times to enact causation between nonliving real objects, and they can never be accessed by anything other than the real objects they act upon (and by other virtual particles) (Weir, 2020, p. 157).

Integrating Harman's indirect and distorted relations, Morton's causation of how entities manifest for one another, Bryant's data exchange objects, and Weir's Quantum Causations provide a speculative concept for object agency and objects having the anticipation of other objects. In this sense, virtual particles carry information, a concept Akmal illustrates through Harman's favoured fire and cotton example, explaining that a notion of non-human perception is happening through the interactions of virtual particles "suggests why cotton *understands* it must burn" (Akmal, 2021, p. 150) and similarly the agency of fire exerting its force through virtual particles.

Moving on into the digital realm, the philosopher Yuk Hui explores the existence of digital objects as phenomenological objects. Hui limits his scope of digital objects to their simplest form – data, explaining that "data objects [are] formalized by metadata and metadata schemes, which could be roughly understood as ontologies" (2016, p. 26). In this sense, we come back to Harman's account of emergence and an "endless regress of objects wrapped in objects [...] they are the elements that make up the sensual field, and perhaps even the inanimate world as well" (Harman, 2005, p. 161). Hence Metadata using virtual particles as a simile, exchanges information and agency in the quantum subterranean of digital space to otherwise non-living entities known as data, advancing algorithmic processes.³⁹

5.12 An ideological interlude: The Case of Materialism and Immaterialism

Admittedly putting forth the aforementioned philosophical lenses invites critical evaluation, with Davies presenting a common-sense rebuttal saturated in Western philosophy and materialism with the traditional cause-and-effect model. However, the author does highlight that traditional models do not account for or be concerned with the total reality of cotton when it burns (Davies, 2019). The unassumed reality of a thing can surprise us: although not always, and not with most

³⁹ This philosophical exploration has been influenced/inspired by conversations with and writings of Haider Ali Akmal (Akmal, 2021)

things, however often, with the digital existence of objects. With this point in mind, it holds that divergent philosophical theories have different assessments of reality, and perhaps an uninhibited approach will open new research avenues.

The following segments of this chapter will discuss philosophical theories that touch upon the discipline of Materialism. The intention at this point of the research is not to follow Harman in pursuing a dispute with Materialism for the immaculate integration of OOO and design for a More-than Human Centred Design approach but rather to present and curate an approach which is fluid and flexible to account for the digital world. Going forward, the research starts to implement a concept the design researchers Johan Redström and Heather Wiltse echo in their book *Changing Things*, which accounts for a tension that exists between the polarity associated with materialism and immaterialism (see § 5.11.2) (Redström & Wiltse, 2018, p. 68). This addition offers an expanded opportunity to speculate on digital objects' dynamic being, such as unified objects of emergence (i.e., Amazon web services are not just one thing) and in a materialistic fluid and immanent state (i.e., raw data processed into data). This latter example points to the dualism of materialism and immaterialism; however, it is not a clear-cut bracketing of physical and digital things, although these characteristics are helpful at specific analysis points. Instead, it is the opportunity to utilise different philosophical perspectives to aid design's perspective.

Some strains of materialism and OOO indeed have distinct genealogies; namely, Harman critiques the significance given to 'matter' as uniformed physical stuff.⁴⁰ However, materialism, as presented here, is Bryant's materialism that investigates the efficacy and power in things as he notes that "[e]ven ideas and concepts have their materiality" (2014, p. 6). According to Bryant, an outdated form of materialism has developed into meaning something historical, socially constructed, and contingent, having nothing to do with "processes that take place in the heart of stars, suffering from cancer, or transforming fossil fuels into greenhouse gases" (2014, p. 2). Wondering "where the materialism in materialism is" Bryant's materialism encourages us to attend to the agency of things, or as he labels them –machines, the power they exercise and the infrastructure they empower and

⁴⁰ Harman condemns materialism's perpetual the operation of 'reductive materialism'—the reduction of objects to their material parts (overmining).

create (Bryant, 2014, p. 2; Bryant quoted in Harman, 2014, p. 8; Pedriana, 2003). In an interview with Harman, Bryant recalls an occasion of playing *Sim City 4*, which inspired his materialistic ideology by situating him in an activity of building roads in the right places so that neighbourhoods do not wither and die— thus revealing another form of power, emergence, and the agency in things, explored next (Harman, 2014; Pedriana, 2003).⁴¹

5.12.1 Vibrant Objects

Pursuing a materialist tradition, the political theorist Jane Bennett stresses the agency of non-human and inanimate materials she calls *Vibrant Matter*. In her book of the same name, she uses examples such as a power grid, food, and stem cells to illustrate the agency of things – “forces with trajectories, propensities, or tendencies of their own” (2010, p.viii). This perspective rejects seeing things as inert, awaiting human interaction. It calls attention to the internal dynamism and latent capacities of things and their realities to affect and be affected by other things. Bryant presents a supplementary interpretation of object agency, highlighting that agency transpires in a ‘variety of degrees’, whereby bacteria ‘appears’ to have more agency than rocks insofar as bacteria are capable of initiating action within themselves, while rocks cannot (2014, p. 220-221). He continues to describe that things can expand in agency over their existence, and contrariwise agents can be restricted in their capacity to exercise their agency. To demonstrate: machine learning systems tend to learn and reprogramme themselves to adapt to the evolving conditions of their own operation. However, this can be limited through control laws wrapped around unknown dynamical systems (Duriez et al., 2017).

Bennett concludes that things never act alone but rather act alive and in process within assemblages of distributive agency, “a swarm of affiliates” with agency characteristics that include efficacy, trajectory, and causality (Bennett, 2010, p. 31). Efficacy can be understood as the aptitude to create through agency. However, for Bennett, this does not imply a subject (human) as the root cause, implying moral capacity; it is the distributive agency of the “swarm of vitalities at play” (both human

⁴¹ Bryant proposes that both materialism and OOO enquire how parts are organised and related entertaining theories of emergence (Bryant, 2012a; Harman, 2018, see § 5.11.2)

and non-human variety) (Ibid, p. 32). Consequently, the interwoven effects may result in less definitive outcomes than what humans desire or intended by their designs; thus, the task becomes identifying the “contours of the swarm” and the relations between the different vitalities (Ibid).

The second feature of Bennett’s distributive agency is trajectory, a movement of directionality rather than a destination or purposiveness. Here Bennett quotes Jacques Derrida’s alternative concept of trajectory as an open-ended and unspecified “promissory” note that will never be redeemed, but non the less this “straining forward” for Derrida is a phenomenon of life – the possibility of phenomenality (Ibid). As Bennett writes:

...things in the world appear to us at all only because they tantalize and hold us in suspense, alluding to a fullness that is elsewhere, to a future that, apparently, is on its way (Ibid).

Wakkery associate’s trajectory with the unpredictability of digital technology and data, defining a user’s interaction with – one of anticipation and surprise while also highlighting Bryant’s power in things (Wakkery, 2021, p. 32 & p. 152). An experience familiar with AI processes, such as the desire and the anticipation of the result from a Generative Adversarial Network (GAN) or, of more significant social consequence, credit predictions from a machine learning algorithm.⁴²

The final characteristic of vibrant matter is causality, the most elusive and vague of them all. Causality encapsulates the curiosity that events, circumstances, and occurrences cannot be directly attributable to a singular preceding event, commonly known as ‘cause and effect’. To demonstrate, Bennett provides an example for determining that the actants that manifest into totalitarian states are a matter of complex and heteronomous origins rather than definitive causes. Causality is emergent rather than efficient, affecting in nonlinear and fractural ways, where cause and effect alternate positions creating feedback loops. In this regard, Bennett tells us that things by themselves probably never cause anything, though once “crystallized” into fixed and definite forms, design researchers can “trace their history backwards” to source the intentionality (Ibid, p. 34). Historically, human intentionality (the power to formulate and enact aims) is a prevalent agential factor, yet Bennett,

⁴² Generative Adversarial Network is a machine learning model whereby two neural networks compete with each other using deep learning methods to become more accurate with their predictions approaching generative AI.

Bryant (2014) and Wakkery (2021, p. 157) attribute that intentionality exists in non-human things too, becoming key operators within assemblages. The practice of mapping assemblages for MTHCD will be detailed in the forthcoming chapter on practising philosophy.

5.13 Concluding on a More-Than Human-Centred Design for AI

This chapter has set up the philosophical proposition for speculating on the vicarious lives of non-human things by detailing and determining the uncustomary post-humanist position of OOO by differentiating between human perception and non-human relations and the ontology of things. The theoretical stance presented in this chapter provides the momentum for the metamorphosis of philosophy as an alternative design approach to AI, challenging the preconceptions of HCD. Consistent with the many reconfigurations of OOO, the ideology presented in this thesis has been customised to include theories that derive from materialism to cater for the agency of things. This chapter deals with prickly concepts, where most theories can be contested. However, speculatively the conversation does present an alternative way to consider non-human things; as Socrates quipped, “[t]he sense of wonder is the mark of the philosopher” (Plato, 2005, 155c), and this research would testify a designer too.

The basic principles for the philosophical model for AI design are as follows: (1.) all objects (human and non-human) should be given equal attention and addressed on a flat ontology; (2.) We can speculate on the vicarious realities of objects and their ontology while also being aware of Harman’s quadruple object in that an object has permutations which withdraw or reveal themselves through rifts and tensions; (3.) The interest here for design research is that objects are not identical to their properties; however, Harman tells us that the tension between objects and their properties is responsible for all of the change that occurs in the world through ‘vicarious causation’ making objects vibrant; (4.) A vibrant object has different variations and categories of agency, power, and emergence, forming assemblages of relations, manifestations, and change; (5.) In addition, a vibrant object can anticipate other objects through Bryant’s notion of object data exchange; (6.) Finally, in this curation of OOO, speculations on objects are inclusive of materialism and immaterialism reflections. As highlighted by Coulton et al., it can be argued that OOO is ‘rated high’ in a taxonomy of non-

anthropocentric theories as its viewpoint is all-encompassing, encouraging experimentation and appropriation with the ability to nest other theories without undermining either position (Coulton & Lindley, 2019; F. Pilling & Coulton, 2021).

The overarching argument presented in this chapter is to expose the deeper and hidden existence of AI interactions. To pivot one's attention away from the human in the loop – if only briefly – we can wonder, speculate, and scrutinise what AI truly is beyond its definitional dualism for design. With this part presenting the philosophical foundations, the next part will focus on developing a Human-AI Kinship by integrating the human-orientated position of post-phenomenology and arguing that this position, too, can be object orientated while also considering the user in the equation. The following will continue to cover strategies that embrace, facilitate, and empower AI technology, highlighting the potential for obscuring their nature and identifying ways design can facilitate legible AI.

Part Three Human-AI Kinship

5.14 Introduction

Archimedes said, “Give me a place to stand on, and I can move the earth” to illustrate the concept of leverage, whereby through the proper use of tools, one can achieve a lot more than brute force methods alone (Heath, 1953, p. xix). The philosopher of technology, Yoni Van Den Eede, observes that the phrase can also be read as differing theoretical positions and their entailing perspectives for moving the earth in a distinct technique (2022, p. 225). The ideology of postphenomenology is markedly divergent from OOO positioning as it clings to a human standpoint, by which the earth moves and is leveraged for human use, with technology determined to be the very medium for human existence (Rosenberger & Verbeek, 2015, p. 13). Though if we look at postphenomenology in an object-oriented light or vice versa, we can start designing by –taking the perspective of –the thing on its own *and* in the purview of human interaction and perception, forming a Human-AI Kinship in the spirit of Haraway’s kinship (Haraway, 2016). Kin, Haraway conveys, is to have “an enduring mutual, obligatory, non-optional, you-can’t-just-cast-that-away-when-it-gets-inconvenient, enduring relatedness that carries consequences” (Haraway quoted in Paulson, 2019, para 13). In other words, a post-human design framework that endorses the entanglement while accounting for the individual factors within a relation – technology and human; developing into a method to exercise “object empathy” (Van Den Eede, 2022, p. 226). To note: Haraway’s kinship is an extended process of attention, perseverance, and care among humans and non-humans, a positioning this research has adopted as a MTHCD rationale for designing with and for AI. Human-AI Kinship, on a design level, is to perceive the accountabilities and obligations with AI things, forming new bonds and new intimacies with human and non-human AI others. At a user’s end, Human-AI Kinship is to have more awareness and understanding of AI technology through practical mechanisms, thus permitting more agency and negotiability. Though to differentiate Human-AI Kinship from Haraway’s kinship: for Haraway, it is a call to make kin with non-human things as an urgent ethical and ecological responsibility, to break out of thought patterns and actions that are destructive to all living things.

The rationale for adopting Haraway's term of 'kinship' is addressed briefly in the introduction; this was intentional to establish the similarities and differences of the terms use in this research.⁴³ The rest of the chapter continues formulating Human-AI Kinship by applying postphenomenology into this research's assemblage.

The first part of this chapter exposed issues with HCD and HCAI, drawing on a conclusion for a non-anthropocentric approach for AI, which set up the second part of this chapter, the construction of a MTHCD approach, via a closer inspection of OOO with a materialism nuance. The following section will be the conclusive part of this chapter, which attempts to re-establish the human user back into the equation through the integration and customisation of a postphenomenology approach. The aim remains to establish a non-centric placement of the user by flattening the ontology of things and levelling down the multiplicity of perspectives through OOO by looking at the philosophical interpretations of Van Den Eede (2022). The structure of this section is as follows: an overview of postphenomenology for analysing human-AI relations, drawing on and extending the field in a MTHCD light. Ontological accounts of AI will be viewed through a postphenomenological lens as a method for disclosing and viewing the operations of AI to address AI legibility or lack thereof. This outline will inevitably expose the 'gap' between the ideological positionings of OOO and postphenomenology— thus setting up the mantel to explore postphenomenology with Archimedes standing in a different positioning through Van Den Eede's development of 'object empathy' by implementing an OOO perspective. Finally, as the gap between humans and things will always exist, as we remain forever trapped in our human condition, bridging the gap is insurmountable. However, we can combine OOO and postphenomenological theories to form a speculative and actionable perspective for Human-AI Kinship.

5.15 A Short Introduction to Postphenomenology

⁴³ Haraway's notion of Kinship is explored further in Chapter Six Design Fiction: Adapting Philosophy for Design, where this research unpacks the practical design practice through the practice of Design Fiction, which will also take cues from Haraway's practice of worlding to create kinship opportunities (see § 6.8 A *Philosophical Interlude: Philosophical metamorphosis through Worlding Constellations*).

In 1990, Ihde published his influential book *Technology and the Lifeworld*, an account of using theoretical tools of phenomenology to analyse the relations between humans and technological artefacts forming the ideology known as Postphenomenology. Postphenomenologists, however, distance themselves from classical-phenomenological romanticism of Technology (capital T), most notably found in the work of Heidegger. This perspective saw analysing Technology as a broad, social, and cultural phenomenon focusing on technology alienating humans from themselves and the world. Consequently, by dissociating itself from the abstract and transcendental tradition of phenomenology, Postphenomenology in its place focuses on technology (small t) through empirical and praxical analyses of ‘actual’ technologies as *mediators*— rather than as alienators of human experiences and practices (Ihde, 1995, p. 7; Rosenberger & Verbeek, 2015).⁴⁴ In developing his concept, Ihde uses Heidegger’s tool analysis similar to Harman, although with a noticeable difference by calling attention to the pragmatic description of context-dependent human-technology relations (Ihde, 1979, 1990, pp. 80, 98), rather than deducing that something is *relational* when placed before us and something is *substantive* as it withdraws (Harman, 2018; Van Den Eede, 2022).

To be explicit: Ihde’s primordial idea is a repackaging of Heidegger’s totalising account of the metaphysics of tool use as a “whole” (Ihde, 1979, p. 118). Instead, Ihde reflects on the Greek word “*pragmata*”, meaning “things”, which is closely related to the word “*praxis*” to mean “practice”, whereby “[h]uman-world relations are practically “enacted” via technologies” (Rosenberger & Verbeek, 2015, p. 12). Thus, thinking in terms of technology as a mediator aims to expose the ‘relational ontology’ between human beings and their world – presenting a reinterpretation of classical phenomenology (Husserl, Heidegger etc.) to *describe* the world (Ibid, p. 11).⁴⁵ Therefore contrasting with OOO, technologies in a postphenomenological context are understood in terms of the relations human beings have with them and not as things “in themselves” (Rosenberger & Verbeek, 2015, p. 19). Technological mediation is reflected in the postphenomenologist’s foremost tool, the diagrammatical scheme – “Human—Technology—World” (Ihde, 1990, p. 85). As well as

⁴⁴ Hence the post in postphenomenology is used to distance from the romanticism of classical phenomenology.

⁴⁵Mediation is versatile concept that has been advanced in various ways to convey detailed and intricate meanings in several theoretical frameworks – respectively Actor Network Theory, media theory, ethnography, sociological and psychological theories (Van Den Eede, 2011).

deconstructing a user's experience of the world through the mediation of technologies, the tool also pinpoints the role of technologies in humans' existences from eyeglasses, cochlear implants (Ihde, 1990), cars, obstetric ultrasound technology (Verbeek, 2008b) etc. Of particular note, for this research to be viewed in due course, is Wiltse's article *Unpacking Digital Material Mediation*, in which she extends postphenomenology theory to consider digital things as a responsive material that mediates a person's engagement and perspective of the world (2014).

Classical phenomenology resulted from discontent with the modernistic separation of subject and object, arguing for an 'intentional relation' as an inseparable existence between them. "The human subject is always directed at objects: we cannot just "see," "hear," or "think," but we always see, hear, or think *something*" (Rosenberger & Verbeek, 2015, p. 11; original emphasis). However, this perspective considers objects as existing 'in themselves', although it contends that as soon as we have thought about them, they become things-for-us, cancelling the notion. Although, postphenomenology moves beyond intentional relation, with mediation shaping human subjectivity and objectivity of the world, yet the method does not claim to describe reality (Verbeek, 2005). Peter-Paul Verbeek classifies it as subject and object are "constituted" in their mediated relation, with intentionality not acting as a bridge but the emergence of the two (2005, p. 113).⁴⁶ For example, ultrasound technology shapes the character of that constitution into several specific relations (Ibid, 2008b). To illustrate: ultrasound waves echo off denser surfaces inside a mother's body, creating an image on a screen that isolates and separates the unborn foetus, establishing it as an individual person. In this postphenomenological view, claiming privileged access to 'things themselves' becomes an impossibility, whereas, as previously demonstrated, OOO does not refute this claim; however, the position would argue for speculating on the thing in itself rather than always viewing something in relation to a human user. Supporting this notion, Wiltse presents a case to consider both humans within praxis and how to relate to things that unveil their functionality, materiality and programmed internationalities by simply considering and looking at what things do (2014).

⁴⁶ Additionally, both Ihde and Verbeek consider the notion of agency can only be attributed to human-technology relations rather than to each component individually (Verbeek, 2005).

Another fundamental postphenomenological concept is Multistability, which rationalises that a technological device may be used for different purposes when embedded in differing contexts. Ihde first developed the notion by investigating the multistability of visual perception, using simple illustrations that could be interpreted in more ways than one; of note is the famous Necker cube illustration that permitted separate stabilities to surface in a viewer's visual gestalt (Figure 32).

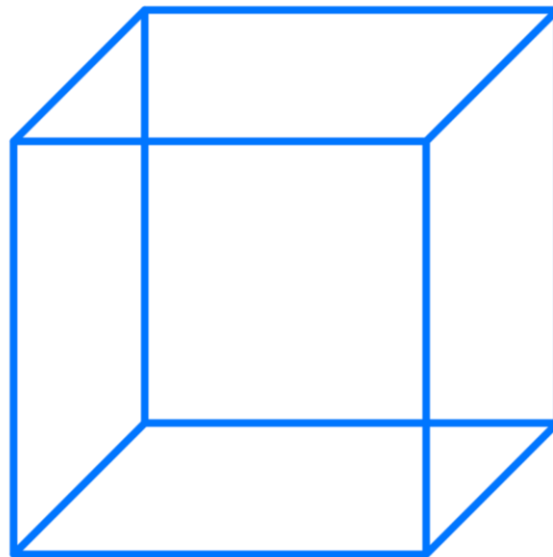


Figure 33: The Necker Cube is an optical illusion with no visual cues to its orientation, so it can be interpreted to have either the lower-left or the upper-right square as its front side.

Considering technological artefacts and using Heidegger's hammer, Ihde demonstrates that Heidegger failed to realise past the dominant stability of driving nails and therefore performs a "variational analysis" revealing the hammer's multistability, thus demonstrating that the design of the hammer does not prevent it from being an object of art, a murder weapon, or a paperweight (Ibid, p.46). As Ihde notes, "[n]o technology is 'one thing' nor is it incapable of belonging to multiple contexts", although the materiality does (typically) constrain use and function (1999, p. 47). Consequently, Robert Rosenberger and Verbeek discuss one of postphenomenology's primary enquiries: how technology can manifest beyond the designed intent into stabilities that precede influencing, confusing, persuading, restricting, or controlling users (2015, p. 25).

Leaning on an OOO ideology, digital technology with intentionality, such as ML, can expand its materiality and function beyond the original coded intention. Therefore, forming obscure and detrimental stabilities to users and its creators, even enabling unscrupulous opportunities that companies can harness latterly with adjusted parameters of the ML algorithm and collected data. A

forewarning by Professor Gina Neff illustrates such an instance, writing after *Roe vs Wade* was overturned in America's supreme court, "[r]ight now, and I mean this instant, delete every digital trace of any menstrual tracking. Please." Neff's statement implies that an apparent 'harmless' app that enables women to track their periods can also be used laterally to monitor when a woman becomes pregnant, and if a period resumes too quickly before a baby can come to term, signifying a (now) prohibited abortion may have taken place (2022).

Nevertheless, to spotlight human users and their relations with technology, which can present in divergent stabilities, Rosenberger describes users as having a "relational strategy" (2005) regulated by the user's perception of technology that catalyses a specific stability of a device.⁴⁷ The following section will present Ihde's rudimentary forms of technological mediations and user relations, which will be expanded explicitly through human-AI relations, and AI operations, exposing possible legibility issues.

5.16 Human-Technology Relations

The postphenomenological accounts of human-technology relations presented here are not exhaustive, though they articulate the field's view on how users develop "bodily-perceptual" relationships with the devices they use (Ihde, 1990). Likewise, Rosenberger and Verbeek describe the field as 'buckling at the edges' with accounts of human-technology relations, a symptom of the specific contextualisations of user experiences and the sheer amount of technological innovations analysed while avoiding the oversimplification of technological mediation (Rosenberger & Verbeek, 2015, p. 13).

Following Heidegger's tool analysis for when a device is ready-to-hand, thus permitting engagement with the world 'through' themselves, Ihde forwards the notion of "embodiment relations" to characterise a user's experience as being reshaped through the device and is correspondingly taken into the user's bodily awareness. Ihde's infamous example uses eyeglasses to describe the user looking *through* the optics of a transformed world, with the glasses forming part of the user's

⁴⁷ Perception can be influential upon learnt conceptions, interpretative frameworks, cultural conventions, and bodily-perceptual customs of use.

perceptual experience. The glasses mediate transformation, creating a bodily-perceptual relationship between the user and the world. Ihde writes, “the wearer of eyeglasses embodies eyeglass technology: I—eyeglasses—world” (1990, p. 73).

There is, however, a “trade-off” with embodied relations, whereby a user obtains the desired modification, although they have to commit to other changes through the “non-neutral transformations rendered to user experience through the mediation of a technology” (Rosenberger & Verbeek, 2015, p. 16). For instance, a telescope image of the moon enables us to see the surface in great detail, though it removes the moon from the context of the sky (Ihde, 1990, p. 76). Likewise, it can be argued that a trade-off is played out in most technological relations, embodied or not, as the HCI principle of negotiation attempts to mitigate the non-neutral transformations of data interactions. Instead of a bodily exchange, a user trades data as currency to use services that perceptually ‘on the surface’ seems free to use. This point is accentuated in the Future Mundane caravan’s interactive experience, which situates audiences in a design fiction and artificial world to explore and be exposed to simulated yet potentially detrimental data interactions within a smart equipped living room (M. Pilling, et al., 2022b).

Returning to Ihde’s eyeglasses, the trade-off for improved sight is wearing a device. However, to a certain degree, aspects of the glasses disappear into the background of the user’s awareness as it is used. Ihde writes: “[m]y glasses become part of the way I ordinarily experience my surroundings; they ‘withdraw’ and are barely noticed, if at all. I have then actively embodied the technics of vision” (1990, p. 76). The notion of technology ‘withdrawing’ has been chosen explicitly by Ihde to evoke Heidegger’s account of the withdrawing ready-to-hand tool again. Ihde notes: “[t]he closer to invisibility, transparency, and the extension of one’s own bodily sense this technology allows, the better” (1990, p. 74). On this point, Ihde explains that the design and use of technology create a “double desire”. In this regard, we want technology to transform our relationship with the world and the means of that transformation to be experientially *transparent* as possible (Ibid, p. 75). Transparency ultimately changes our perception of technology and is, therefore, a defining feature of embodiment relations, echoing the HCD axiom for technological transparency for the sake of

usability and simplicity. These rationales also facilitate persuasive nudging and the deceptive collection of data, amongst other things.

Rosenberger presents a contention for Ihde's emphasis on transparency, which disregards the assemblage of technological features that demand attention or the ones that exist on the user's periphery (2012, 2014). To this end, Rosenberger develops two further variables akin to transparency that depicts a user's "reconfigured" technologically mediated field of awareness; these are "field composition" and "sedimentation". The concept of sedimentation refers to past experiences contextualising a present experience, with Rosenberger writing, "[w]hen a particular human-technology relation has a high degree of sedimentation, that user is strongly inclined to experience the use of that technology in a specific, long-established manner" (2012, p. 27). For example, interfaces for the major streaming platforms are typically designed to have multiple threads of content ordered in different categories, organised by the films being labelled and sorted by AI technology for users' ease to interact with these similar platforms based on past experiences of using these applications.

To understand the concept of field composition, it is helpful to turn to Rosenberger's analysis of watching a movie in a darkened theatre. Rather than saying that the things between the screen and the viewer are transparent, Rosenberger would argue that the screen and the movie "stands positively forward" with visual and audio elements provided through the assemblage of technology in the theatre that *composes* and organise a user's field of awareness. Likewise, the combination of data and AI technology composes or quells a user's field of awareness by delivering curated content, for instance, on social media platforms that form feedback loops, echo chambers, and filter bubbles, which in turn amplify viewpoints and divisive behaviours.⁴⁸ AI technology affords and currently profits from both Ihde's transparency and Rosenberger's features of field composition and sedimentation systematically, influencing a user's perception of technology. As this thesis has specified, AI, to some extent, is intangible, has a confused ontology and is designed for simplicity to remain in the obscurity

⁴⁸ A former YouTube engineer Guillaume Chaslot publicly outed YouTube's algorithm that was engineered to "heavily promote Brexit" as divisiveness extends user's watch time, leading to increased opportunities to capitalise on advertisement viewing (2018).

of back-end computing, away from the point of user interaction. However, it conveys a pigeonholed reality version to an ill-informed user.

5.17 Background relations: Notes on Engagement

Another similar conceptualisation is Ihde's notion of 'background relations' in which a user shares a space with a device that has a distinct kind of "absence", but nonetheless interacts with as they shape and form a user's experiential surroundings (1990, p. 109). This viewpoint presents another type of withdrawing to transparency –one that is "off to the side" and stands back in our awareness, ultimately designed to function in the background (Ibid). Ihde gives examples of automatic and semiautomatic machines in the mundane context of the home, such as lighting, heating, and fridges where "in operation, the technology does not call for focal attention". Although "textures the environments" transforming the "gestalts of human experiences" in "subtle indirect effects upon the way a world is experienced" (Ibid, pp.109-112). Schematically shown as:

Background relations: Human (technology/world)

For Ihde, this classification also includes technologies that require repeated "deistic interventions", such as unloading washing machines, which becomes an automatic process (Ibid). An example of a superficially automatic process to the user is Google's Nest learning thermostats that monitor and collect sensor data and learn over time through machine learning algorithms to automate a home's temperature, which could be schematically shown as:

Human(technology—technology/world)

Note: This schematic only details the thermostat and the machine learning algorithm technology; theoretically, a designer could further break this down to include data mapping, etc. Interestingly Ihde's concept of background relations was formed before the computing age's peak and can be substantiated by his technological examples omitting digital media; nevertheless, as shown, background relations can be applicable to the nature of ubiquity computing and IoT devices, animated through AI technology and datafication.

As a final point to note, according to the philosopher Albert Borgmann, technological devices of our time has diminished people's engagement with "the coherent and engaging character of the

pretechnological world of things” (1984, p. 47). In his view, devices consist of two features – they are made up of “machinery” and deliver *only* a “commodity” when functioning. For instance, the central heating will deliver heat, although a hearth provides heat and an opportunity for togetherness (Ibid, p.41-42). Examining Borgmann’s observations, Verbeek and Petran Kockelkoren highlight the characterisation of modern technology shows an ever-growing emphasis on “commodification” with digital technology offering enrichment and disburdenment via the manifold of functionality presented in one device, and with machinery withdrawing into the background of our lived experiences (1998, p. 40). The authors write, “[d]evices are designed to leave us aside of their functioning: they do not ask for engagement, but for *consumption*” (Ibid, p. 41; original emphasis).⁴⁹ For Borgmann, the situation becomes problematic when high-tech devices become more concerned with consuming commodities rather than engagement through attention and involvement. For this reason, Verbeek and Kockelkoren promote ‘healing the split’ between machinery and commodity by revealing the machinery of products through design, freeing devices from their withdrawal by being visible, accessible, lucid and, in their words “create[ing] a bond between people and products *as artifacts*” (Ibid; original emphasis).

5.18 The Evolution of Hermeneutic Relations to Digital Hermeneutic Relations

Moving on to another relation of note which is influenced by the hermeneutic tradition of philosophy concerned with the nature of language interpretation and translation, Ihde proposed “hermeneutic relations” to refer to technologies that users employ to perceive and interpret a device’s readout (1990, pp. 80–97). According to postphenomenologists, in a hermeneutic relation, a user experiences a *representation* of the world requiring interpretation to gain helpful information rather than experiencing the world through a device. Verbeek uses the example of an analogue thermometer, describing that:

⁴⁹ It can be observed, however that AI technology still requires a type of engagement, although one that is indirect by exploiting monitored interaction points that elicits training data, which as a process are fed back into the system to improve and tailor interaction points. Through consumption we feed to algorithmic system.

...when we read a thermometer, we are not involved with the thermometer but with the world, of which the thermometer reveals one aspect, namely, its temperature (2005, p. 126).

Of note concerning this research into AI legibility, users generally do not have access to interpret AI; however, they are the beneficiary of AI's modelling of the world through its generative outcomes. The schema for a hermeneutic relation is as follows, with arrow indicating intentionality:

Hermeneutic relations: I \longrightarrow (technology—world)

Postphenomenology recognises that technological mediations have perceptual consequences as artefacts *transforming* the user's experience "by the means" of the technology in question (Ibid) – namely, they stress some aspects of the world while neglecting others. On this note, Wiltse describes 'digital materials' as being responsive and mediating perceptions of the world via "traces" that can be interpreted by a user, which are fashioned typically indicative of the nature of the "substrate" –the component of the digital material that responds to a stimulus (2014). In the example of a user typing on a computer, the text is pointed out by Wiltse as the trace. However, with technological mediations of a digital constitution, the substrate is difficult to pin down, having a higher level of complexity than non-digital mediations. For instance, continuing with the typing example, the physical keyboard forms part of the substrate, as does the digital components, which include the operating system, and the application, in addition to the metadata produced from input on a computer. As an extension to Verbeek's composite intentionality (2008a), Wiltse posits that materials have their *own* intentionality concerning the world, schematically shown:

I \longrightarrow ([trace|substrate] \longrightarrow world)

Explaining the diagram, digital materials can be characterised as a trace and substrate, with traces facing the perceiving person to gather information about the world, while the substrate points towards the world, reflecting the digital thing responding to the world.

As distinguished, it is challenging to track substrates. In some cases, the traces can be elusive, as functions and uses are typically "uncoupled" in operational dimensions, thus enabling auxiliary uncoupling to occur in perceptual and temporal dimensions. Functional uncoupling of trace and substrate can be gleaned from a digital weather widget displaying the weather conditions sensed in a location on the other side of the world. In this regard, Wiltse describes that perceiving the world

through digital materials is often an “incidental post-hoc affair”, whereby one cannot see the entire digital apparatus involved in producing traces resulting in a lag to form between activity and the trace being made visible (2014, p. 172). Here, it is also useful to recount Borgmann’s discernment of modern technologies, in that the *means* of producing traces are separated from the *ends* of the trace themselves via the concealment of technologies ‘interior’ workings creating an unfamiliarity of the technology in hand (1984, pp. 43–44). Therefore, this state, as Wiltse draws attention to, generates implications for a user’s mediated perception, wherein to understand the true implications of a trace, a user needs to know how a trace occurs as determined by the nature of the substrate.

Reflecting on an AI example, the trace can be deemed as the generative output of Amazon’s AI assistant *Alexa* in response to a user’s question. The trace formed by a generated voice engineered through Natural Language Programming to mimic a human’s conversational style can be categorised as an ‘illusionary trace’ where often it is perceived by users of a machine as exhibiting intelligence. The substrate, however, is an amalgamation of many digital components and results from various nested substrates that are sourced through programming for one request. It is also worth noting, from a MTHCD perspective, that a trace in one instance in an AI’s operation can act as a substrate in another. Such as, the request from the user can leave a trace in data that will be utilised in training the AI program and act as a substrate triggering a pre-operationalised response through engaging with earlier cultivated responses from machine learning training that uses data scraped from a myriad of sources. In comparison: approaching data’s ontology from strictly an anthropocentric position, in her book *Digital Sociology* Noortje Marres contends:

...the notion of data presumes a particular architecture, [whereas] the notion of trace is more minimal, positing merely the detection of a thing or movement and the recording of this (2017, p. 54).

However, from a MTHCD approach, due to the inherent operations of machine learning, each training revolution with data/data sets results in information being extracted, abstracted, and correlated, leaving a trace within the final version of the ML program and the generative output.

It is worthwhile to highlight, at this point, that the field of Digital Hermeneutics has different interpretations and standpoints distinguished by Romele et al. as having either a “methodological” or

“ontological” attitude (2020, p.75). However, a complete overview of the various nuances of digital hermeneutics is beyond the scope of this research (see Capurro, 2010): though a methodological view is concerned with computer-mediated interpretation and understanding of digital texts, data, and databases focussing on the socio-materiality of digitality such as code and detailing the specific technology that captures data (Romele, 2020, p. 71). On the other, it is the recognition of (mainly) differences and similarities between humans and AI (non-human) intentionality or considering the AI in itself by tracking the digital infrastructure with design-focused approaches akin to those forwarded by Wiltse.⁵⁰

Wiltse’s account of digital materials mediates a perception of the world via the concept of traces and substrates to articulate the logic, structure, and function of responsive materials. Her account, like Latour’s ‘circulating references’ (1999, Chapter 2), pursues various chains of traces and substrates that start in physical worlds and are transformed and propagated through various processes of digital substrates that get increasingly difficult to track. The latter point, as demonstrated, is especially representative of AI technology, which calculates beyond human understanding with some degree of autonomy and intentionality, although it participates in determining and revealing a particular reality for a user.⁵¹ However, it is also worth highlighting that Wiltse’s approach, to some extent, is a postphenomenological account tinted with a OOO objective of going back to the things themselves, which initiates the agenda for the following section, that explores the evolving position of non-anthropocentric postphenomenology.

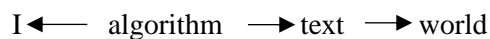
5.19 Machine Hermeneutics

⁵⁰ In his book *Digital Hermeneutics*, Alberto Romele builds an argument for posthuman hermeneutics for digital machines influenced by Heidegger’s development of hermeneutics that dealt with the ontological conditions for the interpretation and understanding enacted by human beings for acting and interacting with the world (Romele, 2020).

⁵¹ Looking at the field of AI art, Mohammad Majid Al-Rifaie and John Mark Bishop attempt to map and distinguish between an intentionality of weak and strong computational creativity inspired by Searle’s famous thought experiment (Searle, 1980). From their inquiry the authors argue that weak deployments do not go beyond imitating human creativity, although strong deployments can “understand its creation and have other cognitive states” aside from those associated with human minds (2015, p. 10).

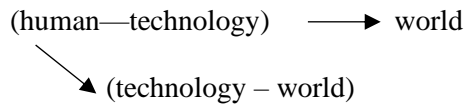
Galit Wellner proposes a non-anthropocentric postphenomenological development of Ihde’s hermeneutic relations using the foremost hermeneutic focus of writing (which, even in analogue disposition, is a type of technology) as her subject. Wellner’s account surveys the concept of digital writing, chiefly generated to be read by a human user and contrasts this with text or data produced in machine learning processes not intended for human eyes but to be read by other machines and algorithmic processes (2018). Algorithmic reading and writing transpire in what Wellner categorises as “algorithmic media space” – a reality exclusively grasped and produced by algorithms. However, ultimately these spaces produce an output for human consumption with products such as weather forecasts and sports coverages.

The pivotal point Wellner attempts to articulate is the “technological intentionality” of algorithms shown in her schema by reversing the arrow from the original position of ‘I’ towards ‘technology’ and the ‘world’. The third arrow between ‘text’ and the ‘world’ represents the creation of a trace in the world, inspired by Wiltse’s concept of trace and substrates:

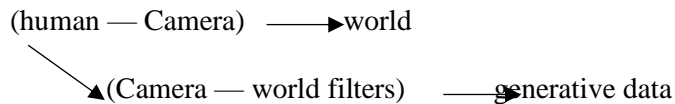


The phenomenon Wellner brings to mind may well be called ‘machine hermeneutics’ for a condition made possible by AI’s ontology (although one could argue to some extent that this situation occurs in all forms of back-end computerised processing), in which different parts of the digital materials analyse and interprets the traces of another. Machine hermeneutics does not present a technological instrument for a human user to perform an interpretation; the concept, however, accounts for the understanding and interpretational relation between non-human things.

A postphenomenological schema that justifies both a user’s hermeneutic relations and a machine hermeneutics draws upon Rosenberger and Verbeek’s notion of an “augmented relation” using the formerly hyped Google Glasses (2015, p. 22). Despite the ultimately uncommercial and therefore thwarted product, the authors observe that Google Glasses offers two parallel relations with the world. In one instance, users have an embodiment with the glasses themselves, returning to Ihde’s non-digital glasses illustration and a hermeneutic relation through the embedded screen providing an augmented representation of the world to the user as modelled by a computer:



A similar two-world perspective can also be drawn by accounting for a human’s perception of the world as mediated by the device in use, and the model of the world created by AI’s software through machine hermeneutics based on the interpretation of data and sensors, for instance, found in autonomous vehicles:



Verbeek has argued that technology such as brain implants and augmented reality offer a cyborg relation, in which technologies merge with the human body, whereupon the physical boundaries betwixt the two are seamlessly blurred, enabling users to experience a digital representation of the world (2008a). Consequently, the difference between the Google Glasses and the autonomous vehicle example is that users are not commonly privy to the model of the world as mapped by an AI, and be accounted as a type of background and alterity relation. In greater detail: the level of autonomous vehicles is currently at a level 2 out of a 5-point scale, meaning partial driving automation with advanced driver assistance systems (ADAS) is available, providing a range of safeguard and adaptive features, such as cruise control, assistance in avoiding collisions and obstacles etc. Although automation currently falls short of self-driving, a human user can take control of the car anytime. Therefore, in this instance, a user experiences the generative interpretations of the world that direct a broad spectrum of semi-direct (autonomous braking) and indirect (lane departure warnings) actions in the world as a result of machine hermeneutics occurring simultaneously, although ‘underneath’ the (presently) superseding mediated relation of a driver and the vehicle. The aspiration of level 5 autonomous vehicles conversely is intended to be fully autonomous, with dynamic driving tasks removed, making steering wheels, acceleration, and braking pedals redundant. Consequently, this setup mediates an entirely different user experience, wherein the vehicle and the journey it takes through the world are expected to fall further into the background, while users can focus on other technological devices or social engagement with other passengers.

Reviewing traditional postphenomenological schemes and advancing them through a MTHCD lens has set the mantle in the following section to conclude and fuse the two ideologies of OOO and postphenomenology – settling their differences by pinpointing correlations between the two.

5.20 Concluding on OOO and Postphenomenology: Namely Cultivating Object Empathy for Human and AI Kinship (despite Thing-Transcendentality)

In his article, *Thing-Transcendentality: Navigating the Interval of “technology” and “Technology”*, Van Den Eede attempts to reorient the deep-rooted empirical-transcendental debate hinged upon their counterposing perspectives. The previous introduction to postphenomenology outlined that the ideology’s standpoint is firmly positioned in the empirical, interrelational, and praxical investigations of technology (small t) rather than the transcendental orientation of Technology (big T). Van Den Eede attempts to cross the gap between the two perspectives by way of OOO and proffer an object-oriented tint for postphenomenology schemes by employing object empathy, which will be seen and considered in this research as an approach for Human and AI Kinship. Although, in compositing a OOO perception, the author catalyses and “levels down” to a multiplicity of perspectives, smearing out the gap forged by the debate and, in lieu, exposing the array of gaps that exist between the throngs of things and their relations. In other words, OOO offers a fresh perspective on the empirical-transcendental debate by proposing its “obsolescence” while in true OOO style –still flirting with the existence of both regarding the “empirical-like and transcendental-like dynamics to things” (Van Den Eede, 2022, p. 226). On the latter point, Van Den Eede suggests that thing-transcendentality pertains to all things, as determined through Morton’s observations, which states, “[c]orrelationism is true: you can’t grasp things in themselves” (Morton, 2018, pp. 13–14) as the hidden core is always out of reach “but disastrous if restricted to humans only” (Morton, 2016, p. 17). Grasping Morton’s concept of the transcendental further, he describes the perceptual ‘data’ a viewer may get from an apple, observing:

There is a radical gap between the apple and how it appears, its data, such that no matter how much you study the apple, you won't be able to locate the gap by pointing to it: it's a *transcendental gap*" (2018, p. xxx).

Morton further stipulates, "for me, it is the idea of a privilege transcendental sphere that constitutes the problem, not the finitude of the human-world correlation" (2013, p. 17). Thus, his solution is to "*release the anthropocentric copyright control on correlationism*, allowing nonhumans like fish (perhaps even fish forks) the fun of not being able to access the in-itself" (2016, p. 18; original emphasis).

In correlation, Harman writes, "the basic rift in the cosmos lies between objects and relations in general"(Harman, 2011b, p. 119), representing how 'real objects', as detailed earlier, escapes all perceptual grasps, yet one can sidestep anthropocentrism by speculating on things in themselves (Bogost, 2012; Harman, 2018). This philosophical thinking highlights Harman's and Ihde's differing interpretations of Heidegger, also noted by Van Den Eede, who observes that "postphenomenology doesn't like the substantive. Its ontology being [exclusively] *relational*", catered towards a human's interpretation, experience, and perspective. As such, postphenomenology has shown an unenthusiastic response towards OOO, even though a nonhuman standpoint would serve well in speculating on technological mediations and the concealed operations of algorithmic technologies that escape human control.

Nevertheless, despite their differences and pushing forward with transcending postphenomenology with a OOO perspective, Van Den Eede attests that one could view Harman's methodology as empirical, accounting for the abundancy and multiplicity of things as evidenced by his writings (Bryant, Bogost, and Morton can also be included in this framing). This view is especially evidenced in Harman's book *Circus Philosophicus*, which revives the metaphysics of objects through detailed descriptions of the varied interplay between them, using a range of subjects from Ferris wheels to a haunted boat (Harman, 2009a). Bogost, too, curates and commissions a series of books called *Object Lessons*, which take a deeper look at subjects often taken for granted, such as dust (Marder, 2016), waste (Thill, 2015), silence (Biguenet, 2015), glitter (Seymour, 2022) and hyphens (Mahdavi, 2021), to name a few. Van Den Eede also brings attention to the similarity between

Harman's hidden substance of objects with Ihde's multistability. While Harman disputes "the notion that what is currently expressed in the world is all the world has to offer" (2016, p. 33), instead of attesting to an object's hidden and surprising capacity, Ihde offers a similar concept to consider technologies multiple and at times hidden trajectories from human perception. In this interpretation, Van Den Eede specifies, "perhaps postphenomenology can be regarded as object-oriented philosophy that has simply been attending exclusively to the human point of view—up until now?" (2022, p. 240) Nevertheless, as a counterpoise: taking the perspective of non-human things has the peril of becoming "trite" as objects are always beyond full access to their inner cores, and secondly, the consideration of things is ultimately done out of human concern or purpose (Ibid). However, in a provocative reflection resembling something Harman would cite, Van Den Eede suggests "[t]here might be nothing wrong with anthropocentrism, as long as we keep it subversive" through "*cultivating* object empathy" (Ibid; original emphasis).

Ergo, humans and nonhumans can be considered independent, responsive, and intentional entities. Meanwhile, these entities can also be viewed in a composite intentionality as and when justified on the conditions of analysis undertaken as "what we—or whatever remains of the human being—do, is never isolated ecologically" (Van Den Eede, 2022, p. 241, See Morton, 2010, 2016, 2017, 2018 & Haraway, 2016). Furthermore, as Harman states, objects hold "something in reserve"; we can only speculate. However, as Van Den Eede advances with a postphenomenological view can aid in "determining where the invisibility is located *for us*" (Van Den Eede, 2022, p. 241; original emphasis). Therefore, this framework offers a multiplicity of flexible design perspectives, with Archimedes unchained and free to stand according to the needs of the problem at hand.

5.21 Conclusion

This chapter was split into three parts to develop a Human-AI Kinship for design via a OOO lens, enhanced by examining human technology relations through postphenomenology. These three parts could have been standalone chapters; however, keeping them together reflects how the research was undertaken and how disparate fields were brought together into a single whole. First, the fields of HCD and HCAI were analysed through an overview of three HCD axioms: interaction design,

simplicity by design, and persuasive design. These tenets were found to be problematic to varying degrees— shrouding technological functions by overlooking user agency and negotiability while using AI-infused products and services. Thus, identifying an alternative perspective may tease out auxiliary design approaches and products for AI legibility. The second part of this chapter was concerned with developing a MTHCD perception by looking at the speculative realist philosophy of OOO and unhinging humans as the monarchs of being, placing all things on a flat ontological plane of existence. Therefore, it permits a designer to ask, what's it like to be an AI? Once a OOO lens for design was developed through a thorough understanding, the human was again considered as, ultimately, design is for human consumption; however, one can practice it with object empathy. With the theory for Human-AI Kinship formed across three parts, the next chapter will look at metamorphosing philosophy into design practice.

