THE UNIVERSITY
of EDINBURGH

EXPECTATIONS AND EXPERTISE IN ARTIFICIAL INTELLIGENCE:

SPECIALIST VIEWS AND HISTORICAL PERSPECTIVES ON CONCEPTUALISATION, PROMISE, AND FUNDING

Vassilis Galanos

PhD Science and Technology Studies

University of Edinburgh

2023

ABSTRACT

Artificial intelligence's (AI) distinctiveness as a technoscientific field that imitates the ability to think went through a resurgence of interest post-2010, attracting a flood of scientific and popular expectations as to its utopian or dystopian transformative consequences. This thesis offers observations about the formation and dynamics of expectations based on documentary material from the previous periods of perceived AI hype (1960-1975 and 1980-1990, including in-between periods of perceived dormancy), and 25 interviews with UK-based AI specialists, directly involved with its development, who commented on the issues during the crucial period of uncertainty (2017-2019) and intense negotiation through which AI gained momentum prior to its regulation and relatively stabilised new rounds of long-term investment (2020-2021). This examination applies and contributes to longitudinal studies in the sociology of expectations (SoE) and studies of experience and expertise (SEE) frameworks, proposing a historical sociology of expertise and expectations framework. The research questions, focusing on the interplay between hype mobilisation and governance, are: (1) What is the relationship between AI practical development and the broader expectational environment, in terms of funding and conceptualisation of AI? (2) To what extent does informal and non-developer assessment of expectations influence formal articulations of foresight? (3) What can historical examinations of AI's conceptual and promissory settings tell about the current rebranding of AI?

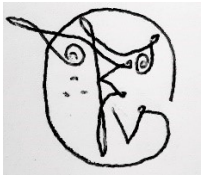
The following contributions are made: (1) I extend SEE by paying greater attention to the interplay between technoscientific experts and wider collective arenas of discourse amongst non-specialists and showing how AI's contemporary research cultures are overwhelmingly influenced by the hype environment but also contribute to it. This further highlights the interaction between competing rationales focusing on exploratory, curiosity-driven scientific research against exploitation-oriented strategies at formal and informal levels. (2) I suggest benefits of examining promissory environments in AI and related technoscientific fields longitudinally, treating contemporary expectations as historical products of sociotechnical trajectories through an authoritative historical reading of AI's shifting conceptualisation and attached expectations as a response to availability of funding and broader national imaginaries. This comes with the benefit of better perceiving technological hype as migrating from social group to social group instead of fading through reductionist cycles of disillusionment; either by rebranding of technical operations, or by the investigation of a given field by non-technical practitioners. It also sensitises to critically examine broader social expectations as factors for shifts in perception about theoretical/basic science research transforming into applied technological fields. Finally, (3) I offer a model for understanding the significance of interplay between conceptualisations, promising, and motivations across groups within competing dynamics of collective and individual expectations and diverse sources of expertise.

Cover image: Alberto Savinio, La Cite Des Promesses [The City of Promises], 1928, Pinacoteca Di Brera, Milano (my photograph). [Machine readable description: a messy arrangement of mostly rectangular and some curved objects resembling wood carved buildings placed on a piece of land sharply cut and removed from below, giving the impression of a floating island; abstract clouds or fog on the background.]

DECLARATION OF OWN WORK

I hereby declare that the present doctoral work has been conducted and composed by myself and that it has not been submitted in any other application for a degree, in whole or in part. Except where acknowledged via reference and citation, the work is my own in its entirety. Parts of this work have been submitted for publication (chapter 6; albeit with substantial changes).

Vassilis Galanos,



Edinburgh, 25-02-2022

LAY SUMMARY

The present research investigates artificial intelligence (AI) from a social science perspective. It is using a combination of historical research and interviews with AI specialists in order to make sense of the influence of imagined futures on contemporary practical decision making and of the different types of experts who are considered to be credible for taking such decisions. Since the late 2000s, AI has often been presented as a new and emerging field with very high promises for various domains of everyday life, from business and entertainment to healthcare and military. Because of its deep association with science fiction and mythological contexts, its adoption and further development evokes multiple future scenarios: will AI become conscious? Will AI take away people's jobs? Will AI solve all our problems? One of the core problems I identified during the early stages of this research was that not so many AI specialists have been involved into public debates which shaped public perception and governmental policy about AI; instead, public influential figures who had no specialisation in AI's technical matters were responsible for a trend which presented AI as a utopian solution to many problems or as a dystopian future nightmare responsible for the end of humanity. In order to understand, therefore, what is to be done with AI's development, I sought to understand what it is, how it came about, and how it is being negotiated. I investigated its history and those who developed it since 1955 and I interviewed 25 specialists in the field. I analysed my findings focusing on how images from the future impact research and development,