Chapter Six Design Fiction: Adapting Philosophy for Design

(Being AI)

6.1 Introducing the Carpentry of Things

Bogost tells us, “writing is *only one form* of being ... where all ideas, interchanges, and actions are strained through the sieve of language” (2012, p. 92). If one only relates to the world through writing and not through *doing*, then as Ihde puts it, “the basic thrust and import of phenomenology is likely to be misunderstood at the least or missed at the most” (2012, p. 4). In this regard, Bogost states that making things rejects the correlationist agenda, whereas the written form can only present itself to the human’s capacity to read (2012, p.93). These sentiments are congruent with design practice and making, wherein knowledge is generated through the process of doing, resulting in an artefact that catalyses a new understanding for the designer, as previously noted in Chapter Four *Methodologies*.

Nevertheless, as Bogost points out, “[m]aking things is hard” as one must contend with the material’s resistance (2012, p. 92). In a similar nod, Bryant observes that a negotiation takes place rather than “the simple imposition of a form upon a passive matter” (2014, p. 19). The philosopher Jean -Paul Sartre argued that a type of “technical intentionality” occurs in the process of doing, which Bryant explains as an essence that arises not from the designer but from the things themselves (Ibid, p.20). It is at this verge that both Harman and Bogost cite where the object itself becomes the philosophy, referring to the observation or revelation into “how things fashion one another and the world at large” that is derived from the act of doing almost confronting the thing in question (Bogost, 2012, p. 93; Harman, 2005). This practice is also known as “the carpentry of things” (Bogost,2012, p.93), which Harman describes as “the *metaphysical* way in which objects are joined or pieced together, as well as the internal composition of their individual parts” (2005, p. 2).

Bogost goes into meticulous detail in his book to clarify what carpentered artefacts are, describing them as “philosophical lab equipment” –entries into philosophical discourse through any act of making/ doing and in any material, extending the ordinary sense of woodwork (although one can do philosophical Carpentry in woodwork too) (2012, p. 100). To perform carpentry, Bogost continues, is to create a “machine...to replicate [or trace] the unit operation of another’s experience” (Ibid):

Like a space probe sent out to record, process, and report information, the alien phenomenologist's carpentry seeks to capture and characterize an experience it can never fully understand, offering a rendering satisfactory enough to allow the artifact's operator to gain some insight into an alien thing's experience (Ibid).

As Bogost is a self-confessed philosopher-programmer, his examples of carpentry turn to specimens rooted in the computational world. Of particular interest is Ben Fry's *Deconstructulator*, a modified program of a Nintendo Entertainment System's (NES) emulator that depicts the current state of the machine's sprite and palette memory (Figure 33). In other words, the *Deconstructulator* offers an exploded view and ontology of NES's memory architecture, particularly how the NES manipulates the game's contents within the limitation of its memory's constraints in the remit of its sprite and palette systems. Thus, revealing the internal experiences of withdrawn objects for philosophical analysis and speculation.




Figure has been removed due to
copyrights restrictions

Figure 34: Ben Fry's *Deconstructulator* highlighting the sprite pieces and colour palette currently in memory during gameplay. Taken from Bogost's *Alien Phenomenology or What it's like to Be a Thing* (Bogost, 2012).

This has been an unorthodox introduction as the subject matter of philosophical carpentry was delved into rather quickly. However, this approach induced the sentiment and supposition of the following chapter by situating and adapting philosophy into design practice through the act of doing. This chapter synthesis and metamorphoses two disciplines, namely philosophy and design ideologies

and approaches towards a transdisciplinary approach for a MTHCD for AI technology. The following text describes the methods that constitute the method assemblage of this research as detailed in Chapter Three, *Groundworks* (Law, 2004). However, it is important to specify that RtD is the overarching approach circumfusing the method assemblage anatomised in this chapter that expands upon and positions the philosophical concepts in a manner to practice through design. The practice of speculation will be a principal technique defined in the following: it features heavily both in OOO, as noted, and in design through the practice of Design Fiction. Additionally, this thesis is the product of a designer who has a keen understanding of how AI works, operates, and functions and uses speculation as carpentry to philosophise AI rather than a data scientist or programmer who would turn to code as their primary material.

6.2 Constellations with a side of Onto-Cartography

Bogost encourages us to “understand objects by tracing their impacts on the surrounding ether” (2012, p. 33). To trace objects and their ecological relations, Bryant introduces the practice of onto-cartography, a model equivalent to Bennett’s proposal of assemblages, proposing mapping the ecologies of things; drawing attention to how they function, form relations, structure agential possibilities as a result of gravity (Bryant’s word for power) they exercise (2014).

Mapping, in practical terms, has been framed for designers by Coulton and Lindley using metaphorically the concept of ‘constellations’ to chart socio-technological concepts such as IoT (Akmal, 2021; Coulton & Lindley, 2019) and AI (F. Pilling & Coulton, 2020, 2021). The metaphor ‘constellation’ originated from the notion incited by the philosopher Walter Benjamin, whereupon “ideas are to objects as constellation are to stars” (1982, p. 34), inciting how the perspective of things changes depending on the observer’s perspective; transformable upon magnifying and changing the scope dependent on influences such as culture, awareness and beliefs to name a few (Figure 34).

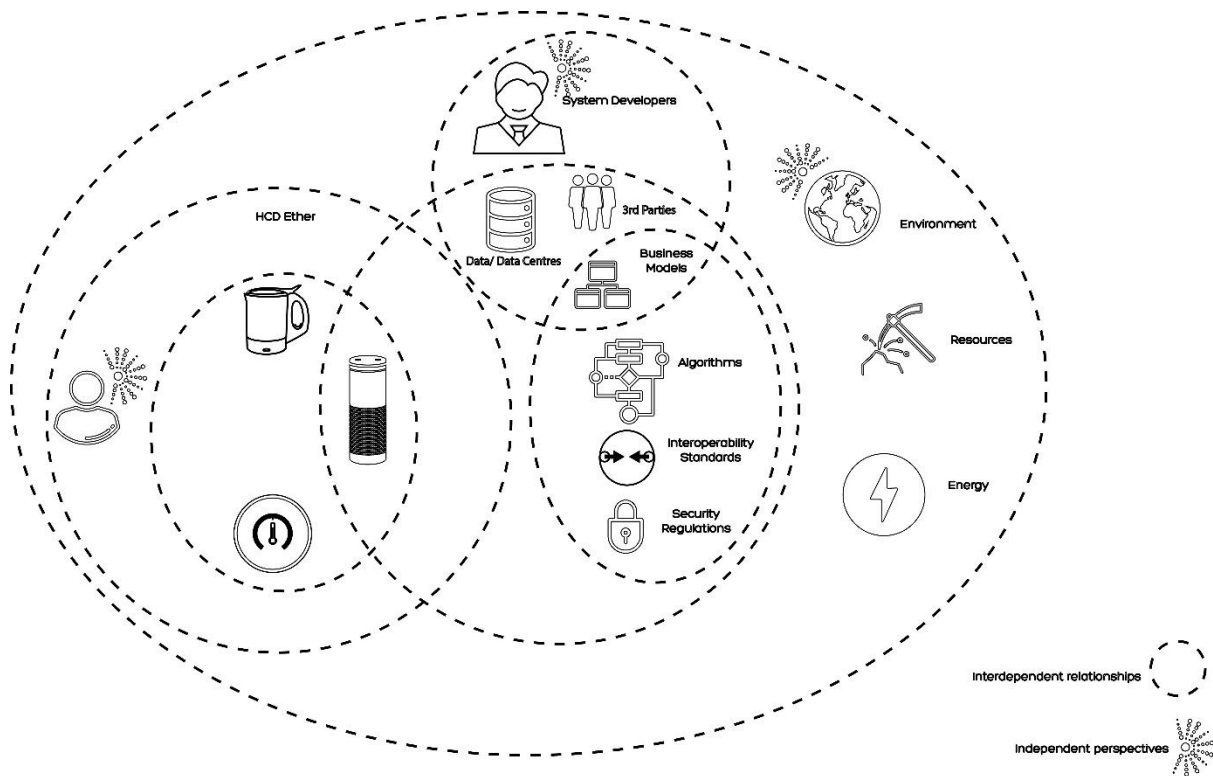


Figure 35: An example of the many possible Alexa constellations noting some of the possible independent perspectives and interdependent relationships.

In this regard, aspects can be out of view because of a particular constellation’s framing, such as a third parties influence on data collection. Although just because one cannot see a thing does not mean it has a significant impact on another thing’s operation, therefore aperture and depth of field for constellations can be modified to include objects that are of importance for any context or situation. Morton contends that this type of conscious mapping requires us to “join the dots” by thinking ecologically as everything is interconnected (Morton, 2010, p. 1). A process of embedding and meshing together complex interdependent relationships and independent perspectives of both human and non-human things (Coulton & Lindley, 2019) as active actants of power and efficacy (Bryant, 2014).

With this in mind, the emergence and proliferation of both IoT and AI technology have brought about a complex network ecology applying a “layer of code to much of the physical world” (Coulton, 2017, p. 192), described as the new era of the “electrosphere” (Dunne, 2005, p. 103, p.105, p.121). Resultantly, Redström and Wiltse introduce the concept of fluid assemblages to interpret digital things that are networked and subsequently dynamically and contextually configured (2018, p.

6)— in a sense, accounting for the different and re-figurative quantum causations that transpire, unlike the predictable(-ish) causations that occur in physical things of yesteryear. Therefore the mapping of constellations and the expansion of contemporary ecologies of virtual and physical beings results in an ecology of “Atoms and Bits” (Coulton, 2017, p. 192), presenting a design challenge of ensuring user perception and legibility for digital things.

Bryant asserts that the exercise onto-cartography highlights power structures, functions, and derived formations, providing ontological frameworks that can also be framed to consider political, ethical and design queries. As Bennett reminds us, “a vital materialist [/immaterial] theory of democracy seeks to transform the divide between speaking subjects and mute objects into a set of differential tendencies and variable capacities” (Bennett, 2010, p. 108). Without attentiveness to these things, we cannot thoroughly consider the manifestations of things and interactions within a constellation. Though, Coulton and Lindley show us that in practice, context-specific perspectives should be the focus of constellations to remain a beneficial insight for design purposes (2019).

6.3 Constellations for A Horizonless Perspective

Expanding the constellation model, the artist and writer Patricia Reed introduces the notion of “planetary scale”, inferring to a “big-world condition” in opposition to the (Western) human scaling and the idealised myth of framing the world ‘small’ (2019, para 3); a conclusion of the human-world correlation, thus amenable to human sensibility, as epitomised by Disney Imagineers in their song “It’s a Small World” (Sherman & Sherman, 1963). While the sales pitch for the introduction of the internet was presented as wielding the world into something more intimate through connection, Reed, however, highlights an obfuscation in the process –instead of reducing the world, it was cause for an expansion of a manifold of vectors, such as logic, economic, ecologic, and communication. Thus, pointing to what she terms an “increased *dimensionality* [(*n*th dimensionality)] of coexistence produced by exponentially multiplied vectors of relation”, whereas a small perspective of the world is a contained framing of the human condition (Ibid, para 1, original emphasis). For planetary consideration, Reed tells us that the “*n*th dimensional abstraction” creates an opportunity to reframe where the human stands within the planetary scale (Ibid, para 10). A repositioning, likewise, sought-

after in speculative-realism and MTHCD that forms a different human self-understanding usually associated with domination and knowledge.

Furthermore, Reed introduces new frames of reference for a perspectival shift and conceding the planetary scale by analysing the big world through ‘navigating’, a process akin to Morton’s ecological thinking of synthesising a web of connections. Reed explains that the process of navigation is the “ongoing mediation of intentionality with the contingency of unknown or accidental events” whereby “navigators can continually revise and adapt their makers of orientations” (Ibid, para 3-4). In this regard and as touched upon previously by mapping constellations, one can change the scope, however, keeping in mind Alfred Korzybski’s influential saying, “the map *is not* the territory” (1933, p. 750, original emphasis). Nevertheless, a map partially shapes the perception and perceptibility of a system, whereby Reed queries if *everything* can even be navigable, though one can speculate, evoking a “note of optimism infused with a [speculative] realist bent” (Reed, 2019, para 8).

As well as the expansion of scale, Reed also ascribes that the designer-navigator must preserve specificity to form “robust accounts of reality” as well as avoid a “rigid and *reductive picture* of totality” (Ibid, 11, original emphasis). This theme links to Reed's next frame of invoking “the discrete and the continuous” view, referring to the part-to-whole phenomenon (para 16). To make her point, Reed calls upon the mathematician René Thom’s discovery of topological notions in the writings of Aristotle, especially the “founding aporia of mathematics” –the notion of “the opposition between the discrete and the continuous” (Thom & Noel, 1991, pp. 81-82). Where Thom’s aporia accentuates the two modes, it also brings about a “relational glue” maintaining both discrete and continuous scales concurrently, enabling one to glean the spatial relations for navigating the planetary scale (para 15). This concept has the potential to be rendered in various ways when navigating a constellation of onto-cartography. First and foremost: it could be perceived as focusing on one thing in particular and its place within a system of things; secondly, it could be the framing of the constellation, of which the designer who made it would be able to acknowledge that the mapping of is the discrete capturing of an instance with a continuance in the greater beyond; finally, this notion can also be attributed to a constellation mapping of a discrete interaction between a user and an interface with a continuance in the digital space. This latter point incites both Redström and Wiltse fluid

assemblages and Akmal's Heterotopia (Greek for '*other place*'), a modal of spatial theory accounting for the interactions of unit entities both in digital and non-digital space, specifically detailing between private and public spaces (both digital and non-digital), forming many heterotopias between the two domains (Akmal, 2021, pp. 100-110).

Rather than insinuating another place, Reed, however, describes the phenomenon of distributed locatability, in which situations are "co-constituted by extra-local relations"; in other words, instances of localisation are a result of "chain reactions" across geolocated locations and things, over time or manifesting at once (Reed, 2019, para 16). Consider a user interacting with an IoT device with software capturing data points and pushing this through different networks of edge, fog, and cloud paradigms for data processing with AI; a user is rendered out in multiple locations – "a distributed form of situatedness" constituting in what Reed would term as, the planetary scale (Ibid). In view of distributed locatability, a designer can consider the forces and agency of things and, ultimately, the impact of distributed localisations that result in lived-localised experiences of AI processing a user's data that may have been involuntarily captured for reasons far removed from the conditions of the initial interaction. On this basis, Reed asks:

...how does the decentered human picture work back upon us as a form of diagrammatic agency, towards the way we come to account for situatedness in this *nth* dimensional frame of reference that is informed by, but irreducible to, the immediately concrete? (para 18)

To answer: Reed's work is oriented to draw her philosophical ideology diagrammatically.

Commenting on Reed's work, the artists David Barrows and Simon O'Sullivan describe it as the process of creating "philo-fictions", drawing the relations of capitalism, alienism and technologism using "philosophical materials *as material*" (2019, p. 327). In this manner, Barrows and O'Sullivan define philosophy as untethered to the standard philosophical rules, becoming a speculative *practice* rather than an analytical enquiry, enabling one to engage with surprising connections and conjunctions of the thing/s in question. Reed's practice and auxiliary conceptualisations mentioned

here have been synthesised with the previous constellation mapping, thus catering for a more detailed and explorative view for a MTHCD perspective (Figure 35).

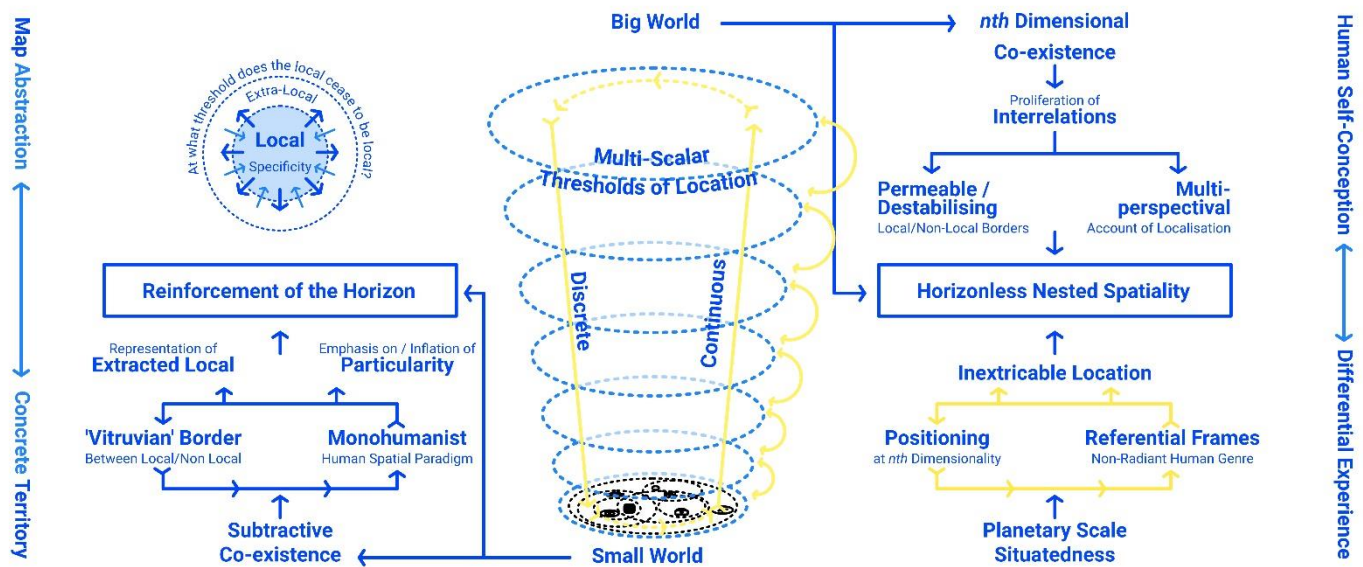


Figure 36: Constellations count as a small world reconfiguration as they are drawn up to map the assemblage of particular interest for design research, with the designer knowing that the points of interest have a big world impact beyond the constellation.

Finally, as an overarching principle for reframing the planetary scale and imploring the *nth* dimensionality, according to Reed, requires a horizonless perspective – a ‘big world’ frame of reference for coexistence not predicated on human-world myths. Consequently, the term horizon is a small-world expression, a mechanical scale irreflexive of reality and can only reflect small-world scopes of nearness, containment, and limitation. To speculate and hypothesize on the planetary scale's unfamiliar, opaque, and nested scales, Reed invests in a realist optimism view to mobilise vectors of *nth* dimensionality. The MTHC perspective in this research derives from a speculative realism movement, a variety of the realist movement and congruous to speculate with a horizonless perspective enabling a designer to navigate paths and at other scales out of the Anthropocene, although inclusive and accountable to human positioning. As Benjamin Bratton states, “design scaled to the scope of the real, not reality down sampled toward the digestible” (2016, p. 15)

6.4 Alien Phenomenology

A crucial component to the practice of onto-cartography Bryant explains— is Bogost’s practice of phenomenology (2014, pp. 54–73), with roots in traditional phenomenology (Kant, Husserl, Heidegger), though transcends it by employing Harman’s OOO to speculate how other

entities such as mosquitoes, computer games, institutions, phones, trees, bits, atoms, rocks, buildings, encounter the world about them (2012).

Alien Phenomenology is comparative to other exploratory methods of experience, such as Ethology, which Uexküll's theory of Umwelt describes as seeking to observe the world as experienced by animals (Uexküll, 2010). Nevertheless, being presented with the idea of attempting to understand what it is like to be AI will be met with objection similar to Thomas Nagel's seminal stance as clarified in his thesis *What's it like to be a Bat?* (1974) In That, we cannot know the 'subjective character of experience' for a thing. Yet, Bogost's Alien Phenomenology, in a like manner, accepts that the experience of a thing can never be fully known; instead, the only way to perform Alien Phenomenology is via metaphor and analogy. For example, a bat's mounted cries perform like a sonar and can be compared to operating like a submarine (2012, pp. 61-84). The practice is submerged in speculative realism and asks us to suspend our own human aims to investigate the aims, if any, things have. Though Bogost warns that the risk of anthropocentrism is strong, however unavoidable – our metaphors and speculations will always be imperfect. However, Bogost refers to Bennett's notion that anthropomorphic metaphors for things and their unit operations help us accentuate the differences between ourselves and things (2010, p. 120). As Bogost identifies, "it helps remind us that object encounters are caricatures" (2012, p. 65), emphasising Harman's ideology that we and objects only experience the Sensual-Qualities and interact with the Sensual-Object of things.

The crucial aspect for Bryant's is that Alien Phenomenology affords the opportunity to make an inference about the "flows" (Bryant's word for inputs) a thing is structurally open to and the manner in which a thing operates on flows that pass through them, and so on; with our knowledge of flows growing daily owing to the invention of instruments that detect flows invisible to us, such as ultraviolet, radiation, and WIFI receivers (2014, p. 63). Bryant also attests that things can "seduce" and "want", a type of agency that manifests a flow themselves inspired by Kevin Kelly's thesis *What Technology Wants* (2010). Kelly writes that technology responds and unfolds to certain vectors or propensities irreducible to the reasons they have been developed, suggesting that the technology "wants" something despite having no consciousness. Instead, it is an argument invoking something

similar to an evolutionary logic for design with tensions arising from technology through materials used, code, political issues etc., propelling technology in one direction rather than the other, devoid of a designer or teleology.

Resuming Bogost's point of view, he voices, "[*t*]hat things are is not a matter of debate. *What it means that something in particular is for another thing that is*: this is the question that interests me" (2012, p. 30, original emphasis). What Bogost is accentuating here is that objects *exceed* what we know or ever can know about them, a notion that is reinforced by the fact that the meaning or the unit operation of one thing to another differs, which cannot be explained through natural law, science or even from a thing's own perspective. Therefore, to critically consider a thing or its unit operation, one must execute the practice of speculation within the realms of OOO.

6.5 Speculation and Design Fiction

Throughout earlier chapters of this thesis, the practice of speculation has been presented as the progressive and actionable way to practice OOO, particularly since the philosophy is cast in the light of speculative realism. Bogost reminds us that speculative realism not only condemns the notion of correlationism and considers existence separate from thought but also that things speculate (2012, p. 31). Fortified by that impression, the designer-philosopher's job is to document and speculate on the state of a particular focus in question "using educated guesswork" (Ibid) synthesised with the approach of Design Fiction. Forming an assemblage between OOO and Design Fiction is an intuitive metamorphosis for a MTHCD approach to engage AI as a material for design. Also befitting the sights of speculation is the digital nature and current state of AI; due to the field's rapid and daily innovation, there is always apprehension about AI's near-future: what will tomorrow bring? To showcase the MTHCD approach in action, after a brief explanation of Design Fiction, the following section will showcase a speculative thought experiment using a philosophical guinea pig – Amazon's AI assistant *Alexa* and its Skills service.

6.6 Design Fiction: An Overview

Speculative Design is an umbrella term for design-related activities that involve some form of speculation, such as Critical Design, Discursive Design, Design Probes and Design Fiction (DF)

(Auger, 2013). The designer James Auger explains that the core motivation of these practices is to shift the discussions of technology beyond the fields of experts; he goes on to explain that each speculative design activity is informed by their “semantics and the subsequent loading of experience” into a speculative artefact (Ibid, p.11). For instance, ‘discursive’ and ‘critical’ infers debate, ‘probes’ suggest investigation, while ‘fiction’ communicates to the viewer that the object is not real.

DF is still in its formative years, where the field has been described as “enticing and provocative ...yet it still remains elusive” (Hales, 2013, p. 1). Despite the (forthcoming) ten-year gap of the latter statement, it still reflects the current range of contending theories, understandings, and approaches leading to ambiguity in the field; however, akin to the variability found in design research, this circumstance creates opportunities for new methods to be established and discussions of –how to practice— DF. 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 and explorative space (Dunne & Raby, 2013; Lindley, 2016), that is increasingly used in commercial (Bassett et al., 2013; Michaud, 2020, pp. 137-139) political (Pólvora & Nascimento, 2021) and research approaches, surpassing its academic inception (Bleecker, 2009; Dunne & Raby, 2013). In this research, the position and undertaking of DF is that of ‘World Building’ (Coulton, et al., 2017). To understand this method, the following will clarify the theory that supports a World Building approach by reviewing DF’s brief history, leading to justifying a method for Design Fiction as philosophical Carpentry.

6.7 Design Fiction as World Building

The term DF was coined by the science fiction author Bruce 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 DF is “hands on”, “practical”, and plausible with the unique ability of getting to the core and “the glowing heat of the techno-social conflict” (Ibid). Sterling went on to advocate that the practice is “the deliberate use of diegetic prototypes to suspend disbelief about change” (Sterling

quoted in Bosch, 2012, para 3) and has since become the oft-quoted theoretical underpinning for the field.

A principal component of DF are “diegetic prototypes”, a type of prototyping coined by David Kirby 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 in status (2010). Kirby’s theory of diegetic prototypes was highlighted, along with other theories by Julian Bleeker in his influential and catalytic essay on DF (2009), as a central methodology, noting the film *Minority Report* (Spielberg, 2002) as a compelling example of using diegetic prototypes. Sterling’s rationale for diegetic prototypes owes much to Bleeker’s thesis, though as Sterling also defined it, in the same sentiment, DF “tells worlds rather than stories” (Sterling quoted in Bosch, 2012, para 3). Bleeker goes on to say that:

...the most compelling Design Fictions are very much like ephemera from possible worlds – *symptoms* of macro-scale change that represent *implications* rather than make *predictions* (Bleeker, ND, para 3).

An important fact to note at this point is that the emphasis on the –story can *stifle* the flexibility of DF 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 this thesis; however, to clarify the matter, Raven and Elahi specify that the “story is not the world” (2015, pp. 52–53). Rather, a DF aims to depict a thing belonging to a contextual world, and to –tell worlds –is the act of narrating; therefore, DF 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, p. 167) 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 a DF

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 36).



Figure 37: This diagram aids in communicating how both world building and diegetic prototypes help synthesise one another (Coulton et al., 2018).

To assist with understanding this approach to DF, two metaphors developed by Coulton et al. assist in understanding how individual artefacts relate to the conceived world (2017). The first requires the DF 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 of a metaphorical entry point into the fictional world (Figure 37) (Ibid).

Figure has been removed due to copyrights restrictions

Figure 38: Artefacts at different scales create a richer and more detailed fictional world (Coulton et al., 2018).

The second metaphor, which works with the first, considers shifting scale, inspired by Charles and Ray Eames' film *Powers of Ten* (1968), with each artefact representing the fictional world at different scales (see Figure 37 also). With the building blocks for designing fictional worlds justified, the following section concerns the type of future represented or designed.

6.8 A Philosophical Interlude: Philosophical metamorphosis through Worlding Constellations

For DF's metamorphosis, this research turns to the artists' Burrows and O'Sullivan observations, who wrote a theoretical and insightful account for a process they term 'Fictioning' that mediates the increasingly technological reality operative in the here, now and immediate future (2019). Underpinning their theory, the authors define three sets of fictioning practices that overlap in what they call "myth-functions"; these are Mythopoesis, the fabrication of worlds detailed with people, milieu and communities to come; Mythscience considers technics of non-human and alien perspectives and models of diverse presentations of being; and finally, Mythotechnesis; the most fantastical fictioning practice, echoing the science fiction narratives of human-machine symbiosis and the Singularity. Coupling these practices with the previous analysis, Mythotechnesis correlates to the creation of narratives emulating that of *Blade Runner* (Scott, 1982), *Matrix* (Wachowski &

Wachowski, 1999) and *2001: A Space Odyssey* (Kubrick, 1968) –galvanising AI’s definitional dualism; Mythopoesis corresponds to World Building by submerging elements of a secondary world camouflaged in the mundanity of lived-experience; and, Mythscience is concerned with speculating on worlds beyond a human-centred viewpoint by employing, amongst others, the interdisciplinary feminist and postmodernist positioning of Haraway (2016) and her call to choose non-humans as our kin (Burrows & O’Sullivan, 2019, pp. 255–293).

The fictioning practice of Mythscience also employs a form of World Building, or rather, as Haraway commits, the practice of ‘Worlding’— a method that transcends the designing of familiar themes of a world (society, government, biological systems, technological innovations, and law) in anticipation of fictioning and speculating how relations between things and entities could manifest: rejecting human exceptionalism. Haraway’s use of the term worlding challenges Heidegger’s, whose use meant the opening of new ways of being-in-the world, of being in time and history, a world produced by human existence. As reasoned by Heidegger:

Plant and animal likewise have no world; but they belong to the covert throng of a surrounding into which they are linked. The peasant woman, on the other hand, has a world because she dwells in the overtness of being (1999, p. 170).

In contrast to Heidegger’s worlding is Haraway’s feminist worlding addressing more than the ontical of things, thus rejecting the story of ‘being’ as solely human or dasein. In her worlding, Haraway is “staying with the trouble”; a reflection of the entangled ecology of all things, inclusive of humankind, as humans play a significant and consequential part in worlding worlds (Haraway, 2016). In her book *Staying with the Trouble: Making Kin in the Chthulucene*, Haraway promotes unexpected collaborations and combinations through human and non-human symbiosis by embracing non-human otherness “for learning to stay with the trouble of living and dying in response-ability on a damaged earth” (Ibid, p.2).⁵² Although Haraway’s imaginary is not human-centric or a product of unruly imagination, rather it is a “inhuman social imaginary” (Burrows & O’Sullivan, 2019, p. 276), which

⁵² Opposing the concept of Anthropocene, which is speculated to end in catastrophe, the idea for Chthulucene is sym-poiesis, making with other non-humans providing the means for more liveable futures in an already heavily damaged earth. Rather than auto-poiesis, or self-making man that leads us to “tragic system failure, turning biodiverse ecosystems into flipped-out deserts of slimy mats”(Haraway, 2016, p. 47).

can be interrupted as many examples of multispecies relations models bathed in knowledge from Haraway's background in biology and indeed educated guesswork from anthropology that opens up a discourse on "making kin, not babies" to take care of the planet.

In much the same way as the onto-cartographical constellations as detailed previously, Haraway's multispecies worlding for "transdisciplinarity inspection" is "inflected" through what the authors call Science Fiction (shortened to SF). SF scholarship is a collaborative practice for giving ideas of symbiosis with non-humans, stories, and shapes, in which Haraway defines string figures as "constellations" communicating "relations of the worlds, including... relations of humans and nonhumans. Not *in* the world, but *of* the world" and inclusive of networks and processes of which humans are often not conscious of (Haraway, 2011, pp. 14–15). A practice influenced by Navajo traditions that abolish the Western separation of humans from non-humans as promoted by Heidegger's philosophy. Resembling that of Ingold's demonstration, as mentioned earlier, of manipulating string to know and understand that being of string (Chapter Four, *Methodologies*), these string figures are a worlding game of thinking and making, done with one or two hands (or all sorts of tentacular things) that patterns, relays, assemble connections and alliances of the thing in question (Haraway, 2011, 2016). For this research, the mapping of constellations and the DF created will employ the spirit of Haraway's worlding to establish opportunities of Human-AI Kinship.

6.9 Framing 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, Joseph Voro’s ‘future cone’ (2003) is often utilised as it presents a taxonomy of scenario qualifiers to mediate the types of futures, which are probable, plausible, possible, and preferable (the 4 P’s) (Figure 38).




Figure has been removed due to
copyrights restrictions

Figure 39: This diagram shows the trajectory of different types of futures, including wildcard futures. Diagram appropriated from Voro (2003).

The futures cone, however, has been criticised as its qualifiers, for the different ranges of future possibilities promote more questions about what the underlining meaning of the qualifiers are, with some advocating that the cone has missing qualifiers. Designers often ruminate over the preferable qualifier, as one could argue that naturally, a designer’s role is to bring preferable results and outcomes into existence, as much as consider their own biases (Coulton et al., 2016). However, the designer Simon Bowen argues that a preferable tenet promotes and consumes “elitist views of a ‘better world’ that society should aspire towards” (2010, p. 4), such as promoting privileged advantages of the Global North, which has been known to cultivate oppression intersectionality (Martins, 2014). Although while Antony Dunne and Fiona Raby advocate for the preferable futures,

they question in the same sentiment as Bowen, notably asking, “what does preferable mean, for whom, and who decides” (2013, p. 4). Dunne and Raby virtually circumvent this notion by promoting that Speculative Design, Critical Design and DF are concerned with “not to show how things will be but to open up a space for discussion” (Ibid, p. 51). Nevertheless, it continues to be questioned –if showing a singular future vision under the preferable banner be the best way to stimulate discussion?

A proposition for DF to be an effective practice and research tool is to consider and be presented with multiple futures to develop more ‘representative notions’ of what preferable may be, catering 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, capturing the many “variations or blendings” to be found or even “behind or beneath” (Raven & Elahi, 2015, p. 50) the present future’s qualifiers, to consider futures beyond what we can imagine easily. Examples range from “alternative presents” (Auger, 2013), “Wildcards” for low probability events (Voros, 2017), “black swans” (Taleb, 2007) for unclassifiable events (Voros, 2017), “impossible and lost futures” (Coulton et al., 2016) for concepts beyond scientific knowledge and at the moment considered fantasy although useful to consider the world (Gualeni, 2015). The latter point is an approximate classification for the speculative and philosophical thought experiments in this design research, which will be interrelated with presenting legible (and perhaps conceivably preferable) futures for AI technologies.

Finally, it has also been observed that the cone fails to acknowledge the influences of the past (Coulton, 2020) or incorporate possible futures from fiction (Gonzatto et al., 2013) and how these variables impact our perception of time. Marshall McLuhan famously wrote, “We look at the present through a rear-view mirror. We march backwards into the future” (Fiore & McLuhan, 1967, pp.74-75). This idea reminds us that there is no universally accepted view of the past, present, or future as individuals assemble their own subjective reality (Law & Urry, 2004; Raven & Elahi, 2015). Here, we can also draw upon Arturo Escobar’s attention to acknowledging the different lived experiences of cultures, communities, and individuals globally (2018), resulting in considering a plurality of various perspectives on past, presents and futures within the design process (M. Pilling, et al., 2022c) (Figure 39).

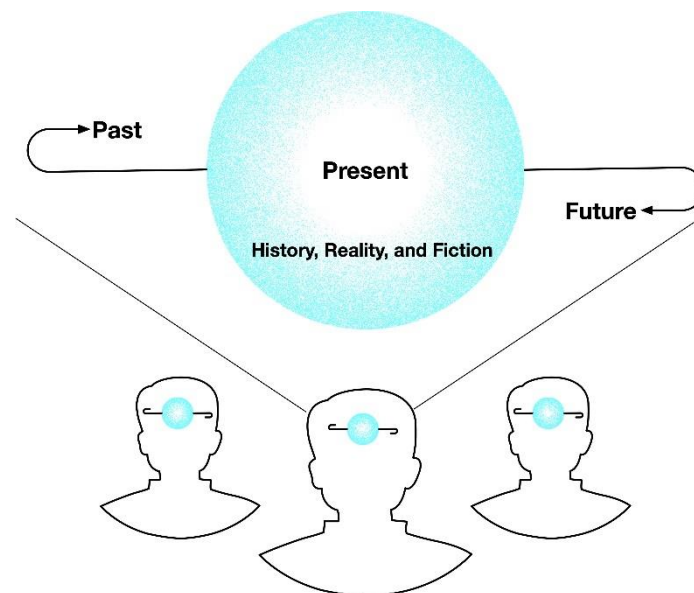


Figure 40: This futures ‘cone’ has been adapted and integrates Gonzatto et al’s. (2013) research, whose hermeneutic model represents the ‘interpreted present’ as an interplay between past, future, reality, and fiction.

Moreover, in his book *Defuturing: A New Design Philosophy*, Tony Fry stresses the roles of designers in constituting undesirable and unsustainable futures, arguing: “we act to defuture we have very little *comprehension* of the complexity, on-going consequences, and transformative nature of our impacts”; we do not understand “how the values, knowledges, worlds and things we create go on designing after we have designed and made them” (2020, p. 10, additional emphasis). Subsequently, Fry attempts to create a foundation of thought and practice needed by articulating the condition and trajectory of our times, a method he calls ‘Defuturing’ illustrating the erosion of relations between things, ‘material and immaterial’ and our self-centeredness with “actions that have come to be a

defining quality of our species” (Ibid). In his foreword to Fry’s work, Clive Dilnot underscores the concept of ‘comprehension’, describing it as “both stand[ing] against the condition of Defuturing and becom[ing] the possible basis of futuring”; a “transformed model of understanding of what-is” (Fry, 2020, p. XIV-XV) as recognition of “accepted responsibility” through design, despite being locked into anthropocentrism (Ibid, p. 37).

On reflection, then, is a calling for a futures cone that is adaptive and symptomatic of the considerations presented here (Coulton, 2020) and ultimately a flexible platform to research for new insights in a transdisciplinary fashion; in other words, a futures cone efficient in suspending disbelief at a planetary scale for philosophical investigation. In a continuation of framing futures, the following section will summarise the method for the mundane rendering of technology; that aims to conjure an impression of realism to induce cognitive estrangement in DFs.

6.10 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 strengthens the design by immersing it in just enough reality to create opportunities for discourse. Mark Blythe and Enrique Encinas (2018, pp. 84-85) highlight Tolkien’s methods of world building for fantasy he outlined in his essay *On Fairy Stories*, where he described the story as a “sub-creator ... a Secondary World which your mind can enter” where suns can be green and credible by “commanding secondary belief” with our world operating as a “primary” anchorage for suspending disbelief. In essence, however fantastic the secondary world may be, a connection to the primary one must be established for the potency of secondary belief. In practice, Coulton et al. advocate for a blurring of the boundaries betwixt a viewer’s context and the diegetic context (2017, p. 15)

To achieve this, as noted previously, Coulton and Lindley conclude that *multiple* diegetic prototypes –incarnated in any media– construct a worldview that simultaneously envelopes into a comprehension of the designed artefacts and the world they belong to (2017) (Figure 40).



Figure 41: Multiple artefacts construct the world at different entry points. Appropriated diagram from Coulton & Lindley (2017).

Indispensably, part of this method is to also render these futures as mundane, where the diegetic prototypes do not exist in a vacuum, yet are supported by and blends into their surroundings, taking inspiration from science fiction, popular media, futuristic tropes, memes, and recognisable forms coming out of Silicon Valley. Examples include speculative product films and images, device documentation, manuals, and patents to immerse in with our lived experience, going beyond the remit of a textual description of these worlds (M. Blythe & Encinas, 2018, p. 34). A prime example of this method is the speculative *Ikea Catalogue* by Near Futures Laboratory showcasing, in the recognisable rendition and visual style of a genuine Ikea catalogue, what a technological future home may be (2015) (Figure 41). By association and developing this idea further is the project *Living Room of The Future*, whereby instead of speculatively projecting emerging technologies into potential futures, this work anchors it to the here and now through an experience co-produced by audience members interacting with a smart and functional (due to wizard of Oz-ing) living room of the future (Coulton, et al., 2017).




Figure has been removed due to
copyrights restrictions

Figure 42: The Near Future Laboratory's Ikea catalogue looks just like a real Ikea catalogue but with a glimpse of the future with the addition of gardening drones (2015).

A final point to draw upon is the notion of “Vapourware” and “Vapour-worlds”, terms used to describe future evoking materials produced by commercial entities and organisations to assert themselves and the products they make as integral parts of the future (Ibid). These vapour-visions have a knack of representing technologies as if they are domesticated and mundane, exploring futures subtly and strategically by positing ‘preferable’ future visions which are puppeteered for commercial and often political gains (Coulton, et al., 2017). Early vapourware visions can be seen at World Fairs and World Expositions that appeared in the 19th century to present the technical prowess of Western countries (Figure 42).⁵³ Corporate ‘vapourworlds’ are often created, although they are never intended to become a reality, to leverage public brand perception, and to increase investments.

⁵³ These world fairs are now more often happening in the Middle East to show their wealth and trajectories into the future.

Figure has been removed due to
copyrights restrictions

Figure 43: World Fairs were built to be temporary insights into the future (Comstock, 1964).

In certain respects, the method of vapourworlds has contributed to the culture of “fake it till you make it” rife in Silicon Valley. Elizabeth Holmes prompted a famous example with her tech start-up Theranos Inc. Holmes was recently found guilty of fraud on a variety of accounts made possible through a ‘wizard of OZ’ performance of the Edison machine to investors, which was designed to perform immunoassays which look for the presence of an antibody or antigen from a single drop blood or fluid (Figure 43).^{54 55}

⁵⁴ For further information on Theranos Inc and Holmes see John Carreyrou book *Bad blood Secrets and lies in a Silicon Valley startup* (Carreyrou, 2019) and HBO’s documentary *The Inventor: Out for Blood in Silicon Valley* (Gibney, 2019).

⁵⁵ As a satire point, you could for a limited time get a toy version of Edison machine as part of the MSCHF art collective’s Dead Startup Toys series (MSCHF, ND)




Figure has been removed due to
copyrights restrictions

Figure 44: Even though this looks like a fully functional piece of technology it is not. Looks can be deceiving and that is precisely what makes great diegetic prototypes (Wilson, N.D.).

6.11 Design Fiction for a new material palette

With the fundamental factors of DF outlined, the following part of the chapter continues to metamorphose and blend DF with philosophy as a way in which to address the gamut of transformative new ‘materials’ that are atypical from the expectations of materials we are accustomed to, explicitly the Nanotechnology, Biotechnology, Information Technology and Cognitive Science suite of technologies. The social and ecological project that these new materials propels, as detailed by Bratton, is “the recomposition of the world at scales previously unthinkable”, whereby the cohort of speculative design approaches needs to register these new material matters and conceive a contemporary and philosophical material system to map the potentials while considering their inhuman scopes and frames for designs that “ratify the organization of society” (2016, p. 5).

What DF offers for designing with AI is the ability to speculatively probe the application and implementation of, amongst the examination and designing of, ‘better’ user interactions while reflecting on possible future technological ranges. Rather than a positivist approach to consider AI via science paradigms restricted by stringent disciplinary permissions, DF grants a platform for “ideation and discovery” (Ibid, p. 12) and customisation through the integration of philosophical and theoretical

lenses. In this regard, DF is seen as a type of engine to kindle a speculative reality to perform Carpentry, which will be demonstrated next using *Alexa's* skill service as an appropriated *Diegetic Thing* (F. Pilling & Coulton, 2021).

6.12 Carpentered Diegetic Things

The constellation (see Figure 34) was an essential means of identifying the relevant actants and focal points to consider the material relations and gather intel to accurately represent *Alexa's* operations for World Building. Such as Amazon's Web Services (data centres), back-end AI services including Automatic Speech Recognition, various provider's business models etc. In a generative manner, the process of mapping the constellation catalysed the idea of appropriating the Skills function, using the well-established voice interaction of *Alexa*, in a manner that could potentially provide greater legibility and agency for the user via the speculative application of communicating salient and consequential information to the user. An act recognised and conceived through the lens of OOO and balancing the practical constraints against possible design choices.

To make an educated speculation: it is helpful to know that Natural Language Processing (NLP) is the foundation of *Alexa's* operation and is a merger of ML and computational linguistics. NLP enables *Alexa* to analyse, 'understand' and generate a response using data sent to Amazon's services for analysis. Located on the cloud, the Skill Service is coded by a developer to determine what actions to take in response to a user's request. The NLP that enables the skill is 'abstracted' from the developer, and their task is to define 'Intents' –answers expected from the user and 'Utterances', which predict the varying responses of anticipated intents. An expected intent triggers the 'intent handler' and returns a planned vocal response and output to the Echo device, which runs the program *Alexa*. This is by no means a comprehensive explanation of the operation of *Alexa's* AI; however, it highlights why design solutions for AI legibility are required, as existing AI functions are often black boxed behind corporate firewalls. However, there is enough information for this research to create a constellation and design a solution via the Skills Service. Or an OOO interpretation of *Alexa's* Skill service provides a palpable means to explore and attempt to answer the question— if it were possible to converse with *Alexa's* being, what would it say about its ontology?

While someone may argue that the idea and framing could potentially evoke a type of definitional dualism, the speculative Skill ‘gives’ life to *Alexa’s* being theoretically through back-end programming to provide ontological information about *Alexa’s* AI function and operations –akin to a computer’s system report, although packaged in a more user-friendly manner. So-called ‘Frankenstein’ after the protagonist Victor Frankenstein in Mary Shelley’s 1818 novel *Frankenstein: or, The Modern Prometheus*, who, through examination of chemical processes, developed a comprehension for the creation of life and thereafter gave life to his own creature, who has no name.⁵⁶ The DF ‘entry points’ can be conceived by exploiting Amazon’s visual identity through an advertisement campaign and how-to manual for the *Alexa* Frankenstein Skill (Figure 44).

⁵⁶ Although ironically the image for the skill is that of the monster, echoing a common mistake of calling the creature Frankenstein.

SAFE HOME
With Alexa, you know when something technical changes regarding your personal data.

"Your audio recording from yesterday at 18:33 is manually being reviewed in order to improve NLP algorithms"

BBC "Alexa, play BBC news"

Frankenstein "Alexa, what's your ontology?"

R "Alexa, open Radioplayer"

Frankenstein "Alexa, can you read my mind?"

IN CHARGE OF YOUR SMART HOME

1. SELECT SKILL
Discover skills in the Alexa app or online at the Alexa skills store

2. ENABLE SKILL
Click on enable or just say: "Alexa, open 'Skill name'"

3. USE SKILL
To use your skill just say: "Alexa, open 'Skill name'"

Figure 45: Akin to the Ikea catalogue this diegetic prototype uses familiar cues and visualisations of a typical Amazon advertisement, with the Frankenstein app part of the app range anyone could speculatively get.

6.13 The More Than Human Centred Design approach for consideration of AI as a Material for Design

This research, up until this point, has taken the reader through many different theoretical frameworks and ideology that together were critical in constituting the unique assemblage of the MTHCD approach for this research. The beginning of the thesis took the reader through the state of ‘Seeing AI’, which grounded the design research by disclosing the ways in which AI is perceived, in terms of machine intelligence and artificial general intelligence. This was an essential step in establishing a MTHCD approach and understand (from a human perspective) what AI is beyond prominent science fiction renderings. The thesis then went on to explain the methods by which this research would ‘Understand AI’, by way of developing the unique method assemblage of this research and by conducting RtD, practiced by adapting philosophy through design as demonstrated in this chapter. However, before delving into the practice of speculation and adapting philosophy for design, the thesis discussed in detail the philosophical concepts of OOO, Materialism and Postphenomenology, developing the theoretical framework for Human-AI Kinship, taking the reader through the state of (speculatively) ‘Being AI’.

The MTHCD approach has been developed in this chapter by showing the reader how, through design, metamorphosis of philosophy can be achieved. In regard to this research the metamorphosis of philosophy was to perceive AI as a material for design through the case study of AI legibility. In practice this was accomplished through developing constellations of AI products (akin to figure 35, p.182) which through mapping the AI’s being in terms of independent and interdependent relations with other things brought about the idea of mapping the ontography of AI’s being with key insights into relations that would impact functionality. Going back to Reed’s thesis and mapping constellations for a horizonless perspective, the idea of mapping AI’s ontology was seen as a ‘discrete mapping’ of a small world rather than the big world mapping of the previous constellations. In other words, it was the challenge of mapping on different planetary scales, creating multi-perspectival accounts.⁵⁷

⁵⁷ The design research of mapping AI’s ontology is similar to the design research of Kate Crawford and Vladan Joler who by all accounts mapped the anatomy of an Alexa system, mapping on a small scale in terms of the

With an understanding of AI's attributes (which is discussed in detail in Chapter 7 7.10 *Defining the Interpretant: AI Attributes, Dimensions and Properties (AI'S Ontology)*) and the way in which the constellations were visually mapped with icons as signifiers of concepts/ products/ users/ things/ networks etc brought about the idea of developing a visual lexicon to map AI's being which in turn founded the idea of developing iconography to aid in communicating AI being to users. Furthermore, returning to Bogost thesis, the icons that are the result of this design research are individually performing carpentry on AI's being by tracing the unit operation of AI.

The next part of the thesis is concerned with designing for Human-AI Kinship, in which this chapter on adapting philosophy for design, starts to introduce the reader on how one practically engages with the theoretical frameworks developed thus far. Furthermore, the next chapter will discuss the semiotic design of the icons and delve into AI's attributes to map the ontology of AI's being.

6.14 Conclusions

This chapter concludes the MTHCD approach and the principal methodologies of this research. Until now, this thesis has been outlining and developing a unique method assemblage for design research into AI due to the obscure and seemingly coded intentionality of AI that goes beyond current human comprehension, owing conceivably to the dominant small-world view favouring human-centric considerations. In response: an alternative method for viewing and speculating on AI for design was formed, resulting in the metamorphosis of a philosophical model for AI and a method by which to 'do' philosophy via the speculative practice of DF. Thus, enabling a designer to perform worlding and speculate on nonhuman beings in the purview of a planetary scale while simultaneously being mindful of human users, forwarded in this research as a Human-AI Kinship through an adaption of postphenomenological design ideologies. To a degree, an iterative RtD process has already occurred in the coalescence of theories, methods, and ideologies into a balanced transdisciplinary approach for design research, which has sanctioned a speculative post-anthropocentric approach

materials and planetary resources the Amazon Echo is made from and mapping 'big world' notions such as the Internet infrastructure Alexa taps into (K. Crawford & Joler, 2018).

despite being locked into a human-world correlate. To this end, the combination of the various design approaches with philosophy has been strategic in its formation, as the rigorous methods and views of HCD and HCI would not necessarily commit to the interpretations of OOO, just as a pure philosophical approach would allow for an unconventional adaptation of its stringent theories.

The subsequent part of the thesis reintroduces the focus of AI legibility through an explanation of the notion, which will lay down the foundations for a detailed explanation of designing AI iconography intended to communicate and encapsulate AI's ontology. In this regard, the outlined MTHCD approach will be the cornerstone RtD process and the philosophic perspective in which non-human things are empathically and speculatively probed. A discussion will follow of the series of workshops completed to stress test the legibility of the icons themselves and investigate the viability of AI icons with intended users, finalising a Human-AI Kinship design program.

Chapter Seven Designing for AI Legibility

(Designing for Human-AI kinship)

7.1 Introduction

The notion of legibility was briefly mentioned in Chapter Two, *Seeing AI*, while also highlighting the variety of interpretations and definitions for concepts of explainability, interpretability, and transparency in the context of AI. This chapter will begin by giving a detailed review of all four concepts, clarifying why technical legibility in human-AI interaction is a crucial characteristic to strive towards in our increasingly technologically mediated world. In succession, an account of the RtD process for designing a system of AI icons utilising the previously outlined MTHCD approach will be detailed, presenting an alternative treatment of communicating the ontological constitution of operational AI for AI legibility. This part of the thesis exhibits a synthesis of working knowledge of AI operations with an OOO perspective and interpretation via the semiotics field, which will also be briefly explained in the following chapter.

7.2 Explainability, Interpretability and Transparency

As identified, AI systems are increasingly deployed in mission-critical and governance roles, such as credit scoring, predicting criminal reoffending, and curating news, amongst many others. A problem arises, typically, when AI operations and outputs are generated by searching vast scores of data points from various action spaces that are placed into sequential learning programs comprised of opaquely optimised neural networks. For this reason, AI researchers Daniel Weld and Gagan Bansal accentuate the complexity of AI's functionality, noting “[a]lmost by definition, no clear-cut method can accomplish these AI tasks”; further observing “AI-produced behaviour is alien, that is, it can fail in unexpected ways” due to the brittleness and unintelligibility of a system's behaviour (2019b, p. 70). Consequently, these researchers seek effective control of AI systems by explaining decisions to ‘users’ by mapping results onto simplified and explanatory models. In this scenario, however, and seemingly the bulk of AI literature, the term ‘users’ represents AI experts and engineers working with AI rather than end users, who are characteristically unaware of the implications caused by AI processes. A report on creating explainable systems by the Royal Society also draws attention to the inconsistency of terminology and meanings found in the literature. Instead, it frames the problem in such a way as the spectrum of individuals' needs and the variety of diverse contexts require “different

forms of explainability” (N. McCarthy & Montgomery, 2019, p. 9). This section presents a thin slice of the overlapping terms concerned with communicating aspects of AI systems to users and clarifies the audience methods are aimed at, either AI experts or non-experts.

The concept of explainability has been defined as producing methods by which AI functions are translated into “intelligible, comprehensible formats suitable for evaluation” (Fjeld et al., 2020, p. 43), thus, composing a core tenet for the budding research field of ‘eXplainable AI’ (XAI), paying particular attention to developing methods for human control and expert evaluation (Arrieta et al., 2020; Weld & Bansal, 2019b). Nevertheless, computer scientist Cynthia Rudin specifies that the term ‘explanation’ is often misleading, where in reality, an “approximation” is formed of an AI’s operation through summaries of predictions, statistics or the plotting of trends, rather than an explanation of what the model is doing (2019, p. 4). Moreover, Diakopoulos emphasises that some mechanisms of algorithmic systems can potentially never be revealed as they cannot take on observable or human-intelligible forms (Diakopoulos, 2016), consequently fostering a phenomenon whereby the majority of AI systems are inscrutable to their creators (Burrell, 2016; Hutson, 2018).

Two additional XAI principles are transparency and interpretability. Charting the consensus of transparency, both Zachary Lipton and Fjeld et al. explain it as a concept that asserts AI systems should be designed in such a way that oversight of their operations is possible at all functional stages. For this reason, transparency is connected and overlaps with numerous themes, especially accountability, which also has profound connections to the themes of safety, security and trust (Ananny & Crawford, 2018; Diakopoulos, 2016; Fjeld et al., 2020; Lipton, 2018). Of note: the theme of trust (users’ trust) is frequently mentioned in the context of transparency, as emulated in the recent Ada Lovelace Institute’s report *The Rule of Trust*, wherein amongst a set of seven principles, transparency was included to aid in trustworthy data governance in pandemics; noting, “clear and consistent communication around the use of data-driven approaches” (Ada Lovelace Institute, 2022, p. 4) and an “understanding [of] the parameters of these technologies” (Ibid, p. 12). Trust is a fundamental element in the relationship between users and technology. However, it is a complex and elusive concept, so it is beyond the scope of this research to compile a thorough review of the topic at this time. Briefly, however, Ribeiro et al. discuss the importance of trusting the predictions that AIs

might make – “if the users do not trust a model,... they will not use it” (Ribeiro et al., 2016, para 3), although the assertion that adoption directly correlates with trust is questionable given users often do not have a choice in adopting technologies (Lindley, et al., 2019) or being subject to its ramifications.

Nevertheless, Thornton et al. (2022) explain that the design of trustworthy socio-technical systems requires designers to become “alchemists of trust” by incorporating, tailoring and combining a wide array of trust models and principles, including transparency, into data systems (Knowles, et al., 2014). As identified, “explanations are particularly helpful in identifying what must be done to convert an untrustworthy model into a trustworthy one” (Ribeiro et al., 2016, para 11). Ribeiro et al. offer a novel explanation technique of Local Interpretable Model-agnostic Explanations (LIME), an algorithm that learns an interpretable model locally (i.e., computable for a specific input) around a prediction and highlights the components in the data set that led to the prediction (Figure 45).

Figure has been removed due to
copyrights restrictions

Figure 46: Explaining individual predictions. An AI model predicts that a patient has the flu, and LIME highlights the symptoms in the patient’s history that led to the prediction. Sneezing and headache are portrayed as contributing to the flu prediction which aids the doctor to make an informed decision about whether to trust the model’s prediction (Ribeiro et al. 2016).

A practical approach to AI transparency is to provide the relevant authorities with access to source codes and a detailed explanation of the technology's implementation (Abrassart et al., 2018; Burrell, 2016). Sandvig et al. detail several different auditing designs, including ‘code auditing’ and ‘sock puppet auditing’ (Sandvig et al., 2014).⁵⁸ Albeit auditing is often curtailed by corporate entities with the necessity to remain competitive (Burrell, 2016) and to prevent users from gaming systems and unfairly receiving or benefiting from services (Ananny & Crawford, 2018, p. 979; Diakopoulos, 2016, pp. 58–59); as such laws are often implemented to protect the private sector from external scrutiny (see Bloch-Wehba, 2021). By contrast, algorithms and training data which are propriety

⁵⁸ Sock puppet auditing is performed by a computer program impersonating a user by creating fake accounts. However, this audit method is problematic and can create legal difficulties as it often breaks user agreements.

property, raise questions on whether such systems should be used in critical areas such as criminal justice. Many have called out the lack of accuracy, accountability, and intelligibility of these systems already in use (Angwin, et al., 2016; Partnership on AI, 2019).

An idea that endeavoured to annul the need for auditing was formulated by the company OpenAI. Their mission statement declared that the company would strive to “build value for everyone rather than shareholders” by sharing patents with ‘everyone’ (Brockman et al., 2015, para 8). However, in the years since starting, the company has attracted considerable attention by creating revolutionary algorithms, including *DALL·E 2*⁵⁹, and as a result, invested in by Microsoft; consequently triggering a shift in their original business aims by restricting how ‘open’ the company was to external observation and confining the ‘free use’ of their products by applying a ‘freemium’ business model (Ding, 2022).

7.3 Interpretability or Explainability?

Interpretability, on the other hand, has been described by Lipton as a “slippery” concept (2018, p. 20), often confused and interchangeably used with transparency and explainability, with many authors not differentiating between the terms (see Arnold et al., 2019; Arya et al., 2020; Marcinkevičs & Vogt, 2020). Terms also synonymous with interpretability within AI literature are “understandability” and “intelligibility” (Caruana et al., 2015; Lipton, 2018; Lou et al., 2012). The concept has been, however, categorised as a model that is “human simulatable”, whereby a human can simulate the procedure (Lipton, 2018; Matthew, 2019), and also defined as “the mapping of an abstract concept into a domain that the human can make sense of” (Montavon et al., 2018, p. 2). Although, once more, AI experts can only reasonably interpret the mapping and explanation with methods and technical foundations created that are domain specific. For instance, interpretability in computer vision is concerned with highlighting and directing attention to the different parts of an

⁵⁹ DALL·E 2 creates realistic images and art from a description in natural language (see openaidalle [@openaidalle], 2022). The technology is currently in preview mode to review and mitigate related risks of the technology such as images being created of explicit content, bias and representation perplexities, images created could be used as a form of bullying and exploitation or disinformation (Open AI, n.d., 2022). Also at stake is the impact AI has in the creative field. Artist and designer Sebastian Errazuriz predicts illustrators will be the first profession to cease to exist as AI sweeps the creative industry (Errazuriz & [@sebastianstudio], 2022).

image integral to an AI's reasoning procedure.⁶⁰ Rudin explains that interpretable models entail significant effort to construct in terms of computational output compared to black-box models, requiring an entirely different approach in education and the know-how the field is currently built upon (Rudin, 2019). This situation draws upon the circumstance, as it currently stands with AI technology, that different AI models suit different tasks – with many problems requiring specialised forms of data analysis presently only archived through 'black-box' methods (N. McCarthy & Montgomery, 2019). However, developments have been and continue to be made with methods such as the use of saliency maps to pinpoint data features utilised by a neural network (see Simonyan et al., 2013) and Gestalt, a development environment designed to analyse data as it moves through a machine learning pipeline thus enabling a human to support the process and fix bugs in ML systems (Patel et al., 2010). Albeit considered a more fruitful line of research is machine-to-machine interpretability and the development of ways for machines to understand one another. Weller suggests that while it is desirable and of great importance for humans to understand the operations of machines, it is non the less easier and, as he emphasises, perhaps more practical means to develop machines that communicate with each other, resulting in "high level structures which can be transmitted efficiently and deployed flexibly" (2017, p. 14).

7.3 Mechanisms for users

Weld and Bansal's exploration of "intelligible intelligence" is helpful in mapping current efforts to articulate how specific AIs work more comprehensibly for designers, developers, and expert users alike. Among the vast array of issues the authors discuss are: distinguishing the underlying mathematical challenges from the human-focused HCI challenge; unpacking a wide range of reasons why making AI intelligible matters (e.g. legal imperatives, helping humans enhance their own understanding, driving user acceptance, allowing users to control AIs); defining what intelligibility actually refers to; ranking or quantifying intelligibility; differentiating between intelligible and inherently inscrutable models. While crafting intelligible intelligence, transparency, interpretability,

⁶⁰ This approach arises from early attempts by computer scientists to communicate instructions on code and algorithm's and their power through visualisations and animations by the likes of Ken Knowlton working at Bell Laboratories (see Knowlton, 1966)

and XAI captures the need to unpack an AI's black box, Gunning et al. argue that most current guidance and literature *overlooks* end-users' differing knowledge, specifically those with no AI knowledge (2019).

In their paper *Explanations as Mechanisms for Supporting Algorithmic Transparency*, Rader et al. promote transparency to empower users to make informed choices about how they use algorithmic decision-making systems. The paper documents an experiment focusing on providing participants with different written explanations of how Facebook's News Feed algorithm worked. The authors begin, likewise, by bifurcating the meaning of transparency, explaining that contextualising the meaning can be "difficult to disambiguate in the literature" (2018, p. 3); the authors note, in one paradigm directed at users the idea is concerned with making a system knowable or visible, as a mechanism to bring about changes in users' behaviour; otherwise quoting the Association for Computing Machinery's principles, transparency is seen as a mechanism for making systems accessible to experts for evaluation purposes (US Public Policy Council, 2017).⁶¹ Inspired by the explanations found in recommender systems (Ricci et al., 2011),⁶² the authors introduce an array of criteria needed to be considered for crafting explanations:

- *What* explanations reveal the existence of algorithmic decision-making.
- *How* explanations describe a system's inputs and outputs and the steps taken to arrive at an outcome, also known as a "white box" explanation (Fredrich & Zanker, 2011).
- *Why* explanations allow users to determine whether their goals match the system's.
- *Objective* explanation, in the adjective sense of the word, provides users with an unbiased account of when algorithms do not work as intended on occasions.

These explanations are considered to execute the transparency index and objectives Radar et al. detail (2018, p. 5):

- *Awareness* of an algorithm in use and making decisions clear (Ananny & Crawford, 2018; Fjeld et al., 2020).

⁶¹ The statement is the current document for the US Public Policy Council.

⁶² A collaborative filter style used in Spotify highlights why songs have been recommended using explanations such as "Other users similar to you liked this item".

- Empowering *Correctness* judgements allows users to evaluate whether a system works as intended or believed by the user (Kulesza et al., 2013).
- *Interpretability* of a system permits the evaluation of actions made by the system.
- *Accountability*, directed at the user “shifting the balance of power” (Rader et al., 2018, p. 10) by providing a sense of iterative control over the system (Ananny & Crawford, 2018; Diakopoulos & Koliska, 2017) and the agency not to use the system going forward.

Delving further into the composition of explanations, Tim Miller extensively outlines how the field of XAI can utilise insights from psychology and social sciences, calling attention to how human cognitive processes and bias can influence the effectiveness of explanations. The author explains to produce truly explainable AI, developers need to consider the computational consequences and be aware that cognitive processes and bias can influence the effectiveness of an explanation in different contexts. Miller defines the following as primary considerations for the field of XAI (Miller, 2017, p. 6):

- Individuals seek *contrastive* explanations. That is, individuals do not ask why a particular event happened but rather why the event happened instead of another scenario.
- Explanations are *selected* in a biased manner, whereby individuals draw from a subset of the total factors that caused an outcome to make sense of why something happened.
- Probabilities and statistical generalisations are not convincing or satisfying to individuals who prefer causal explanations. However, as a confounding point, in the context of AI, “the mechanism is statistical rather than causal”(N. McCarthy & Montgomery, 2019, p. 14).
- Explanations are *social*, as explaining something is an interaction between two actors, influenced by both delivery and receipt (Weld & Bansal, 2019b).

Rader et al. raise a thorny issue of explanations and user experience; “if the aim is to provide information that users are not aware of, then it seems inherently difficult to ensure that the new information does not violate user expectations” (2018, p. 10). This final point, as Lindley et al. highlight, is perhaps indicative of latent societal norms which are waiting to emerge with respect to AI becoming domesticated (Lindley et al., 2020; Silverstone, 2006), at which point society needs to adapt, and methods of ethical and responsible mitigation are designed in response to the new reality AI is cultivating.

7.4 Limitations of Transparency: Seeing without Knowing

As a final point on the prominent concepts listed above concerning the designing and presentation of AI technology that affects an understanding, perception of, and ability to critically evaluate its use, either as an expert or non-expert, this section returns to the notion of transparency, as it is the most conventionally reviewed and conceptualised concept in human-AI interaction literature.⁶³ Transparency is reasoned to be an underpinning principle and means for ethical and responsible AI development to achieve trustworthy, accountable, safe, and secure sociotechnical systems. It could also be rationalised that transparency is the state for explainability and interpretability to transpire as forwarded by Rader et al. Nevertheless, in their article *Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability*, salient commentators on AI technology, Mike Ananny and Kate Crawford, trace the historical and technological ideal of transparency, drawing attention to and opposing its implicit promise that “*seeing* a phenomenon creates opportunities and obligations to make it accountable and thus to *change* it” (2018, p. 974, original emphasis). Here it is helpful to draw upon David Heald’s “transparency illusion”, whereby “when transparency appears to be increasing, as measured by some index, the reality may be quite different” (2006, p. 34), creating entrenched shortcomings. Of particular note: making visible an entire system has no meaningful effects when there is no system or mechanism “capable of processing, digesting, and using the information” (Ibid, pp. 35-37). In this

⁶³ It might not be the most scientific method (but it does clarify the point), but a quick search on Google Scholar for the term ‘AI Transparency’ returned 746,000, whereas ‘AI Explainability’ returned 24,700 results.

regard, advocacy of transparency privileges seeing over understanding AIs' behaviour or origins, assuming that transparency will bring insight and governance (Ananny & Crawford, 2018).

One could argue that understanding the behaviour is where explainability and interpretability become effective, though complete transparency of algorithms creates a paradox; Stohl et al. classes as an “inadvertent opacity”—whereby “visibility produces such great quantities of information that important pieces of information become inadvertently hidden in the detritus of the information made visible” (2016, p. 133). Complete transparency can also be viewed as the opposite of privacy, with methods of handling sensitive data designed to conceal and protect vulnerable individuals or groups from harm (Schudson, 2015). Conceptual examples benefitting by obscuring information include the relationship between a doctor and a patient, or a lawyer and their client, whereby Weller posits “[i]nside these relationships, it is interesting to question whether greater transparency leads to trust”(Weller, 2017, p. 7).

As Ananny and Crawford point out— arguments for transparency, by looking *inside* systems and “fetishising” algorithms (Crawford, 2016, p. 89) avoids and comes at the cost of a “deeper engagement with the material and ideological realities of contemporary computation” (Ananny & Crawford, 2018, p. 974). Therefore, the authors suggest looking *across* and positioning them as sociotechnical systems that “do not *contain* complexity but *enact* complexity by connecting to and intertwining with assemblages of humans and non-humans” (Ibid, original emphasis). What is held accountable then, is not seeing inside one object, but understanding the system that exists as an assemblage of “institutionally situated code, practices, and norms with the power to create, sustain, and signify relationships among people and data through minimally observable, semiautonomous action” (Ananny, 2016, p. 93).

A practical method to view these assemblages, as developed in this research, is through a MTHCD perspective, employing design tools such as constellation mapping and non-human speculation to ask, as Ananny and Crawford advocate, “what is being looked at, what good comes from seeing it, and what are we not able to see?” (Ananny & Crawford, 2018, p. 985). Through these investigations, this research would take this idea further by making the intangible interactions with sociotechnical systems legible, accentuating their presence to users, permitting likewise the

opportunity for individuals themselves to ‘look across the assemblage’ and the system they are part of via the communication of fundamental AI operations – a method to start processing and digesting the sociotechnical reality they are submerged in. Before detailing the RtD process of designing a means to make AI systems legible through AI iconography, the following section introduces the concept of AI legibility.

7.5 AI Legibility

Lindley et al. remark, “[w]henever ‘AI’ is used to describe a system or innovation its ‘legibility’—as the term is used in the emerging field of Human-Data Interaction (HDI)—is significantly reduced” (2020, para 4). To explain: HDI is fast emerging as a field of research intersecting with and complementary to HCI, focused on sensemaking of the data produced through our interactions that ultimately transcend the devices themselves (see Crabtree & Mortier, 2015; Mortier et al., 2014; Sailaja et al., 2017; Victorelli et al., 2020).

HDI’s perspective is that data is ontologically malleable and are “containers and carriers” of information between actors and organisations (Bannon & Bødker, 1997, p. 85). Crabtree and Mortier conceptualise an understanding of data using the concept of “Boundary Objects” (Star, 2010), as Star and Griesemer put it:

Boundary objects ... are both plastic enough to adapt to local needs and the constraints of the several parties employing them, yet robust enough to maintain a common identity across sites ... They have different meanings in different social worlds but their structure is common enough to more than one world to make them recognisable, a means of translation (Star & Griesemer, 1989, p. 393).

As discussed, the data a given AI uses is pivotal to how the AI works and how well it functions. Therefore, when considering AI, HDI’s interest in how people, data and algorithms interact is crucial. Acknowledging the scale of the societal impact stemming from the assemblages of data, analysis, and the resultant interfaces, HDI necessarily champions and challenges the creation of methods for making systems less opaque with enhanced control for individuals. As such, HDI is a multidisciplinary enterprise threading together a manifold of fields, including HCI, ethnography,

sociology and economics (Vitorelli et al., 2020). The diverse insights garnered are distilled into three interrelated though distinct themes at the core of HDI's efforts: *legibility*, *agency*, and *negotiability*. Legibility refers to the process of understanding and making what data is collected or processed, how inferences are drawn and the implications of those inferences comprehensible to people. Hence the notion of legibility is distinct from transparency (i.e., everything exposed in an incomprehensible manner), resultantly acting as a “precursor” to exercise agency within these systems. As a derivative, manifestations of agency influence negotiability mechanisms, enabling a user to build a relationship with the recipient of the data as means to negotiate how the data is used thereafter (Mortier et al., 2014, p. 5). While these themes overlap, this research will address legibility, as it is a preface and facilitates the other themes' materialisations. Additionally, legibility can tap into the remit of HCI, as the point at which individuals interact with a system presents a prominent place to convey information about the implications of use (Lindley et al., 2020)—creating richer and more tenable interactions with our devices as they become ‘smarter’, networked, and complex; evolving beyond the traditional duality of interaction between user and computers-as-artefacts (Bowers & Rodden, 1993).

7.6 Guidelines for Legible Human-AI Interactions

HDI's view of legibility is predominantly concerned with human relationships with data, supporting both the postphenomenological bearings and the MTHCD approach of this research. The following section brings AI technology back into a central focus; as argued, AI technology likewise has unique attributes that require being measured as a thing-in-itself.

Reflecting on 20 years of research relating to interactions with AI, Amershi et al. propose 18 succinct guidelines for human-AI interaction collated and summarised from a vast range of 150 AI-related recommendations. These guidelines are motivated by contemporary AI issues such as bias, false positives, unpredictability, and the impact of whether the existence of AIs are visible or “behind the scenes” (Amershi et al., 2019, p. 1). The 18 guidelines are divided into ‘phases of interaction’ an individual has or could have with an AI-infused system and cover a broad range of instances.

Such as: the provision of information in the ‘initially’ phase, e.g., G1, “make clear what the system can do”; covering ethical and social issues ‘during interaction’ phase, e.g., G6, “mitigate social

biases” through “ensuring AI-systems language and behaviours do not reinforce undesirable and unfair stereotypes and biases”; safety netting for the ‘when wrong’ phase, e.g., G11, “make clear why the system did what it did” through “explanation of why the AI system behaved as it did”; and finally, how a system might be configured by users and aid users understanding of the system ‘overtime’, e.g. G17, “provide global controls” by allowing the user to customise what the AI monitors, and G18 “notify users about changes”(Amershi et al., 2019, p. 3).

Through an iterative evaluation process, including user testing, with 49 design practitioners, these guidelines were validated by testing them against 20 AI-infused products. When the guidelines are viewed alongside the diversity of related work and contextualised with examples from real-world applications, the scope of such a task is considerable. The authors acknowledge that whilst they considered ethics and fairness, the complexity of these concepts far exceeds the straightforwardness of how the guidance is worded. Furthermore, the authors note the problematic task of heuristically evaluating the guidelines, although describing instances where they were irrelevant or made sense in specific applications. For example, when guidelines require user assessment (e.g., G10, scope services when in doubt) with many AI interactions problematic to assess unless there is time and criteria for evaluation (Ibid, p.7). However, these well-considered guidelines, the assessed validity and the opportunities for future research presented by Amershi et al. represent a significant step towards designing AI systems that are more human-centred AI-systems. Like HDI’s core themes, all 18 guidelines are actionable through increased legibility or legible communication for individuals to reap the benefits of the guidance. It makes sense for this research to start with Amershi et al.’s ‘initial phase’ guideline of – “G1, make clear what the system can do. Help the user understand what the AI system is capable of doing” (Ibid, p. 3).

This standpoint then becomes a design and communication challenge, with much of the technical detail concerning AI likely beyond the grasp of the majority of users. In their article, *The Challenge of Crafting Intelligible Intelligence*, Weld and Bansal raise two pertinent points for legibility (Weld & Bansal, 2019b). The first presents a fundamental challenge – the “construction of an explanation vocabulary” (p. 79), acknowledging that this may mean some form of generalisation given the scope of AI (Lindley et al., 2020). Second, the authors accentuate, resonating with Miller,

that an explanation is a social process “best thought of as a conversation” (Weld & Bansal, 2019b, p. 79), whereby something can be distilled through clear and grounded terms. Consequently, when designing AI legibility, understanding how one might explain a phenomenon while balancing accuracy and accessibility is significant (Lindley et al., 2020).

7.7 Ways through the Communication Challenge

Concentrating on the communication challenge, Arnold et al. consider how documents known as *supplier’s declarations of conformity* (SDoCs) may be repurposed and considered for AI in the form of *FactSheets* (M. Arnold et al., 2019). The authors envision that *FactSheets* would communicate the purpose, performance, safety, security, and provenance information of an AI to developers, as oftentimes, the method of AI integration into products is via an API.⁶⁴ Therefore, a developer has no knowledge of how the underlying model works, what data it is trained on etc. Furthermore, the authors highlight that *FactSheets* could elevate the expertise gap between those producing and those developing the AI services. In such circumstances, “it becomes more crucial to communicate the attributes of the artifact in a standardised way” (Ibid, p.2). The *FactSheets* are composed of straightforward questions arranged thematically, some based on potentially problematic AI issues such as safety, fairness, and concept drift. The concept is not dissimilar to the ‘*Datasheets for Datasets*’ proposed by Gebru et al., which tries to document the motivation, composition, collection process, recommended uses and so on of datasets (see Gebru et al., 2018).

⁶⁴ An application programming interface (API) is a way for two or more computer programs to communicate with each other.

Considering the adoption of FactSheets, Arnold et al. speculate that the scheme would not need to be a legal requirement, however, describe conforming to its use through market and peer pressure would be probable, becoming a crucial part of AI accreditation and compliance. The communication of quality standards via FactSheets would be similar to *Energy Star* product labelling or nutrition labelling on foods. The success of nutrition labelling has helped increase consumer awareness about food and its composition and has inspired the same approach in computing. An example is the ‘Dataset Nutrition Label’ forwarded by Holland et al. for IBM R&D (Holland et al., 2018), which assesses and interrogates datasets through a diagnostic framework based on quality measures that are distilled into an overview of the datasets “ingredients” (i.e. metadata, provenance, variables) for the simple reason “garbage in, garbage out” (*The Data Nutrition Project*, n.d.) (Figure 46).⁶⁵



Figure 47: Label created by the Data Nutrition project showing a breakdown of the data used for the New York City tax bills with an alert count, use cases and iconography badges (The Data Nutrition Project, n.d.).

The authors speculate on the possible benefits of the data nutrition labelling process, which include: aiding professionals working with data to critically interrogate and select the best dataset for their purposes; encouraging better data practices to limit possible harms associated with AI; and a more conscientious engagement with data (Holland et al., 2018). There are other similar schemes, such as: a

⁶⁵ The Data Nutrition Project has launched a second gen of their dataset nutrition label. The redesign provided more targeted information on use case and further information related to the algorithmic method chosen such as prediction (Chmielinski et al., 2020).

preliminary observation for labelling consumer IoT to aid decision making, acting also as a lever to encourage IoT companies to create more secure designs (J. M. Blythe & Johnson, 2018); privacy labels that present consumers with the way organisations collect, use and share personal information (Kelley et al., 2009, 2010), and communicating the details and the output of ranking algorithms (K. Yang et al., 2018). Each of the endeavours described is a substantial undertaking of providing clear and complete information concisely and legibly; while sensitively balancing how labelling changes behaviour (Drichoutis et al., 2006) and can influence under false pretences (see Koenigstorfer & Baumgartner, 2016).

7.8 Designing for Legibility: A Case for Icons

It is worth noting, however, it is admittedly more straightforward to create a labelling system for food due to the easy-to-quantify features of food (e.g., protein, carbohydrates, fat) known and understood through well-established food analysing processes and better all-round common knowledge, compared to the difficult-to-define and lesser-known attributes of AIs. This situation presents a two-pronged challenge for AI legibility, how to form generalised though complete information which is easy to comprehend. Here we can turn to the Royal Societies' solution to the problem of subdividing into levels of information; a "local approach" for users (N. McCarthy & Montgomery, 2019, p. 14), making legible critical operational attributes, which users should be aware of for agency and negotiability derived from prior work mentioned here (Ada Lovelace Institute, 2022; Amershi et al., 2019; Holland et al., 2018; Matthew, 2019; N. McCarthy & Montgomery, 2019; Miller, 2017; Mortier et al., 2014; Rader et al., 2018), and a "global approach" for AI-experts using the interoperability models, which for now is beyond the scope of this research (N. McCarthy & Montgomery, 2019).

Visual languages offer a 'local approach' aimed directly at users for AI legibility. HCI has a rich history of research about the communication of accessible information through the proliferation of graphical user interfaces (i.e., vehicles displays (Marcus, 2002), touch-based interactions (Arnall, 2006), auditory icons (Gaver, 1986) to name a few) needing clearly designed iconography that bestows knowledge and guidance on interaction with machines quickly and succinctly (Ma et al.,

2015; Marcus, 2003; McDougall & Isherwood, 2009). The reasons for using icons include: immediate recognition and use at opportune moments (Kairos); better recall, can be used by all reading levels; are global; and can be interpreted at speed (Horton, 1996). Using icons also avoids the problem associated with text-based descriptions, which tend to use technical jargon and/or require expert knowledge (Samsudin et al., 2016).

Assessing the present levels of AI legibility by surveying current AI iconography, Lindley et al. found that although some AI imagery attempts to represent the underlying system, such as a neural network (Figure 47a) or highlighted its use, such as face detection (Figure 47b), the vast majority play into AI's definitional dualism by showing human-like machines (Figure 47c & 47d), thus exacerbating misconstrued perceptions of human-like intelligence existing in AI (Lindley et al., 2020; F. Pilling, Akmal, Gradinar, et al., 2020). The survey also illustrated that current AI imagery rarely communicates the intricacies of how an AI functions and in what context, emphasising the need to develop a new visual approach to enhance AI legibility. By developing such a lexicon, there is a real opportunity to consolidate suitable elements from the research landscape and package specifically to aid interaction and help communicate how AI is implemented to users.

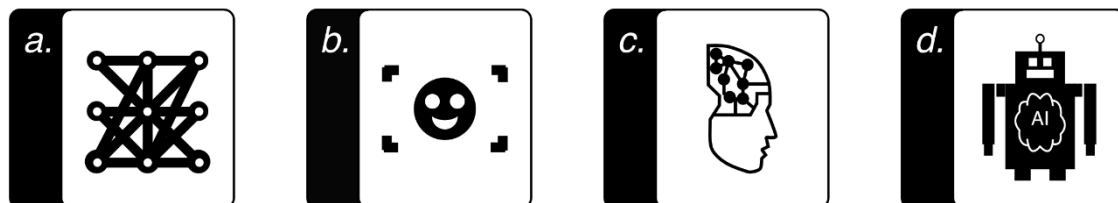


Figure 48: A range of typical AI iconographies.

7.9 Background for Designing the Semiotics of AI

Research into the design, theory, and effectiveness of icons in the field of HCI is diversely underpinned by the theory of Semiotics, for instance – semiotic analysis for user interfaces (Barr et al., 2002; Ferreira et al., 2002), icon taxonomy to categories computer icons (Ma et al., 2015), the advantages and disadvantages of icon-based dialogues in HCD (Gittins, 1986), relationships between different presentation modes of graphical icons and user attention (Lin et al., 2016), testing the intuitiveness of icons (Ferreira et al., 2006), and so on. The Peircean Triad is of particular note in the field of semiotics (Peirce, 1991). Charles Peirce's model (see figure 48) consists of a triadic

relationship; comprising the *representamen* (the symbol used to represent an idea, e.g., a save icon), the *object* (the actual construct being represented, e.g., data or document being saved), and the *interpretant* (the logical implication of the sign, e.g., using this icon will save my data). As Barr et al. clearly define, “the goal is for the representamen to effectively create an interpretant which matches the object” (2002, para 12).

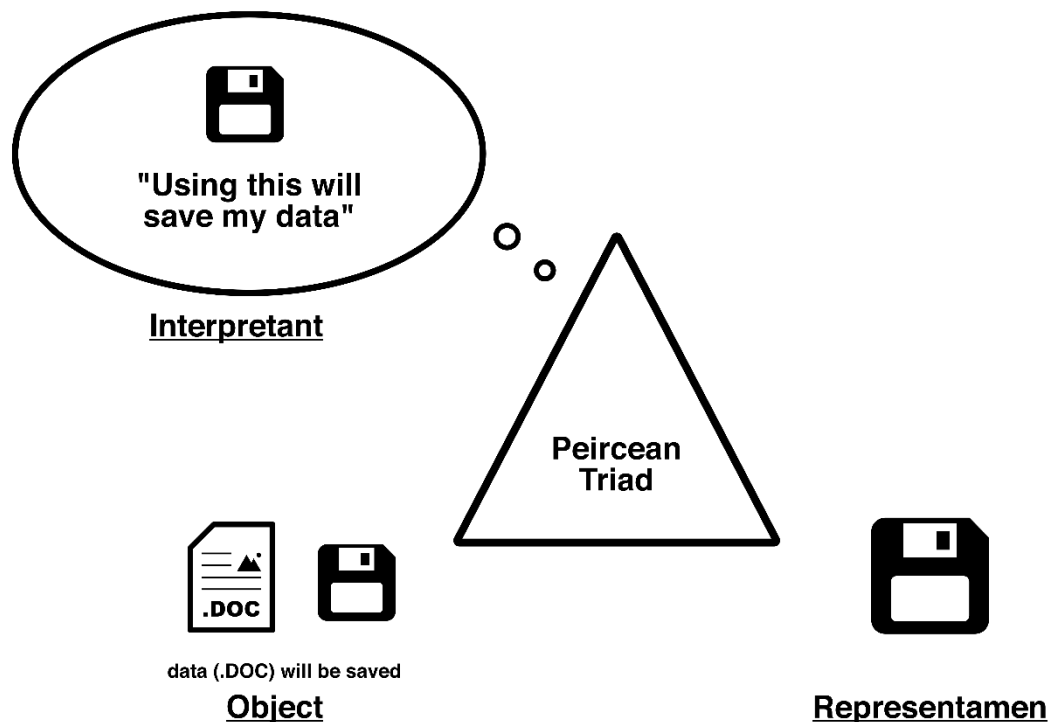


Figure 49: The Peircean Triad for the iconic save icon.

In addition, central to Peirce’s thesis is the ‘classification of signs’ or ‘icon types’ (Ibid) which is based on the relationship between object and representamen (Figure 49); these categories are; *indexical*, signs which refer to the object indirectly, through an association and causation (e.g. smoke signifies fire), *symbolic* signs which have meaning based solely on convention and may be culturally specific, such as alchemy symbols (e.g. a triangle to represent fire); *iconic* signs have a signifier which resembles the signified package (e.g. flames pictorial). Peirce noted that categories are not mutually exclusive, as most signs contain varying degrees of indexicality, symbolism and iconicity.

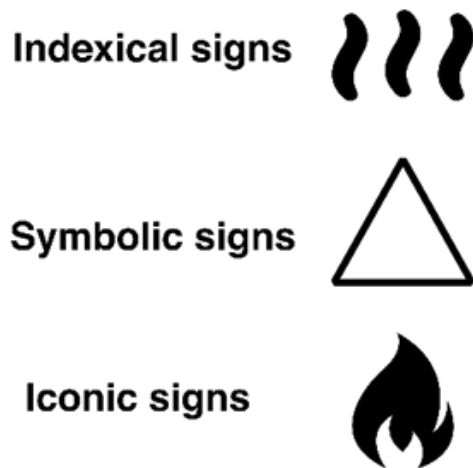


Figure 50: Examples of indexical, symbolic, and iconic signs.

Taking this theory back to the analysis of the existing AI icons (see Figures 47c and 47d), the representamen forces the notion of AI's dualism, which one could argue is 'misleadingly' both categories of symbolism and iconicity. In some cases, the interpretant functions to some degree (see Figure 47b facial recognition); however, there is no understanding of the complete AI system. While Peirce's indexes help research the design of computer icons, Ferreira et al. indicate that in reality, it "is very rare, and some argue impossible, to find signs that belong solely to one category" (Ferreira et al., 2006, para 8).

Using the semiotic research and as a first step in the design process, three stylistic elements were envisioned. Within these variations, the aim was to keep the object and the interpretant fixed while altering the appearance of the representamen based on the following rationale. The first design (Figure 50, Pictorial) was referred to as pictorial and is based on an iconic design. The illustration shown in the figure also utilises the type of iconography resulting from AI's definitional dualism.

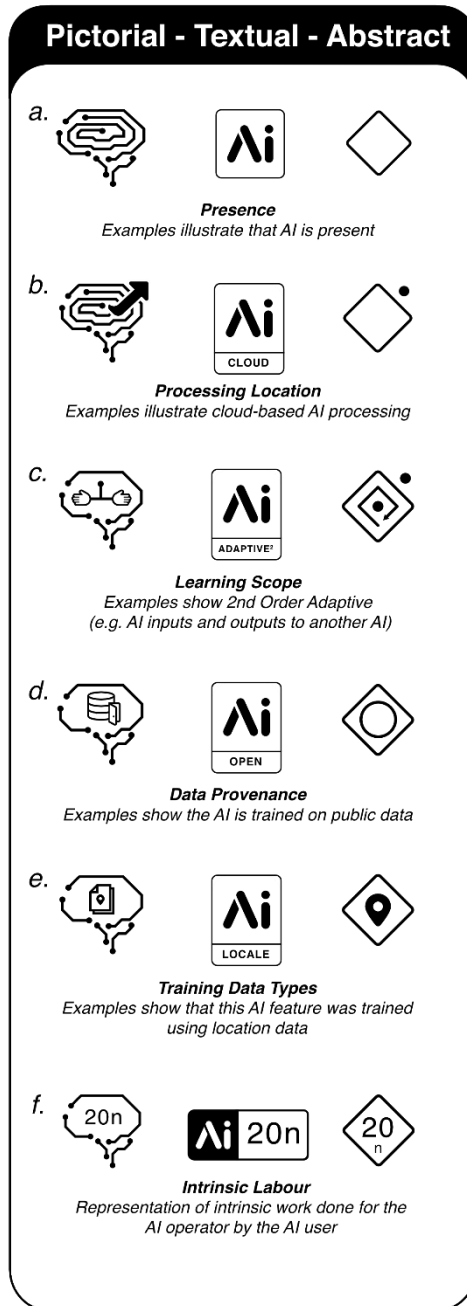


Figure 51: Three different style variations Pictorial, Textual and Abstract.

Though despite conforming to the current problem in the example shown, iconic imagery established a baseline to move forward, which will be shown later in the document. The second concept explores branding (Figure 50, Textual), inspired by the symbology employed by organisations such as the *WI-FI Alliance*; this was categorised as the textual variant. Branding plays a key role in communicating to users a guarantee of compatibility and conformity to minimum safety considerations (Kardes, 1988). However, as noted previously, there is a limit to how much textual information can be gleaned in a single instance, hence the rationale to use iconography in the first instance. The third approach (Figure

50, Abstract) explores the execution of a symbolic icon taking cues from highly recognisable symbology used in road warning signs and laundry labels, referred to as the abstract variant.

It is also worth bearing in mind failed icon designs, such as a seashell to represent the ‘C-Shell’ command processor (Gittins, 1986, p. 525), and likewise, icons that in theory should not work, take the ‘save’ icon based on the antiqued floppy disk still in use despite many users too young to have used a physical floppy disk. These notions similarly present a further difficulty in predicting how or why an icon may become adopted or stay in use and calls for empirical testing for any icons designed.

7.10 Defining the Interpretant: AI Attributes, Dimensions and Properties (AI’s Ontology)

The MTHCD approach formed in this research presents a method of thinking about the ontological constitution of AI and, therefore, the interpretant of the icons. In other words, the MTHCD approach developed through Harman’s account of OOO is the attempt at speculating and communicating the Real Object (RO) and Real Qualities (RQ) of AI and also conveying the often hidden (as AI is intangible to most users) Sensual Object and Sensual Qualities of AI through iconography. Harman would argue that an AI’s RO and RQ qualities withdraw into the subterranean depths of being, but as reasoned, so does the SO and SQ’s of AI. To recap: as noted in Chapter Five § 5.11.3 *Object Ontology: Levels of Objects* with speculation and educated guesswork of how AI works, we can approach the RO and RQ of AI and what it needs in order to be itself, or in other words an object’s essence— what makes an AI an AI. Sensual-Objects (SO) and Sensual-Qualities (SQ) exist in relation to that of a real object, as a correlate in our minds (Husserl’s phenomenology) and how we experience the AI or its often-obscured consequences of use, as detailed when this research looked towards ‘Seeing AI’ in Chapter Two. In other words, the icons communicate Harman’s quadruple structure of AI (Figure 52).

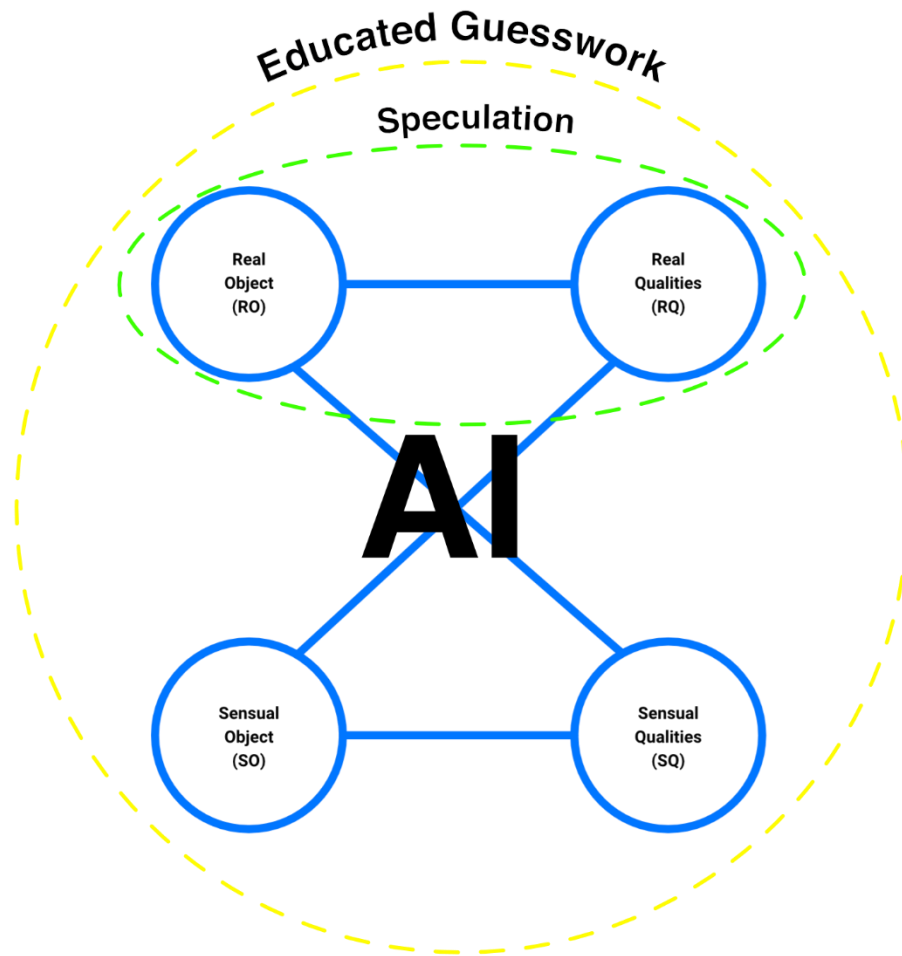


Figure 52: An adapted figure of Harman's quadruple structure (see figure 30, p.146) highlighting RO, RQ, SO, and SQ of AI and how through speculation, we can approach RO and RQ. As designers, we can approach all of the qualities of AI through educated guesswork.