Indeed, AI is not as new as it is often presented in mass media. Since its conception in 1955, it has attracted investor and public attention several times. This attention, or hype, was followed by periods of disillusionment when grandiose promises were proven to be unfulfilled or unrealisable, and investors lost trust to the technical communities who promised great results. Such periods are known among specialist communities as "AI winters" and according to AI history they have occurred twice thus far, once in the mid-1970s and once in the early 1990s. In this dissertation, I show that this notion of the AI winters is problematic for mainly two reasons: on the one hand, it does not take into account that AI is not a single technology but a bundle of applications which continue to evolve even during these periods of lower trust and funding in AI projects; otherwise, there would not have been a current AI resurgence. On the other hand, the metaphor of the AI winter suggests that it is going to be repeated, in a seasonal manner; through the following ways, I show that the problem is much more complicated than that.

First, by reviewing the history of AI, I show how several social factors, from personal interests and academic rivalry to political games, international competition, and social justice awareness, were responsible for slowly transforming AI from a scientific vision to create machines that think like humans (or in some cases nonhuman animals) into a much more restricted technological application. Simply put, those who wanted to invest in AI development did not care about creating artificial brains, but they were be interested in creating practical industrial benefit or producing technologies directly beneficial to military applications. This happened in response to visionary researchers who exaggerated their short-term promises. Right now, contemporary advancements in AI have led to a series of policies regulating how AI is developed and disseminated, and currently, governments understand AI as a very narrow technical field with little or nothing to do with the grand vision of replicating intelligence.

Based on this historical examination, interviewed specialists offered their views based on four central themes: how they understand AI? What do they think about promises? How do they develop their funding strategies? These questions are interconnected and led to the following results. AI specialists agree that they disagree as to what AI is, or even that AI can ever have a definition. This can be seen both as a problem of non-specialist bodies of influence dominating what is considered to be AI which is at odds with their own experience, but also their own backgrounds suggesting alternative approaches. On the topic of promises, they report that hype may harm their research because people might expect or fear more than what they can deliver, but many of them also admit they often have to make slightly exaggerating promises

in order to convince investors and secure funds. Thus, depending on the approach or mind-set of every researcher, they might belong to a “romantic” visionary type who wishes to advance AI as a science in the traditional way, and those who use the hype as an opportunity to attract funds, often tailoring their interests and research proposals to the trends of governmental strategies for science and technology funding.

To offer a model that can help future social analysts of technology to think about different sources of expertise and influence in projecting expectations, I suggest to look simultaneously at how (a) the ways a technology is understood is based on (b) the underlying motivations for developing it and (c) the strategies people use to convince about its significance. This model takes into account the various experts from within who will suggest different options for technical choices as well as legal, and other political experts, but also the other sources of influence, from media to mythological fascinations and fears. The thesis ends with a proposal for future applications of this framework in more niche settings such as medical or art uses of AI.

ACKNOWLEDGEMENTS AND DEDICATIONS

This thesis would not exist if it were not for the support of numerous individuals and circumstances. I would first want extend my warmest gratitude to my supervisors Dr Gill Haddow and Prof Robin Williams; it was an honour to study under their guidance and benefitted enormously during what will always appear to me as the relationship of a Zen student to the master. I am also grateful to Prof Shannon Vallor and Prof Kornelia Konrad for being my respective internal and external thesis examiners and who offered detailed and constructive feedback.

I am indebted to my interviewees and I express my apologies for being unable to name them hereby, especially those who were happy to, or insisted that should, be named. I was astounded by their kind willingness to participate in my research, generously offering hours taken out of their precious working time, following up through emails and secondary meetings.

Additional thanks to my MSc Thesis supervisor Prof Jens-Erik Mai and my BSc Thesis supervisor Prof Stella Korobili, as well as Dr Spyros Pierros for offering enormously important guidance and believing in my academic skills at earlier stages of my career, supporting me in numerous ways. Prof Stuart Anderson and Dr Karen Gregory for being my PhD progression board advisors and Catherine Heeney for being my MSc by Research supervisor. Dr James Stewart, Dr Morgan Curry, Dr Sarah Parry, Dr Lawrence Dritsas, Dr Liz McFall, and Dr Oliver Escobar for kindly offering work opportunities during the PhD.

My biological relatives Aikaterini Galanou, Anna Fagoura, Zacharoula Fagoura, Nikos Fagouras, Giannis Hadjianagnostou, Dimitra Hadjianagnostou, and Dimitris Kazakis for offering moral and economic support to a self-funded doctoral candidate.

My logical (i.e. non-biological) siblings, my sister-of-choice Chrysa Anagnostou, Dimitris Michalas (Doom) and his family Maggie Gianakopoulou and the Doombots, and Dora Chatziioannou for always being there when needed.

I was blessed to be surrounded by amazing colleagues at our AI Ethics and Society group: SJ Bennett, Benedetta Catanzariti, and Yazmin Morlet Corti. More than colleagues, you are friends and this is rare and valuable. Sending hugs to the group of late night peer support: Sophie Stone, Simone Sambento, and Yuchen Lin. Indebted to Erik Børve Rasmussen for the lovely discussions, drinks, and disruptions during his stay in Edinburgh and for inviting me to OsloMet for my first invited talk. To Aida Ponce Del Castillo for all the chat and chocolate, and for inviting me to offer my service to ETUI. Stephen Harwood, Lukas Engelmann, Rick Woodward, and Vasilis Tsiatouras for co-organising our 2017-8 Cybernetics reading group. Numerous close friends and colleagues have been gifting me books while I was conducting this research, many of which made it to the reference list. Thank you, Chris Ecclestone for Strehl's book (now a paper published with *Interfaces*), Anna Kuslits for Deleuze's book on Foucault, Thomas Tsakalakis for your own book on political correctness (and for being the big bro in Douglas Acadams's understandings of the number 42), James Gardiner for J.G. Ballard's *High Rise* and for keeping me sane during lockdown writing times. All the following people for blurring the boundaries between friendly chat, scholarly conversation, academic feedback, invitation to conferences, co-authorship of papers, and more: Rachel Simpson, Daniel Thorpe, James Fleck, Andrey M. Elizondo, Andrez Dominguez, Francesco Michele Noera, Mary Reisel, Oscar Moreno, José David Gómez-Urrego , Rhodri Leng, Yu-Sheng Chang,

Katherine Stephen, Asli Ates, Cassius Smith Frazer, Xiao Yang, Barbara Hof, Moses Boudourides, Takaharu Oda, Carole Cusack, Anna Henschel, Rita Nikolaidou, Nathan D. Horowitz, Agnieszka Krzeminska, David Brook, Desdemona Van Tent (I am yet to finish *From Hell*), Nikolis Palikaros and John Mollindris, Phil Pantos and Dimitris Bardakas, Ray Radford, Liam Simmonds, Josh Ostrup Shires, Clare Button, Guergana Pavlova and Kostaa FlyKites, Everet Zachary Laroca, Apostolos Keratidis, Nikos Kouzinos, Panagiota Nikomani, Eirini (Eye Reene), Maria Koumarianou, Sofia Papadima, Eirini Armaou, Oliver Doherty, Carmensita Spue, Sara Manero Perez, Olga Anastasiadou, Kostas Kalhas, Ioanna Giannikopoulou, Dora Levakis, Marianna Michail, Thanos Pappas, Sarah Elizabeth Demarest, Malena Muller, Margarita Alexopoulou, Marina Palaisti, Vaggelis Koumparoudis, Dena Arya, Vaso Kriara, Andrea Hrckova, Kristi Lazlo, Melina Tsentemeidou (for our longstanding friendship, but also for inspiring me to be a vegetarian and introducing me to Situationism), Vasilis Valatsos, Andreas Lougakos (Tωm Unit), Despoina Greka, Dimitra Mavrodima, Dionysis Grekas, Eftychia Zoumpouli, Tasos Zoumpoulis, Rosie and Neil Buckland, Marianna Michail, Marianna Vasilianou, Rebecca Pitkin, Helena Lyhme, Ana Luiza,, Jarmo De Vries. Mario Alvarado, Nicolas Malo, Alexandros Gougousis (those IELTS books were very useful!), Carina Assunção, Essi Mäkelä, Alejandro Boucabeille, Takis Efstathiou, Maria Argyropoulou, Liam Sutherland, Colin Sanderson, Fiona Coyle, Yogesh Dwivedi, Jonathan Smellie, Fanis Pistokoulos, my great flatmates, Jason Zhou, Anelhy Kleeblatt, Ishak Beno, Samer Wael Hamadneh, and my landlady in Edinburgh for five years Jane Meredith, and my current landlord Rory McKenzie.