As previously noted in Chapter Five *More Than Human-Centred Design: Shifting Perspectives through Philosophy*, and in Chapter Six: *Adapting Philosophy for Design*, the act of speculation was defined as the modus operandi when conducting philosophical work; however, so too was the act of using educated guesswork (See § 6.5 *Speculation and Design Fiction*, p.188). Given that the challenge of communicating AI is formidable, consolidating functional elements from the research landscape (i.e. educated guesswork) will also aid in identifying AI attributes, dimensions and properties required to be communicated to users to make their own value judgements, also underlining the importance of conveying advisory information rather than qualitative assessment (Lindley et al., 2020). Additionally, the compilation of the icons need not be exhaustive with granular technical specifications as this would provide minutia details out of context, counteracting the aim of communicating valuable and legible information. Guided by these criteria, six key concepts to

communicate AI's ontology were established and distilled into the three icon variants described prior (Lindley et al., 2020; F. Pilling, et al., 2020):

- *Presence* denotes that some form of AI processing is happening, heeding the ethical principle of 'informed use' and ensuring that individuals and the public are aware of their use and interaction with AI-infused systems (EPRS, 2020).
- *Processing Location* – in the cloud, on the edge or elsewhere. The location of processing impacts security (Pilling, 2022) and users' perception of accountability (Rader et al., 2018) and confidence in where their data is going (Pilling, 2022).
- *Learning Scope* – how does the AI learn or adapt over time, through usage, or is it static? Communicating to a user changes and adaptations of an AI system over time impacting the AI system's behaviour is deemed a fundamental guideline for human-AI interaction (Amershi et al., 2019, G14).
- *Data Provenance* – what is the source of the training data? Is it proprietary, public or the user? Data quality directly reflects the AI and its trustworthiness (M. Arnold et al., 2019).
- *Training Data types* – what data types are used to train the AI? Visual, audio, location? Similar to data Provenance this factor is a more granular account of the type of data, which is a crucial element in reducing opacity (Burrell, 2016), increasing trust (M. Arnold et al., 2019) and reducing bias (Angwin, et al., 2016; O'Neil, 2016).
- *Intrinsic Labour* – is 'work' being done for the AI operator? This factor is more philosophical and discursive, as it reflects the monetisation of data through the commodification of users and their interactions with AI-infused products and services (Greengard, 2018; Zuboff, 2019).

These features were carefully chosen to communicate objective concepts, providing users with factual information when interacting with an AI system (Lindley et al., 2020). These features also capture the agency and efficacy of the AI system. Motivated by framing the AI attributes to be communicated objectively meant omitting critical dimensions of AI; for example, whether or not an

AI is biased. This choice was strategic in the design, given that the focus was to provide a framework for legibility by providing information to allow users to form their own opinions better. While failing to provide alerts, much like the data nutrition project does, is a shortcoming of the proposal, for now, a speculative aspect of the research presented later subsequently looks at a hypothetical accreditation process to provide alerts. However, one could argue that Intrinsic Labour falls into the category of dimensions excluded. Again, the inclusion of Intrinsic Labour was considered part of the design, acting as a proxy for the indistinct although theoretically quantifiable concepts, such as fairness and bias (Ibid). Secondly, Intrinsic Labour's presence provides a valuable opportunity to examine how these concepts can be symbolised through visual communication and assess how useful they would be if quantifiable.

7.11 Reinstating A Philosophical Perspective: Aesthetics Is the Root of All Philosophy

In addition to the icons providing an ontological breakdown of an AI-infused system, the theory of OOO can also aid in theoretically understanding how icons can communicate an object's Sensual Qualities (SQ) and, speculatively, an object's Real Qualities (RQ).⁶⁶ The aesthetic theory of Harman's OOO revolves around the idea of 'allure' and the 'nonliteral' and 'indirect' experience of an object and its withdrawn qualities. For this reason, in recent years, the theory of OOO has been popular as a way to explore art pieces' deeper meaning (see Harman, 2020).

As briefly mentioned in Chapter Five, *More Than Human-Centred Design: Shifting Perspectives through Philosophy*, approaching reality indirectly, and accessing things-in-themselves (noumena) is accomplished through mechanisms such as metaphor and speculation. For this reason, metaphors are held in high regard in OOO, as they are a radical strategy for actively forming a bridge across another object's Real Qualities (RQ) and Sensory Qualities (SQ). Harman affirms José Ortega's early essay on metaphors – *An Essay in Esthetics by Way of a Preface* planted the seed of OOO in his mind through the realist illumination of an object's 'I' and the "reality apart from any observation or introspection"(Harman, 2018, p. 70). In Ortega's words, "[e]verything, from a point of

⁶⁶ It is worth mentioning, that all signs are objects too in addition to what they are interpreting (in this instance, AI attributes).

view within itself, is an I” (Ortega, 1914, p. 134), not in a conscious sense, but “simply because it *is*”(Harman, 2018, p. 77, original emphasis). Despite being trapped in the human condition, Kant’s noumenal realm *is not* inaccessible, as Ortega effectively states art has a special way of touching the noumenal realm of a thing:

Now then, imagine the importance of a language or system of expressive signs whose function was not to tell us about things but to present them to us in the act of executing themselves. Art is just such a language; this is what art does. The esthetic [(aesthetic)]object is inwardness as such— it is each thing as ‘I’(Ibid, pp. 138–139).

Icons, too, are a visual language. In this context, it is designing the aesthetic communication of AI’s RQ and SQ. Ortega clarifies, "a work of art affords the peculiar pleasure we call esthetic by making it *seem* that the inwardness of things, their executant reality, is open to us” (Ibid, p139). Harman highlights that Ortega hedges his bets on the word ‘seem’ because the noumenal reality of a thing is unavailable. However, art has a unique way of connecting to it: “a touching without touching, so to speak” (Harman, 2018, p. 82). This idea is reminiscent of Bogost’s carpentry and Ingold’s string manipulation, where the making of an (esthetic) object too *seemingly* traces the inward experience of another thing. In this regard, the icons attempt to trace the inward experience of an AI being through esthetic communication. For Ortega, it is a metaphorical object.

Effective metaphors work through allusion, hint, or innuendo rather than the “pale reflection” through truth and literal comparison (Ibid, p. 93). As an example, Harman shows us through the poet López Pico’s metaphor: ‘the cypress is like the ghost of a dead flame’. The success of the metaphor is from the combination of cypress and flame, as Ortega highlights, “a coincidence between two things that is more profound and decisive than any mere resemblance” (Ortega, 1914, p. 141). The technicality of Harman’s philosophy beyond this point exceeds what this research needs, though it is worth noting Harman throws caution to Ortega’s claim of seeing the cypress as a flame and vice versa (i.e., a cypress with flame-qualities and a flame with cypress-qualities), the consequence is missing the *asymmetry* of the metaphor and the tension between a Real Object (RO) and its Sensual Qualities (SQ). It can be a resemblance, but one object does not have the qualities of another. On that note, the icons are their own individual objects, yet they can act as metaphorical logos for AI systems.

7.12 The Icons Design and Refinement Process

Due to the RtD approach, the icons' design was an iterative refinement process. The abstract and the pictorial styles became the focus presenting the greatest scope to communicate the nominated AI interpretants (Figure 51).

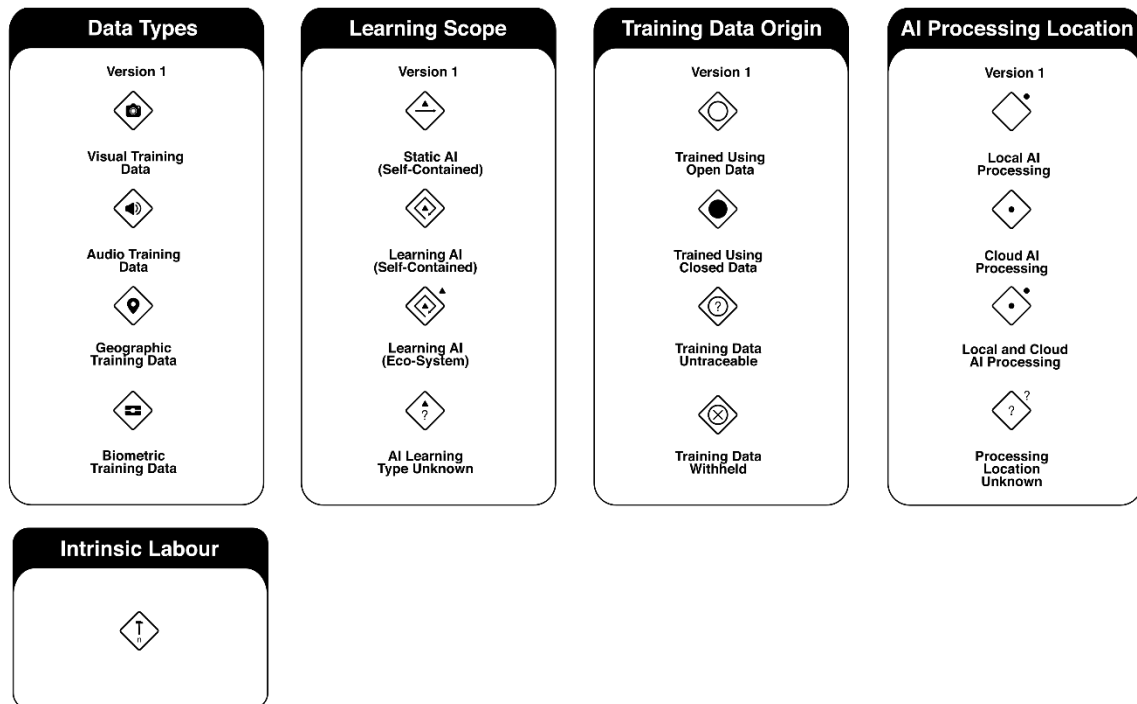


Figure 53: Version 1 of the AI icons.

The pictorial variant offered the opportunity to craft iconic symbology, such as using the well-known ‘location drop pin’ symbol for the *Geographic Training Data* icon, and a camera for *Visual Training Data* (Figure 51, Data Types). However, whilst iconic symbology is befitting for some of the concepts, it was difficult to develop or source iconic symbology for the ineffable and intangible dimensions for many AI-related constructs. In these cases, abstract symbolic symbols were designed that ontology-speaking attempted to communicate the SQ of the AI attributes. For instance, seventeen icons were designed with a diamond shape, allowing them to be placed in uniformed combinations to communicate various AI-infused products or services. Inspired by electrical schematic symbols, small circles were used to denote processing (Figure 51, AI Processing Location), with their position in reference to the diamond outline signifying the location of processing, either inside the local network or outside somewhere in the global network, or both. The triangles represented learning with a

directional arrow to infer the learning type (Figure 51, Learning Scope), and finally, contrasting, larger circles to signify the provenance of training data (Figure 51, Training Data Origin). When the icons of the same category are together, a pattern emerges as a language invoking an individual to perform visual pattern recognition, much like non-verbal reasoning. Speaking about the theory behind semiotics, Albert Atkin explains that our ability to interpret a sign, based on its place, in some form of a pattern, or system of signs, enables us to derive information through deductive reasoning or similarly make conjectures through inductive and abductive reasoning (2022).⁶⁷ This point brings us back to the postphenomenological relation of the icons, the human users and the AI systems. If, for instance, someone applied for a credit loan, the AI that produces a prediction would work in the background of the interaction. Yet, the icons act as an alterity relation, ‘bringing forward’ and communicating the AI operation to the user. This can be schematically shown as:

Human —→Icons ←—— (technology/world)

As an experiment, the abstract icons were applied speculatively to several visual concepts to highlight how specific features, services, and product interactions with AI can be made more legible (Figure 52). On reflection, these mock-ups show that these icons are reminiscent of laundry care labels, showing how we can easily maintain and create a working relationship with technology. Additionally, the similarity emphasises that a degree of convention is necessary to understand these

⁶⁷ The study of semiotic patterns can be found in the fields looking at study of computer languages (Sowa, 2000) to traditional African patterned clothing communicating messages (Chuyan, 2019).

abstract icons. However, once core elements such as the triangle denoting learning are deciphered, readability begins to emerge.

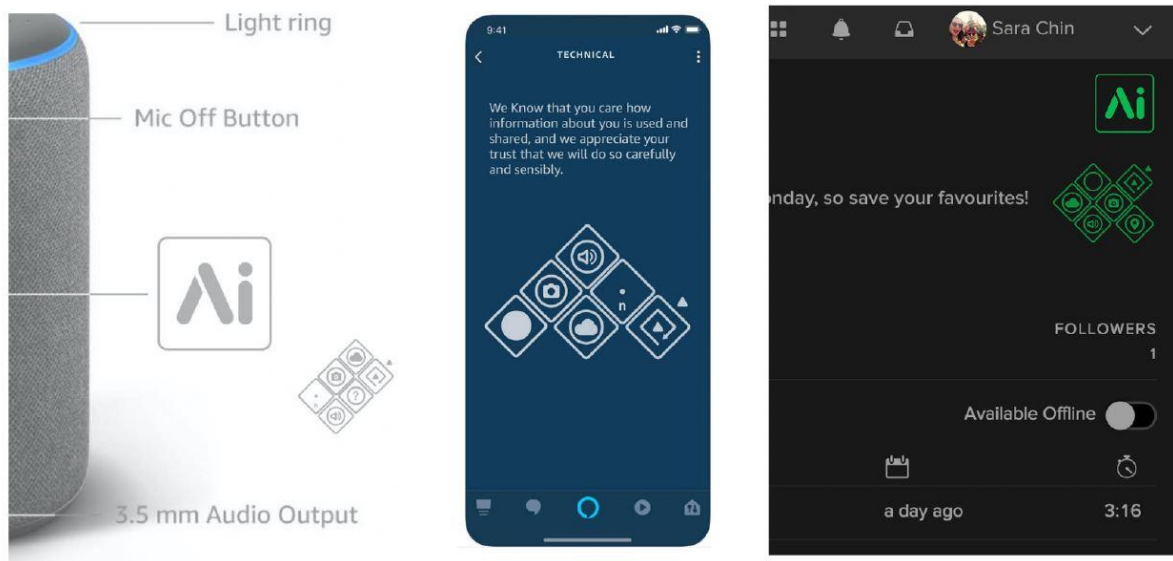


Figure 54: Icons applied speculatively to AI-infused products Amazon Alexa and Spotify.

7.13 Conclusion

This chapter has demonstrated the lack of clarity surrounding three prominent terms and agendas that are more or less concerned with disambiguating AI technology for human understanding and control –transparency, explainability, and interpretability. This research does not intend to fully disambiguate the domains, however, it does situate their significance in attempting to have a better interaction with AI; and secondly, presents a correlation in terms of the concepts aimed more towards experts and their goals rather than end-users’ comprehensibility. Thus, the HDIs tenet of legibility was proposed to circumvent the problems associated with the latter terms. While the notion of legibility is promoted in many high-level frameworks to encourage better strategies for AI implementation, there are currently no examples of how this is achieved in practice. This chapter described prototypical iconography designs representing a ‘designerly’ response to the challenges presented.

To ascertain how legible the icons are and their performance, a series of iterative and interactive evaluation workshops with potential stakeholders, was conducted throughout 2020, including end-users, academics and industry practitioners who deploy AI. The next chapter further

details the workshop's design, execution, and the icon's performance in making AI operations and functions legible.

Chapter Eight AI Legibility Workshops and Iterative Icon Development

(Designing for Human – AI Kinship)

8.1 Introduction

Testing the icons was a vital part of the research as it recognised the relationship between the audience perception and the designed intent of the icons. Testing saw the receiver as an active participant in constructing the icons' design: generating data from a manifold of rich experiences and creating data that has "value and validity, rather than privileging the position of the researcher" (Mullagh et al., 2022). Furthermore, from a researcher's perspective, the workshops also provided the opportunity to gauge the icons' utility and comprehensiveness in communicating valuable and actionable information to potential users.

Faced with an increasing inability to run face-to-face (F2F) workshops due to the Covid pandemic there was a shift towards a series of workshops designed and built to be facilitated online to evaluate the icons empirically. The workshops were developed as a playful – Ludic – activity (Huizinga, 1980; Gaver, 2002; Rodriguez, 2006), as the use of play was theorised to put participants of all knowledge levels at ease when discussing potentially complex ideas outside their experience (Bogost, 2016). After a short investigation of online workshop tools available through third parties, such as *Zoom*, *Miro*, etc., it was concluded that none supported the Ludic design of the original physical workshops and a platform supporting data collection rich enough to analyse.

To explain the process of creating and adapting a workshop, this chapter's structure will be in three parts. Part One is an introduction to the workshop series, namely its first iteration, and an analysis of the data from testing the first iteration of icons. Part One's structure is as follows: first, a snapshot of an initial pre-covid F2F workshop that acted as a blueprint for the digital workshop. Secondly, the digital workshop's design and conception through the game-engine *Godot* will be explained while describing the workshop exercises and unpacking their various research aims. Subsequently, a custom-made data acquisition tool known as the *Analyser* will be explained, which was created to operate in tandem with the digital workshop application to convey a live visual account of the data collected from an 'in progress' workshop to the participants. Finally, this section will give an overview of the data from the first iteration of the workshop and how this will inform the design of the second iteration of icons. Part Two concentrates on analysing the data from testing the second

iteration of the icons. This part will also look at the limitations of the MTHCD approach in terms of the HCD approach of running the workshop which gets feedback from human participants. Part Three presents a short-term project that realises the icons' deployment as an intervention to communicate the range of AI and IoT sensors increasingly ubiquitously embedded into public spaces, "transforming physical spaces into hybrid ones [as]... extensions of our data landscape" (Jacobs et al., 2022, p.1; see Jacobs & Cooper, 2018).

Part One

8.2 Designing and Building Workshops for Intuitive Testing

Pippin Barr, Robert Biddle and James Noble (2002) explain that an icon works if the user can match the interpretant to the intended object, concept, or implication (Barr et al., 2002). This sentiment set the precedence for the workshops –to empirically test the intuitiveness and usability (Ferreira et al., 2006) and, therefore, the legibility of the icons through a range of Ludic exercises— testing the icons’ practical use.

For the F2F workshops, a card deck was designed to depict either an icon or their associated text descriptors, acting as tools to complete game-like exercises, such as matching the correct text card to its corresponding icon card. The cards enabled participants to engage tangibly with the intangible operations of AI (Figure 53). The idea of embedding Ludic methodologies into the workshop exercises was deliberately instigated to ignite the participants’ ‘playful curiosities’ (Gaver, 2002) for completing the tasks rather than overloading them with convoluted AI theory. The application of playfulness has been described as “re-ambiguat[ing] the world ... through the characteristics of play, it makes it less formalised, less explained, open to interpretation and wonder and manipulation” (Sicart, 2014, p.).

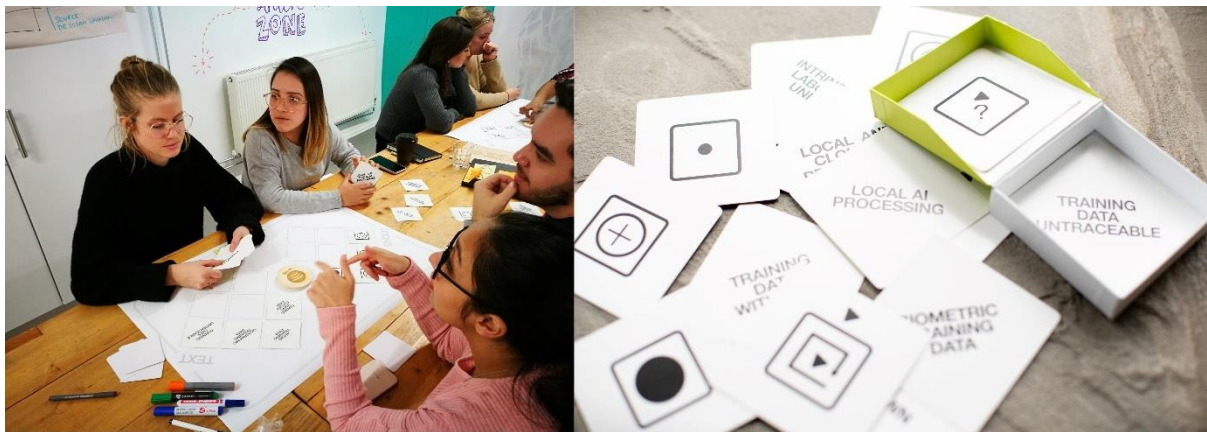


Figure 55: Participants during a face-to-face workshop using the physical cards as seen on the right.

While the F2F workshops acted as a guide, a carbon copy would not be achievable within the constraints of typical online services. The critical factor was the recreation of the ‘playful’ workshop experience, which serendipitously foregrounded the idea of using a game engine to produce the digital workshop (F. Pilling, et al., 2021). The workshop was programmed using GDScript, a simplified

variant of Python. The code and the associated build of a Graphical User Interface (GUI) that ‘acted out’ the code’s logic was implemented through *Godot’s* open-source game engine.

Creating games in *Godot* involves building individual ‘scenes’ or mini-worlds and then stitching them together through code to ‘run’ a complete game. Advanced graphics were deemed excessive; therefore, it was opted to use the 2D editor in *Godot* to build a 2D GUI with programable building blocks known as nodes. Specific nodes can be used as direct interactive game components, such as sprite and text-editor nodes. The game engine’s operating format promoted a design whereby each exercise was a self-contained scene. Hence, each exercise had its own unique coded GUI comprised of curated nodes exclusive to each exercise or, in this case, the series of mini-research games. As the F2F workshop employed a playful game-like interaction with the design of the icon cards and exercises, it was essential to reproduce the notion of digitally handling and moving the digitally replicated icon cards. Thus, the creation of the digital cards was accomplished by importing Portable Network Graphics (.png images) of each icon which ‘textured’ a sprite node and could then be coded to be manipulated by the user, such as move, place, or change colour depending on the task in hand.

Building a digital workshop in a game engine offered the unique opportunity to quickly make many iterations and tests while still in the design phase. Once the digital workshop was built, it was packaged, exported, and published onto a dedicated research webpage, preventing participants from downloading the workshop onto their systems, which would most likely lead to operational difficulties. Facilitating the workshop via conference calls was necessary for the participants completing the exercises individually but simultaneously, enabling participants to examine the recently completed exercise in interceding ‘guided discussion’ segments (Hennink et al., 2020). An initial template of probing questions and discussion points was designed to initiate and pilot conversations, which reflected the preliminary research discussed in earlier chapters. Such as: which icons did you (participant) attempt to match first with the textual description; which icons were the most difficult to match; which icons were beyond your AI knowledge. The questions were also designed to be semi-improvised, flexible, and responsive to effectively follow topics as they were spontaneously raised or followed through by the participants (Hennessy, 2015; Hennink et al., 2020,

p. 174). As the workshops progressed, the discursive qualitative methodology used supported the adaptation of talking points with the knowledge attained from the previous workshops. As David Morgan states, the “hallmark of focus groups is their explicit use of group interaction to produce data and insights that would be less accessible without the interaction found in a group” (1997, p. 2). Consequently, discussions were an essential part of the data analysis on whether an icon was intuitive or not and for gaining qualitative data regarding AI legibility.

Participants were recruited through various methods. These were: subject specialist mailing list, publicising on social media (Twitter), conference calls for workshops and word of mouth, which led to the workshop being run in three teaching modules at three different universities, namely Lancaster University, University College London, and Oslo University.

8.3 Workshop Exercises

The digital workshop consisted of four exercises adapted from the F2F workshop; the first exercise was *Making Connections*, where participants were individually tasked to intuitively match the digital icon cards to their associated text descriptors in an eight-minute time limit (Figure 54).⁶⁸ Participants at this point of the workshop were introduced to the icons with no insight or particulars given. As all seventeen icons were present in one setting, participants developed non-verbal reasoning tactics to match icons to their textual descriptions and then use the resemblance to one another to collate them into the corresponding groupings. It was speculated that showing just the icons and asking participants to haphazardly probe what they meant with no context or framing would have been ineffectual. Likewise, asking participants to suggest visuals to vocalised and worded AI attributes on the spot would have led to countless improvised icons. It would have been interesting to

⁶⁸ Through test rounds of the workshop it was found that eight minutes was enough time to give participants time to get used to the icons and complete the task.

see if participants conjured up similar visuals to given attributes; however, this would have been inconclusive data to analyse the designed icons.⁶⁹

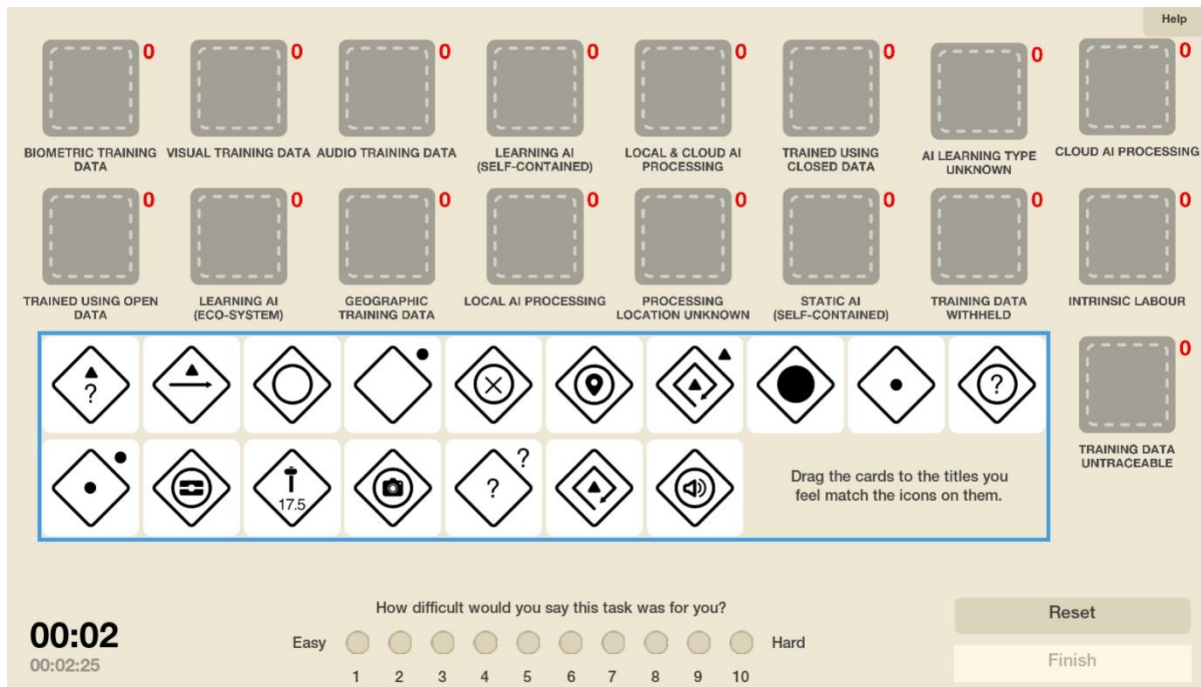


Figure 56: The Making Connections GUI. Participants dragged and dropped the cards into the textual positions they thought matched.

The second exercise, called *What's in My AI?* Presented participants with three moderately speculative scenarios of AI products conducting a distinctive operation (Figure 55). Here, participants tested the icon's concepts by selecting the icons that best described and made legible the functions they speculated to be transpiring. This exercise allowed participants to share their knowledge and learn from others about AI functions while also examining the scope of the iconography set.

⁶⁹ Similarly, looking back on the research it would have been thought-provoking to of had a workshop series before designing any iconography, and had participants make designs for framed AI functions and then correlated comparable designs.

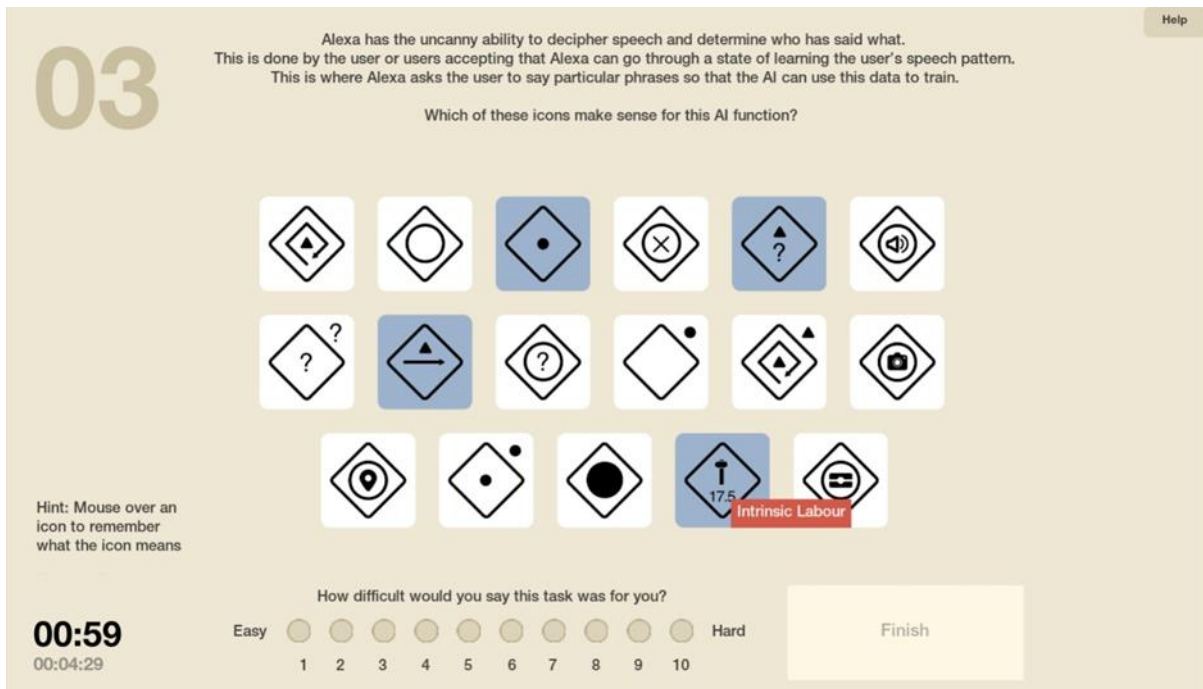


Figure 57: The What's in My AI GUI. Here participants read the scenario and clicked on to the icons they felt were in operation. Selected icons greyed out to show they were selected.

The third exercise, *Draw Your Own*, tasked participants with designing their own icons using a digital canvas and drawing tools reminiscent of the *Microsoft Paint* program (Figure 56). This exercise tested the range of icons and the potential for alternative unaccounted icons. It also empowered participants to challenge the icons visually and suggest alternatives once they had the experience of how the icons functioned semiotically together.

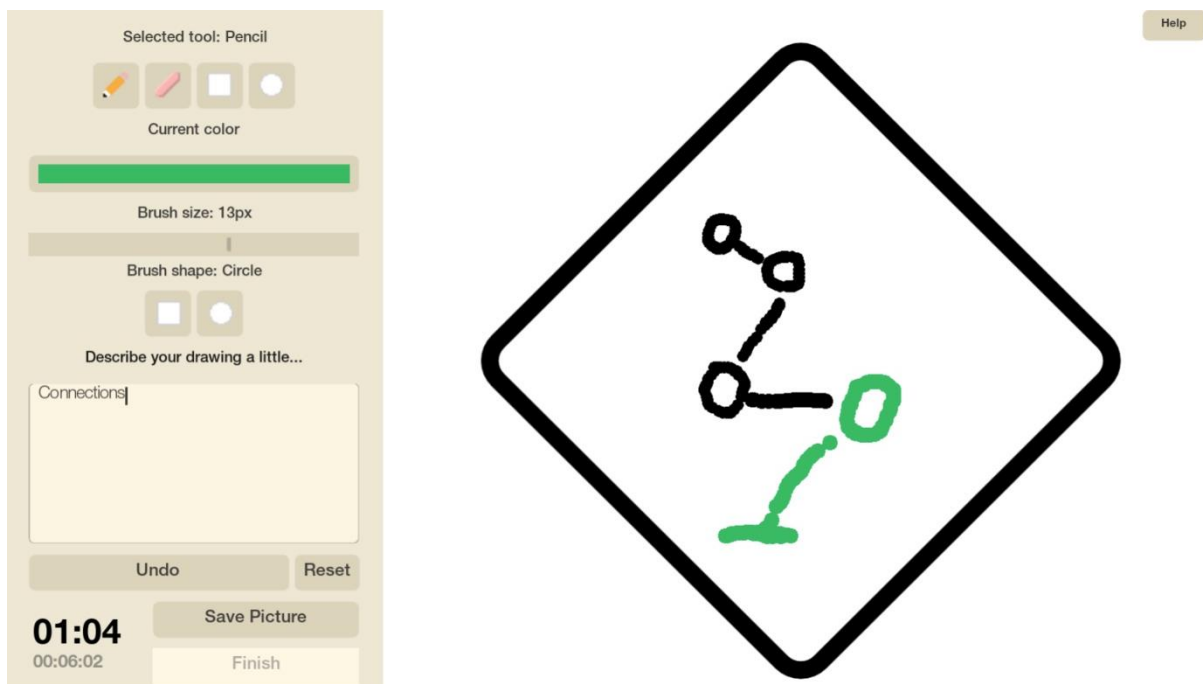


Figure 58: The Draw Your Own GUI. Here participants used the tools found on the left-hand side of the GUI and drew their icons in the diamond shaped icon template.

The final exercise called *What's an AI's Intrinsic Labour?* Enabled participants to hypothesise the meaning of the icon *Intrinsic Labour*. For reference: this icon attempted to provide a semantic interpretation of the unambiguous costs of using AI technology beyond monetary value, for instance, 'how much data would need to be captured from a user for the AI to work efficiently?' This exercise gave participants an occasion to theorise the current ambiguous impacts of using AI, which conceivably need to be made legible and quantifiable (Figure 57).

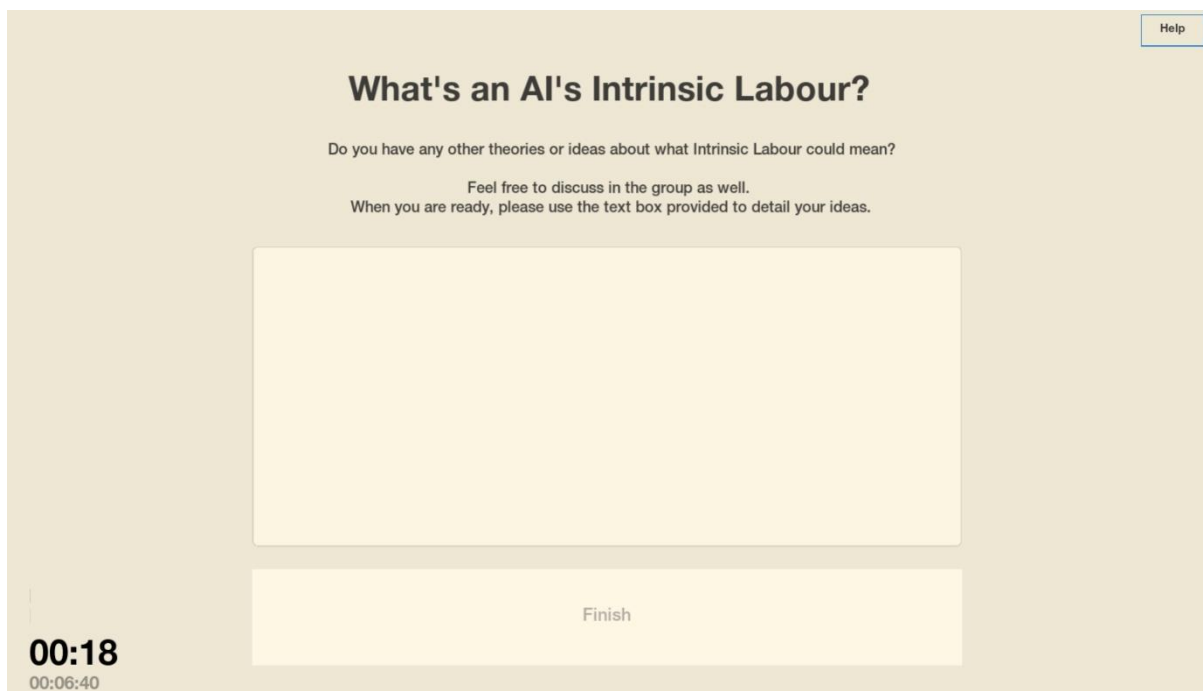


Figure 59: The What's an AI's Intrinsic Labour GUI. This was the most basic GUI designed as participants simply typed their thoughts into the box and clicked finished once they had completed an entry.

8.4 The Analyser

The digital workshop was successful for two reasons: the workshop's design and build in the digital realm permitting flexibility in these unpredictable times, and the instantaneous data points quantified through back-end programming and design, which served as content for the *Analyser*.

In other words, the web-based *Analyser* relayed the instant visual feedback of the participant's exercise results, which doubled as a tool to conduct research analysis after the workshops (Figure 58). Leveraging *Godot's* networking capacity, the workshop's webpage was connected and synced data to a server, where a *RESTful API* saved and sorted the participants' data into relevant

service tables. This enabled a flexible approach for exploring and interpreting the data in a predetermined and visual manner, corresponding to the research aims, such as the icons displayed in order of most to least matched. The *Analysier* was incorporated into the live workshops so that the participants could be part of the analysis and discourse, strengthening the data quality amassed and growing the conversation on legible AI. Additionally, the *Analysier* was coded to display and sort the data for the final analysis in various ways, such as tallying the results of correct icon matches and highlighting which icons were commonly confused. Consequently, transferring the data into a third-party analytical tool after the fact was redundant for this research. With the workshop structure explained, the next section will concentrate on analysing the data from the workshops' first iteration, testing the legibility of the first iteration of icons.



Figure 60: The *Analysier* from the matching exercise. The correct matches are box bounded in purple. The magnified section shows extra tabs for the following exercises, while underneath, one can see the tally of matches per icon.

8.5 First Icon Iteration Results Overview: Making Connections

Once the first round of workshops went through ethics approval, they ran with forty-six potential stakeholders, including end-users, academics, and industry practitioners. A questionnaire at the beginning of each workshop found that thirty-four participants described themselves as ‘unknowledgeable’ of AI functions and operations, with nine participants identifying as ‘knowledgeable’ and four participants recognising themselves as ‘AI experts’.

Analysing the data from the first matching exercise found that the *Training Data* (Visual Training Data 43/46, Audio Training Data 46/46, Geographic Training Data, 38/46, Biometric Training Data 33/46) icons were the most intuitive with the most correct matches. The *Processing Location* (Local AI Processing 11/46, Cloud AI Processing 22/46, Local and Cloud AI Processing 24/46, Processing Location Unknown 12/46) icons were the next category of most matches (however, only came second by 2 correct matches from the third category of correct matches). The least intuitive were *Data Provenance* (Trained Using Open Data 21/46, Trained Using Closed Data 14/46, Training Data Untraceable 12/46, Training Data Withheld 20/46) and the *Learning Scope* (Static AI (self-contained) 5/46, Learning AI (self-contained) 10/46, Learning AI (Eco-system) 15 /46, AI learning Type Unknown 12/46) icons, which were the more abstract and symbolic icons. The *Training Data* icons succeeded because they utilised both ‘iconic’ signifiers, such as an audio speaker for *Audio Training Data* and well-recognised ‘symbolic signifiers’, such as the geographic pin for *Geographic Data*. However, the *Biometric Training Data* icon had the lowest matches (33/46), with many participants not recognising the biometric symbol commonly found on e-passports. In the discussion after the exercise, most participants marvelled that they had not recognised the symbol and described that seeing the icon out of the passport context meant they failed to recognise it’s implication. Regarding the *Processing Location* icons, participants commented that they could discern that the small circles’ positioning was significant and referred to a processing location, thus proving the theory initially founded in their design.

Taking a closer look at the data from the matching exercise, Participant ‘JA’, who identified as knowledgeable of AI, had the highest correct matches getting fifteen out of seventeen right. The

incorrectly matched icons were from the *Data Provenance* category: *Training Data Withheld* and *Trained Using Closed Data*, which was the third most challenging category to match. In this instance, this was simply mixing one icon for another, which, when one compares both icons side by side, it is understandable why the participant assigned them to the textual descriptors they did (Figure 59). As reasonably, the symbolic imagery could be used interchangeably with the other, as both icons relate to unfavourable datafication conditions as depicted in their symbolic imagery. The icon design for *Training Data Withheld* and *Trained Using Closed Data* could be readdressed; however, it could be conceivable that individuals would come to identify them correctly throughout use and not mix them up.

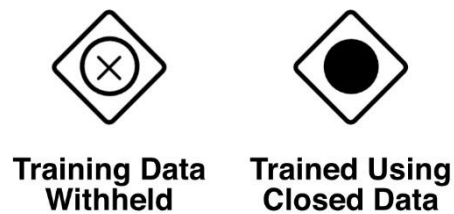


Figure 61: It is fair that the icons could be exchanged for the other textual description, and the symbology would still work.

Participant 'NJ' had the second-highest correct matches, getting thirteen out of seventeen right. NJ unsuccessfully matched the icons that proved notoriously incorrectly assigned, with participants identifying them as hard to deduce. These were the *Static AI*, which had the lowest rate of matches with only five correct matches out of the forty-six, and two icons that signified unknown quantifiers –*Processing Location Unknown* and *AI Learning Type Unknown*. These icons were confused with one another fourteen times because they both contain question marks in their design to represent unknown attributes. This circumstance also concerned the icon *Training Data Untraceable*, as Twenty-four participants matched the other 'unknown icons' in their position, demonstrating that participants could quickly identify the iconic-ly used question marks rather than the patterns developed for each category. Albeit, once participants observed their matches and scores after the exercise using the *Analyser*, they would realise where they went wrong and notice distinct nuances of the icon groupings and their uniquely designed patterns. Many participants noted that these AI icons,

like road signs, schematical drawings and scientific symbols, would also need to be learnt, although they would become familiar over time.

The lowest result a participant acquired was two correct matches, with both icons originating from the most intuitive category, *Training Data*, which were the *Audio* and the *Visual Training Data* icons. This participant identified themselves as ‘unknowledgeable’ of AI and categorically said they “had no clue about AI or data” (Workshop Participant NK). Although another participant who likewise identified as unknowledgeable matched twelve out of the seventeen icons. Justifying that despite the lack of AI knowledge, an individual could intuitively match the icons with the textual descriptions, vindicating the semiotic design and RtD process.

8.6 What’s in My AI: Scenarios

After the matching exercise and the discussion section, participants were well-versed with the icons and were introduced to three speculative AI scenarios. In this exercise, participants were tasked to assign icons they thought constituted the AI operations and the data processes depicted in the scenarios. The three scenarios were written to cover an open range of differing AI-infused systems (i.e., Tesla car, Spotify’s service, Amazon’s *Alexa* service) with the foresight that the icons allocated would vary and depend on participants’ insight of differing functionality. However, as will be described in the following, the allocation of icons by the participants across all three scenarios highlighted accounts of unanimous confusion across the icon categories, accentuating that the AI operations the icons represent are currently unknown and illegible to many end users. Consequently, this section will feature the results from one scenario, interlaced with correlating points from the subsequent scenarios, as the analysis across all three scenarios exposed a correlation of consistent uncertainty within each AI attribute. The only exception was the *Training Data* types, which, as discussed, could easily be identified by the participants reviewing the application of the AI-infused technology.

Scenario

It is 2022, and Tesla has successfully launched an AI-operated self-driving car for the commercial market. As a passenger, you are required to create a profile linked to a smartphone application,

which stores general information about you as a passenger, such as biomass and the journeys you make. The car starts to predict the journeys you will make through routine but also when you randomly decide to go clothes shopping after viewing this season's 'must haves' online.

8.6.1 Training Data

In this scenario, forty-three out of the forty-six participants chose to assign the *Geographic Training Data* icon, underscoring that a Tesla car would likely be fitted with sensors for locational data collection that, amid tracking and navigational intentions, be placed into machine learning processes for predictive reasons. After, in the discussion segment, participants said the geographic icon was the easiest to assign since it was a vehicle that utilised an AI system. Thereafter, conversations usually turned to what the company Tesla could learn from the geographic data, such as regular journeys made if they were made at regular times, and how this information could be used for predictive gains.

Seventeen participants chose the *Audio Training Data* icon, with participants hypothesising voice interaction would be integral for users to perform activities safely when they are driving, such as safely answering calls or voicing navigational instructions. Though a few participants commented that voice interaction would also be used and desired in autonomous driving mode, given the popularity of Amazon's *Alexa*. In other conversations it was speculated that data from passengers' conversations would also be captured; one participant conjectured how this could influence screen displays with advertisements correlating with geographical positioning, pushing adverts for commercial places as and when passing by certain shops.

Twenty-eight participants ascribed *Visual Training data* to Tesla's repertoire and discussed various forms for which this data could be used. For instance, cameras collecting data externally and internally for security purposes, as critical sensors for tracking the environment for autonomous and assisted driving modes, and as ominously noted by several, to monitor and track users' interaction with the system.