Those who know me are aware of music's influence in my overall performance. Therefore, I want to extend my gratitude to Lyrkoss (Pas Cruz Varela) and Pleiades Cluster (Theodora Poimenidou) for running the warmest YouTube folk rock music channels, my homies at the Deeper Than Underground Facebook community, my league of weeping wailers at the Wide World of Sad Old Songs Facebook group, and all mouth harp communities across different platforms. Moreover, to all good and gentle folk I have met during my endless hours on Instagram and Flickr, being my visual content companions, and my last resort in hoping for a better digital world. I have to thank all strangers I have met in my travels for teaching me that knowledge and expertise is truly ubiquitous. All nonhuman entities I have bonded with intentionally or not. My favourite bookstores and recordstores. The anarchist squats and CouchSurfing friends. Rob Liefeld and Fabian Nicieza for creating Deadpool. Due to a variety of reasons, constraints, and personal choice, this thesis has been written in various places and I am thankful to the (literal and metaphorical) gatekeepers who allowed this to happen. Breakout rooms and the computer microlab at the Crystal MacMillan Building in Edinburgh; my desk at office 1.06 in Old Surgeons' Hall, The Angel Coffee café, Bar 50 by Cowgate street, the Edinburgh Airport, office 1.09 at 27-28 George Square, table 42 at the historical Caley Picture Gallery pub, the Gasteig's Stadtbibliothek in Munich, among others.

I have dedicated my BSc and MSc dissertations to the memory of my favourite philosopher Vilém Flusser, and I so do with the present dissertation. However, during the years required to complete this work, figures which I held dear, either personally or in spirit departed from the present spatiotemporal plane, and I wish to dedicate this thesis further to the inspiration I will constantly draw from Stan Lee, Antonis Manousakis (Bdelygma), Stavroula Veloni, Angeliki Tsioli, and Daniel Dumile (MF DOOM).

While writing up, I proposed and got married to a silkily nymph. Hence, this final and warmest dedication, being an expression of eternal love, belongs to Effrosyni Antoniou.

TABLE OF CONTENTS

"[…] nowadays, all energy, all attention, is claim'd by Futurity"

Thomas Pynchon (Pynchon 1997: 369)

"What one can do but speculate, speculate, until one hits on the happy speculation?"

Samuel Beckett (Beckett 1958: 363)

"And since the Angels are not able to conceive of time, they also have a different idea of eternity than earthly people do; they understand by it an infinite state, not an infinite time."

Emanuel Swedenborg (Musil 1978: 1311)

CHAPTER 1: INTRODUCTION AND BACKGROUND

I want to begin this introduction by highlighting the interplay of imagined futures and technological developments in the everyday life. By reflecting on the following picture, I am making a case that the complex dynamics behind such everyday instances are façades of deeper and longstanding complex social processes involving governments, militaries, industries, academics, of which, I will mostly focus on the academic-governmental nexus.

Figure 1 "Future: Computer and Things Shop," Athens, Greece, October 21, 2019.

I took this picture during a visit to post-financial crisis Greece, a few months since I began putting together this dissertation's puzzle pieces. It is indicative of the many businesses that closed down during the financial crisis and no company or individual managed to rent or buy the property, replacing the old signage. I took the specific picture because it further indicates the theme of this doctoral thesis: the use of technological "future" as a currency of negotiation between different social groups. I tried to find an online reference for the shop on the picture; I could only identify a single advertisement on a digitalised version of a 1987 popular Greek tech magazine (Pixel 1987: 82). Assuming that this shop launched around that era, it is further indicative of the way computer-related "futures" have been used and hyped repeatedly in the history of computing across the world. In this thesis, I focus on what is considered a particular branch of computer science, namely artificial intelligence (AI[1]), and the contests of promise and expertise that are

[1] Ironically, while most contemporary descriptions of AI consider it a branch of computer science, the term "computer science" was first proposed in 1956 (Tedre 2014: 220), one year after its AI's initial coinage as a term – more on this, shortly.

shaping it. What is the role of expectations in the development and understanding of technologies as (at least seemingly) impactful as AI? Can there be individuals credible enough to generate such expectations and promise certain results or protect AI from unintended hazards?