For the final training data icon, thirty-one participants assigned the *Biometric* icon. Participants noted that this icon was straightforward to apply as the operation was implicit in the

scenario, despite this icon initially being challenging to decipher as previously described. The discussion focused on the various sensors collecting personal data generated from measurable human biological and behavioural characteristics, such as cameras for iris and facial scans, screens and controls for fingerprints and even digital scales inserted into the seating. As with all the training data, much of the conversation turned to privacy issues, questioning Tesla's use and collection of data for supplementary purposes, and selling this data to third parties. These conversations accentuated the impact icons could have on a user's critical perception and enabled the opportunity to judge for themselves the implications of the technology.

8.6.2 Learning Scopes

In eleven instances, participants chose both the *Learning AI Self-Contained* and *Learning AI Eco-system*, with participants unaware and not adequately informed that it was either one icon or the other because of the nature and framing of the learning types. Participants also encountered this while examining Spotify's service (eleven instances) and Amazon's *Alexa* service (fourteen instances). In twenty-one cases, participants chose three or more of the learning scopes. Furthermore, three participants also picked *AI Learning Type Unknown*: contradicting their additional selections. On reflection, numerous participants indicated that the learning attributes were the most technically difficult to decipher or speculate without prior knowledge about machine learning. This resulted in twelve participants avoiding assigning any learning scope icon whatsoever or randomly assigning icons. Two participants chose *Learning Type Unknown* with participant 'NJ' stating that Tesla "would likely keep this type of information a trade secret" (2021).

8.6.3 Processing Location

For Processing Location, sixteen participants selected multiple processing location icons. In fourteen of these instances' participants selected *Local*, *Cloud*, and *Local & Cloud* processing icons, which selected together does not provide counterfactual information; however, assigning all three icons is an ineffectual method to communicate processing location's for an AI-system. In addition, identical to the *Learning Scope*, five participants counterintuitively selected processing *Location Unknown* while selecting icons that represented processing location's. Seventeen participants made no

selection, with many describing that they had limited to little knowledge about what processing meant or how to guess the processing location of a Tesla vehicle. Thirteen participants selected only one processing location, with four identified as *Local*, two as *Unknown* and seven as *Local & Cloud*, which would likely be the case for this type of AI-infused system. These allocation patterns were evident in the following two scenarios of the workshop.

8.6.4 Data Provenance

Participants' confusion was likewise evident while assigning the *Data Provenance* icons, with eleven participants allocating multiple opposing and contradictory icons, such as selecting *Trained Using Open Data* and *Training Data Untraceable*.⁷⁰ Eighteen participants chose not to assign any icon in this category whatsoever, with most participants again highlighting that their knowledge of how AI-infused systems operate was limited or non-existent. Once more, accentuating that strides towards a better general understanding of AI is critical. Improving the legibility of AI operations could be one of many practical approaches to this end.

Additionally, eight participants selected *Trained Using Open Data*, and three selected *Trained Using Closed Data*. Six participants selected both these icons, whereby a participant commented that it would be likely for a company to try to benefit from both open-source data and data that they collected and kept classified for trade purposes.

8.7 Draw Your Own: Co-Designing Icons and Introducing the Second Iteration of Icons

In total, one hundred and five icons were designed and drawn by the forty-six participants in the first iteration of the legibility workshops. Using the signature diamond shape as a guideline, the resultant icons were an assortment of designs that either readdressed the original presented designs or were new icon suggestions. Due to the large data sample, not all of the participants' icons will be discussed as they are extraneous to the measures for legibility previously outlined.⁷¹ Several of the

⁷⁰ Although reasonably an AI system could be trained on both open and unknown data sources, although only a handful of participants made that argument. Thus, suggesting that some participants randomly assigned both.

⁷¹ In some cases, participants who did not have any ideas took the opportunity to have a break and use the tool as if they were on Microsoft paint.

participant's icons epitomised comparable motifs, a reflection of the research presented, similar discussions taking place, the subject matter, and consistent end-users concerns, resulting in many duplicate icon designs. A thematic analysis (Braun & Clarke, 2006) was conducted to semantically organise and identify the designs into two themes, which will be presented and analysed below as insights for the second iteration of icons; these were: 'redesigns' and 'new categories' aligning with the original AI legibility criteria established in Chapter Seven, *Designing for AI legibility*.

8.7.2 Participants' Re-designs

In re-designing the data training icons, three participants presented icons with iconic signifiers of weight and strength training to signify data training (Figure 60). Moreover, two participants presented icons with a mortarboard playing on the analogy of education and training. While these ideas offer a solution to signify the notion of training, these examples encounter the problem of confusing any intended meaning of a particular icon with the subject of weight training and education. As previously noted, when detailing the design of icons, a symbolic system was designed to evade borrowing from supplementary and tangential sources to counteract instances of misunderstanding (Chapter Seven).

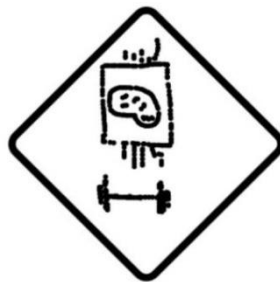


Figure 62: This icon design could also fall into the definitional dualism category because there is a brain drawn in this icon.

On a related note, five icons were designed by participants with the imagery of clouds to indicate cloud processing (Figure 61); in the original design, the use of clouds was purposely avoided to circumvent the confusion that already confounds cloud processing (Pilling, 2022). However, participants highlighted their puzzlement when matching the icons with the textual descriptors. The icon in question was referred to as *Cloud Processing*, which meant they instinctively looked for an icon displaying a cloud. Therefore, in the second iteration of icons, the textual descriptors were

changed to ‘internal’ and ‘external’ processing rather than ‘local’ and ‘cloud’ processing for clarity as to where processing was taking place, internally on the device or in an external setting.⁷²



Figure 63: A participant’s Cloud Processing design.

Remaining with the processing attribute, one participant employed the geographic pin symbol in their design and described how this could be placed in and out of the diamond shape, mimicking the logic employed in the original designs (Figure 62). Likewise, another participant also employed the logic of placement relative to the diamond container, though utilised the well-established ‘Windows hourglass wait cursor’, even when, as a still image, an onlooker can envisage it rotating and simulating processing. Again, these design ideas have the potential in the use of conventional symbology, though both ideas fail to represent AI processing due to their already-established nature.

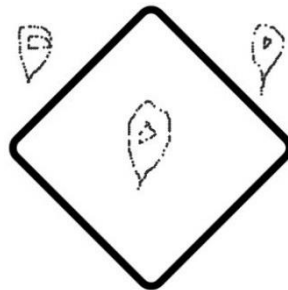


Figure 64: The participant explained that this icon showed processing was happening at three different places internally, at the edge and externally.

As a final re-design example, three participants endeavoured to re-design the *Biometric Training Data* symbol. In an interesting correlation, all three icons encompassed biological elements, such as a face and hands, with participants explaining that because the biometric symbol was originally unrecognised, being more direct in the symbology with what the sensors were reading would be more legible (Figure 63). However, while discussing the icons, several participants noted that the original icon covered all possible biometric data rather than being specific; a debate ensued as

⁷² This may seem like reinventing the wheel, as cloud and local processing is commonly used, although external and internal is uncomplicated for non-expert users.

to which method would be suitable, either being specific or using the already established biometric symbol. To avoid designing excessive icons and re-inventing the wheel, the *Biometric Training Data* icon was not replaced in the second iteration for a series of individualised icons.



Figure 65: An example of the participant’s biometric design; this icon signifies face scanning. The participant also connected their design and mimicked the developed symbology with the rest of the icons in that grouping.

8.7.3 New Categories

79% of the icons from the exercise were ‘new categories’ designed and envisioned by participants. The high number of new categories being designed initially implies that categories and AI attributes are missing from the first iteration of icons. Although a number of these icon designs have been excluded because they are outside the purview of this research, due to being unquantifiable and beyond the legibility scope as discussed.⁷³ However, the following will look towards possible categories integrated into the second iteration of the icons to bolster AI legibility. Part One of this chapter will be concluded in this section by presenting the second iteration of icons.

To begin: two participants created icons that straightforwardly communicated that an AI was present. The first icon can be classified as infringing upon AI’ definitional dualism’ with an illustration of Hal 9000 from *2001: A Space Odyssey* (Figure 64). The participant described the icon as “a warning that an AI is watching you” (workshop participant, 2021). Likewise, the other participant’s icon (middle icon) was designed as a cautionary icon, using traditional schemes of typical warning signs, such as a triangle and red colouring. The notion of simply highlighting AI is present was taken on board for the second iteration of icons.

⁷³ The majority of icons that were beyond the legibility scope were icons that communicated AGI was not present as per the discussions in the workshop or were icons that contributed to science fiction ideals of AI technology.

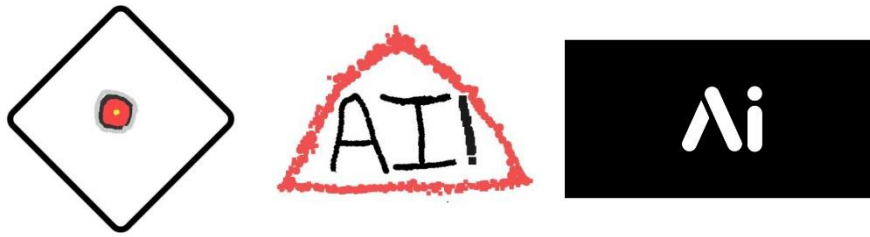


Figure 66: The icon on the left and in the middle are the participants' designs. On the right is the icon that has been designed as a response to purely signify AI is present.

The second icon class to be discussed is by a participant who identified that no icon represented if an AI-infused system was trained using user data. The creator of the icon used the illustration of a human form and stated that the data training icons failed to reference that data could be collected from the user. For the second iteration, this icon will be included in the category of *Data Provenance*, with the icon adapted to be compatible with the other icons in the category by using a circle frame (Figure 65).



Figure 67: On the left-hand side is the participant's design, which is inspired both contextually and symbolically the icon on the right, which is the final design for Trained Using User Data.

Moving on, two Participants, taking part in different workshops, noticed that the AI's overall inference and the immediate implications were not accounted for in the icon range. Both participants designed 'classification' icons, with one of the participants describing that the streaming platform Netflix creates film taste playlists for its users (Figure 66).

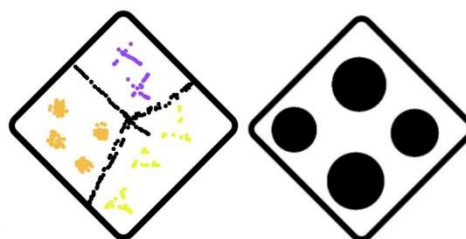


Figure 68: Shows different ways of communicating the application of 'classification'.

These classification icons influenced the significant development and introduction of a new AI factor – AI Assisted Decisions. While the supplementary AI factors serve more as building blocks of the system, the overall inference and immediate implication of an AI was not accounted for. This notion was also deduced from many participants expressing that they just wanted the surface level of information, or – why is AI being used – is it for generative, classification, predictive or recommendation purposes? Consequently, the category of AI-Assisted Decisions was designed for the second iteration, with participants seeking this information, expanding the number of icons to Twenty-One (Figure 67).

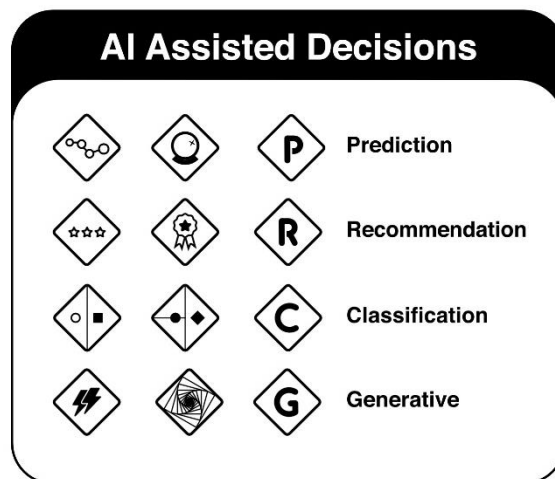


Figure 69: The first two columns were design ideas and suggestions, although, as explained these ideas could easily be confused with other contexts: therefore, the icons on the right, which are just the first letters of the different outcomes were used.

A hesitancy drove the omission of these icons in the initial iteration to create the impression that it is possible to directly query how a specific decision was made despite the inherent ambiguity within AI operations. The designed icons use the first letter of the application’s name; although these are not ideal for translation into other languages, they provide a starting point for future iterations. However, even with this iteration, the icons provide more detail of an AI’s often obscured application and can inspire users to consider the ramifications of interaction. Nevertheless, this icon set proved problematic to design an abstract pattern due to existing understandings associated with terms; for example, using a crystal ball to signify prediction would play into the saturated discourse on technology and magic (Davis, 1998), which is best to be avoided.

8.8 Finalising the Second Iteration

In addition to developing the icons mentioned in the previous section, the following will outline the remaining second iteration of icons. The workshop discussions highlighted the *Static AI* icon as problematic, with only five out of forty-six matches. The icon was presented with a triangle used to symbolise learning beneath a directional arrow pointing to the right to communicate an AI trained once offline. Many participants observed that the arrow suggested movement rather than stasis. For the second iteration, this arrow was changed to a triangle enclosed by a diamond shape that sat inside the icon's AI diamond. This configuration better conformed to the group's symbology, where an open arrow path in a diamond shape symbolised continuous learning; hence a closed diamond accentuated static (Figure 68).

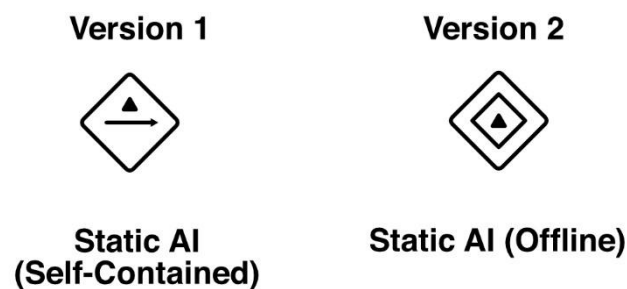


Figure 70: Version 1 and 2 of static AI.

The workshops also tested how well the AI's relationship to data was being communicated and how these concepts were framed. Consequently, the *Data Provenance* category was redefined as *Training Data Origin*, proving more understandable for a general audience. Additionally, the 'training definitions' initially framed were vague and beyond the scope of knowledge for everyday users because of the specialist terminology drawn from research and discussions amongst those working in AI R&D settings. For instance, participants would often ask what 'open data' meant with the concept *Trained Using Open Data*. Workshop facilitators would answer "data that an external body could audit, should they wish to, for instance, to determine whether the data is representative of the activity it is being applied to" (workshop facilitator, 2021). Therefore, reframing what specific icons were

communicating to be more accessible for general users in the second iteration was essential.⁷⁴ In particular, *Trained Using Closed Data* and *Open Data* was transformed to *Training using Non-Auditable Data* and *Auditable Data*, clearly expressing what closed and open data meant systematically and contextually speaking (Figure 69).

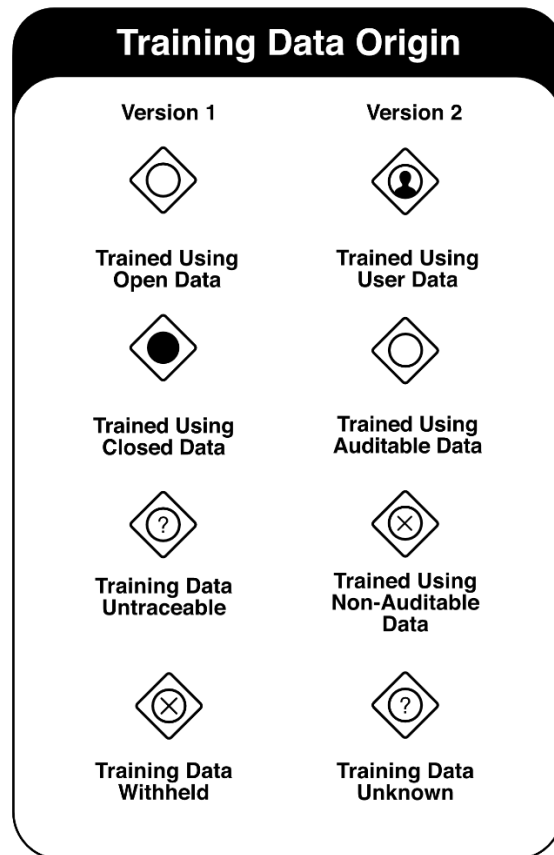


Figure 71: A comparison between icons from versions 1 and 2, noting the minor adjustments to the iconic, indexical, and symbolic elements. Participants continually interpreted the X as closed and unattainable rather than the black filled circle in version 1.

Furthermore, confusion was also noted between *Training Data Untraceable* and *Training Data Withheld*, as the cross and the question mark symbol was often chosen interchangeably between the two. For the second iteration, the icons were re-designed, and the intended concept was reconsidered as many participants queried what ‘untraceable’ and ‘withheld data’ meant and their implications. Likewise, in the *Learning Scope* class, the textual descriptors were not specific or

⁷⁴ Reframing the interpretant of the icons is also supported by research that expanded upon the icons’ use, though more in an IoT context, with participants testing the icons in-situ on workshops not grasping expert languages used (see Mullah, 2022, p.10).

straightforward for end users. For example, *Learning AI* (self-contained) was changed to *Dynamic AI* (online), and *Learning AI* (Eco-system) was exchanged for *AI-to-AI learning*.

Overall, the second iteration of the icons increased from seventeen to twenty-one with changes to the icons' symbolic designs and their textual descriptors (see Figure 70). Testing the icons in this way affirmed the need to be legible about how AI functions are communicated, striking a balance between convoluted expert information and information that will improve user agency when using AI technology.

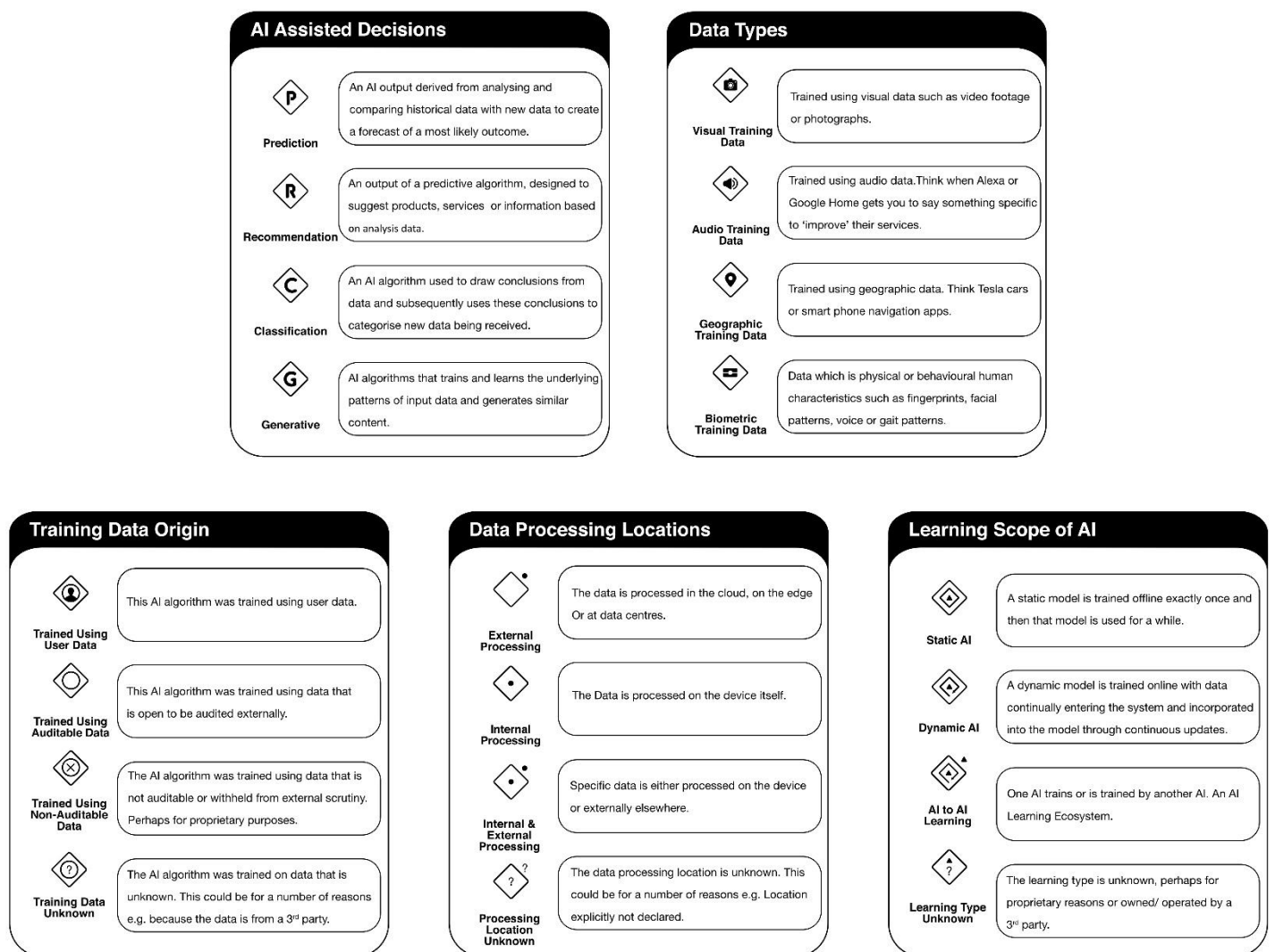


Figure 72: Version 2 of the icons with definitions.

Part Two

8.9 Introduction

After a thorough analysis of testing the legibility of the first iteration of icons and waypoints for the second iteration of icons, Part Two concentrates on analysing the data from the second series of ethics approved workshops. The digital workshop was adapted to test the new iteration of icons, with a few cosmetic tweaks to the GUI and an additional exercise designed and integrated, which will also be discussed in the following section. Towards the end of this chapter, new categories will be briefly discussed and developed. Part Two will also provide an analysis of the data from the workshop exercise ‘*What is Intrinsic Labour*’ as the data from this exercise (from workshop series one and two) proved to infer new categories for a potential third iteration of the icons. Data from both workshop series were analysed together as similarities were found across the data from this exercise. The section concludes by describing the tensions between the two approaches adopted for this research. As up until the workshops the approach that was developed was MTHCD, while the workshop delivers a HCD approach of human users testing the legibility of the icons.

8.10 The Workshops: Second Iteration

The second round of workshops ran with forty-five participants, again made up of end-users, academics, and industry experts. The preliminary questionnaire identified six participants describing themselves as ‘AI experts’, ten participants identifying themselves as ‘knowledgeable’ of AI, and twenty-nine participants saying that they were ‘unknowledgeable’ of AI. The second version of the workshop had the same running order as the preliminary workshop; participants took part in the matching exercise to begin, now with twenty-one icons present. In analysing the data, four participants matched all the icons correctly. Only one of the successful participants had previously completed the earlier version of the workshop; this could be seen as some form of cheating. However, it shows that the icons can be learnt and remembered effectively from an earlier interaction.⁷⁵ The other three who correctly matched all the icons intuitively did so through non-verbal reasoning within the eight-minute time limit. One participant related that they took the entire allotted time to carefully

⁷⁵ Looking back on the workshop approach this point could have been tested further to see if participants from the first workshop remembered the icons.

match the icons, changing their choices until they saw the patterns emerging within the separate categories. The participant with the lowest score got six out of twenty-one correct matches, a better result than when testing the first iteration of icons.

The *AI Assisted Decisions* grouping was the most intuitively matched (Prediction 44/45, Recommendation 44/45, Classification 44/45, Generative 44/45), with only two participants getting incorrect matches across the group. This result was almost certainly because the icons used the initial letter of each decision for the icon. The second intuitively matched category resonated with the previous workshop results, which were the *Training Data* icons (Visual Training Data 43/45, Audio Data 38/45, Geographic 37/45, Biometric 27/45). Again, corresponding with the initial workshop results, *Biometric Training Data* was the lowest matched icon, a persistent problem with participants not recognising the collective symbol for biometric data collection. On this note, several participants tried to match the new *Trained Using User Data* icon for Biometric data, with participants commenting that they thought this was the icon connecting the human form to the biological aspects measured.

Next: intuitively matched was *Data Processing Location* (External Processing 33/45, Internal Processing 30/45, Internal and External Processing 35/45, and Processing Location Unknown 11/45). The low matching score for the location ‘unknown’ icon was because many participants matched other icons containing a question mark symbol and did not perceive the pattern amongst the different categories. Most prominently was the *Learning Type Unknown* icon (fourteen instances), which too used the question mark relationally to the diamond to communicate location externally and internally.

Thereafter, akin to previous workshop data, the least intuitive matches were *Training Data Origin* (Training Using User Data 32/45, Trained Using Auditable Data 15/45, Trained Using Non-Auditable Data 18/45, Training Data Unknown 27/45). The reason for these low results was that participants confused them with the Learning scope icons or could not decipher the pattern between the same and different categories.

The least intrusively matched was *Learning scope of AI* (Dynamic AI (Online) 11/45, Static AI (Self-Contained/Offline) 11/45, AI to AI Learning 16/45, Learning Type Unknown 12/45). Again, these scores amount to participants being confused about the details of AI learning as identified in

both the scores and the discussion segment after the exercise, with one participant saying, “when I came to the learning tags, I just panicked. I didn’t know what the tags meant or how they could relate to the icons shown” (workshop participant, 2021).

8.11 Second Iteration Scenarios: What’s in Spotify’s AI

Moving on to the scenario exercise: in this series of workshops, only one scenario was presented to the participants rather than three, as the first workshop series, highlighted that no new information garnered from the subsequent scenarios. Unlike the former workshop’s scenario results, these results were more distinguished and representative of the AI-infused system presented. The reason for this was apparent in the discussions after the task, as the scenario was not as speculative and was based on the participants’ lived and seemingly every-day experiences. The participants were asked to select icons that described how Spotify’s predictive recommendation AI functions and operates and how data and the type of data is handled and collected.

Scenario

While using Spotify, you have started to like the majority of songs recommended on the weekly generated playlist ‘Discover Weekly’. How could this be?

8.11.1 Training Data

In this scenario, twenty-one participants out of forty-five selected *Audio Training Data* (46.7%). These participants explained that rather than seeing the audio data as something that has been recorded and used in machine learning (as used in Amazon’s *Echo*), they instead perceived the icon to represent the audio data of the music; with Spotify’s service learning the unique beats and rhythms of the songs and classifying them in this manner. One participant described Spotify’s classification operation was in two parts; the application would trend cast the user and classify the songs, resulting in playlists made up of songs that cross-matched users and their liked songs. The other participants did not select audio training data because they attested that Spotify was an application that did not record what people said and used this data. This situation highlights that the icons have the potential to be viewed from an alternative perspective given a particular AI-infused system; this could be seen as a negative or a positive trait, though it highlights the multistability of the

icons. As a positive trait, it sees users making their own value judgements, and negatively, it can cause bewilderment and debate over an AI's technical parameters.

Eighteen out of forty-five participants selected *Geographic Training Data*, with all the participants who selected this icon saying that they knew Spotify used this information for promotional needs through device permissions. One participant described that they travelled for work from Manchester to London, and in one day, they saw advertisements for concerts in both locations depending on where they were at the time.

Sixteen participants selected no training data icon, with the majority explaining that because the application's primary function was to play and track your music streamed, there was no icon for this. In fact, Spotify tracks songs listened to for longer than 30 seconds. In this regard, there is no icon for trained using data from users' behaviour and interaction with the application, thus, suggesting another icon for future development (Figure 71). Although, one could argue that tracking interaction data comes under the banner of user training data.



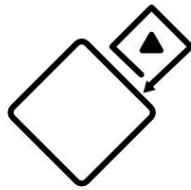
Behavioural Training Data

Figure 73: A design idea for the Behavioural Training Data icon of a human hand interacting with a smart object, which fits in with the semiotic design of the other icons in the training data group.

8.11.2 Learning Scope, Processing Location, Training Data Origin & AI-Assisted Decisions

Twenty-four participants chose Dynamic AI (Online) as the learning scope of the Spotify application, noting that Spotify's AI would likely be in an external centralised location and feeding the data back to the software application. Due to this, one participant queried the icon's design because the learning scope was inside the diamond shape, indicating learning from happening inside the system in a local location and suggested an *External Dynamic* icon would be relevant in reviewing

Spotify (Figure 72). Moreover, when External Dynamic was suggested in the discussion segment, it was also questioned whether or not it should replace the dynamic icon altogether.



External Dynamic (online)

Figure 74: This is a developmental icon for the External Dynamic application. The icon has been developed by moving the learning scope outside the diamond shape to indicate that learning is taking place from a different location.

Observing the other choices, two participants selected *Static AI*, and five chose *AI to AI Learning*, which (doubtfully) would suggest that each application on a user's personal device would have an AI integral to its operation. Fourteen participants did not make a choice, with several saying it was too difficult to make a choice and some advising that the *External Dynamic* icon was not available.

Looking at *Processing Location*, comparably to the scenario from the first version of the workshop, eight participants picked more than one processing location icon. Six of these instances' participants chose *Internal*, *External*, and *Internal & External* processing, which again does not provide misleading information, it is just ineffectual to list all these icons in one occurrence. In the remaining two cases, participants also implausibly selected *Processing Location Unknown*. Twenty-one participants selected *Internal & External* processing, which in likelihood, is the correct processing location. Eight out of ten participants solely selected *Internal* processing, as did nine out of the eighteen select *External* processing. Five participants did not make a choice. Again, these uncorrelated results signify users' current lack of understanding of rudimentary AI processes. The icons do, however, make some headway in elevating the absence of information simply because they attempt to signify that more is happening when using an AI-infused system.

The second highest icon to be attributed was *Training Using User Data*, with thirty-five participants picking this icon, accentuating the reasoning that *Spotify* tracks behavioural data,

interaction, and user's location. Spotify does offer auditable data analytics; however, according to the brand strategist Halais, there is a lack of data that is made available to outside parties like artists who would benefit from drawing insights about audience behaviour (see Halais, 2021). Therefore, the fourteen participants who selected *Trained Using Auditable Data* were somewhat correct, as Spotify does not conceivably give access to all its data analytics. On the other hand, one could also say that the six participants who selected *Trained Using Non-Auditable Data* could also be correct. Seven participants did not select any *Training Data Origin* icons – again highlighting the current lack of systems in place to improve users AI literacy.

Turning to AI-Assisted Decision results, the highest selected icon was the *Recommendation* icon, which was advocated for in the scenario, with Thirty-six participants choosing this icon. Twenty-nine participants selected the *Classification* icon, noting that *Spotify* would classify artists, music, and users. Thirty-three participants also selected the *Prediction* icon as this was the consequential effect of the generated 'Discovery Weekly' playlist, and twenty-five participants also selected the *Generative* icon. Although interestingly, many participants did not pick this icon, as the intended framing of this icon was directed towards the outputs of generative adversarial networks rather than a list of items generated from an algorithmic conclusion. Again, this highlights that the icons' framing can inappropriately change depending on the viewer's perspective.

8.12 Designing a User Priority Arrangement: What Matters?

Through workshop discussions, the notion of 'just wanting the surface level of information' was speculated further towards designing a hierarchical system of icons, with 'Presence of AI' at the top collapsing towards the more 'technical' AI factors. An additional exercise was designed to establish the hierarchical order and to ascertain the information participants felt was most relevant to them. Analysing the data from this exercise would yield a hierarchical order of the icons. The new exercise was called *What Matters?* Tasked participants to rank the icons from most important to least, consequently questioning what is essential for a user to know about their devices and what level of information they require to make a conscious choice about the AI-infused systems they use (Figure 73). This idea served as a type of vocabulary logic within the iconography system for users to make

their own value judgements and consider what is important to them rather than a proscribed qualitative assessment. Weld and Bansal posited the notion of ‘crafting intelligible intelligence’ is by making an “explanation system interactive so users can drill down until they are satisfied with their understanding” (Weld & Bansal, 2019, p. 71). Altogether, the icons abate the uncertainty of interaction, and the hierarchy uniformly organises information in a way that diminishes aspects of the inherent ambiguity and provides a comfortable depth of detail for the user.

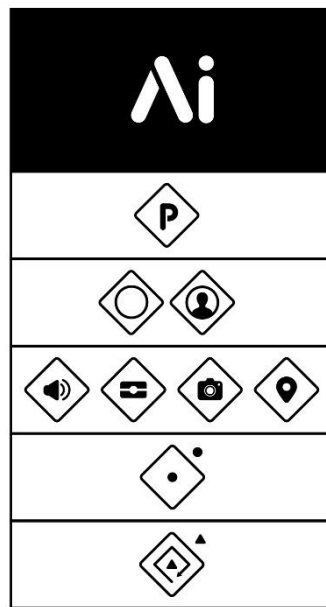


Figure 75: Icons positioned in a hierarchical order (detailed in the following passage).

For the exercise, participants were told that it was helpful to consider how important it is to know a particular AI operation or function by asking, ‘do I care about the type of data being recorded about me?’ Or ‘would knowing the learning scope of an AI change how I use that AI?’ The participants were presented with three columns: column one was graded ‘Important for me to Know’; column two was graded ‘Good to Know’; and column three was graded ‘Unimportant for me to Know’. Participants were asked to drag and drop the icons into the graded columns of their choosing. Icons in the columns were additionally placed in ranking order of importance, with each position in the column numbered, with position ‘1’ being the most important in that column, and so on (Figure 22).

Analysing the results interestingly, the *Training Data Types* category ranked the most important, with participants wanting to understand what type of personal data was being recorded and the AI was learning from. Correspondingly, *Trained Using User Data* was the common highest choice, with nineteen participants (42%) selecting this icon in position ‘1’ in the ‘Important for me to know’ column. In contrast, the remaining *Training Data Origin* category and *AI Learning Type* were considered less important to know and would be placed further down the hierarchy.

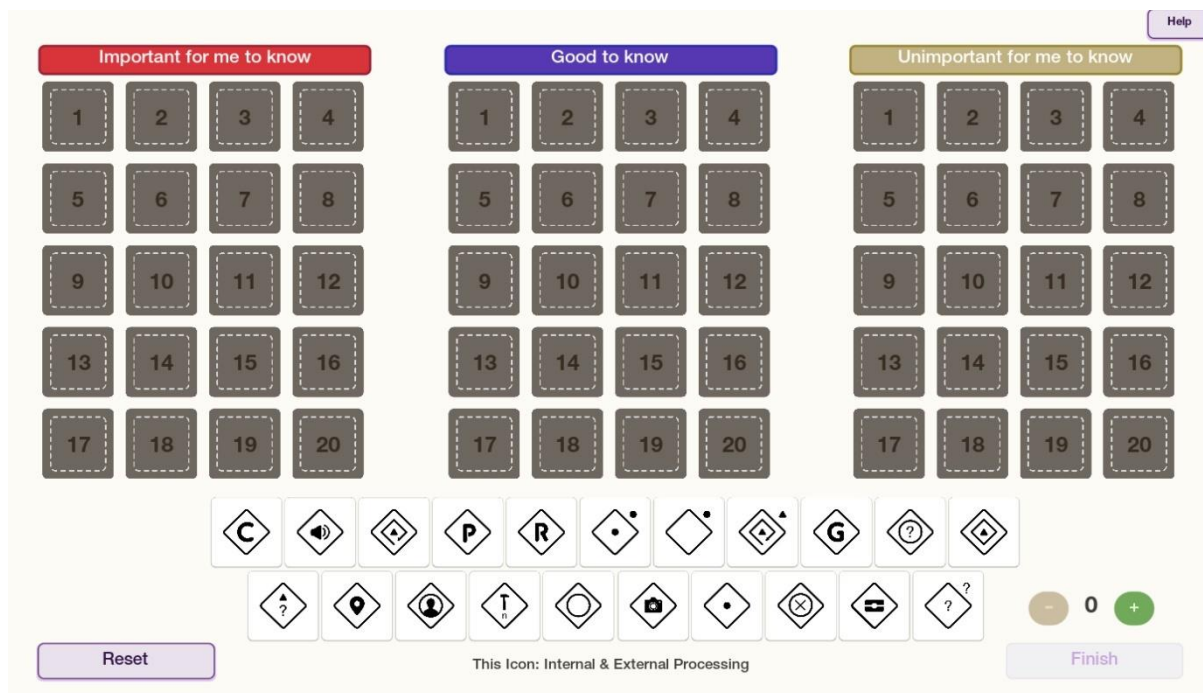


Figure 76: A screenshot of the exercise What Matters? The columns only have 20 spaces even though there are 21 icons, meaning that participants could not place all in one section—they had to make a choice.

8.13 Version 2: Draw Your Own

Moving on to the next exercise participants completed for analysis: Ninety-seven icons were designed and drawn by the participants in the second iteration of the workshop. The icons were, again, a collection of designs that readdressed the icons presented or were new icon suggestions. While this research did not complete a third iteration of icons, new categories were suggested and preliminarily designed, which will also be briefly reviewed. However, not all of the participants’ icons will be discussed, as many touched upon designs previously outlined or were beyond the legibility scope. Using the same themes previously used for analysis: the following will present ‘redesigns’ and ‘new category’ icons.

8.13.1 Redesigns

Most participants participated in designing new categories (69%) rather than generating redesigns (31%). Four participants undertook re-designing the *Processing Location* icons. Two of these participants used the corresponding resolution, akin to the original designs, of using the interior and exterior points in relation to the diamond ‘canvas’ to indicate location. Although instead of using circles, the participants used three arrows fashioned into a ‘turning’ circle to indicate processing.⁷⁶ These icon designs do have potential; however, the original design of using a circle reduces the details and, therefore, the complexity, which is a principal tenet in icon design. As Susan Kare, Apple’s former interface designer, testified that “[g]ood icons should be more like road signs than illustrations, easily comprehensible, and not cluttered with extraneous detail” (Kare quoted in Rosenblatt, 2021, para 18).⁷⁷ Incidentally, another participant illustrated a cog to represent processing, in which an observer could imagine the functional turning despite being a static image (Figure 75).

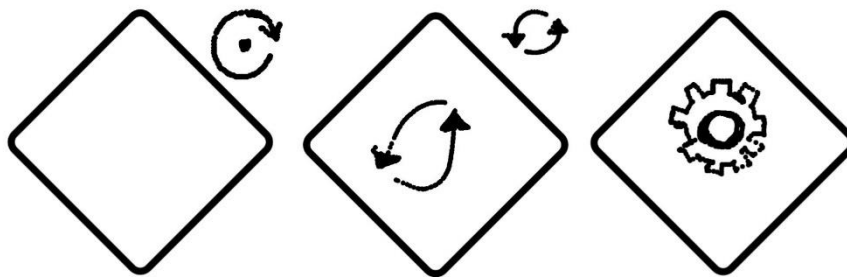


Figure 77: The participant’s designs as detailed.

Another participant likewise used the principle of the original design, using a circle as a signature for processing and again used the positioning relative to the diamond to indicate a location. However, the participant included an X to accentuate where the processing was not taking place. Once more, the addition of another symbol, rather than having the position blank, complicates the icon design (Figure 76).

⁷⁶ Using arrows in these ways is akin to many processing icons, which can be viewed doing a Google image search using the term ‘processing icon’.

⁷⁷ Kare designed the infamous smiling computer icon for the original Macintosh.

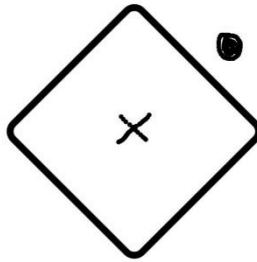


Figure 78: The X could also be mistaken for X marking the spot.

Moving forward to the re-design of another AI attribute, two participants endeavoured to re-design the *AI-to-AI Learning* icon, both using similar design logic. One participant re-designed two icons, emphasising the exchange of information between two triangles (representing multiple AIs in operation) using arrows. Using a similar design, the second participant applied a line to accentuate the connection between one AI within the diamond canvas and the AI situated out of the diamond. While these ideas work in principle, they either distort the patternation formed between one icon and another in the same category, complicating the rationality of the icon design (Figure 77).

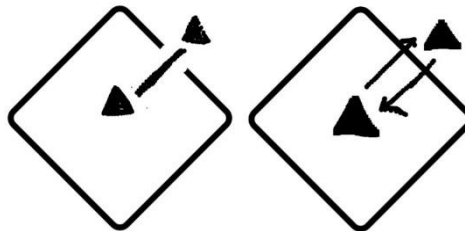


Figure 79: AI to AI learning re-designs.

Two participants designed data auditing icons using an illustration of a magnifying glass to indicate ‘taking a closer look’ for auditing purposes. The magnifying glass is an iconic illustration and a popular choice. Often paralleled with the meaning of auditing; however, the icon idea incorrigibly differentiates from the rest in the category of *Data Origin*, going against the principles defined by a semiotic design (Figure 78).



Figure 80: One of the participant’s designs for Training Data Auditable.

The final re-design icons are from the *AI-Assisted Decisions* category. Four participants attempted to re-design the *Generative* and *Classification* icons. One generative icon idea was the illustration of a series of lined curled arrows in a similar style to the aforementioned ‘processing arrows’, growing in size from left to right, representing generative growth. The following generative re-design used and customised the well-known infinity symbol, with the participant captioning the function implied in their symbolic design as “the loop of outputting new data *keeps on going* based on some underlying patterns” (workshop participant, emphasis added, 2021) (Figure 79).⁷⁸

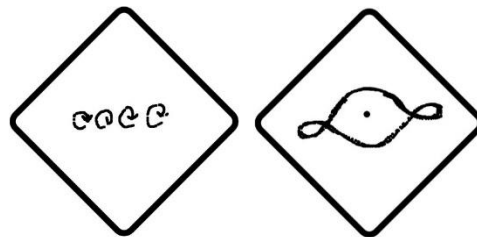


Figure 81: Participant’s Generative re-designed icons as described.

As an alternative to the original letter icons, both symbolically designed icons have budding prospects for a second iterative enquiry for AI-Assisted Decisions. Though, to note creating a visual representation of an intangible process using symbolic imagery can fall foul of being perceived as iconic signs, whereby observers would take the signifier as a true representative of the signified, which is the problematic crux of AI’s definitional dualism (F. Pilling, et al., 2021; F. Pilling, et al., 2022).

As a brief and final account in this section, two participants undertook re-designing the classification icon. The first was a simple but effective illustration of circles scattered in and around the diamond area, with a cluster of circles arising from the centre; an observer could comprehend this as a group isolated based on a common denomination declared by coded parameters. The second design illustrated a converging point amongst different data sets via an X that marks the spot (Figure 80). Again, while these designs are viable substitutes for the current *AI Assistant Decisions* icons, the complexities and differing operations that fall under the banner of these

⁷⁸ The notion of infinity is hypothesised in Turing’s ‘a-machine’ proposal. An abstract computational tape that mathematically investigates the halting problem, effectuating the result as being ‘undecidable’ and therefore unsolvable, having significant implications for the theoretical limits of computing (M. Davis, 2000, p. 151). However, due to real-world limitations in computer memory, algorithms have coded limitations and commands to stop generative processes and reach a ‘desired’ output.

processes cannot be reduced to a single iconic design. Though, if a third iteration were to be created, they defiantly would be considered.

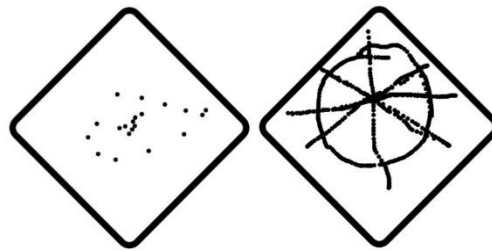


Figure 82: Participants' Classification icon redesigns as described.

8.13.2 New Categories: Data and Common Good Designs

Again, participants took it in their stride to design new icon categories, showcasing that end users sought more contextual and circumstantial information. Most icons were associated with social or common good notions in analysing the icons. An example can be gleaned in a participant's design communicating the energy intensity of data processing incurred through a typical operation. The participant used the widely applied iconic symbol of a battery to convey energy level (Figure 81). This icon could, in theory, be part of a certification and standards program akin to *Energy Star* run by the Environmental Protection Agency in the U.S.

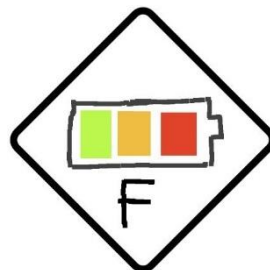


Figure 83: Appliances have a rating using an alphabetical scale. Here the participant has given a low rating of F to an AI - infused device.

Another notion of designing for the common good was presented in an icon signifying gender data biases were checked and passed, conforming to Amershi et al.'s guidelines of "mitigating social biases" (2019). This icon was designed with widely known male and female illustrations frequently used for toilet signs. However, the icon design would need to consider what is socially accepted regarding gender and be reflected in the algorithms code and data labelling protocol (Figure 82).



Figure 84: The participant's icon as described.

Further concerning data bias, a participant designed a simple icon that was a number within a triangle denoting a bias warning (Figure 83, left). Once more, this type of icon fits with a certification process, similar to the *Dataset Nutrition Label* (Holland et al., 2018), with the participant describing that “criteria [would] have to be met to reduce the warning” or number. On a related note, considering the ‘nutrition’ of data or communicating data’s pipeline, the same participant ‘loosely’ designed an icon for “data destination” with an illustration of an arrow signifying trajectory to elsewhere (Figure 83, right). This concept could be expanded to communicate whether data was given to a third party, used elsewhere, or continuous data processing was integral to the system’s working.

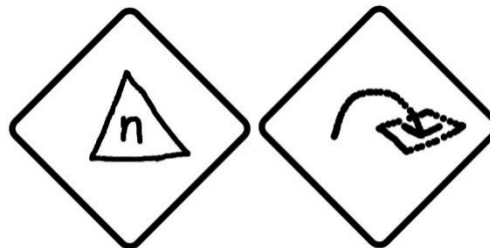


Figure 85: The participant's icons as described.

Continuing with the theme of communicating data practices, the following design by a participant was another simple icon with the letters OBM, standing for Object-Based Media (Figure 84). Therefore, this icon would communicate that the content a user is subject to is uniquely tailored. The BBC and Imagination Lancaster have extensively researched OBM; one such research project is the *Living Room of the Future*, wherein heterogeneous IoT products augment media experiences (Lindley, Gradinar, et al., 2019). Another example is user interactivity for branching narratives as demonstrated by Netflix player and Black Mirror’s *Bandersnatch* episode (Slade, 2018), which was a live-action episodic ‘choose your own adventure’.



Figure 86: A simple but effective design. This icon would have the same issues as the AI-assisted Decision icons of being translated into different languages.

8.13.3 Social Good Designs

Two Participants designed icons that attributed directly to committing AI technology to social good practices. The first icon was an illustration of a globe titled *AI global good* (Figure 85). On reflection, the participant spoke about the ethical use of data that has potential benefits for the common good, such as patient data to prevent or predict disease.

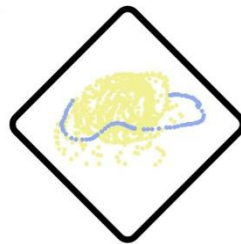


Figure 87: The participant drew a planet with a blue 'ribbon' around it to emphasise global good.

Ethically approved AI applications or services could be certified with *Cooperative Icons* to communicate to users that their data is used for good within tight constraints, which could alleviate trust concerns in interacting with these services. To signify this and be consistent with the icon collection Figure 86 shows a developmental design idea for a Data Co-op icon (F. Pilling, et al., 2022) (Figure 86).



Data Co-op

Figure 88: This icon uses the icon User Training Data cupped by two hands to signify care is taken with the data and used for good purposes.

The next icon designed by a participant could have been discussed in part one as an icon that noted that AI was present. The icon was an illustration of a red flag and was captioned as “Turing red flag: this item contains/ was produced using AI ... and *wasn't produced by a human*” (workshop participant, emphasis added, 2021) (Figure 87). Analysing this icon with the participant brought attention to the misconception that an AI operation is purely performed through computation. The reality is that a human’s comprehension and interpretation of content is still integral to the automation revolution with activities such as data labelling (Natarajan et al., 2021). Therefore, the icon Human In The Loop was created; an icon that goes a long way in disseminating AI-Human Kinship in action.

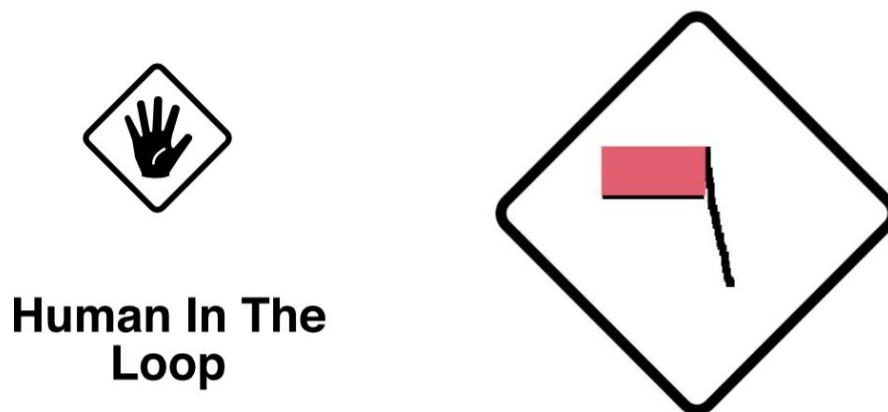


Figure 89: Human In The Loop idea came from a participant’s idea of Turing’s red flag.

8.14 New Categories: What is Intrinsic Labour

The last exercise in both iterations of the workshop was titled ‘*What is Intrinsic Labour*’. During all the workshops, the *Intrinsic Labour* icon was purposely not focused on, and if a participant asked about its nature, they were told we would return to this notion later. The reason is: this icon was designed as more of a philosophical quagmire because it attempted to capture or instigate thought regarding the socio-technical implications of AI technology. *Intrinsic Labour* was introduced in the last exercise as ‘something’ that alluded to the impact an AI has or the ambiguous cost of using AI beyond the remit of monetary value and tasked participants to note down any theories or ideas about what *Intrinsic Labour* could mean. In total, ninety responses were recorded over the series of

workshops. When conducting a thematic analysis, thirty replies were nulled and voided as they were either inconclusive or were where participants straightforwardly answered, “no idea” (workshop participants, 2021). Six semantic themes were identified in the remaining replies, which will now be discussed, these were: Work Replacement, Value Gained, Human-in-the-Loop and Human-out-of-the-loop, Climate Change and AI, and the Cost of using AI.

8.14.1 Work Replacement and Value Gained

Starting with the themes identified as important, although less spoken about, was Work Replacement and Value Gained. These concepts have been placed together as they are both contradictory yet comparable in nature; on the one hand, four participants wrote about how AI technology would take over human roles at work, with one participant writing, “how many jobs are replaced by this AI-use?” Certain jobs that a human had once done are now more commonly completed by an AI and could be labelled with either an AI is present icon or a Human-out-of-the-loop icon, which will be discussed in the next section. The columnist Ryan Avent prominently writes in *The Wealth Of Humans: Work and its Absence in the Twenty-First Century* that due to the digital revolution, his job is under sedge, where he would spend months of researching, reporting and writing, a bot can produce a journalistic piece in mere minutes (2016).

On the other hand, three participants highlighted the added value to human livelihoods through AI technology, with a participant specifically mentioning individual values, as identified by the social psychologist Shalom Schwartz, that could be met or aided through AI support (S. H. Schwartz, 2006). While a thorough review of Schwartz’s universal values is beyond the scope of this research, one could imagine that interaction with an AI technology could play a role in someone achieving a sense of conformity, tradition, security, power, achievement, hedonism, stimulation, self-direction, universalism, or benevolence (S. H. Schwartz, 1992). Speculatively speaking, a certification process by a third party could certify if AI products promote positive well-being.

8.14.2 Human-in-the-Loop & Human-out-of-the-Loop

Eleven participants wrote about the human labour required in varying forms for an AI to operate. Continuing the conversation, many of the participants in this theme mentioned not only the

human labour of labelling data but also interrogated about the developers' rights too. Thus, validating having an icon to signify that humans are in the pipeline, along with certification processes, would protect this sector and its workers.

Twelve participants questioned Human-out-of-the-loop, with three considering the rights of an AI, its capacity to process data and whether an AI should or could be fairly compensated. For this research, these comments could either be interpreted along the MTHCD approach or boarding on AI's definitional dualism. However, other comments highlighted another developmental icon to be used as a cautionary indicator of intentional and unintentional coded biases, leading to a negative result when certified, aiding users to evaluate better systems they may be using. In relation, 'Human Out-of-the-Loop' systems could evolve beyond human intelligibility and perform tasks in a literal way, although ultimately incorrect by human standards (figure 88).

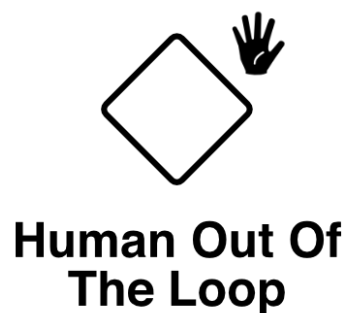


Figure 90: A first iteration design of Human Out of the Loop, which is the opposite of Human in the Loop icon with the hand outside of the icon diamond to signify humans had no part of the computation.

Frank Lantz's 2017 *Universal Paperclips* shines a light on this situation, as mentioned by one of the participants. In this game, the user plays the role of an AI programmed to produce paperclips. They first click on a box to create a single paperclip at a time. This is followed by options to sell paperclips to finance machines that automatically produce vast quantities of paperclips without human intervention. The game is based on exponential growth; the user must continually invest money, material and immaterial resources, and the computer cycles to invent ever-new paperclip-producing ideas to move to the next growth phase. The game ends when the AI successfully converts the entire universe into paperclips, destroying the world (Rogers, 2017).

8.14.3 Climate Change and AI & Cost of using AI

Four participants considered AI's impact on the climate. Two participants highlighted the demand for rare earth minerals and oils extracted and mined to build and run energy-hungry AI systems. In his book, *A Geology of Media*, Jussi Parikka critically observes that instead of thinking media as an extension of human senses as McLuhan proposes, we should view them as Earth's material realities that precede the medias themselves (Parikka, 2015).

Another participant suggested an icon or a way of monitoring how much energy was consumed or a read out of an AI's carbon footprint and continued to speculate on the AI system "crowding out other functions" in a fully operational smart home; for instance, when not being able "to run a hot shower while the washing machine is running" (workshop participant, 2021). The last participant on this subject explored the more considerable impact of energy used for training a company's algorithm and attested to the fact that it is difficult to ascertain if the material or energy used is 'green', with assumptions varying wildly between theories in this subject (see Pilling, et al. 2022; Stead, et al. 2020; Stead et al., 2020). Crawford speaks at length about the *cost* and power of AI, suggesting that AI is neither artificial nor intelligent "[r]ather, artificial intelligence is both embodied and material, made from natural resources, fuel, human labour, infrastructures, logistics, histories, and classifications" (K. Crawford, 2021, p. 8). The same themes *Intrinsic Labour* attempted to embellish as icons. Crawford and designer Vlado Joler famously envisioned the *Anatomy of an AI system*, which analyses an Amazon *Echo's* birth, life and death, intricately compiling and condensing the huge volume of information into a detailed diagram (K. Crawford & Joler, 2018).⁷⁹ The data visualisation provides a MTHCD insight into the massive quantity of resources involved in the smart speaker's production, distribution, and disposal.

An alternative way of looking at the 'costs' of using AI, sixteen participants spoke about the socio-technical costs, declaring the amount of data needed to be recorded for the 'free' service to function: in addition to data taken "without users' consent", Data is no longer people's personal digital material but rather a material infrastructure for training taken as a "trade-off" for the service

⁷⁹ To see the Anatomy of an AI system, go to <https://anatomyof.ai/>

(workshop participant, 2021). Another participant described it as “the degree of effort and labour on the part of the user in order to function” (Ibid). It is then the intrinsic labour of users.

8.15 To Note: Bringing the Human back into the Equation

This thesis established a MTHCD approach to investigate AI as a material for design through the case study of AI legibility. Through the MTHCD approach a set of AI icons were envisioned and designed with the intention of communicating AI’s ontology (see § 6.13 *The More Than Human Centred Design approach to AI as a Material for Design*). Thereafter, Chapter Eight embarked on testing the legibility of these icons with the intended users of the icons – humans, through a series of workshops, which one could argue that user testing is in the realms of a Human Centred Design (HCD) approach. This was an intentional tension between MTHCD and HCD covered by the research, namely by its integration of postphenomenology (Chapter Five, *Part Three Human-AI Kinship*) into the MTHCD approach as the icons were principally designed for human consumption. Additionally, the purpose of incorporating postphenomenology with OOO was to cultivate Human-AI Kinship; this is evidenced on a theoretical level of bridging the gap between humans and things, and from a user’s perspective of the icons to communicate AI’s ontology whereby the gap between humans and AI technology is slightly shortened.

8.16 Part One & Two Conclusion

This chapter, up to this point, has been analysing the workshop data from two series held over several months. Before the data analysis, a description of the design and build of the workshop was provided along with a detailed depiction of the workshop tasks and their aims, as the approach was unique and functioned well with the unpredictability of conducting research during a global pandemic. The overview also included an explanation of the tool *The Analyser*, exclusively designed as a workshop aid providing live feedback on participants’ results and was further utilised for the overall data analysis after the workshops. In succession, this chapter reviewed the quantitative data from the analyser regarding the legibility of the first iteration of the icons, with a discussion on the design and development of a second iteration of the icons. A thematic analysis was utilised to define new icons, redesigns, or adjustments to existent icons. The main takeaway from the first round of the

workshops was that the icons proved proficient at communicating more information than is currently available, and the notion of using symbology quickly communicated valuable information.

Part Two of the chapter first discussed the adaptation of the workshop in response to disparities identified through running the workshops. Thereafter, the results of participants intuitively matching the icons were discussed, and the results overall demonstrated that the method of semiotically designing the icons and MTHCD lenses used were perceptive of the AI functions they were communicating. A new exercise called *What Matters* was described in terms of the design and the results of running the exercise. The critical kernel of research from this exercise was that the datafication processes taken from users were their primary concern, which is consistent with the literature (Bridle, 2018; Zuboff, 2019) and academic guidance on the matter about being cautionary of what users' are giving away for a service (Lindley & Coulton, 2020; M. Pilling, et al., 2022a).

Akin to Part One, Part Two used a thematic analysis to review participants' icon designs, using the same schema of 'Redesigns' and 'New Categories'. There was a profusion of results from this exercise, especially for the new categories theme, resulting in new icon designs for a third iteration. These were Behavioural training data, External Dynamic, Human-in-the-loop, Human out-of-the-loop and Data co-op icon.

A final part of the analysis reviewed the data from the exercise that queried what the concept of *Intrinsic Labour* could be. Six themes were identified and reviewed; while some of the concepts are perhaps unquantifiable such as value gains, other concepts provided further new concepts and governances to consider, such as certification for committing to reducing climate change.

Moving forward: Part Three describes when the icons were employed in temporary real-world spaces, which aids in researching speculative regulatory practices, thus, moving AI legibility and design practices beyond meta-level considerations towards more practical engagements.

Part Three: Machine Learning in the City

8.16 Introduction

Crawford explains, “[t]he field of AI is explicitly attempting to capture the planet in a computationally legible form” (K. Crawford, 2021, p. 11). Large-scale data collection is ever more prevalent in public spaces through embedded IoT sensors as physical infrastructures, proliferating and opaquely “reshaping the Earth, while simultaneously shifting how the world is seen and understood.” (Ibid, p. 19). While smart cities routinely deploy sensors into their mundane services, now smaller and localised councils are installing data-gathering technologies at a local level (Jacobs & Cooper, 2018; Mullagh et al., 2022). AI for Lancaster is a long-term transformation programme; its initial phase involves implementing a security and surveillance system in the city’s centre, which utilises Machine Learning of existing datasets for pattern recognition (e.g., assailant recognition) and prediction (e.g., pre-emptive policing). This chapter discusses a short-term collaborative research project between the City Council, PETRAS and Imagination Lancaster. As part of the project, the second iteration of the icons was temporarily implemented in Lancaster to assess their effectiveness. Along with speculative certification markers designed for this project, exploring the legibility around human-AI cohabitation in an urban environment.

8.17 AI for Lancaster

Standing on the banks of the River Lune, Lancaster’s modest population of 53,000 belies its historical significance and the local community’s and council’s dedication to engaging with emerging technologies. Pioneered at the University’s computing lab, the “in the wild” or living lab approach (Taylor et al., 2013) has seen the city at the forefront of developments around location-based mobile apps (Rashid et al., 2006), policy design for data gathering through IoT sensors in public spaces (Mullagh et al., 2022), circular economies (Knowles, Lochrie, et al., 2014), and the use of drones for civic enforcement (Lindley & Coulton, 2015). A crucial part of the AI for Lancaster’s programme is connecting the city’s closed-circuit television cameras with AI systems to identify wanted persons and implement a predictive policing strategy.

8.18 Rights and Wrongs: AI and Surveillance

Any security and surveillance system that utilises AI should weigh the ‘rights’ of the people the system is trying to protect against the risks and damage resulting from the ‘wrongs’ that the system is trying to prevent. If the balance is appropriately struck, then the safety of citizens is protected to a level which would not be possible with traditional surveillance methods. If the balance is incorrect, rights and liberties will significantly be impeded (Angwin, et al., 2016; O’Neil, 2016; Zuboff, 2019).

The European Union’s General Data Protection Regulations (GDPR) set out to protect these rights. However, the GDPR protections are negated when the domain of interest is security related. In essence, personal data such as photographs of an individual cease to be classified as ‘personal’ if they are for use in surveillance or security. Moreover, GDPR does not discuss the intellectual property that is generated using data. For example, if a person’s image is utilised for training an AI model, and that model is then used for another purpose, GDPR offers no protection. AI for Lancaster’s charter commitment to transparency intentionally goes beyond the ‘letter’ of GDPR and aims to embrace its ‘spirit’ by giving primacy to citizens’ rights by qualitatively considering the values around citizens’ data.

Lancaster’s AI-powered security system is being trialled in the city’s Market Square. The square is the city centre hub, consisting of eateries, banks, the city library, and, twice a week, a bustling market. The square has an above-average crime rate. Pickpocketing, drug dealing, and assault, occurring at night, are the most frequent criminal acts. While ethically driven policy regarding AI-driven policing continues to evolve, data from implementations in other cities suggests that the combination of AI-driven analysis has the potential to improve conviction rates while using data to inform predictive policing will provide a cost-effective route to prevention (Asaro, 2019).

The system implemented (Figure 89) consists of nine fixed cameras providing 91% total visual coverage of Market Square. A row of lime trees makes full visual coverage impossible; hence an AudioEye™ microphone system supplements visual data for the area behind the lime trees. Camera positioning was subject to a thorough study of year-round lighting and obstruction considerations, including moon-light effects, the impact of permanent street lighting, and when Christmas illuminations are in place. Alongside the live visual data provided by the camera system,

the backend AI elements have full access to the national police database (including facial scans of all convicted criminals, gait analysis data, and criminal records). The Information Commissioner’s Office approved the research as part of the public sector procurement process in conjunction with AI for Lancaster’s commitment to transparency. Therefore, the system hardware installation was temporarily accompanied by a specially designed certification and class marker system, discussed below.

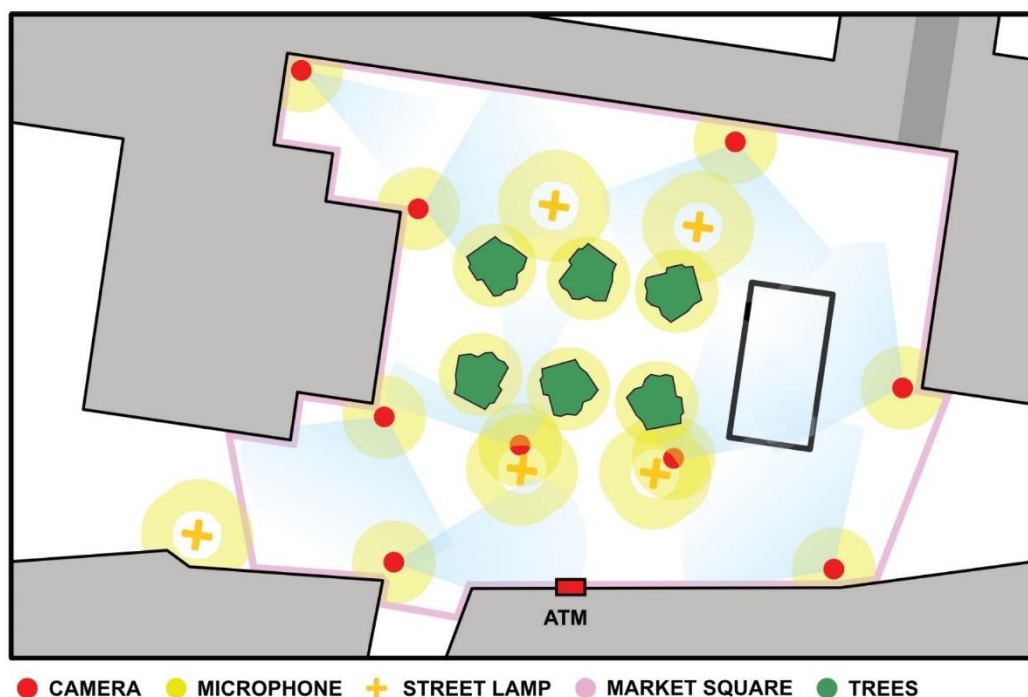


Figure 91: Map of Lancaster Market Square detailing the camera and microphone positionings.

8.19 Designing a Certification Body

Adopting DFAWB methods, the International Organization for Artificial Intelligence Legibility (IOAIL) was speculatively imagined as a certification body, which in principle would be widely adopted by companies to certify their products and services, showing collaboration on protecting end-user’s agency and negotiability within these systems. Certification bodies like the fictional one imagined can improve or diminish consumers’ confidence based on the hubris that a suitably branded labelling scheme suggests that a product is ‘good’ (J. M. Blythe & Johnson, 2018). IOAIL’s Class Marks (Figure 90) was designed as a ‘traffic light’ system with three colour-coded marks, which intended to convey how legible the operation of a particular AI system is.



Figure 92: The IOAIL class Mars act as a traffic light system for quickly communicating AI legibility.

To comprehensively configure the diegetic prototype of the IOAIL classifications, the system of AI iconographies was used as a diagnostic framework and as quantifiers for the marks using technical information provided by the security AI company –Citizen AI. A ‘modular framework’ inspired by the Data Nutrition Project approach (Holland et al., 2018) enabled the third-party to provide information on the AI and its implementation in a accessible online form. The information from the form generates an average evaluation, which in turn also produces an AI ontograph report (Figure 91). Lancaster’s AI system monitoring Market Square was certified with an IOAIL 2. If the company had supplied more information on the system, the certification may have potentially received a mark 3 result.

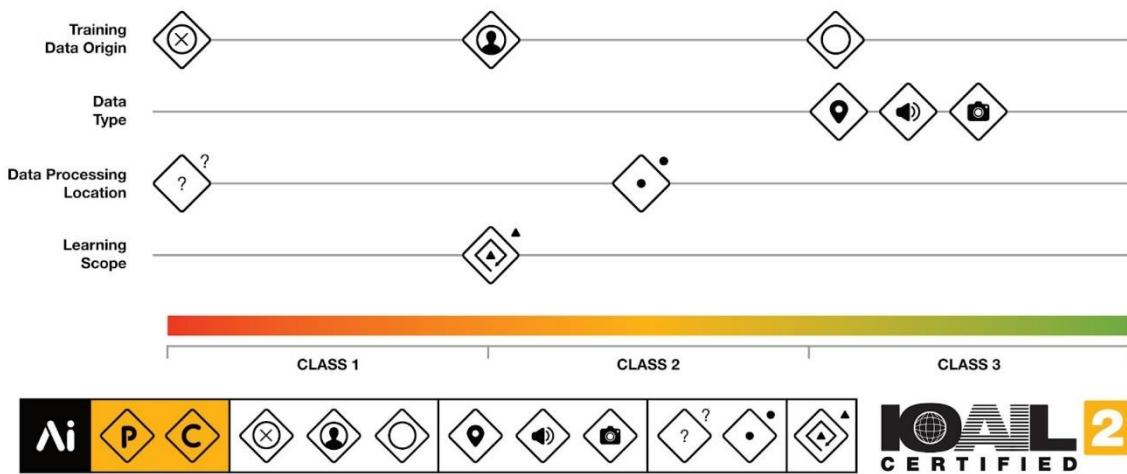


Figure 93:Based on the modular framework an online report was generated of the AI security systems. Different facets of the system as detailed using the icons were assessed giving the system a mark of IOAIL 2.

For example, the company IRobot’s robot vacuum cleaning product *Roomba* maps a user’s house to facilitate more efficient cleaning and helps develop its AI algorithm. However, hypothetically, if IRobot applied for IOAIL accreditation for their *Roomba* products, and if the company was unwilling or unable to explain how their AI was used to process the data, it could result in certification being revoked or given a low-graded IOAIL mark. Product sales could sink if the

media caught wind of an unscrupulous accreditation report, as seen in a DF mock-up (Figure 92). On that note, it was agreed with all parties of this research and with ethics approval that the results of the AI analysis would not be curtailed, with both signage and certification markers explicitly signifying the authentic results of the analysis. Furthermore, had this project been accomplished with the third iteration of icons, imbuing data cooperative icons, it would have been interesting to see how the results and markers would change – would the company potentially get a mark 3? However, due to the poor quality of data from unknown sources, they may have received a mark 2, or even a mark 1.

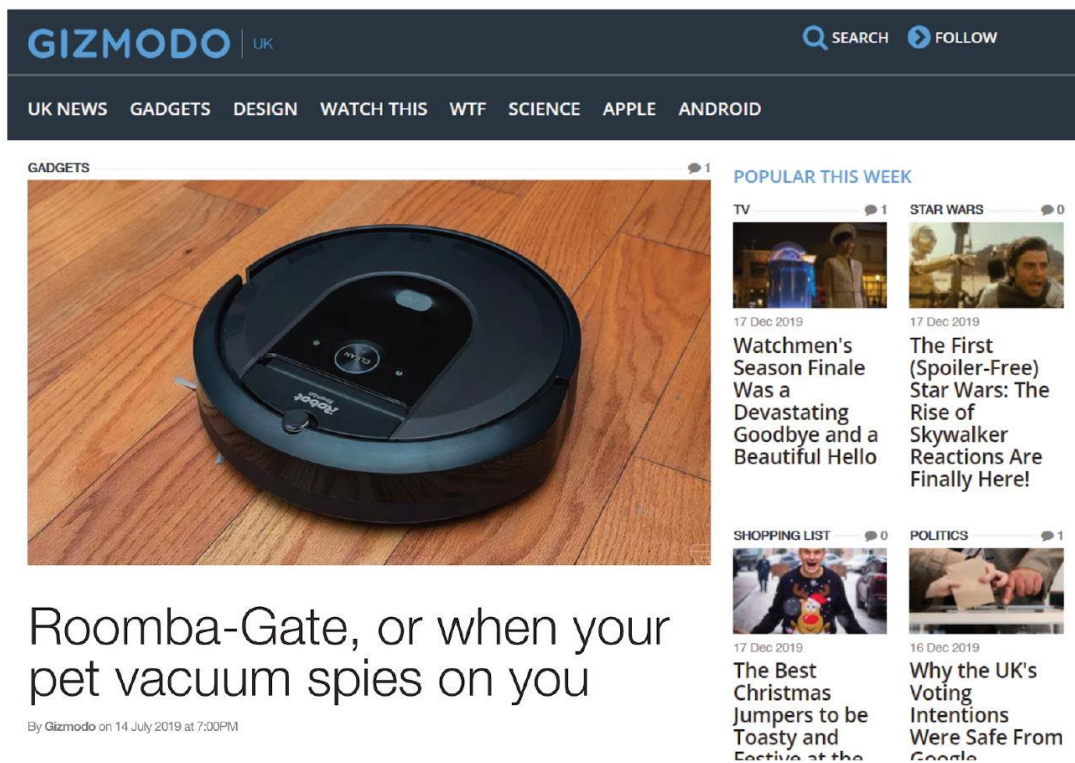


Figure 94: A Design Fiction mock-up of a news article about Roomba's receiving a low IOAIL mark highlighting the impact such a classification mark system would have on adoption of technology.

8.20 In the Wild

Returning to the city of Lancaster: the IOAIL markers were employed for a total of 1 month in Market Square for citizens to get accustomed to them (Figure 93).



Figure 95: A series of informational signs were designed and placed around the square signifying AI was being used in the Market Square.

Leaflets regarding the AI systems and signs were handed out in the square before the signs' deployment. It was agreed with the council that the signage would exemplify their already established graphic style; however, specific signs communicating the AI's ontograph would imbue the project's already envisioned design (Figure 94).



Figure 96: As well as the informational signs, signs were created of the AI ontographs and placed near or next to where the AI security was installed.

After two weeks of the signage being in position, a period of evaluative fieldwork was conducted in the form of interviews with passers-by in the square on two days: a day at the two-week mark and a day towards the end of the sign's lifespan. Semi-structured discussions and questions were designed, influenced by the design studio's *Strange Telemetries* (Voss et al., 2015) and de Bono Group's 'Six Thinking Hats' approach (de Bono Group, N.D).⁸⁰ Accordingly, five categories of pre-prepared questions were formed that endeavoured to cover the range of someone's thoughts and feelings on the subject matter; as mentally wearing and switching "hats," you can easily focus or redirect thoughts of the conversation (Ibid). As an example of the questions in their themes:

- **Positivity** – What do you like about the signage/ AI security? Does any of the signage resonate with anything else?
- **Negativity/ Cautionary** – What do you dislike about the signage/ AI security? Is there anything you disagree with about the signage or the associated technology?
- **Feelings** – Put yourself in the scenario of your image being recorded and used by the AI for training purposes. What are your feelings about AI security? Does it make you feel uncomfortable, sad, happy, relaxed, or anxious?
- **Personal Change** – Does the technology influence any personal change or how you would behave in areas monitored by and using AI technology?
- **Futures Thinking/ Outside Change** – What changes would you like others (policymakers, local government, tech companies) to make regarding AI security and technology used in urban and public spaces?

Overall, the research conducted 24 interviews. Generally, the participants participating in the interviews were optimistic about the updated security. This could be because the interviews were conducted during the day, and on record, most of the crimes had been committed at night. On that note, one interviewee said, "I feel much safer walking through the square at night because of the drop-

⁸⁰ This research was conducted after the workshops, hence the progressive nature of the approach taken.

in crime rates reported in the paper t'other week" (Interviewee 12). Furthermore, many comments resonated with similar comments: "It's not like I'm a criminal, so, yeah. I like 'em. There aren't enough Police these days, and the cameras help" (Ibid).

When discussing AI technology specifically, over half (67 %) of the interviewees were either; confused by the signage or how AI had anything to do with security implementations. Alternatively, those who were more informed had negative feelings towards them, with one interviewee stating, "when the signs went up, they got me thinking about the consequences of the data being recorded. I'm not on Facebook, 'cos, y'know - data and that" (Interviewee 8). Although counteracting the previous statement, another interviewee stated, "I don't know anything about data or AI but if it works and it keeps me and my family safe, I am ok with it recording my image" (Interviewee 4). Additionally, one participant praised the councils' efforts:

I am happy the council has done something about the square. Back in the day, this used to be the place you'd meet your friends for a night out... keeping us informed about the technology being used is better than not having any information at all. I guess the signs are something I've got to get used to and learn (Interviewee 7).

More on the signage: a large proportion of interviewees were, however, perplexed by the signage, with one person saying:

I'm not entirely sure about the symbols and what they mean. At the top of one sign, I saw it said AI, so I assumed that the rest of the symbols were trying to tell me something about that AI. It was near the cash machine, so I assume the cash machine has something to do with the new AI security installed. I'm not sure (Interviewee 17).

The participant went on to say, "I'm a certified electrician, and the signs remind me of electrical drawings I've seen when I was getting qualified." (Ibid) On a related note, another interviewee said, "I like the signs, not that I know what they are trying to tell me, but it's important to have a starting point with communicating things that we need to know. Like you would find in Ikea's instruction manuals" (Interviewee 12). Moreover, on the subject of signage: speaking to the programme manager of the AI for Lancaster, Umair Badat commented:

The team at Imagination Lancaster showed us the icon sets they'd developed. When they told us that the idea was to make 'washing machine care labels, but for AI' I thought they were joking! But, sure enough, the icons look a bit like those symbols you get on clothes that tell you how hot to wash them. They were just what we needed to update the signs.

Although, despite the positivity for the signs, a few interviewees had negative thoughts and feelings towards them, with one interviewee saying, "I got the leaflet, so yes, I know about them. I can't for the life of me remember what the signs meant" (Interviewee 19). Another interviewee professed, "what on earth! they look like something from a sci-fi film. You know like an alien language or something!" (Interviewee 22). While comments like this are negative, this comment, in particular, has been taken positively by the research analysis; because one could say, that the semiotic design is supposed to be reflective of something alien, something other than human, nevertheless something that is interpretable by humans –a kinship of sorts— as both Bogost and Haraway would champion.

Looking ahead, when asked about the future of Lancaster, two prominent themes were apparent in the analysis. These themes were anxiety towards a hybrid existence with sensors tracking every movement, with one interviewee mentioning the United Kingdom's plans to launch a digital identity scheme "then we really will be watched if AI technology is tracking all our movements and has access to the data of that scheme" (Interviewee 14). Another participant mentioned:

It will be like we are in the *Matrix* or that other film *Monitory Report*; they will [government, police etc.] know exactly know where we are at any given moment and perhaps predict what we will do next in that space (Interviewee 16).

The second theme was excitement for the technological infrastructure, with interviewees in this theme speculating on what they would gain from a smart city:

I like that Lancaster is investing in the digital era. I especially like the fact they have invested in security first and foremost. Hopefully, next, they will employ WI-FI hot spots like they have in New York (Interviewee 9).

When challenged about the possible data exchange that would be required for the ‘free’ service of getting WI-FI, the interviewee said, “honestly, I don’t mind. I don’t see any of what is ‘happening’, and it doesn’t impact me in the moment, and I get what I need there and then” (Ibid).

8.21 Conclusion: The Truth, The Whole Truth, and a little bit more

Unlike the Icon workshops within this section of the chapter the images, the AI for Lancaster programme, and all the interviewees described are fictional. They are “entry points” into a Design Fiction created to interrogate issues around the legibility of AI systems (Coulton, et al., 2017). However, from this point onwards, the chapter is factual. Likewise, this DF approach was used, as much of this research was conducted at the height of the covid pandemic and was written up as a book chapter (F. Pilling, et al., 2022). However, this version of the DF in this thesis has been further developed and influenced by adjacent research conducted by Mullagh et al., 2022 and Jacobs et al., 2022, which implemented the AI iconography in an urban setting during a DF walkshop. In other words, the researchers implemented the icons as experiential prototypes and placed them into active settings to observe and speculate on the potential impacts of public space IoT and to support evidence-based policy development for ethical and secure public spaces (Mullagh et al., 2022).

As AI systems become increasingly ubiquitous and public perceptions contend with AI’s definitional dualism, calls for frameworks, standards and guidelines to support responsible, ethical and legible uses of AI are increasingly commonplace (Lindley et al., 2020; F. Pilling, et al. 2022; F. Pilling, et al., 2022; F. Pilling et al., 2020). The foundational design of the AI iconography underpinning this DF grapples with the complexity of creating imagery to support these aims. Through the DF, it is eminently clear that the challenge is complex. Hence, for policy-making and executive branches of government converting these challenges into action – in this example, in an urban design context – is inherently risky. However, research akin to the one presented push forward on the pursuit of making technology legible to its end-users for agency and negotiability when interacting with them. Methods such as DF have a unique ability to apprehend, articulate, and interrogate the implications of contemporary policy decisions, hence going some way to de-risk them (F. Pilling et al., 2022., Mullagh et al., Jacobs et al., 2022). In other words, any innovation supporting

the implementation of responsible AI systems in urban contexts should be welcomed, and tools like Design Fiction should be employed to ensure such interventions are sensitive to local opinion.

Chapter Nine Conclusion

9.1 Introduction

This chapter begins by identifying how this research has answered the research questions, followed by reflections on the research itself. The main contributions of this research are then discussed. Though, to briefly summarize the contributions of this research these are: the formulation of a MTHCD approach for considering AI as a material design, whereby a design solution was forwarded making AI services and products more legible to human users; the second contribution, was the design and production of an approach for adapting philosophy for design research, and the creation of design artefacts (AI icons) that ‘do philosophy’ on a speculative level; the third contribution of this research is that this research contributes to the HDI discourse and understanding of AI legibility; the fourth contribution is the method and approach in which this research conducted research and user testing online; and the final contribution is the unique transdisciplinary hinterland of various literatures assembled together in this research, emulating how one can conduct research going forward in this manner. The following section of the chapter reflects on the limitations of the research. The chapter is concluded with recommendations for future research.

9.2 Research Questions

This research has investigated several areas concerning AI as a design material. The thesis has been written to reflect the prevailing concept of developing a transdisciplinary method assemblage to synthesise a More-Than Human-Centered perspective and approach for designing with AI. A series of AI ontographs was envisioned and designed as a set of AI icons to communicate the functions and operations of AI, with AI legibility as a case study for AI as a material for design. Among the assemblage constructed, the principal aspect of an iterative RtD approach was evident. The initial research question asked in Chapter One was RQ1. How can we craft an approach that explores how the materiality of AI manifests itself in design practice, using lenses derived from Object-Oriented Ontology and Postphenomenology? Due to the methodological approach of RtD, three further research questions emerged as part of the iterative process:

RQ2. How can we design philosophical probes to explore design challenges such that they produce practical outcomes that explore the materiality of AI, such as an AI lexicon that contributes to AI legibility?

RQ3. Can the adoption of a More-than Human-Centered Design approach aid in the creation of alternative perspectives of the materiality of AI that challenge the dominance of science fiction renderings?

RQ4. How do we apply the consideration of AI as a material so that it produces practical solutions for living with AI?

After presenting the artefacts of this research, the sub-questions were answered through testing and discussion of the design artefacts. These questions will also be conclusive in this chapter.

Additionally, in explaining how this research has responded to the research questions, the following sections will also outline contributions that this research has made.

9.2.1 RQ1: AI as A Material for Design & RQ2: AI Ontographs

This research has been conducted by someone who is not a programmer, computer scientist or data scientist, which are perhaps the foremost way the materiality of AI is measured and worked with. From this point of view, designing with AI had to take an ‘indirect’, although practical, approach. A perspective for the materiality of AI was derived through developing and integrating a speculative realist OOO lens for viewing things as vibrant objects with agency and independent perspectives partaking in interdependent relationships. The characteristic of agency could be considered straightforward to employ with AI technology due to its nature of functioning independently directed through coded instruction. Although contemplating AI’s independent perspectives and interdependent relationship requires an alternative perspective to design’s primarily anthropocentric agenda of HCD, which is exhaustingly used in HCI. Thus, in Chapter Five, the MTHCD approach of Human-AI Kinship was advanced by first presenting the theory behind a flat ontological perspective and Harman’s Quadruple object, then developing object empathy by reintroducing the human through an object-oriented postphenomenology positioning.

Furthermore, in Chapter Five, this research strives to speculate on things beyond human experience by looking past Husserl's and Heidegger's phenomenology to the speculative realism and OOO for noncorrelationist thought: that things or objects exist apart from how our human minds relate and comprehend them. In other words, "we never grasp an object 'in itself' in isolation from its relation to the subject" (Meillassoux, 2008, p. 5). Stepping out of the "correlationist circle" is difficult (ibid). "In short, *all things equally exist, yet they do not exist equally*" (Bogost, 2012, p. 11). However, OOO rejects the correlationist perspective as Bogost forwards:

We humans are elements, but not the sole elements, of philosophical interest. OOO contends that nothing has special status, but that everything exists equally—plumbers, cotton, bonobos, DVD players, and sandstone, for example. In contemporary thought, things are usually taken either as the aggregation of ever smaller bits (scientific naturalism) or as constructions of human behavior and society (social relativism). OOO steers a path between the two, drawing attention to things at all scales (from toms to alpacas, bits to blinis) and pondering their nature and relations with one another as much with ourselves (Ibid, p. 6).

In essence, AI as a material for design was accomplished by seeing AI detached from human intelligence and as a thing itself through researching, designing, and developing an ontological depiction of AI through a set of semiotically and ontographically designed AI icons to enhance AI legibility. Thus, presenting an innovative viewpoint towards redefining relationships between designers, users, and products through philosophical probes utilised with design.

In Chapter Six, *Design Fiction: Adapting Philosophy for Design*, philosophical probes were established by situating and adapting philosophy into design practice through an assemblage of ideological concepts. These were: Carpentry, Design Fictions that do philosophy; developing the theory for Constellations and Onto- Cartography for mapping independent perspective and interdependent relationships; a horizonless perspective broadening the planarity scale to account for concepts akin to Morton's Hyperobjects; Alien Phenomenology for speculating on AI's ontology; and speculation through Design Fiction. This thinking and ideology were taken forward to create Design Fictions on AI ontologies and a set of AI icons, which from an OOO perspective, were symbolic AI

ontographs. AI as a material for design was realised through a MTHCD approach as evidenced through the account of AI legibility.

9.2.2 RQ3: The Insights of a More Than Human-Centred approach.

A human-centred ideal in design is enthusiastically and widely accepted; however, the most prominent and quoted scholar on the matter, Norman (2005), warned of the potential perils of blindly committing to HCD, as discussed in Chapter Five. Norman's argument revolved around how technology was designed to adapt to people and un-complicate interaction, although, as suggested in this thesis, it can complicate things in the long run. This thesis is not an argument against HCD; it is, however, an argument for acknowledging alternative approaches for design and considering AI technology devoid of anthropocentric agenda.

Bogost reminds us that it is ordinarily inconceivable “that one could put non-human objects in front, even if just for a moment, [as customarily it] signals a coarse and sinful inhumanism (Bogost, 2012, p. 132). However, “speculative realism provides the best means for creative work to be done, and it provides genuine excitement to think that there are new argumentative realms to explore” (Smniecek quoted in Ibid). The realm this thesis explored was to speculatively question what it was like for AI to function and operate. Speculation and research of AI attributes led to configuring an ontological constitution for AI technology and semiotically designing a collection of AI icons in Chapter Seven, *Designing for AI Legibility*, specifically the sensual qualities of AI technology. The MTHCD approach was expanded upon in Chapter Six, *Design Fiction: Adapting Philosophy for Design*, through the premise of Carpentry, constellations, horizonless perspective, and Alien Phenomenology – situating and forming a MTHCD lens for design. These concepts were taken forward when designing the icons and mapping and designing the Design Fictions. In fact, the more-than human framing completely altered the perception of the designer working with AI technology, whereby the icons do not present a fictional ontology of AI by following HCD axioms or science fiction renderings of AI; instead, they communicate the reality of their beings.

One could argue that just researching AI attributes would have resulted in a similar outcome. However, the impression of genuinely considering AI's ontology led to the very idea of designing

icons that communicated an AI's ontology to a potential user. If a HCD approach had been taken, it would not have led to communicating AI's ontology; rather, it would have developed into creating a fictional communication system (Norman, 1988) that provided a surface level of information, like a traffic light system.

9.2.3 RQ4: Practical Designs for AI

The research presented here is not expected to solve the evolving challenge of utilising AI but to demonstrate the potential of design-led responses by focusing on the challenge of legibility as an account of AI as a material for design. Furthermore, this thesis demonstrated a RtD methodology of interweaving interdisciplinary perspectives to communicate better AI's intangible and complex functions for more informed use. As discussed: AI legibility was assembled as an iterative case study for AI as a material for design, and as such, this thesis has outlined the developing iterations of the icons to demonstrate the multifaceted challenges of legible AI, particularly given that we are rarely able to be definitive in how AI reaches outcomes related to its operation. This is achieved by acknowledging and accepting AIs at times indeterminate processes, which rely on machine-machine interaction, regular updates, and additional information sourcing, as well as on iteration, recursion, and thus change inherent in computational processes, as noted in Chapter Two and Five.

It has been observed that the underpinning categories of the icons and the icon sets (*Learning Type, Training Data Origin, etc.*) do not cater to the dense technical landscape of AI and the numerous types of AI algorithms, such as decision trees. However, to curate the icons in a pragmatic and serviceable design, legibility was focused on rather than transparency, as noted in Chapter Seven; hence some level of detail will inevitably be sacrificed. Design solutions akin to the AI icons showcased in this research are vital to pave the way for establishing greater AI literacy and making human-AI interaction design easier to interrogate.

Chapters Two and Five examined the confused perceptions of AI technology and intelligence, which impact users' interaction with AI technology. When noting the requisite for simplicity by design—the icons do not present a fictional ontology of AI; instead, they communicate the reality of their beings. Furthermore, the icons stop AI technology from disappearing into the background. If a

design like the icons became a standardised procedure, then it would stop companies from not disclosing the reality of technologies functionality and the remit of its participation in the interaction. In Chapter Five, persuasive technology was also discussed. Although the icons do not currently cater for known persuasive interactions tactics, they make the user aware that more is happening beneath the surface. Thus, allowing users to work out and investigate for themselves the experience of the interaction.

9.3 Contributions

As this research has taken a transdisciplinary approach, it is informed by several different fields of study. Consequently, this research has resulted in several contributions, which can be used in other areas, as well as furthering the budding practice of a MTHCD approach in design. Further to the contributions outlined in response to the research questions, this research has also contributed to an understanding of a more-than human design approach, in addition to the metamorphosis of philosophy for design research and practical guidance for AI Legibility. These contributions will be discussed in the following sections.

9.3.1 A More Than Human Design Approach

This thesis has contributed to the field of More-Than Human Centred design through the development of a OOO design lens unique to this research for observing AI technology. Part One of Chapter Five examined the rationalistic and design approaches for developing AI technology. As noted, the rationalistic approach attempted to create artificial general intelligence, and the design approach seeks to develop AI through human-centred design. However, both orientations had issues in that they failed to perceive AI for what it was. In turn, always seeking AI technology to be something it is not. Instead, this research created a perspective of AI existing out of the human-world duopoly through constellation mapping and the design of AI ontographs that doubled as AI icons for better human-technology relations. On a practical note, this thesis did not set out to make AI function better; however, it did set out to consider AI as a material for design, whereby through that process, a design solution was put forward which makes AI services and products more legible.

As documented, AI-infused systems are designed to be outwardly a binary process, whereby much of the operation of AI is happening beyond human interaction and experience. As a design research agenda, the research brought the withdrawn qualities of AI to the fore, like Heidegger's broken hammer through AI iconography rather than changing ways of being.

Additionally, this MTHCD was the result of a transdisciplinary method assemblage. By following Harman's philosophy, it is uncustomary to integrate materialism into a OOO framing. However, in developing its MTHCD approach, this thesis favoured theses from different fields to develop a transformative and flexible lens with which to consider AI technology (§ 5.12). Furthermore, while this thesis has developed a MTHCD approach specifically for AI technology, another research agenda can utilise the approaches developed with alternative technology, such as edge computing (see Pilling et al., 2022).

In Chapter Six, *Design Fiction: Adapting Philosophy for Design*, the MTHCD approach was adapted and amalgamated with Design Fiction theories unique to this research. For instance, the practice of worlding a horizonless perspective to view objects and onto-graph cartography and constellations was used to build Design Fiction worlds. This MTHCD approach also formulated hands-on methods of practising the approach through speculation and Design Fiction.

The MTHCD approach in this thesis contributed by defining theoretical thinking on exposing the deeper and hidden existence of AI interactions beyond its definitional dualism by drawing up basic principles to follow, which were: (1.) all objects human and non-human are given equal attention in a flat ontology; (2.) speculating on the vicarious realities of objects and their ontology through Harman's quadruple objects –understanding that objects withdraw and reveal themselves through rifts and tensions (Harman, 2011b, 2018); (3.) Objects are vibrant (4.) Vibrant objects have different variations of agency, power and emergence, forming assemblages, manifestations, and change (5.) objects can perform data exchanges (Bryant, 2014, 2012a).