1.1 Towards an Everyday Understanding of AI

An introductory chapter to a treatise on technology should be expected to offer a definition of the given technology. In the present research's case, however, AI's definition, its many interpretations and lack of terminological consensus is examined as part of the very problem formation and its historical unfolding (see Chapter 3). I want to approach the core by circling the periphery that, in my view, consists of everyday perceptions about AI that are fuelled by and fuel the large expectations associated with it. Thinking with Henri Lefebvre, instead of defining AI technically, one should look first at its everyday instances, as the

> "[…] everyday can be defined as a set of functions which connect and join together systems that might appear as distinct, Thus defined, the everyday is a product, the most general of products […]. The everyday is there for the most universal and the most unique condition, the most social and the most individuated, the most obvious and the best hidden" (Lefebvre 1987: 8).

Thus, the above picture is a snapshot of this everydayness, which I will introduce as a complex mix of theological and mythological narrative enabling "grand" visions of AI expressed in newspapers and popular science, in turn influencing policy.

1.2 The Theological Undertones of AI and their Implications for Media, Policy, and Everyday Perception

> *"Science, in one aspect, is ordered technique; in another, it is rationalized mythology" (Bernal 1954: 3).*

But what is AI and why is it important to study it? AI, although deeply associated with the idea of an abstract future, becomes so embedded in its application in everyday life, that it is available for and deserves close examination. Moreover, while the media vernacular implies that AI will shape the future of humanity, the present thesis aims at examining the same question from an almost opposite direction: what is the force that this imagined future imposes upon the shaping of AI, and consecutively, technologies of similar magnitude? Who shapes such futuristic images and what are the implications of these negotiations of promises?

The fascination with comprehending the meaning of thinking and with its mechanical replication is not new. "As if driven by some invisible hand, humans have always yearned to understand what makes them think, feel, and be, and have tried to re-create that interior life artificially" (Crevier 1993: 1-2). Literature, mythology, and religion (often difficult to separate) from around the world and from various moments in time have captured this excitement. Pygmalion's will to vivify his beloved sculpture of pseudo-Galatea[2], the "self-activated automata" created by god Hephaestus, the copper gigantic patrol automaton Talos created by Daedalus in Crete, god Ilmarinen's bride made of gold and silver in the

[2] A name given to the mythological entity by Jean Jacques Rousseau in his theatrical play based on the myth. The original story does not give a name to the sexualised feminine artefact (which some would consider precursor to contemporary sexbots; Aylett and Vargas 2021).

ancient Finnish epic story *Kalevala*, the story of wooden puppet Pinocchio and his creator Gepetto and of Mary Shelley's Dr Frankenstein and his monster (ibid: 1-2) are some well-known cases. Another similar religious creature is the golem, a product of Jewish mysticism often related to the early human fascination with artificial creatures, which, incidentally, Science and Technology Studies (STS) scholars Collins and Pinch have used as a metaphor to speak about science and technology:

> "[…] a humanoid made by man [sic] from clay and water, with incantations and spells. It is powerful. It grows a little more powerful every day. It will follow orders, do your work, and protect you from the ever threatening enemy. But it is clumsy and dangerous. Without control a golem may destroy its masters with its flailing vigour; […] A golem, in the way we intend it, is not an evil creature but it is a little daft. Golem Science [and technology] is not to be blamed for its mistakes; they are our mistakes. A golem cannot be blamed if it is doing its best. But we must not expect too much" (Collins and Pinch 1998: 1).