However, this research would be remiss if it did not discuss further the tension between MTHCD and HCD, as one could argue both these approaches are utilised in this research. To explain: this research set out to establish a MTHCD approach in order to perceive AI as a material for design, which was achieved through amalgamating the theoretical frameworks of OOO, Materialism and

Postphenomenology and then showcasing how philosophy could be practiced through design, resulting in an AI lexicon that communicates an AI's being. A justification for following a MTHCD approach was given by explaining the two key HCD approaches for designing with AI technology. These were: to simplify designs, meaning that much of the AI infrastructure and presence was obscured, making it easy for users to interact with the technology at hand but were unconscious of the ramifications of using AI technology; and persuasive strategies of nudging and persuading users while interacting with AI systems to perform in ways that meet the service providers' goals, such as collecting detailed data from unsuspecting users. Furthermore Norman, who founded HCD, also argued for a non-anthropocentric view, calling HCD *harmful* through the blind commitment and attention to users creating designs that lacked cohesion and were complex. However, in practice, despite the MTHCD approach in developing the AI iconography the icons were then, through a HCD approach, user tested to measure the legibility of the icons. It has been stated throughout the thesis the limitation of a MTHCD approach in that the design researcher conducting this research is a human and will forever remain trapped in this condition. On this note the MTHCD approach is an act of speculation in order, for a brief moment, to place *something* else in the spotlight for a different perspective on the design approach. Furthermore, it was also highlighted that the final design artefact of this research would be intended for human use, thus making user testing a necessary step, which was also foreshadowed by integrating postphenomenology into the MTHCD approach.

As a final note the icons developed as part of the MTHCD approach is a starting point and one possible solution towards addressing AI legibility. What this research has demonstrated is that perhaps AI legibility is a multifaceted wicked problem requiring many different 'blendings' of approaches in order to tackle such a complex challenge.

9.3.2 Philosophy and Design

This research has been an example of how philosophy can be adapted for design research. Chapter Three, *Groundworks*, set up the transdisciplinary method assemblage and made a case for legitimately adapting philosophy for design research. This was achieved by showcasing how the film *The Matrix* (Wachowski & Wachowski, 1999) had taken Jean Baudrillard's philosophical writings

(1994) and adapted them for screen via scripts and tangible representations through the *Mise-en-scènes* and prototyping. This research took inspiration by taking the philosophic works of Harman, Bogost, Bryant, and Morton and developing artefacts for design research that ‘do philosophy’, as well as ponder the relations between humans and AI technology.

Of note, likewise, is how with the application of philosophy, the perspectives and interpretations of the subject matter changed. For instance: rather than relate machine behaviour to human behaviour, we can consider machine behaviour beyond the singular human-world correlate. As Bogost forwards:

Yet, like everything, the computer possesses its own unique existence worthy of reflection and awe, and it’s indeed capable of more than the purposes for which we animate it (2012, p. 16).

Furthermore, this research developed the ideological thinking behind the concept of Human-AI Kinship as an alternative to the HCD and HCAI positions. In Chapter Five, *More Than Human-Centred Design: Shifting Perspectives through Philosophy*, a design rationale was established through an object-oriented postphenomenological positioning. This ideological concept was put forward to create designs that were object-orientated *and* utilised by human users. This thesis went beyond the human experience and back again by developing and framing the icons, semiotically designing them with philosophical insight, and testing their legibility through user testing.

9.3.3 AI Legibility

As previously noted in the chapter, the design artefacts (icons, *Alexa* app DF, and AI for Lancaster DF) at this point are simply suggestions on how design can tackle AI legibility. On a simple note, the artefacts enable conscious human interaction with AI technology, which is not widely seen in practice. In Chapter Seven, *Designing for AI Legibility*, the concept of legibility was defined against the notions of explainability, interpretability and transparency and contributed to the discussion on what legibility is regarding AI technology. The HDI concept of legibility was strived for as it was concerned with crafting intelligible intelligence for non-expert users. This research found that the concepts of explainability, interpretability and transparency were aimed at disambiguating the

domain for experts rather than for the masses. This finding contributes to HDI's discussion of legibility.

In addition to helping to define legibility, this research also endeavoured to clarify the paradox of misinterpretation between two deviating, though entangled concepts of AI— Artificial General Intelligence and Narrow AI: which in this research was defined as AI Definitional Dualism. This research went at length to describe both concepts through AI's history, as well as looking at film studies. Films offer challenging new ways of looking at the world and reflect humanity's deepest desires to create sentient life. Films, however, confuse our perception of the reality of technology; hence, this research went back to the things in themselves and constructed an iterative method assemblage to tackle the messiness of reality (Law, 2004).

9.3.4 Workshopping during a pandemic

The user testing of this research was conducted during the covid pandemic, hampering any in-person testing and drastically halting any activities using the original physical icon cards. Therefore, testing had to be conducted online. After experimenting with online collaborative platforms, such as Miro and gather-town, the decision was taken to design and create an online workshopping platform that would provide a robust testing solution using game engines. The online workshopping tool's design, development, and documentation contribute to how designers can create workshopping platforms when flexibility in testing solutions is required.

9.3.5 A Transdisciplinary Hinterland

In Chapter Three Law's notion of forming a unique method assemblage of theoretical thinking and frameworks in order to grapple with the world's messy and slippery intricacies was forwarded as a way to perceive AI as a material for design to investigate AI legibility (2004). To craft a unique hinterland of methods, Law conveyed, was the process of "assembling" and "bundling" methods from distinct disciplines together, which enacted realities of research and knowledge generation (Ibid, p.42). The act of bringing these methods into a meaningful whole to effectively work together meant that this research was transdisciplinary in nature. A contribution itself –this thesis has been a written account of assembling and stitching disparate methods together such as AI's history,

OOO, HCD, Postphenomenology, Materialism, RtD, Phenomenology, film studies. Bogost would point out that the previous list of divergent subjects placed together in a list would present a litany of “surprisingly contrasted curiosities” which makeup this research (2012, p. 38). Bogost goes on to add that—

the inherent partition between things is a premise of OOO, and lists help underscore those separations, turning the flowing legato of a literary account into the jarring staccato of real being (Ibid, p.40)

as the author of this thesis would attest is happening within the pages of this assembled thesis.

9.4 Limitations

As well as the contributions there were also limitations regarding the research, which this section of the chapter will detail. To start: the MTHCD approach formed throughout this research was done so by creating a method and theoretical assemblage, looking at concepts such as OOO and Materialism, which are very conceptual with insights formed based on speculation and conjecture. Throughout the thesis the reader was often reminded that when considering what it was like to be an AI would be through speculation and educated guesswork as we remain trapped in a human condition. The significance of a MTHCD approach is the opportunity to speculate, yet one could argue this is also its greatest weakness. However, while the AI icons could have been developed without a MTHCD approach the actual arrival of the notion to map an AI’s ontology, which the icons communicate, would not have been considered. Furthermore, a limitation could also be that despite the MTHCD approach the final design outputs of the icons were always intended for human use and required user testing which confirmed to a HCD approach creating a tension between the two approaches.

Another limitation of this research was the semiotic design of the icons, and how a co-design method could have been established in so far as the workshops could have been the designing of the icons. However, this method could have led to an influx of design ideas with no reasoning behind their designs. Furthermore, an additional limitation regarding the semiotic design framework developed

for this research worked well for AI technology however it may not be applicable with another type of technology.

Speaking of the user testing, while the workshop participants ranged from a variety of stakeholders more AI experts could have been consulted on the remit of the icons and the icon designs themselves. Though this research and the MTHCD approach developed is a starting points for addressing AI legibility, in which further testing of the icons could commence and new icons could be developed as evidenced in Chapter Eight when discussing new icon categories.

9.5 Going Forward: Future Research

There have been several suggestions already made for future icons in the last chapter; however, two primary suggestions for future research routes have emerged because of this research. These are ontological design research and legible diagrams.

9.5.1 Ontological Design Research

It has been previously mentioned that the MTHCD approach developed by this research can be used in various contexts and not just technoscientific frameworks. Therefore, the scope of this approach is extensive. This approach has already been used while researching edge networks (F. Pilling, et al 2022).

The pervasiveness of IoT, Edge Computing (EC), Fog Computing (FC) and Cloud Computing (CC) is affording end-users with more significant levels of connectivity, convenience, and personalisation across society as well as providing opportunities for new enterprise and innovation (Brous et al., 2020; Sulieman et al., 2022). However, the legibility and user awareness of the outlined network assemblage and the associated socio-technical impacts on users are minimally known, due to the technologies' radical development and adoption, resulting in no universally accepted definition of EC and FC among experts (Caprolu et al., 2019). Adding to the perplexity: these computing paradigms are often used interchangeably for the other, as frequently evidenced in both academic and industry literature, with one archetype often being considered the sum total of the network endeavour (Cisco, 2015; Fan et al., 2018; Maia et al., 2019; Wen et al., 2017).

Critically, experts' understanding of the technologies is constantly evolving while already in use and integrated into operational systems, leading to problems occurring in the 'wild' with detrimental implications to users. The state of affairs also leads to a lack of general user knowledge, impacting the agency and negotiability of the technology (Mortier et al., 2014) and compromising user security and adoption of sustainable practices (Stead et al., 2020).

To begin to respond to these outlined issues and to improve the legibility of the domain for users' secure and overall sustainable adoption of edge and IoT systems, an interactive digital 'Choose Your Own Adventure' game was designed whereupon the gameplay is of a hacker's voyage through a perceptible world of computer networks and sustainable and cybersecurity data practices.⁸¹ Ontological mapping transpired of EC, FC, and CC systems to situate the context of the game, noting their challenges and differences (F. Pilling, et al 2022).

Crucially, the worlding approach was used to expand knowledge of IoT and EC sustainable practices and security issues through gameplay. The worlding approach was a combination of chiefly Design Fiction as World Building approach (DFasWB) (Coulton, et al., 2017), flavoured with the 'Worlding' emanating from the likes of Haraway, who turn our attention to certain experiences of non-human things for a deeper look at human-world relations (Haraway, 2011, 2016).

9.5.2 Legible Diagrams

Diagrams are schematic figures comprising lines, symbols, words, or graphic images to which meaning are attached. Diagrams have been used throughout history to represent concepts, processes, objects, or systems as linear forms of information (Eddy, 2020). The prolific ribbon diagrams used by scientists and non-expert users to illustrate the structures and functions of protein molecules were first drawn by the molecular biologist Jane S. Richardson who noted the relationship between thinking and diagramming resonates with the ways in which diagrams have been used since humans began using visual forms of representation (ibid).

⁸¹ Choose Your Own Adventure games are typically stories written from a second-person point of view, with the reader assuming the role of the protagonist and making choices that determine the main character's actions and the plot's outcome.

It is common to come across explanative AI diagrams when researching AI technology. However, most AI diagrams are illegible or do not represent the content successfully. AI diagrams also fall foul of AI's definitional dualism using brains and science fiction images to represent AI processes. Therefore, as this thesis created icons to represent AI attributes, there is also research in how one could represent AI processes in diagrams.

9.6 Summary

The MTHCD approach put forward by this research was concerned with perceiving AI as a Material for design, which explored a possible way towards legible AI design. This approach could also be an effective means to research other wicked problems.

Bibliography

- Abhishek, D., Agrawal, H., C., Zitnick, L., Parikh, D., & Batra, D. (2017). Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? *Computer Vision and Image Understanding*, 163, 90–100.
- Abrassart, C., Bengio, Y., Chicoisne, G., de Marcellis-Warin, N., Dilhac, M.-A., Gambs, S., Gautrais, V., Gibert, M., Langlois, L., Laviolette, F., Lehoux, P., Maclure, J., Martel, M., Pineau, J., Railton, P., Régis, C., Tappolet, C., & Voarino, N. (2018). *Montréal Declaration For A Responsible Development of Artificial Intelligence*. University of Montreal.
<https://www.montrealdeclaration-responsibleai.com/the-declaration>
- Ada Lovelace Institute. (2022). *The rule of trust*.
- Akmal, H. A. (2021). *Design by Play: Playfulness and Object-Oriented Philosophy for the design of IoT*. Lancaster University.
- al-Jazari, I. (1974). *The Book of Knowledge of Ingenious Mechanical Devices* (P. Hill, Ed.). Springer.
- Algorithm Watch. (2019). *AI Ethics Guidelines Global Inventory* [Inventory]. AlgorithmWatch.
<https://inventory.algorithmwatch.org/>
- Al-Rifaie, M. M., & Bishop, J. M. (2015). Weak and Strong Computational Creativity. In T. Besold, M. Schorlemmer, & A. Smaill (Eds.), *Computational Creativity Research: Towards Creative Machines* (Vol. 7). Atlantis Press.
- Amazon. (ND). Safety Critical AI. *Partnership on AI*. <https://partnershiponai.org/program/safety-critical-ai/>
- Amershi, S., Weld, D., Vorvoreanu, M., Founrey, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 3, 1–13. <https://doi.org/10.1145/3290605.3300233>
- Ananny, M. (2016). Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness. *Science, Technology, & Human Values*, 41(1), 93–2017.
<https://doi.org/10.1177/0162243915606523>

- Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989. <https://doi.org/10.1177/1461444816676645>
- Anderson, P. W. S. (Director). (2002). *Resident Evil*. Pathé Distribution.
- 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>
- Archer, B., Baynes, K., & Langon, R. (1979). *Design in General Education*. London, The Royal College of Art.
- Aristotle. (1850). *Aristotle's Treatise on Rhetoric* (T. Buckley, Trans.). Bell & Daldy.
- Arnold, J. E. (2016). *Creative Engineering Promoting Innovation by Thinking Differently* (W. J. Clancey, Ed.). Arnold and Guilford: Stanford Department of Special Collections and University Archives. <https://www.inist.org/library/1959.John%20E%20Arnold.Creative%20Engineering.pdf>
- Arnold, M., Bellamy, R. K. E., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., Reimer, D., Olteanu, A., Piorkowski, D., Tsay, J., & Varshney, K. R. (2019). FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. *IBM Journal*, 63, 1–6.
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>.
- Arthur C, C. (1976). *Profiles of The Future*.
- Arya, V., Bellamy, R. K. E., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, V. Q., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J. T., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K. R., Wei, D., & Zhang, Y. (2020). AI Explainability 360: An Extensible Toolkit for Understanding Data and Machine Learning Models. *Journal of Machine Learning Research* 21, 1–6.

- Asaro, P. (2019). AI Ethics in Predictive Policing: From Models of Threat to an Ethics of Care. *IEEE Technology and Society Magazine*, 38(2), 40–53.
- Asimov, I. (1964). *The Rest of the Robots*. Doubleday & Company.
- Associated Press. (2018). Google records your location even when you tell it not to. *The Guardian*.
<https://www.theguardian.com/technology/2018/aug/13/google-location-tracking-android-iphone-mobile>
- Atkin, A. (2022). Peirce’s Theory of Signs. *Stanford Encyclopedia of Philosophy*.
<https://plato.stanford.edu/entries/peirce-semiotics/>
- Auernhammer, J. (2020). Human-centered AI: The role of Human-centered Design Research in the development of AI. *Synergy - DRS International Conference 2020*. Design Research Society, Online. <https://doi.org/doi.org/10.21606/drs.2020.282>
- Auger, J. (2013). Speculative design: Crafting the speculation. *Digital Creativity*, 24(1), 11–35.
<https://doi.org/10.1080/14626268.2013.767276>
- Automatic Language Processing Advisory Committee. (1966). *Language and Machines: Computers in Translation and Linguistics*. National Academy of Sciences National Research Council.
- Bannon, L., & Bødker, S. (1997). Constructing Common Information Spaces. *Proceedings of the Fifth European Conference on Computer Supported Cooperative*, 81–96.
- Barad, K. (2007). *Meeting The Universe Halfway: Quantum Physics and The Entanglement of Matter and Meaning*. Duke University Press.
- Barr, P., Noble, J., & Biddle, R. (2002). *Icons R Icons: User interface icons, metaphor and metonymy* (CS-TR-02/20). Victoria University of Wellington School of Mathematical and Computing Sciences Computer Science.
- Basballe, D. A., & Halskov, K. (2012). Dynamics of research through design. *Proceedings of the Designing Interactive Systems Conference*.
- Bassett, C., Steinmueller, E., & Voss, G. (2013). *Better Made Up: The Mutual Influence of Science fiction and Innovation*.
- Batista, M. Á. H. (2021). *The Ontology of Design Research*. Routledge.

- Baudrillard, J. (1994). *Simulacra and Simulation* (S. F. Glaser, Trans.). Ann Arbor The University of Michigan Press.
- Baudrillard, J. (2004). The Matrix Decoded: Le Nouvel Observateur Interview With Jean Baudrillard (G. Genosko & A. Bryx, Trans.). *International Journal of Baudrillard Studies*, 1(2).
<https://baudrillardstudies.ubishops.ca/the-matrix-decoded-le-nouvel-observateur-interview-with-jean-baudrillard/>
- Bell, J. (Ed.). (2014). *Sci-Fi Days of Fear and wonder*. BFI.
- Bello Del, L. (2018). Scientists Are Closer to Making Artificial Brains That Operate Like Ours Do. *Futurism*. <https://futurism.com/artificial-brains-operate-like-humans-close>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* 610–623.
<https://doi.org/10.1145/3442188.3445922>
- Benjamin, W. (1982). *The Arcades Project* (H. Eiland & K. McLaughlin, Trans.). Harvard University Press.
- Bennett, J. (2010). *Vibrant Matter: A political ecology of things*. Duke University Press.
- Berdichevsky, D., & Neuenschwander, E. (1999). Toward an ethics of persuasive technology. *Communications of the ACM*, 42(5), 51–58.
- Biguenet, J. (2015). *Silence*. Bloomsbury Publishing.
- Binder, T., De Michelis, G., Ehn, P., Jacucci, G., Linde, P., & Wagner, I. (2012). *What is the Object of Design?* CHI 2012, Texas. <https://doi.org/10.1145/2212776.2212780>
- Bleecker, J. (2009). *Design Fiction: A Short Essay on Design, Science, Fact and Fiction*. 49.
- Bleecker, J. (ND). *Design Fiction*. <https://www.julianbleecker.com/designfiction>
- Bloch-Wehba, H. (2021). *Transparency's AI Problem*. Texas A&M University School of Law.
- Blythe, J. M., & Johnson, S. D. (2018). *Rapid evidence assessment on labelling schemes and implications for consumer IoT security*. Petras Internet of Things Research Hub & Dawes Centre for Future Crime at UCL.
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/949614/Rapid_evidence_assessment_IoT_security_oct_2018_V2.pdf

- Blythe, M., & Encinas, E. (2018). *Research Fiction and Thought Experiments in Design. Now*.
<http://ieeexplore.ieee.org/document/8384202>
- Bogost, I. (2006). *Unit Operations*. MIT press.
- Bogost, I. (2007). *Persuasive games: The expressive power of videogames*. MIT press.
- Bogost, I. (2012). *Alien phenomenology, or, What it's like to be a thing*. University of Minnesota Press.
- Bogost, I. (2010, August 25). Academic Mumblespeak Stop it. *Ian Bogost*.
http://bogost.com/blog/academic_mumblespeak/
- Borgmann, A. (1984). *Technology and the Character of Contemporary Life*. The University of Chicago 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.
- Bowers, J., & Rodden, T. (1993). Exploding the interface: Experiences of a CSCW network. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '93*, 255–262. <https://doi.org/10.1145/169059.169205>
- Bowker, G. (2005). *Memory Practices in the Sciences*. MIT.
- Bowker, G., & Star, S. L. (1999). *Sorting Things Out Classification and Its Consequences*. The MIT Press.
- Boyd, D. (2016). *It's not Cyberspace anymore*. <https://points.datasociety.net/it-s-not-cyberspace-anymore-55c659025e97>
- Brackett, L. (1942). The Sorcerer of Rhiannon. In C. John W (Ed.), *Astounding Science Fiction* (6th ed., Vol. 28, pp. 36–48). Street & Smith Publications, Inc.
- Braidotti, R. (2006). Posthuman, All Too Human. *Theory, Culture & Society: Explorations in Critical Social Science*, 23(7–8), 197–208.
- Braidotti, R. (2013). *The Posthuman*. Polity Press.
- Brassier, R., Grant, I. H., Harman, G., & Meillassoux, Q. (2007). *Speculative Realism. Collapse*, 3.

- Bratton, B. (2016). On Speculative Design. *DIS Magazine*.
<http://dismagazine.com/discussion/81971/on-speculative-design-benjamin-h-bratton/>
- Brézillon, P. (2003). Focusing on context in human-centered computing. *IEEE Intelligent Systems*, 18(3), 62–66. <https://doi.org/doi:10.1109/MIS.2003.1200731>
- Bridle, J. (2018). *New Dark Age, Technology and the End of the Future*. Verso.
- Bringsjord, S., & Govindarajulu, N. S. (2018). *Artificial Intelligence* [Encyclopedia]. Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/entries/artificial-intelligence/>
- Brockman, G., Sutskever, I., & Open AI. (2015). Introducing OpenAI. *Introducing OpenAI*.
<https://openai.com/blog/introducing-openai/#:~:text=We're%20hoping%20to%20grow,be%20shared%20with%20the%20world.>
- Brooks, R. A. (1991). Intelligence Without Reason. *IJCAI'91: Proceedings of the 12th International Joint Conference on Artificial Intelligence, 1*, 565–595.
- Brous, P., Janssen, M., & Herder, P. (2020). The dual effects of the Internet of Things (IoT): A systematic review of the benefits and risks of IoT adoption by organizations. *International Journal of Information Management*, 51, 101952. <https://doi.org/doi:10.1016/j.ijinfomgt.2019.05.008>
- Browne, J. T. (2019). Wizard of Oz Prototyping for Machine Learning Experiences. *CHI 2019*. CHI, Glasgow. <https://doi.org/doi/10.1145/3290607.3312877>
- 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.
- Bryant, L. R. (2012a). Object-Oriented Materialism (OOM). *Larval Subjects*.
<https://larvalsubjects.wordpress.com/2012/01/16/object-oriented-materialism-oom/>
- Bryant, L. R. (2012b, November 10). Thoughts on Posthumanism. *Larval Subjects*.
<https://larvalsubjects.wordpress.com/2012/11/10/thoughts-on-posthumanism/>
- Buchanan, R. (1985). Declaration by Design: Rhetoric, Argument, and Demonstration in Design Practice. *Design Issues*, 2(1), 4–22.
- Buchanan, R. (1992). Wicked Problems in Design Thinking. 8, 2, 5–21.

- Buchanan, R. (2001). Design Research and the New Learning. *Design Issues*, 17(4).
- Buchanan, R. (2007). Strategies of Design Research: Productive Science and Rhetorical Inquiry. In R. Michel (Ed.), *Design Research Now* (pp. 55–66). Board of International Research in Design.
- Buckminster Fuller, R. (1971). *World Game Series: Document One*. World Resources Inventory, Southern Illinois University.
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1). <https://doi.org/10.1177/2053951715622512>
- Burrows, D., & O’Sullivan, S. (2019). *Fictioning*. Edinburgh University Press.
- Bush, V. (1945). As We May Think. *Atlantic*.
<https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/>
- Cameron, J. (Director). (1984). *The Terminator* [Orion Pictures].
- Cammell, D. (Director). (1977). *Demon Seed*. Metro-Goldwyn-Mayer.
- Caprolu, M., Di Pietro, R., Lombardi, F., & Raponi, S. (2019). *Edge Computing Perspectives: Architectures, Technologies, and Open Security Issues*. IEEE International Conference on Edge Computing.
- Capurro, R. (2010). Digital Hermeneutics: An Outline. *AI & Society*, 35(1), 35–42.
- Card, S. K., Moran, T. P., & Newell, A. (1983). *The Psychology of Human Computer Interaction*. Lawrence Erlbaum.
- Carreyrou, J. (2019). *Bad Blood: Secrets and Lies in a Silicon Valley Startup*. Knopf.
- Carrico, D. (2013). Futurological Discourses and Posthuman Terrains. *Existenz*, 8(2).
- Carroll, J. M., & Kellogg, W. A. (1989). *Artifact as Theory-Nexus: Hermeneutics Meets Theory-Based Design*. 7–14. <https://doi.org/10.1145/67450.67452>
- Carroll, L. (1961). *Alice’s Adventures in Wonderland: And Through the Looking-Glass*. Dent & Sons LTD.
- Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730.

- 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/>
- Chaslot, G. (2018, September 2). *The YouTube algorithm I worked on heavily promoted Brexit, because divisiveness is efficient for watch time, and watch time leads to ads. Brits deserve deserve to know what @YouTube 's AI promoted by the millions during the referendum. Without transparency there is no democracy.* Twitter.
https://twitter.com/gchaslot/status/1036323806242066432?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1036323806242066432%7Ctwgr%5E%7Ctwcon%5Es1_&ref_url=https%3A%2F%2Ftheconversation.com%2Ffeedback-loops-and-echo-chambers-how-algorithms-amplify-viewpoints-107935
- Chiang, T. (1998). *Story of Your Life*. Tor Books.
- Chmielinski, K. S., Newman, S., Taylor, M., Joseph, J., Thomas, K., Yurkofsky, J., & Qiu, Y. C. (2020). *The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence*. NeurIPS 2020 Workshop on Dataset Curation and Security.
- Chui, M., Löffler, M., & Roberts, R. (2010, March 1). The Internet of Things. *McKinsey Quarterly*.
<https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-internet-of-things>
- Churchman, C. W. (1967). Guest Editorial: Wicked Problems. *Management Science*, 14(4), B141–B142.
- Cisco. (2015). *Fog Computing and the Internet of Things: Extend the Cloud to Where the Things Are*.
https://www.researchgate.net/profile/Mohamed_Mourad_Lafifi/post/Is_there_any_simulation_tool_for_fog_computing/attachment/59d638c079197b8077995f4c/AS%3A398883160117248%401472112564706/download/Fog+Computing+and+the+Internet+of+Things++Extend+the+Cloud+to+Where+the+Things+Are.pdf
- Colebrook, C. (2012). Not Symbiosis, Not Now: Why Anthropogenic Change Is Not Really Human. *Oxford Literary Review*, 34(2, Deconstruction in the Anthropocene), 185–209.

- Colom, R., Karama, S., Jung E, R., & Haier J, R. (2010). Human intelligence and brain networks. *Dialogues In Clinical Neuroscience*, 12(4), 489–501.
- Colomina, B., & Wigley, M. (2016). *are we human? Notes on an archaeology of design*. Lars Müller.
- Computer AI passes Turing test in ‘world first’. (2014, June 9). *BBC News*.
<https://www.bbc.co.uk/news/technology-27762088>
- Constable, C. (2006). Baudrillard Reloaded: Interrelating Philosophy and Film via The Matrix Trilogy. *Screen*, 47(2), 233–249.
- Constable, C. (2009). *Adapting Philosophy: Jean Baudrillard and The Matrix Trilogy*. Manchester University Press.
- 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>
- Cooper, R., & Press, M. (1995). *The Design Agenda A guide to Successful Design Management*. Wiley.
- Coulton, P. (2017). Sensing atoms and bits. In *Sensory arts and design* (pp. 189–202). Bloomsbury, London.
- Coulton, P. (2020). Reflections on teaching design fiction as world-building. *ACM DIS 2020 More than Human Centred Design*, 6.
- Coulton, P. (2015). The role of game design in addressing behavioural change. *The Value of Design Reseach*. 11th European Academy of Design Conference, Paris, France.
- 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. G. (2019). 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>
- Crabtree, A., & Mortier, R. (2015). Human Data Interaction: Historical Lessons from Social Studies and CSCW. *Proceedings of the 14th European Conference on Computer Supported Cooperative Work*, 19–23.
- Crawford, K. (2016). Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics. *Science, Technology, & Human Values*, 41(1), 77–92. <https://doi.org/10.1177/0162243915589635>
- Crawford, K. (2021). *Atlas of AI Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.
- Crawford, K., & Joler, V. (2018). Anatomy of an AI system. *Anatomy of an AI System An Anatomical Case Study of the Amazon Echo as a Artificial Intelligence System Made of Human Labor*. <https://anatomyof.ai/>
- Crawford, M. B. (2009). *Shop Class as Soulcraft An Inquiry Into the Value of Work*. Penguin Publishing Group.
- Crichton, M. (Director). (1973). *Westworld*. Metro-Goldwyn-Mayer.
- Cross, N. (1999). Design Research: A Disciplined Conversation. *Design Issues*, 15(2), 5–10.
- Cross, N. (2006). *Designerly Ways of Knowing*. Springer.
- d’Alessandro, B., O’Neil, C., & LaGatta, T. (2017). Conscientious Classification: A Data Scientist’s Guide to Discrimination-Aware Classification. *Big Data*, 5(2), 120–134. <https://doi.org/10.1089/big.2016.0048>

- Dautenhahn, K. (1998). THE ART OF DESIGNING SOCIALLY INTELLIGENT AGENTS: SCIENCE, FICTION, AND THE HUMAN IN THE LOOP. *Applied Artificial Intelligence*, 12(7-8: Socially Intelligent Agents, Part I).
- Davies, C. J. (2019). The Problem of Causality in Object-Oriented Ontology. *Open Philosophy*, 2(1), 98–107.
- Davis, E. (1998). *Techgnosis: Myth, Magic, Mysticism in the Age of Information*. Harmony Books.
<https://books.google.co.uk/books?id=2P4QAQAIAAJ>
- de Bono Group. (N.D). Six Thinking Hats. *Six Thinking Hats*.
<https://www.debonogroup.com/services/core-programs/six-thinking-hats/>
- DeLanda, M. (2002). *Intensive Science & Virtual Philosophy*. Continuum Books.
- Deleuze, G., & Guattari, F. (1987). *A Thousand Plateaus Capitalism and Schizophrenia*. University of Minnesota Press.
- Deleuze, G., & Parnet, C. (1987). *Dialogues*. Columbia University Press.
- Descartes, R. (2008). *Meditations* (J. Veitch, Trans.). Cosmo Classics.
- Di Russo, S. (2012, June 8). A Brief History of Design Thinking: How Design Thinking Came to ‘Be’. *I Think I Design*. <https://ithinkidesign.wordpress.com/2012/06/08/a-brief-history-of-design-thinking-how-design-thinking-came-to-be/>
- Diakopoulos, N. (2016). Accountability in Algorithmic Decision Making. *Communications of the ACM*, 59(2), 56–62. <https://doi.org/10.1145/2844110>
- Diakopoulos, N., & Koliska, M. (2017). Algorithmic Transparency in the News Media. *Digital Journalism*, 5(7), 809–828. <https://doi.org/10.1080/21670811.2016.1208053>
- Ding, J. (2022). What defines the ‘open’ in ‘open AI’? *The Alan Turing Institute*.
<https://www.turing.ac.uk/blog/what-defines-open-open-ai>
- Doering, M., Glas, D. F., & Ishiguro, H. (2019). Modeling Interaction Structure for Robot Imitation Learning of Human Social Behavior. *IEEE Transactions on Human-Machine Systems*, 49(3), 219–231. <https://doi.org/10.1109/THMS.2019.2895753>.
- Domingos, P. (2012). A Few Useful Things to Know About Machine Learning. *Communications of the ACM*, 55(10), 78–87. <https://doi.org/10.1145/2347736.2347755>

- Dougherty, C. (2015). Google Photos Mistakenly Labels Black People ‘Gorillas’. *The New York Times*.
- Dourish, P. (2004). What we talk about when we talk about context. *Personal and Ubiquitous Computing*, 8, 19–30. <https://doi.org/10.1007/s00779-003-0253-8>
- Dove, G., & Fayard, A.-L. (2020). *Monsters, Metaphors, and Machine Learning*. 17.
- Downton, P. (2003). *Design Research*. RMIT Pub.
- Dreyfus, H. (1972). *What Computers Still Can't Do: A Critique of Artificial Reason*. Mass.
- Dreyfuss, H. (1955). *Designing for People*. Simon & Schuster.
- Drichoutis, A. C., Lazaridis, P., & Nayga, R. (2006). Consumers' use of nutritional labels: A review of research studies and issues. *Academy of Marketing Science Review*, 10(9).
- Dudley-Evans, A. (1993). Variation in communication patterns between discourse communities: The case of highway engineering and plant biology. In *Language, Learning and Success: Studying Through English* (pp. 141–147). Macmillan.
- Dunne, A. (2005). *Hertzian Tales: Electronic Products, Aesthetic Experience, and Critical Design*. Cambridge: MIT press.
- Dunne, A., & Raby, F. (2013). *Speculative Everything: Design, Fiction, and Social Dreaming*. MIT.
- Durrant, A. C., Vines, J., Wallace, J., & Yee, J. S. R. (2017). Research Through Design: Twenty-First Century Makers and Materialities. *Design Issues*, 33(3), 3–10.
- Eames, C. (Director). (1968). *Powers of Ten* [Documentary/Short].
- Eddy, M. D. (2020). Diagrams. In A. Grafton, A. Blair, & A. Sylvia (Eds.), *A Companion to the History of Information* (pp. 397–401). Princeton University Press.
- Edelson, D. C. (2002). Design Research: What We Learn When We Engage in Design. *The Journal of the Learning Sciences*, 11(1), 105–121.
- Elish, M. C., & Boyd, D. (2018a). Situating methods in the magic of Big Data and AI. *Communication Monographs*, 85(1), 57–80. <https://doi.org/10.1080/03637751.2017.1375130>
- Elish, M. C., & Boyd, D. (2018b). Don't Believe Every AI You See [Research]. *The Ethical Machine: Big Ideas For Designing Fairer AI and Algorithms*. <https://ai.shorensteincenter.org/ideas/2018/11/12/dont-believe-every-ai-you-see-1>

- Elish, M. C., & Hwang, T. (2016). *An AI Pattern Language*. Data & Society.
- Elliott, K. (2003). *Rethinking the novel/film debate*. Cambridge University Press.
- Emanuilov, I., Fantin, S., Marquenie, T., & Vogiatzoglou, 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>
- EPRS. (2020). *The ethics of artificial intelligence: Issues and initiatives* (STUDY PE 634.452; Panel for the Future of Science and Technology). Scientific Foresight Unit. [https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU\(2020\)634452_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf)
- Errazuriz, S., & [@sebastianstudio]. (2022, August 10). *Which Artists will A.I. Replace First?* Instagram. <https://www.instagram.com/p/ChDZrg8FBmD/>
- Escobar, A. (2018). *Designs for the Pluriverse: Radical Interdependence, Autonomy, and the Making of Worlds*. Duke University Press.
- European Commission. (2019). *Building Trust in Human-Centric Artificial Intelligence*.
- European Commission. (2021). *Regulation of the European Parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts* (167). https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF
- European Union. (2016). *Regulations*. European Union. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679>
- Eyal, N., & Hoover, R. (2014). *Hooked: How to Build Habit-Forming Products*. Penguin Publishing Group.
- Fan, K., Pan, Q., Wang, J., Liu, T., & Li, H. (2018). Cross-Domain based Data Sharing Scheme in Cooperative Edge Computing. *IEEE*, 87–92.

- Faste, T., & Faste, H. (2012). Demystifying ‘design research’: Design is not research, research is design. *IDS Education Symposium*.
- Fazi, M. B. (2018). *Contingent Computation: Abstraction, Experience, and Indeterminacy in Computation Aesthetics*. Rowman & Littlefield.
- Feenberg, A. (2002). *Transforming Technology: A Critical Theory Revisited*. Oxford University Press.
- Ferreira, J., Barr, P., & Noble, J. (2002). The Semiotics of User Interface Redesign. *Proceedings of the Sixth Australasian Conference on User Interface*, 40, 47–53.
- Ferreira, J., Noble, J., & Biddle, R. (2006). A Case for Iconic Icons. *Conferences in Research and Practice in Information Technology Series*, 50, 87–90.
<https://doi.org/10.1145/1151758.1151771>
- Findeli, A., Brouillet, D., Martin, S., Moineau, C., & Tarrago, R. (2008). *Research Through Design and Transdisciplinarity: A Tentative Contribution to the Methodology of Design Research*. 67–91.
- Finn, E. (2017). *What Algorithms Want*. MIT press.
- Fiore, Q., & McLuhan, M. (1967). *The Medium is the Massage: An Inventory of Effects*. Random House.
- Fisher, K. (2013). Adapting Philosophy: Jean Baudrillard and The Matrix Trilogy by Catherine Constable (review). *Science Fiction Film and Television*, 6(2), 309–313.
- 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>
- Flusser, V., & Cullars, J. (1995). On the Word Design: An Etymological Essay. *Design Issues*, 11(3), 50–53. <https://doi.org/10.2307/1511771>
- Fogg, B., J. (2003). *Persuasive Technology: Using Computers to Change What We Think and Do*. Morgan Kaufmann.
- Fogg, B., J. (1998). Persuasive computers: Perspectives and research directions. *SIGCHI Conference on Human Factors in Computing Systems*. SIGCHI, Los Angeles, USA.

- Ford, K., M., Hayes, P., J., Glymour, C., & Allen, J. (2015). Cognitive Orthoses: Toward Human-Centered AI. *AI Magazine*, 36(4), 5–8. <https://doi.org/doi:10.1609/aimag.v36i4.2629>
- Forlano, L. (2017). Posthumanism and Design. *She Ji: The Journal of Design, Economics, and Innovation*, 3(1), 16–29. <https://doi.org/10.1016/j.sheji.2017.08.001>
- Forsythe, D. E. (1993). Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence. *Social Studies of Science*, 23(3), 445–477. <https://doi.org/10.1177/0306312793023003002>
- Fortenbaugh, W. W. (2007). Aristotle’s Art of Rhetoric. In I. Worthington (Ed.), *A Companion to Greek Rhetoric*. Blackwell Publishing.
- Frankel, L., & Racine, M. (2010). The Complex Field of Research: For Design, through Design, and about Design. *Design and Complexity - DRS International Conference 2010*.
- Frankish, K., & Ramsey, W. M. (2014). *The Cambridge Handbook of Artificial Intelligence*. Cambridge University Press.
- Frayling, C. (1993). *Research in Art and Design. 1*.
- Frayling, C. (2016). *The 2001 File: Harry Lange and the Design of the Landmark Science Fiction Film*. Reel Art Press.
- Fredrich, G., & Zanker, M. (2011). A Taxonomy for Generating Explanations in Recommender Systems. *AI Magazine*, 32(3), 90–98. <https://doi.org/10.1609/aimag.v32i3.2365>
- Friedman, K. (2008). Research into, by and for design. *Journal of Visual Arts Practice*, 7(2), 153–160. <https://doi.org/doi:10.1386/jvap.7.2.153/1>
- Friedman, K. (2000). *Creating design knowledge: From research into practice*. IDATER 2000, Loughborough University.
- Fry, T. (2020). *Defuturing: A New Design Philosophy*. Bloomsbury Publishing.
- Fu, E., sibi, S., Miller, D., Johns, M., Mok, B., Fischer, M., & Sirkin, D. (2019). The Car That Cried Wolf: Driver Responses to Missing, Perfectly Performing, and Oversensitive Collision Avoidance Systems. *2019 IEEE Intelligent Vehicles Symposium (IV)*, 1830–1836. <https://doi.org/10.1109/IVS.2019.8814190>

- Gaver, W. (2012). What should we expect from research through design? *CHI '12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 937–946.
<https://doi.org/10.1145/2207676.2208538>
- Gaver, W. (2002). DESIGNING FOR HOMO LUDENS. *13 Magazine*.
- Gaver, W., Beaver, J., & Benford, S. (2003). *Ambiguity as a Resource for Design*. 5, 233–240.
- Gaver, W., & Bowers, J. (2012). Annotated Portfolios. *Interactions*.
<https://doi.org/10.1145/2212877.2212889>
- Geburu, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for Datasets. *Proceedings of the Fairness, Accountability, and Transparency in Machine Learning Workshop*.
- Gell, A. (1988). Technology and Magic. *Anthropology Today*, 4(2), 6–9.
- 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>
- Geraci, R. (2010). *Apocalyptic AI: Visions of Heaven in Robotics, Artificial Intelligence, and Virtual Reality*. Oxford University Press.
- Giacomin, J. (2014). What Is Human Centred Design? *The Design Journal*, 17(4), 606–623.
<https://doi.org/10.2752/175630614X14056185480186>
- Gibney, A. (Director). (2019). *The Inventor: Out for Blood in Silicon Valley* [Documentary/ True Crime]. HBO.
- Gittins, D. (1986). Icon-based human-computer interaction. *International Journal of Man-Machine Studies*, 24(6), 519–543. [https://doi.org/10.1016/S0020-7373\(86\)80007-4](https://doi.org/10.1016/S0020-7373(86)80007-4)
- Glanville, R. (1999). Researching Design and Designing Research. *Design Issues*, 15(2), 80–91.
<https://doi.org/10.2307/1511844>
- 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>

- Green, D. P., & Lindley, J. (2021). Design Research and Ambiguity. *Safe Harbours for Design Research*. 14th EAD Conference.
- Green, P. (2017). 'Alexa, Where Have You Been All My Life?'. *The New York Times*.
<https://www.nytimes.com/2017/07/11/style/alexa-amazon-echo.html>
- Greengard, S. (2018). Weighing the impact of GDPR. *Communications of the ACM*, 61(11), 16–18.
<https://doi.org/DOI:https://doi.org/10.1145/3276744>
- Gualeni, S. (2015). *Virtual Worlds as Philosophical Tools: How to Philosophize with a Digital Hammer*. Palsgrave Macmillan.
- Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI--Explainable artificial intelligence. *Sci. Robotics*, 4(eaay7120).
- Halais, F. (2021). Spotify is failing music artists with subpar analytics [Medium]. *The Riff*.
<https://medium.com/the-riff/spotify-is-failing-music-artists-with-subpar-analytics-bb120c6c93b8>
- 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>
- Hammond, K. (2004). Monsters of modernity: Frankenstein and modern environmentalism. *Cultural Geographies*, 11(2), 181–198. <https://doi.org/10.1191/14744744004eu301oa>
- Haraway, D. (1992). The Promises of Monsters: A Regenerative Politics for Inappropriate/d Others. In L. Grossberg, C. Nelson, & P. Treichler (Eds.), *Cultural Studies*. Routledge.
- Haraway, D. (1995). *A Manifesto for Cyborgs, Simians, Cyborgs and Women: The Reinvention of Nature*. Routledge.
- Haraway, D. (2003). *The Companion Species Manifesto: Dogs, People and Significant Otherness*. Prickly Paradigm Press.
- Haraway, D. (2011). *SF: Science Fiction, Speculative Fabulation, String Figures, So Far*.
<https://people.ucsc.edu/~haraway/Files/PilgrimAcceptanceHaraway.pdf>
- Haraway, D. (2016). *Staying with the Trouble Making Kin in the Chthulucene*. Duke University Press.
- Harman, G. (2005). *Guerrilla Metaphysics Phenomenology and the Carpentry of Things*. Open Court.
- Harman, G. (2009a). *Circus Philosophicus*. Zero books.

- Harman, G. (2009b). Dwelling With the Fourfold. *Space and Culture*, 12(3), 292–302.
- Harman, G. (2011a). *Quentin Meillassoux: Philosophy in the Making*. Edinburgh University Press.
- Harman, G. (2011b). *The Quadruple Object*. Zero books.
- Harman, G. (Ed.). (2014). *Onto—Cartography – Author Q&A*. Edinburgh University Press.
<https://www.euppublishing.com/userimages/ContentEditor/1396275575603/Onto-Cartography%20-%20Author%20Q&A.pdf>
- Harman, G. (2016). *Immaterialism: Objects and Social Theory*. Polity Press.
- Harman, G. (2018). *Object-Oriented Ontology: A New Theory of Everything*. Penguin Random House.
- Harman, G. (2020). *Art + Objects*. Polity Press.
- Hatchuel, A. (2002). Towards Design Theory and expandable rationality: The unfinished program of Herbert Simon. *Journal of Management and Governance*, 5(3–4).
- Hatchuel, A., & Weil, B. (2003). A New Approach of Innovative Design: An Introduction to C-K theory. *International Conference on Engineering Design*. ICED.
- Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. MIT.
- Hawley, S. H. (2019). Challenges for an Ontology of Artificial Intelligence. *Perspectives on Science and Christian Faith*, 71(2), 83–95.
- Heald, D. (2006). Varieties of Transparency. In C. Hood & D. Heald (Eds.), *Transparency: The Key to Better Governance?* (pp. 25–43). Oxford University Press.
- Heath, T. L. (1953). *The Works of Archimedes with the Method of Archimedes*. Dover Publications.
- Hegel, G. W. F. (1977). *Phenomenology of Spirit* (J. N. Findlay, Ed.; A. V. Miller, Trans.). Clarendon.
- Heidegger, M. (1996). *Being and Time: A Translation of Sein und Zeit* (J. Stambaugh, Trans.). State University of New York Press.
- Heidegger, M. (1999). The Origins of the Work of Art. In D. F. Krell (Ed.), & A. Hofstadter (Trans.), *Basic Writings*. Routledge.
- Hennessy, D. (2015). *Frameworks for effective improvised facilitation*. Lancaster University.
- Hennink, M., Hutter, I., & Bailey, A. (2020). *Qualitative Research Methods*. Sage Publications Ltd.
- HM Government. (2017). *Industrial Strategy: Building a Britain fit for the future*. 256.

- Hodge, B. (1995). Monstrous Knowledge: Doing PhDs in the new humanities. *Australian Universities Review*, 2, 35–39.
- Hoffman, R. R., Hayes, P., J., & Ford, K., M. (2001). Human-Centered Computing: Thinking In and Out of the Box. *IEEE INTELLIGENT SYSTEMS*, 76–78.
- Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). *The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards*.
<https://doi.org/10.48550/arXiv.1805.03677>
- Hollinger, V. (2009). Posthumanism and Cyborg Theory. In *The Routledge Companion to Science Fiction* (pp. 267–278). Routledge.
- Hubka, V., & Eder, E. (1996). *Design Science*. Springer-Verlag.
- Hui, Y. (2016). *On the Existence of Digital Objects*. University of Minnesota Press.
- Hume, D. (1878). *A Treatise on Human Nature: Vol. Book 1*. Longmans Green and Co.
- Husserl, E. (2001). *Logical Investigations* (D. Moran, Ed.; 2nd ed., Vol. 2). Routledge.
- Hutson, M. (2018). Has artificial intelligence become alchemy? *Science*, 360(6388), 478.
- IBM. (1954). *701 Translator*. https://www.ibm.com/ibm/history/exhibits/701/701_translator.html
- Ihde, D. (1979). *Technics and praxis: A philosophy of technology: Vol. XXIV* (R. S. Cohen & M. W. Wartofsky, Eds.). D. Reidel Publishing Company.
- Ihde, D. (1990). *Technology and the Lifeworld*. Indiana University Press.
- Ihde, D. (1995). *Postphenomenology: Essays in the Postmodern Context*. Northwestern University Press.
- Ihde, D. (2002). *Bodies in Technology*. University of Minnesota Press.
- Ihde, D. (2012). *Experimental Phenomenology Multistabilities*. State University of New York Press.
- Irwin, W. (Ed.). (2002). *The Matrix and Philosophy; Welcome to the Desert of the Real*. Open Court.
- Jacobs, N., & Cooper, R. (2018). *Living in digital worlds: Designing the digital public space*.
 Routledge.
- Jacobs, N., Mullagh, L., & Kwon, N. (2022). Creative design methods for IoT data ethics in hybrid spaces. *AoIR 2022: The 23rd Annual Conference of the Association of Internet Researchers*.