Interestingly, in their book, Collins and Pinch do not make any association between their golem metaphor and AI or robots[3], while the golem and the robot are notions deeply connected, expressing humanity's desire to generate artificial creatures (as shown for example in Contrada 1995). AI's increasing technical development has led to its recent transformation from a specific area of computing into the centre of attention for a great number of legislative (Edwards, Schäfer, and Harbinja 2020), philosophical (Vallor 2016), or journalistic (O'Connell 2017) discourses. Governmental documents are concerned with the future of AI (European Parliament 2017; House of Commons. Science and Technology Committee 2017), more recently fixed in more stable forms of policy (European Commission 2021). Various thinkers investigate the potential ethical and ontological questions raised by humanity's cohabitation with mechanical forms of advanced intelligence in a trend which followed the transformation of "information ethics" in computing or information science, to "roboethics," "data ethics," and "AI ethics" (Veruggio and Operto 2008; Hagendorff 2020). Public influential figures associated with science and technology make attention-capturing statements about the existential threats imposed by AI and intelligent robots (as in the cases of cosmologist Stephen Hawking and business magnate Elon Musk, Hern 2014; Cellan-Jones 2014), and a number of risk and futurist studies are dedicated to such potential implications (for example, Warwick 2000; Vinge 2008; Bostrom 2014; Ord 2020) with authors responsible for such studies being further associated with an emerging trend in establishing institutes for the study of existential risk, with AI as one of their primary foci[4]. An overarching pattern noticed when looking at this large orchestration of AI commentary from various disciplines in various domains of the everyday life (from newspapers to popular science and accessible philosophy of technology) is the impressive lack of voices from those who conduct the everyday tasks of technically advancing such AI technologies; let us call them technical AI specialists,

[3] Collins and Pinch, in their book, examine case studies relating to missiles, nuclear fuel flasks, the *Challenger* explosion, the Chernobyl accident, and AIDS, among others.

[4] A good example is Nick Bostrom who has been associated with (and initiated one of the) three well-known institutes for research about potential implications of AI: the Future of Humanity Institute (FHI, of which he is the founder and director), the Future of Life Institute (FLI, in which he is member of the Scientific Advisory Board), the Leverhulme Centre for the Future of Intelligence (CFI, in which he is member; it should be noted that in the opening ceremony of which Stephen Hawking gave a public warning about AI's catastrophic future), as well as the Centre for the Study of Existential Risk (CSER, member of the Scientific Advisory Board), according to the websites of all these institutes. (The very emergence of these institutes and centres is sociologically very telling of "future" becoming a new academic currency; I explore these issues *in tandem* with the evolving narratives about superintelligence, ultraintelligence, and singularity arguments in Galanos (2022a).)

well-versed in coding, programming, calibrating software, tinkering with interplays between algorithmic and robotic functions; for simplicity: AI specialists.

Like the cases studied by Collins and Pinch (see footnote 3), AI can be viewed as an example of how technological outcomes cannot be foreseen during preparatory experimentation: "distance lends enchantment," and positive and negative aspects (expectations or outcomes) of technologies are always context-based (Collins and Pinch 1998: 2-3). Several of the intermixed domains of AI non-practitioner commentaries are concerned with its potential future implications, which, as expressed below, appears loaded with polarised positive or negative connotations, giving the impression that technology bears an inherent ethically or unethically transformative nature. This resurgence of AI as a term, after several years of dormancy (more on its long history later), attracting governmental, intellectual, and media attention, is based on a number of undeniably important innovations related to AI methodologies which allow non-specialists to speculate based on extrapolations. These have already been implemented or are in the process of rapid development. Examples can be found in the laboratories but also in the everyday life, with AI becoming an umbrella term that is not a single, easily definable piece of technology. Aspects of it may include applied technologies such as decision-making algorithms, recommendation and matching systems, facial and speech recognition, autonomous systems such as navigation systems, robotic vehicles, weapons, and more; all this, seen as a whole, retaining much of the aforementioned hopeful/fearful golem religiosity or the godly redeemer/condemner. To quote Ian Bogost:

> "Here's an exercise: The next time you hear someone talking about algorithms, replace the term with "God" and ask yourself if the meaning changes. Our supposedly algorithmic culture is not a material phenomenon so much as a devotional one, a supplication made to the computers people have allowed to replace gods in their minds, even as they simultaneously claim that science has made us impervious to religion. […] Data has become just as theologized as algorithms, especially 'big data,' whose name is meant to elevate information to the level of celestial infinity." (Bogost 2015: n.p.).