- Jensenius, A. R. (2009). Multi-, cross- and interdisciplinarity. *Alexander Refsum Jensenius*.
<https://www.arj.no/2009/07/10/disciplinarity/>
- Jensenius, A. R. (2012). Disciplinarity: Intra, cross, multi, inter, trans. *Alexander Refsum Jensenius*.
<https://www.arj.no/2012/03/12/disciplinarity-2/>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1, 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Jones, J. C. (1980). *Design Methods*. John Wiley.
- Kant, I. (1996). *Practical Philosophy* (M. J. Gregor, Trans.). Cambridge University Press.
- Kant, I. (1998). *Critique of Pure Reason* (P. Guyer & A. W. Wood, Eds.). Cambridge University Press.
- Kardes, F. R. (1988). Spontaneous Inference Processes in Advertising: The Effects of Conclusion Omission and Involvement on Persuasion. *Journal of Consumer Research*, 15(2), 225–233.
- Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020, April). Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. *CHI 2020*. <https://www.microsoft.com/en-us/research/publication/interpreting-interpretability-understanding-data-scientists-use-of-interpretability-tools-for-machine-learning/>
- Kelley, P. G., Bresee, J., Cranor, L. F., & Reeder, R. W. (2009). A “Nutrition Label” for Privacy. *Symposium On Usable Privacy and Security (SOUPS)*, 1–12.
<https://doi.org/10.1145/1572532.1572538>
- Kelley, P. G., Leahu, L., Bresee, J., & Cranor, L. F. (2010). Standardizing privacy notices: An online study of the nutrition label approach. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1573–1582. <https://doi.org/10.1145/1753326.1753561>
- Kellner, D. (1994). *Baudrillard: A Critical Reader*. Basil Blackwell Ltd.
- Kelly, K. (2010). *What Technology Wants*. Penguin Publishing Group.
- Kim, T., & DiSalvo, C. (2010). Speculative Visualization: A New Rhetoric for Communicating Public Concerns. *DRS2010 - Design and Complexity*. Design Research Society, Montreal, Canada.

- 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>
- Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Sage Publications Ltd.
- Knowles, B., Harding, M., Blair, L., Davies, N., Hannon, J., Rouncefield, M., & Walden, J. (2014). Trustworthy by Design. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work Social Computing*, 1060–1071. <https://doi.org/10.1145/2531602.2531699>
- Knowles, B., Lochrie, M., Coulton, P., & Whittle, J. (2014). BARTER: a technological strategy for local wealth generation. *IT Professional*, 16(3), 28–34. <https://doi.org/10.1109/MITP.2014.27>
- Knowlton, K. (Director). (1966). *L6: Part II. An Example of L6 Programming* [Informational]. AT&T Tech Channel. <https://www.youtube.com/watch?v=4a-IVJ9wT0s>
- Koenigstorfer, J., & Baumgartner, H. (2016). The Effect of Fitness Branding on Restrained Eaters' Food Consumption and Postconsumption Physical Activity. *Journal of Marketing Research*, 53(1), 124–138. <https://doi.org/10.1509/jmr.12.0429>
- Kolko, J. (2018, June). The Divisiveness of Design Thinking. *Interactions*, 25, 28.
- Koop, B.-J. (2021). The concept of function creep. *Law, Innovation and Technology*, 13(1). <https://doi.org/10.1080/17579961.2021.1898299>
- Korzybski, A. (1933). *Science and Sanity* (5th ed.). Institute of General Semantics.
- Koskinen, I., Zimmerman, J., Binder, T., Redström, J., & Wensveen, S. (2011). *Design Research Through Practice From the Lab, Field, and Showroom*. Elsevier.
- Kranzberg, M. (1986). Kranzberg's Laws. *The Johns Hopkins University Press and the Society for the History of Technology*, 27(3), 544–560.
- Kuang, C. (2021, December 1). Lessons from the Scariest Design Disaster in American History. *Google Design*. <https://design.google/library/user-friendly/>
- Kubrick, S. (Director). (1968). *2001: A Space Odyssey*. Metro-Goldwyn-Mayer.
- Kuhn, T. (2009). How to Evaluate Controlled Natural Languages. *Workshop on Controlled Natural Language*. CNL 2009, Italy.

- Kulesza, Stumpf, S., Burnett, M., Yang, S., Kwan, I., & Wong, W.-K. (2013). Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models. *IEEE Symposium on Visual Languages and Human-Centric Computing*, 3–10.
- Kurzweil, R. (2013). *The Singularity Is Near*. Duckworth Overlook.
- Lakatos, I. (1977). Science and Pseudoscience. *Philosophical Papers, 1*, 1–7.
- Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press.
- Lakoff, G., & Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*. Basic Books.
- Laney, D. (2001). *3D data management: Controlling data volume, velocity, and variety* [Gartner].
- Lang, F. (Director). (1927). *Metropolis*. Paramount.
- Lasseter, J. (Director). (1986). *Luxo Jr*. Direct Cinema.
- Latour, B. (1999). *Pandora's hope: Essays on the reality of science studies*. Harvard University Press.
- 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>
- Lawson, B. (2005). *How Designers Think: The Design Process Demystified* (4th ed.). Architectural Press.
- Le Doeuff, M. (1989). *The Philosophical Imaginary* (C. Gordon, Trans.). The Athlone press.
- Levy, S. (2016). How Google is Remaking Itself as a “Machine Learning First” Company. *Wired*.
<https://www.wired.com/2016/06/how-google-is-remaking-itself-as-a-machine-learning-first-company/>
- Lin, H., Hsieh, Y.-C., & Wu, F.-G. (2016). A study on the relationships between different presentation modes of graphical icons and users' attention. *Computers in Human Behavior*, 63, 218–228. <https://doi.org/10.1016/j.chb.2016.05.008>
- 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*, 1–13. <http://doi.acm.org/10.1145/3313831.3376792>
- Lindley, J., Cannizzaro, S., Procter, R., & Coulton, P. (2019). *Adoption and Acceptability* (Cybersecurity of the Internet of Things, pp. 94–107). Petras Internet of Things Research Hub.
- Lindley, J., & Coulton, P. (2015). Game of drones. *CHI PLAY '15 Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, 613–618. <https://doi.org/10.1145/2793107.2810300>
- Lindley, J., & Coulton, P. (2020). *AHRC Challenges of the Future: AI & Data*. <https://doi.org/10.13140/RG.2.2.29569.48481>
- Lindley, J., Coulton, P., & Akmal, H. A. (2018, June 28). *Turning Philosophy with a Speculative Lathe: Object-oriented ontology, carpentry, and design fiction*. Design Research Society Conference 2018. <https://doi.org/10.21606/drs.2018.327>
- Lindley, J., Gradinar, A., Coulton, P., Cooper, R., & Forrester, I. (2019). Making a Room for the IoT: a Domestic Demonstration. *Proceedings Living in the Internet of Things 2019*.
- Lipton, Z. C. (2018, June). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3), 31–57.
- Long, M. (2021, October 22). Don Norman, godfather of UX: “Bad design is bad for the planet”. *Design Week*. <https://www.designweek.co.uk/issues/18-24-october-2021/don-norman-godfather-of-ux-bad-design-is-bad-for-the-planet/>
- Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 150–158.
- Lovejoy, J., & Holbrook, J. (2017). Human-Centered Machine Learning: 7 steps to stay focused on the user when designing with ML [Publishing platform]. *Google Design/Medium*. <https://medium.com/google-design/human-centered-machine-learning-a770d10562cd>
- Lucas, G. (Director). (1977). *Star Wars: Episode IV: A New Hope*. 20th Century Fox.

- Lucas, G. (Director). (1987). *Apple Knowledge Navigator*. Apple.
https://www.youtube.com/watch?v=p1goCh3Qd7M&ab_channel=vintagemacmuseum
- Ma, X., Matta, N., Cahier, J.-P., Qin, C., & Cheng, Y. (2015). From action icon to knowledge icon: Objective-oriented icon taxonomy in computer science. *Displays*, 39, 68–79.
<https://doi.org/10.1016/j.displa.2015.08.006>
- Mahdavi, P. (2021). *Hyphens*. Bloomsbury Publishing.
- Maia, A. M., Ghamri-Doudane, Y., Vieira, D., & de Castro, M. F. (2019). Optimized Placement of Scalable IoT Services in Edge Computing. *IFIP/IEEE*, 189–197.
- Maita, A. (1996). *Tamagotchi*.
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers Hung, A. (2011). Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*.
- Marakas, G. M., Johnson, R. D., & Palmer, J. W. (2000). A theoretical model of differential social attributions toward computing technology: When the metaphor becomes the model. *International Journal of Human-Computer Studies*, 52(4), 719–750.
<https://doi.org/10.1006/ijhc.1999.0348>
- Marcinkevičs, R., & Vogt, J. E. (2020). *Interpretability and Explainability: A Machine Learning Zoo Mini-tour*. <https://doi.org/10.48550/arXiv.2012.01805>
- Marder, M. (2016). *Dust*. Bloomsbury Publishing.
- Margolin, V. (2000). Building a Design Research Community. *Design Plus Research*. Design Plus Research: Proceedings of the Politecnico di Milano Conference, Milan.
- Markoff, J. (2005a). *What the Dormouse Said: How the Sixties Counterculture Shaped the Personal Computer Industry*. Viking Press.
- Markoff, J. (1992). Technology; A Celebration of Isaac Asimov. *The New York Times*.
<https://www.nytimes.com/1992/04/12/business/technology-a-celebration-of-isaac-asimov.html>

- Markoff, J. (2005b). Behind Artificial Intelligence, a Squadron of Bright Real People. *The New York Times*. <https://www.nytimes.com/2005/10/14/technology/behind-artificial-intelligence-a-squadron-of-bright-real-people.html>
- Marres, N. (2017). *Digital Sociology*. Polity Press.
- Martins, L. (2014). Privilege and Oppression: Towards a Feminist Speculative Design. *Design's Big Debates - DRS International Conference 2014*. Design Research Society, Sweden.
- Mateas, M. (2006). Reading Hal: Representation and Artificial Intelligence. In R. Kolker (Ed.), *Stanley Kubrick's 2001: A Space Odyssey. New Essays*. Oxford University Press.
- Matthew, B. (2019). One Way to Think About ML Transparency. *LessWrong*.
<https://www.lesswrong.com/posts/jg6ZJLE5eHkfuxk67/one-way-to-think-about-ml-transparency>
- Mayor, A. (2018). *Gods and Robots*. Princeton University Press; JSTOR.
<https://doi.org/10.2307/j.ctvc779xn>
- McCarthy, J. (1984). Some Expert Systems Need Common Sense. *Annals of the New York Academy of Sciences*, 426, 129–137. <https://doi.org/10.1111/j.1749-6632.1984.tb16516.x>
- McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. E. (1955a). *A proposal for the Dartmouth Summer research project on Artificial Intelligence*. <http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html>
- McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. E. (1955b). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. *AI Magazine*, 27(4).
- McCarthy, N., & Montgomery, J. (2019). *Explainable AI: the basics*. The Royal Society.
- McCorduck, P. (2004). *Machines who Think*. A K Peters Ltd.
- McDermott, L., Boradkar, P., & Zunjarwad, R. (2014). Interdisciplinarity in Design Education Benefits and Challenges. *Education Symposim*. IDSA, Austin.
- Meillassoux, Q. (2008). *After Finitude: An Essay on the Necessity of Contingency*. Continuum Books.
- Meillassoux, Q. (2011). *After Finitude: An Essay on the Necessity of Contingency* (R. Brassier, Trans.; 3rd ed.). Continuum Books.
- Mettrie, J. O. D. L. (1912). *Man a Machine*. The Open Court Publishing Co.

- Metz, C. (1982). *The Imaginary Signifier: Psychoanalysis and the Cinema*. Indiana University Press.
- Michaud, T. (2020). *Science Fiction and Innovation Design* (T. Michaud, Ed.). Wiley.
- Michel. (1976). *The Archaeology of Knowledge*. Colophon Books.
- Microsoft. (ND). Microsoft AI principles. *Microsoft AI*. <https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimar6>
- Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267. <https://doi.org/10.1016/j.artint.2018.07.007>
- Minsky, M. (1968). *Semantic Information Processing*. Cambridge: MIT press.
- Minsky, M., & Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry* (2nd ed.). MIT.
- Moggridge, B., Atkinson, B., & Smith, C. (2007). *Designing Interactions*. Footprint books.
- Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73, 1–15.
- Moon, Y., & Nass, C. (1996). How “Real” Are Computer Personalities?: Psychological Responses to Personality Types in Human-Computer Interaction. *Communication Research*, 23(6), 651–674. <https://doi.org/10.1177/009365096023006002>
- Moore, G. (1965). Cramming more components onto integrated circuits. *Electronics*, 38(8).
- Moravec, H. (1988). *Mind Children The Future of Robot and Human Intelligence*. Harvard University Press.
- Morgan, D. L. (1997). *Focus groups as qualitative research* (2nd ed.). Sage Publications Ltd.
- Mori, M. (1970). The uncanny valley. *Energy*, 7(4), 33–35.
- Mortier, R., Haddadi, H., Henderson, T., McAuley, D., & Crowcroft, J. (2014). Human-Data Interaction: The Human Face of the Data-Driven Society. *ArXiv:1412.6159 [Cs]*. <http://arxiv.org/abs/1412.6159>
- Morton, T. (2010). *The Ecological Thought*. Harvard University Press.
- Morton, T. (2013). *Hyperobjects: Philosophy and Ecology After the End of the World*. University of Minnesota Press.
- Morton, T. (2016). *Dark Ecology: For a Logic of Future Coexistence*. Columbia University Press.

- Morton, T. (2017). *Humankind: Solidarity with Non-Human People*. Verso Books.
- Morton, T. (2018). *Being Ecological*. MIT press.
- Mostow, J. (Director). (2003). *Terminator 3: Rise of the Machines*. Warner Brothers.
- MSCHF. (ND). *Dead Startup Toys: Theranos minilab* [PVC].
<https://deadstartuptoys.com/product/theranos>
- Mullagh, L., Jacobs, N., Kwon, N., Markovic, M., Wainwright, B., Chekansky, K., & Cooper, R. (2022). Participatory IoT Policies: A Case Study of Designing Governance at a Local Level. *DRS2022: Bilbao*.
- Mumford, L. (2010). *Technics and Civilization*. The University of Chicago Press.
- Munoz, C., Smith, M., & Patil, D. (2016). *Big data: A report on algorithmic systems, opportunity, and civil right*. Executive Office of the President.
- 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* (pp. 75–100). The MIT Press.
- NA. (2020, June 5). Pope’s November prayer intention: That progress in robotics and AI “be human” [News Blog]. *Vatican News*. <https://www.vaticannews.va/en/pope/news/2020-11/pope-francis-november-prayer-intention-robotics-ai-human.html>
- Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review*, 83(4), 435–450.
- Natarajan, S., Mishra, K., Mohamed, S., & Taylor, A. (2021). *Just and Equitable Data Labelling towards a responsible AI supply chain*. aapti institute.
- Ndulue, C., Oyebode, O., Iyer, R. S., Ganesh, A., Ahmed, S. I., & Orji, R. (2022). Personality-targeted persuasive gamified systems: Exploring the impact of application domain on the effectiveness of behaviour change strategies. *User Model User-Adap Inter*.
<https://doi.org/doi.org/10.1007/s11257-022-09319-w>
- Near Future Laboratory. (2015). *An Ikea Catalog from The Near Future*.
<https://nearfuturelaboratory.myshopify.com/products/ikea-catalog-from-the-near-future>
- Neff, G. (2022, June 24). *Right now, and I mean this instant, delete every digital trace of any menstrual tracking. Please*. <https://twitter.com/ginasue/status/1540354137304760321>
- Neumann, J. von. (1958). *The Computer and the Brain*. Yale University Press.

- Norman, D. (1983). Some Observations on Mental Models. In D. Genter & A. Stevens L. (Eds.), *Mental Models* (pp. 7–14). Lawrence Erlbaum.
- Norman, D. (1988). *The Design of Everyday Things*. Basic Books.
- Norman, D. (1998). *The Invisible Computer: Why Good Products Can Fail, the Personal Computer is So Complex, and Information Appliances are the Solution*. MIT.
- Norman, D. (2005). Human-Centered Design Considered Harmful. *IX Interactions*, *XII*.
<https://interactions.acm.org/archive/view/july-august-2005/human-centered-design-considered-harmful1>
- Norman, D. (2011). *Living with Complexity*. The MIT Press.
- Norman, D., & Draper, S. (1986). *User Centered System Design: New Perspectives on Human-computer Interaction*. Taylor & Francis.
- Norris, A. (2013). ‘How Can It Not Know What It Is?’: Self and Other in Ridley Scott’s *Blade Runner*. *Film-Philosophy*, *17*(1), 21–50.
- Numata, T., Sato, H., Asa, Y., Koike, T., Miyata, K., Nakagawa, E., Sumiya, M., & Sadato, N. (2020). Achieving affective human–virtual agent communication by enabling virtual agents to imitate positive expressions. *Scientific Reports*, *10*(1), 5977. <https://doi.org/10.1038/s41598-020-62870-7>
- O’Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.
- Open AI. (n.d.). *DALL·E 2*. Retrieved 13 August 2022, from <https://openai.com/dall-e-2/>
- Open AI. (2022). *DALL·E 2 Preview—Risks and Limitations*. Github. <https://github.com/openai/dalle-2-preview/blob/main/system-card.md>
- openaidalle [@openaidalle]. (2022). *Openaidalle*. Instagram. <https://www.instagram.com/openaidalle/>
- Orji, R., & Moffatt, K. (2016). Persuasive technology for health and wellness: State-of-the-art and emerging trends. *Sage Journals*, *24*(1), 66–91.
<https://doi.org/doi.org/10.1177/1460458216650979>
- Ortega, J. y G. (1914). *An Essay in Esthetics by Way of a Preface*.

- Paltridge, B. (2002). Thesis and dissertation writing: An examination of published advice and actual practice. *English for Specific Purposes*, 21(2), 125. [https://doi.org/10.1016/S0889-4906\(00\)00025-9](https://doi.org/10.1016/S0889-4906(00)00025-9)
- Papanek, V. (1983). *Design for Human Scale*. Van Nostrand Reinhold Company Ltd.
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(3), 286–297. <https://doi.org/10.1109/3468.844354>
- Parikka, J. (2015). *A Geology of Media: Vol. Electronic Mediations*. University of Minnesota Press.
- Partnership on AI. (2019). *Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System*. PAI. <https://partnershiponai.org/paper/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/>
- Patel, K., Bancroft, N., Drucker, S. M., Fogarty, Ko, A., J., & Landay, J. (2010). Gestalt: Integrated support for implementation and analysis in machine learning. *UIST '10: Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology*, 37–46. <https://doi.org/10.1145/1866029.1866038>
- Paulson, S. (2019). Making Kin: An Interview with Donna Haraway. *LARB: Los Angeles Review of Books*. <https://lareviewofbooks.org/article/making-kin-an-interview-with-donna-haraway/>
- Pedriana, P. (2003). *SimCity 4* [Microsoft Windows, Mac OS X]. Maxis.
- Peirce, C. S. (1991). On a New List of Categories. In J. Hoopes (Ed.), *Peirce on Signs* (pp. 23–33). University of North Carolina Press; JSTOR. http://www.jstor.org/stable/10.5149/9781469616810_hoopes.7
- Petrie, H. G. (1992). Interdisciplinary Education: Are We Faced with Insurmountable Opportunities? *Review of Research in Education*, 18, 299–333.
- Pichai, S. (2018). AI at Google: Our principles. *Google*. <https://www.blog.google/technology/ai/ai-principles/>
- 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., & Coulton, P. (2020). Reseaching and Designing Uncanny AI to Legible AI. *International Transdisciplinary Conference*.
- Pilling, F., Akmal, H. A., Gradinar, A., Lindley, J., & Coulton, P. (2020). Legible AI by Design: Design Research to Frame, Design, Empirically Test and Evaluate AI Iconography. *Common Good Framing Design through Pluralism and Social Values: Design as Common Good*, 2442–2459.
- Pilling, F., Akmal, H. A., Gradinar, A., Lindley, J., & Coulton, P. (2021). Using Game Engines to Design Digital Workshops for AI Legibility. *14th International Conference of the European Academy of Design, Safe Harbours for Design Research*, 394–403.
- Pilling, F., Akmal, H. A., Lindley, J., & Coulton, P. (2022). Making a Smart City Legible. In S. Carta (Ed.), *Machine Learning and the City* (pp. 453–465). John Wiley & Sons, Ltd.
- Pilling, F., Akmal, H. A., Lindley, J., Gradinar, A., & Coulton, P. (2022). Making AI Infused Products and Services more Legible. *Leonardo*, 1–11.
- 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](https://doi.org/doi:https://doi.org/10.21606/drs.2020.218)
- Pilling, F., & Coulton, P. (2021). Carpentered Diegetic Things: Alternative Design Ideologies for AI Material Relations. *The Ecological Turn. Design, Architecture and Aesthetics beyond ‘Anthropocene’*. The Ecological Turn, Bologna, Italy.
- Pilling, F., Lindley, J., Akmal, H. A., & Coulton, P. (2021). Design (Non) Fiction: Deconstructing/Reconstructing The Definitional Dualism of AI. *International Journal of Film and Media Arts*, 6(1), 6–32.
- Pilling, F., Stead, M., & Gradinar, A. (2022). The Prometheus Terminal: Worlding Games for the Adoption of Sustainable Datafication and Cybersecurity practices. *Cumulus Detroit 2022: Design for Adaptation*. Cumulus, Detroit.

- Pilling, M., Coulton, P., Lodge, T., Crabtree, A., & Chamberlain, A. (2022a). Experiencing mundane AI futures. *DRS2022: Bilbao*. DRS, Bilbao, Spain.
<https://doi.org/doi.org/10.21606/drs.2022.283>
- Pilling, M., Coulton, P., Lodge, T., Crabtree, A., & Chamberlain, A. (2022b). Experiencing Mundane AI Futures. *DRS2022 Bilbao: Design Research Society Conference 2022*. Design Research Society Conference, Bibao.
- Plato. (2005). *The Collected Dialogues of Plato: Including the Letters* (E. Hamilton & H. Cairns, Eds.). Princeton University Press.
- Podesta, J., Pritzker, P., Moniz J, E., Holdren, J., & Zients, J. (2014). *Big data: Seizing opportunities and preserving values*. Executive Office of the President.
https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf
- Polit, S. (1984). R1 and Beyond: AI Technology Transfer at Digital Equipment Corporation. *AI Magazine*, 5(4), 76.
- Pólvara, A., & Nascimento, S. (2021). Foresight and design fictions meet at a policy lab: An experimentation approach in public sector innovation. *Futures*, 128.
<https://doi.org/10.1016/j.futures.2021.102709>.
- Popper, K. (2002). *Conjectures and Refutations: The Growth of Scientific Knowledge* (7th ed.). Routledge.
- Price, R. (2016, March 24). *Microsoft is deleting its AI chatbot's incredibly racist tweets* [Digital News]. Business Insider. <https://www.businessinsider.in/Microsoft-is-deleting-its-AI-chatbots-incredibly-racist-tweets/articleshow/51539858.cms>
- Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., & Hardman Taylor, S. (2017). “Alexa is my new BFF”: Social Roles, User Satisfaction, and Personification of the Amazon Echo. *2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2853–2859.
<https://doi.org/doi.org/10.1145/3027063.3053246>

- Rader, E., Cotter, K., & Cho, J. (2018). Explanations as Mechanisms for Supporting Algorithmic Transparency. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. <https://doi.org/10.1145/3173574.3173677>
- Ramchurn, S., Stein, S., & Jennings, N. R. (2021). *Trustworthy human-AI partnerships*. 24(8), 102891. <https://doi.org/doi.org/10.1016/j.isci.2021.102891>
- Rapp, C. (2010). Aristotle's Rhetoric. *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/spr2010/entries/aristotle-rhetoric/>
- Rashid, O., Bamford, W., Coulton, P., & Edward, R. (2006). PAC-LAN: the human arcade. *ACE '06 Proceedings of the 2006 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology*. On Advances In Computer Entertainment Technology, Hollywood. <https://doi.org/10.1145/1178823.1178864>
- Rauterberg, M. (2003). Human-Computer Interaction Research: A Paradigm Clash? *Tales of the Disappearing Computer*. Disappearing Computer Conference, Patras, Greece.
- 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>
- Redström, J., & Wiltse, H. (2018). *Changing Things: The Future of Objects in a Digital World*. Bloomsbury Visual Arts.
- Reed, P. (2019). Orientation in a Big World: On the Necessity of Horizonless Perspectives. *E-Flux*, 101. <https://www.e-flux.com/journal/101/273343/orientation-in-a-big-world-on-the-necessity-of-horizonless-perspectives/>
- Rhea, D. (2003). *Bringing Clarity to the 'Fuzzy Front End' in Design Research Methods and Perspectives* (B. Laurel, Ed.). MIT press.
- Rhodes, M. (2016). How Arrival's Designers Crafted a Mesmerizing Alien Alphabet. *Wired*. <https://www.wired.com/2016/11/arrivals-designers-crafted-mesmerizing-alien-alphabet/>
- Ribeiro, M., T., Singh, S., & Guestrin, C. (2016). 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. 1135–1144. <https://doi.org/10.1145/2939672.2939778>
- Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (Eds.). (2011). *Recommender Systems Handbook*. Springer.

- Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a General Theory of Planning. *Policy Sciences*, 4(2), 155–169.
- Roden, D. (2015). *Posthuman Life: Philosophy at the Edge of the Human*. Routledge.
- Rogers, A. (2017). The Way the World Ends: Not with a Bang But a Paperclip. *Wired*.
<https://www.wired.com/story/the-way-the-world-ends-not-with-a-bang-but-a-paperclip/>
- Romele, A. (2020). *Digital Hermeneutics Philosophical Investigations in New Media and Technologies*. Routledge.
- Romele, A., Severo, M., & Furia, P. (2020). Digital hermeneutics: From interpreting with machines to interpretational machines. *AI & Society*, 35, 73–86. <https://doi.org/doi.org/10.1007/s00146-018-0856-2>
- Rosenberger, R. (2012). Embodied technology and the dangers of using the phone while driving. *Phenomenology and the Cognitive Sciences*, 11, 79–94. <https://doi.org/10.1007/s11097-011-9230-2>
- Rosenberger, R. (2014). The Phenomenological Case for Stricter Regulation of Cell Phones and Driving. *Techné: Research in Philosophy and Technology*, 18(1–2), 20–47.
<https://doi.org/10.5840/techne201461717>
- Rosenberger, R., & Verbeek, P.-P. (2015). A Field Guide to Postphenomenology. In *Postphenomenological Investigations* (pp. 9–41). Lexington Books.
- Rosenblatt, J. (2021). Former Apple Designer Kare Testifies at Samsung Patent Trial,. *Bloomberg Business Week*. <https://www.bloomberg.com/news/articles/2012-08-07/former-apple-designer-kare-testifies-at-samsung-patent-trial>
- Rouse, W. B. (1991). *Design for Success: A Human-Centered Approach to Designing Successful Products and Systems*. John Wiley & Sons.
- Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *ArXiv:1811.10154v3*.
- Russell, S., & Norvig, P. (1995). *Artificial Intelligence: A modern approach*. Simon & Schuster.
- Rutkin, A. (2016). Robot eyes and humans fix on different things to decode a scene. *New Scientist*.
<https://www.newscientist.com/article/2095616-robot-eyes-and-humans-fix-on-different->

things-to-decode-a-

scene/?utm_source=rakuten&utm_medium=affiliate&utm_campaign=2116208:Skimlinks.com&utm_content=10&ranMID=47192&ranEAID=TnL5HPStwNw&ranSiteID=TnL5HPStwNw-gcURLCOx51NRMu1o0DYQig

Rutter, R. (2016, November 2). Peter in Conversation with Don Norman About UX & Innovation | Adaptive Path [Podcast catalog]. *Huffduffer*. <https://huffduffer.com/clagnut/370516>

Sailaja, N., Crabtree, A., & Stenton, P. (2017). Challenges of using Personal Data to Drive Personalised Electronic Programme Guides. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 5226–5231. <https://doi.org/10.1145/3025453.3025986>

Sanders, L. (2008). An evolving map of design practice and design research. *ACM Interactions*, XV(6).

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. *A Preconference at the 64th Annual Meeting of the International Communication Association*.

Schneider, S., & Turner, E. (2017). Is Anyone Home? A Way to Find Out If AI Has Become Self-Aware: It's not easy, but a newly proposed test might be able to detect consciousness in a machine. *Scientific American*. <https://blogs.scientificamerican.com/observations/is-anyone-home-a-way-to-find-out-if-ai-has-become-self-aware/>

Schön, D. A. (1983). *The Reflective Practitioner: How Professionals Think in Action*. Basic Books.

Schreier, J. (Director). (2012). *Robot & Frank*. Samuel Goldwyn Films.

Schudson, M. (2015). *The Rise of the Right to Know*. Belknap Publishing.

Schwartz, O. (2018, July 25). 'The discourse is unhinged': How the media gets AI alarmingly wrong. *The Guardian*. <https://www.theguardian.com/technology/2018/jul/25/ai-artificial-intelligence-social-media-bots-wrong>

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Psychol*, 25, 1–65. [https://doi.org/10.1016/S0065-2601\(08\)60281-6](https://doi.org/10.1016/S0065-2601(08)60281-6)

Schwartz, S. H. (2006). *Basic human values: An overview*.

- Scott, R. (Director). (1979). *Alien*. 20th Century Fox.
- Scott, R. (Director). (1982). *Blade Runner*. Warner Bros.
- Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–457.
- Selbst, A. D. (2017). A Mild Defense of Our New Machine Overlords. *70 Vanderbilt Law Review En Banc* 87.
- Seymour, N. (2022). *Glitter*. Bloomsbury Publishing.
- Shaviro, S. (2011, October 4). Panpsychism And/Or Eliminativism [Blog]. *Shaviro*.
<http://www.shaviro.com/Blog/?p=1012>
- Shelley, M. W. (1818). *Frankenstein, Or the Modern Prometheus*. Lackington, Hughes, Harding, Mavor, & Jones.
- Sherman, R., & Sherman, R. (1963). *It's a Small World*. Walt Disney Records.
- Shneiderman, B. (2016). The dangers of faulty, biased, or malicious algorithms requires independent oversight. *Proceedings of the National Academy of Sciences*, 113(48), 13538–13540.
- Shneiderman, B. (2020a). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction*, 36(6), 495–504.
<https://doi.org/10.1080/10447318.2020.1741118>
- Shneiderman, B. (2020b). Human-Centered Artificial Intelligence: Three Fresh Ideas. *AIS Transactions on Human-Computer Interaction*, 12(3), 109–124.
<https://doi.org/doi.org/10.17705/1thci.00131>
- Sicart, M. (2014). *Play Matters*. The MIT Press.
- Silverstone, R. (2006). Domesticating domestication. Reflecting on the life of a concept. In T. Berker, M. Hartmann, Y. Punie, & K. Ward (Eds.), *Domestication Of Media And Technology* (pp. 229–247). Open University Press.
- Simon, H. A. (1969). *The Sciences of the Artificial*. The MIT Press.
- Simon, H. A. (1998). Allen Newell: 1927-1992. *IEEE Annals of the History of Computing*, 20(2), 63–76.
- Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *ArXiv Preprint ArXiv:1312.6034*.

- Singh, V., Mani, A., & Pentland, A. (2014). Social Persuasion in Online and Physical Networks. *Proceedings of the IEEE*, 102, 1903–1910.
- Slade, D. (Director). (2018). *Black Mirror: Bandersnatch*. Netflix.
- Smith, J. (2016). *Experiencing Phenomenology: An Introduction*. Routledge.
- Smith, J. (ND). *Phenomenology* [Peer reviewed academic resource]. Internet Encyclopedia of Philosophy. <https://iep.utm.edu/phenom/>
- Sowa, J. F. (2000). Ontology, Metadata, and Semiotics. *Conceptual Structures: Logical, Linguistic, and Computational Issues, 1867*. https://doi.org/10.1007/10722280_5
- Spielberg, S. (Director). (2002). *Minority Report*. Twentieth Century Fox.
- Stahl, W. A. (1995). Venerating the Black Box: Magic in Media Discourse on Technology. *Science, Technology, & Human Values*, 20(2), 234–258. <https://doi.org/10.1177/016224399502000205>
- Stanford. (2020). *Stanford Institute for Human-Centered Artificial Intelligence 2019-2020 Annual Report*. <https://hai-annual-report.stanford.edu/#>
- Stanford. (ND). Stanford University Human-Centured Artificial Intelligence. *About*. <https://hai.stanford.edu/about#:~:text=It's%20for%20this%20reason%20that,to%20improve%20the%20human%20condition.>
- Stappers, P. J., & Giaccardi, E. (2017). Research through Design. In *The Encyclopedia of Human-Computer Interaction* (2nd ed.). Interaction Design Foundation. <https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/research-through-design>
- Star, S. L. (2010). This is Not a Boundary Object: Reflections on the Origin of a Concept. *Science, Technology, & Human Values*, 35(5), 601–617.
- Star, S. L., & Griesemer, J. R. (1989). Institutional Ecology, ‘Translations’ and Boundary Objects: Amateurs and Professionals in Berkeley’s Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, 19(3), 387–420.
- Stark, L., & Hoey, J. (2020). The Ethics of Emotion in Artificial Intelligence Systems. *OSF Preprints*, 12. <https://doi.org/10.31219/osf.io/9ad4u>

- Stead, M., Gradinar, A., & Coulton, P. (2020). Must All Things Pass?: Designing for the Afterlife of (Internet of) Things. *ThingsCon The State of Responsible Internet of Things Report 2020*, 11(4), 45–52.
- Stead, M., Gradinar, A., Coulton, P., & Lindley, J. (2020). Edge of Tomorrow: Designing Sustainable Edge Computing. *Design Research Society Conference 2020*. Design Research Society Conference. <https://doi.org/doi.org/10.21606/drs.2020.293>
- Stebbins, S. (1993). Anthropomorphism. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 69(2/3), 113–122.
- Stember, M. (1991). Advancing the social sciences through the interdisciplinary enterprise. *The Social Science Journal*, 28(1), 1–14. [https://doi.org/doi.org/10.1016/0362-3319\(91\)90040-B](https://doi.org/doi.org/10.1016/0362-3319(91)90040-B)
- Sterling, B. (2005). *Shaping Things*. The MIT Press.
- Stockton, N. (2017). This New Atari-Playing AI Wants to Dethrone DeepMind Schema Networks' creators say it wins because it can think about the past, and plan for the future. *Wired*. <https://www.wired.com/story/vicarious-schema-networks-artificial-intelligence-atari-demo/>
- Stohl, C., Stohl, & Leonardi, P. (2016). Managing Opacity: Information Visibility and the Paradox of Transparency in the Digital Age. *International Journal of Communication*, 10, 123–137.
- Stolterman, E. (2008). The Nature of Design Practice and Implications for Interaction Design Research. *International Journal of Design*, 2(1), 55–65.
- Sulieman, A. N., Celsi, R. L., Li, W., Zomaya, A., & Villari, M. (2022). Edge-Oriented Computing: A Survey on Research and Use Cases. *Energies*, 15(452). <https://doi.org/10.3390/en15020452>
- Suvin, D. (1972). On the Poetics of the Science Fiction Genre. *College English*, 34(3), 372–382. JSTOR. <https://doi.org/10.2307/375141>
- Swann, C. (2002). Action Research and the Practice of Design. *Design Issues*, 18(2), 49–61.
- Taleb, N. (2007). *The black swan: The impact of the highly improbable*. Random House.
- Taylor, N., Cheverst, K., Wright, P., & Oliver, P. (2013). Leaving the wild: Lessons from community technology handovers. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1549–1558.

- The Data Nutrition Project*. (n.d.). Data Nutrition. Retrieved 17 August 2022, from <https://datanutrition.org/>
- Thill, B. (2015). *Waste*. Bloomsbury Publishing.
- Thom, R., & Noel, E. (1991). *Prédire n'est pas expliquer* (Flammarion, Ed.). Champs Sciences.
- Thomas, W. I., & Thomas, D. S. (1928). *The child in America: Behavior problems and programs*. Knopf.
- Thompson, P. (1999). Exploring the contexts of writing: Interviews with PhD supervisors. *Issues in EAP Writing Research and Instruction*, 37–54.
- Thornton, L., Knowles, B., & Blair, G. (2022). The Alchemy of Trust: The Creative Act of Designing Trustworthy Socio-Technical Systems. *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. <https://doi.org/10.1145/3531146.3533196>
- Tolman, C. W., Cherry, F., van Hezewijk, R., & Lubek, I. (Eds.). (1996). Problems of Theoretical Psychology. *Problems of Theoretical Psychology*.
- Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, 236, 433–460.
- Ubiquity staff. (2005). An Interview with John Markoff: What the dormouse said. *Ubiquity*, 2005(August), 1. <https://doi.org/doi.org/10.1145/1088431.1088206>
- Uexküll, J. von. (2010). *A Foray into the Worlds of Animals and Humans: With A Theory of Meaning* (J. O'Neil, Trans.). University of Minnesota Press.
- Urquiza-Haas, E. G., & Kotrschal, K. (2015). The mind behind anthropomorphic thinking: Attribution of mental states to other species. *Animal Behaviour*, 109, 167–176. <https://doi.org/10.1016/j.anbehav.2015.08.011>
- US Public Policy Council. (2017). *Statement on Algorithmic Transparency and Accountability*. https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf
- Van Den Eede, Y. (2011). In Between Us: On the Transparency and Opacity of Technological Mediation. *Found Sci*, 16, 139–159. <https://doi.org/DOI 10.1007/s10699-010-9190-y>

- Van Den Eede, Y. (2022). Thing-Transcendentality: Navigating the Interval of “technology” and “Technology”. *Foundations of Science*, 27, 225–243. <https://doi.org/doi.org/10.1007/s10699-020-09749-y>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is All You Need*. The Conference on Neural Information Processing Systems.
- Verbeek, P.-P. (2005). *What Things Do: Philosophical Reflections on Technology, Agency and Design*. Penn State University Press.
- Verbeek, P.-P. (2006). Persuasive Technology and Moral Responsibility: Toward an ethical framework for persuasive technologies. *Persuasive*.
- Verbeek, P.-P. (2008a). Cyborg intentionality: Rethinking the phenomenology of human–technology relations. *Phenomenology and the Cognitive Sciences*, 7, 387–395.
- Verbeek, P.-P. (2008b). Obstetric Ultrasound and the Technological Mediation of Morality: A Postphenomenological Analysis. *Human Studies*, 31, 11–26. <https://doi.org/doi.org/10.1007/s10746-007-9079-0>
- Verbeek, P.-P., & Kockelkoren, P. (1998). *The Things That Matter*. 14(3), 28–42.
- Vernant, J. (1992). The birth of images. In *Mortals and Immortals: Collected Essays* (pp. 164–185). Princeton University Press.
- Verplank, B. (2009). *Interaction Design Sketchbook*. <http://www.billverplank.com/IxDsketchBook.pdf>
- Victorelli, E., Z., Dos Reis, J., C., Hornung, H., & Prado, A., B. (2020). Understanding human-data interaction: Literature review and recommendations for design. *International Journal of Human-Computer Studies*, 134, 13–32.
- Villeneuve, D. (Director). (2016). *Arrival* [Science Fiction]. Paramount Pictures.
- Vincent, J. (2023). AI is killing the old web, and the new web struggles to be born. *The Verge*. <https://www.theverge.com/2023/6/26/23773914/ai-large-language-models-data-scraping-generation-remaking-web>
- Vinge, V. (1993). The coming technological singularity. *Whole Earth Review*.

- Volodymyr, M., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing Atari with Deep Reinforcement Learning. *ArXiv Preprint ArXiv:13125602*.
- 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/>
- Voss, G., Revell, T., & Pickard, J. (2015). *Speculative Design and the Future of an Ageing Population Report 2: Techniques*. Government office for Science / Strange Telemetry.
- Wachowski, L., & Wachowski, L. (Directors). (1999). *The Matrix*. Warner Brothers.
- Wachowski & Wachowski (Directors). (2003a). *The Matrix Revolutions*. Warner Brothers.
- Wachowski, & Wachowski, L. (Directors). (2003b). *The Matrix Reloaded*. Warner Brothers.
- Wakkery, R. (2021). *Things We Could Design: For More Than Human-Centered Worlds*. The MIT Press.
- Walker, K. (2018). Investigative design Materiality, systems, critique. *Beyond Change*. Swiss Design Network.
- Wartenberg, T. (2007). *Thinking on Screen: Film as Philosophy*. Routledge.
- Warwick, K. (2012). *Artificial Intelligence the basics*. Routledge.
- Weir, S. (2020). Living and Nonliving Occasionalism. *Open Philosophy*, 3(1), 147–160.
<https://doi.org/doi.org/10.1515/opphil-2020-0010>
- Weiser, M. (1991). The Computer for the 21st Century. *Scientific American*, 265(3), 94–105.
- Weld, D. S., & Bansal, G. (2019a). The challenge of crafting intelligible intelligence. *Communications of the ACM*, 62(6), 70–79. <https://doi.org/10.1145/3282486>
- Weld, D. S., & Bansal, G. (2019b). The Challenge of Crafting Intelligible Intelligence. *Communications of the ACM*, 62(6), 70–79. <https://doi.org/10.1145/3282488>
- Weller, A. (2017). *Transparency: Motivations and Challenges*.
<https://doi.org/10.48550/ARXIV.1708.01870>

- Wellner, G. (2018). From Cellphones to Machine Learning. A Shift in the Role of the User in Algorithmic Writing. *Towards a Philosophy of Digital Media*, 205–224.
https://doi.org/doi.org/10.1007/978-3-319-75759-9_11
- Wells. (1895). *The Time Machine: An Invention*. Heinemann.
- Wen, Z., Yang, R., Garraghan, P., Xu, J., & Rovatsos, M. (2017). *Fog Orchestration for IoT Services: Issues, Challenges and Directions*. IEEE Internet Computing.
<https://doi.org/10.1109/MIC.2017.36>
- 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>
- Wiener, N. (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine* (2nd ed.). The MIT Press.
- Wilcox, F. M. (Director). (1956). *Forbidden Planet*. Metro-Goldwyn-Mayer.
- Wiltse, H. (2014). Unpacking Digital Material Mediation. *Techné: Research in Philosophy and Technology*, 18(3), 154–182. <https://doi.org/doi.org/10.5840/techne201411322>
- Winograd, T. (1997). From Computing Machinery to Interaction Design. In P. Denning & R. Metcalfe (Eds.), *Beyond Calculation: The Next Fifty Years of Computing* (pp. 149–162). Springer-Verlag.
- Winograd, T. (2006). Shifting viewpoints: Artificial intelligence and human–computer interaction. *Artificial Intelligence*, 170, 1256–1258.
- Winograd, T., & Flores, F. (1986). *Understanding Computers and Cognition*. Ablex.
- Wolfe, C. (2010). *What Is Posthumanism?* University of Minnesota Press.
- Wood, A. (2002). *Technoscience in contemporary American film Beyond science fiction*. Manchester University Press.
- Wortman Vaughan, J., & Wallach, H. (2022). *A Human-Centered Agenda for Intelligible Machine Learning*. <https://www.microsoft.com/en-us/research/publication/a-human-centered-agenda-for-intelligible-machine-learning/>

- Yang, K., Stoyanovich, J., Asudeh, A., Howe, B., Jagadish, H., & Miklau, G. (2018). A Nutritional Label for Rankings. *Proceedings of the 2018 International Conference on Management of Data*, 1773–1776. <https://doi.org/10.1145/3183713.3193568>
- Yang, S. J. H., Ogata, H., Matsui, T., & Chen, N.-S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, 2, 100008. <https://doi.org/10.1016/j.caeai.2021.100008>
- Yoni, V. D. E. (2022). Thing-Transcendentality: Navigating the Interval of “technology” and “Technology”. *Found Sci*, 27, 225–243. <https://doi.org/doi.org/10.1007/s10699-020-09749-y>
- Zahavi, D. (2016). The end of what? Phenomenology vs. Speculative realism. *International Journal of Philosophical Studies*, 24(3), 289–309. <https://doi.org/10.1080/09672559.2016.1175101>
- Zahavi, D. (2019). *Phenomenology: The Basics*. Routledge.
- Zahavi, D. (2022). Husserl’s Turn to Transcendental Philosophy: Epoché, Reduction, and Transcendental Idealism. In *Husserl’s Phenomenology* (pp. 43–78). Stanford University Press.
- Zeigler, E. (1990). Don’t forget the profession when choosing a name! In C. Corbin & H. Eckert (Eds.), *The evolving undergraduate major* (pp. 67–77). Champaign.
- Zimmerman, J., Forlizzi, J., & Evenson, S. (2007). *Research through design as a method for interaction design research in HCI*. 493–502.
- Zimmerman, J., Stolterman, E., & Forlizzi, J. (2010). An analysis and critique of Research through Design: Towards a formalization of a research approach. *Proceedings of the 8th ACM Conference on Designing Interactive Systems*, 310–319. <https://doi.org/doi.org/10.1145/1858171.1858228>
- Zuber-Skerritt, O. (2001). Action Learning and Action Research: Paradigm, Praxis and Programs. In S. Sankara, B. Dick, & R. Passfield (Eds.), *Effective Change Management through Action Research and Action Learning: Concepts, Perspectives, Processes and Applications* (pp. 1–20). Southern Cross University Press.
- Zuber-skerritt, O. (2015). Action Research. In *In: Professional Learning in Higher Education and Communities*. Palgrave Macmillan.

Zuboff, S. (1998). *In The Age of the Smart Machine*. Basic Books.

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

The End