I suggest that the same exercise would carry the same results by adding "AI," "machine learning," or "intelligent robotics." Reflecting on this passage by Bogost in conjunction to Lefebvre's everydayness, the high degree of theological construction of ideas associated with AI is responsible for a profound alienation from the world of AI technological production. In his novel *Gravity's Rainbow*, concerned with the V2 rockets in World War II, Thomas Pynchon uses his characters to brilliantly make a case for and against technology's deification by means of reference to the role of funding:

> "It means this War was never political at all, the politics was all theatre, all just to keep the people distracted . . . secretly, it was being dictated instead by the needs of technology . . . by a conspiracy between human beings and techniques, by something that needed the energy-burst of war, crying, "Money be damned, the very life of [insert name of Nation] is at stake," but meaning, most likely, *dawn is nearly here, I need my night's blood, my funding, funding, ahh more, more*. . . . The real crises were crises of allocation and priority, not among firms—it was only staged to look that way—but among the different Technologies, Plastics, Electronics, Aircraft, and their needs which are understood only by the ruling elite . . . Yes but Technology only responds '[…] Go ahead, capitalize the T on technology, deify it if it'll make you feel less responsible." (Pynchon 1973: 521).

This passage is used as a preview of what is going to be explored empirically in chapters 4 and 5, and the way in which deification of AI funding is possibly more crucial a question than deification of AI.

1.3 AI Technical State-of-the-Art and Challenges

The initial drafting of proposals for this PhD project happened at a time between 2015 and 2016, when AI as a term gained momentum in mass media, when it seemed that a promising technology with a long history was revived because of increased computational power and proliferation of available user-generated data, which enabled the use of algorithmic techniques to extract meaningful patterns which could be used *for predicting user behaviour*. By 2016, an array of different technologies gained popularity through the news, some of them being adopted as part of everyday technologies, allowing journalists and user minds to speculate about and *predict the futures of AI*. AI consists of many technical features, themselves treated as technologies. Separately examined, such "component technologies" (Williams, Stewart, and Slack 2005) are easily ruled out as non-AI systems (e.g. one can claim, "this is not AI, that's just an algorithm," often implying that AI would mean artificial human-level intelligence; in contradiction to everyday parlance about AI). Hence, what is "AI" about an algorithm, about large datasets, or an online recommendation system? Probably nothing, at least, without them acquiring meaning through their users and existence through their developers. The notion that a combination of all these aspects might be a golem-like sum which is greater than its parts, however, seems to make "AI" a keyword/buzzword that researchers, companies, journalists, and other interested parties might use to attract funds, sponsorship, justify PhD research, or simply receive some attention. To give the context of such trends and to describe the current state-of-the-art of AI without much technical jargon, one can think of applications in marketing, healthcare, public services, military, finance, law/civil rights, entertainment, transport, education, human resources/consultation, security, or robotic automation that might make use of chatbots/customer interaction/virtual agents, reuse of massively produced data, transcriptions/translations, predictive analytics based on pattern recognition and insights, emotion/facial recognition/detection, encryption, biometrics, peer-to-peer networking, and various problem solving methods for finding the shortest optimal solutions. Such successes in computer science often make the newspaper headlines, usually followed up by sensationalist speculations about the societal or even existential implications of AI[5]:

- Housekeeping assistants like Google Home or Amazon Alexa, computer assistants like Apple Siri and Microsoft Cortana, making everyday management of information, communication, and entertainment faster and more efficient (Nickinson, 2016). If these devices understand human language, will they develop emotions? If they remain emotionless, will they turn humans to mechanistic creatures too? If they are trained according to specific languages and dialects, will this enforce marginalisation of linguistic groups?
- Deep AI algorithms used for predicting user behaviour based on previous traits, allowing users to connect and find more relevant information based on their habits (social media such as Facebook, Novet, 2015; also online machine translation). If these algorithms have access and process all this online information, isn't it probable that a networked intelligent creature

[5] The list of AI applications stems from my own observations, however, Mitchell (2019) is a great introduction to current AI capabilities and methods for general audiences, written from an expert's perspective. Given that speculations are going to be discussed in further detail later, I will not provide references for the speculations; they can be treated as long-term observations from exposure to relevant sources which have become somewhat commonplace. Yet, two basic sources the reader can consult on the existentialist risk or social implications ends of speculations are Bostrom (2014) and Zuboff (2018), respectively.

will be created, surpassing the intelligence of human species? Will these huge networks of collected information transform humanity into a manipulatable mass?

- Machine learning robotics with applications in social and medical robotics assisting surgery or children education (like long-term object learning in iCub robot's embodied cognition, Keller & Lohan 2016; Metta et al 2008). If such applications exist now, isn't it probable that such creatures will develop into more sophisticated ones and that this might imply that they should acquire legal rights such as personhood? Are there implications for human psychology due to the heavy anthropomorphisation and zoomorphisation of robots in terms of either attachment to the machine, or the treatment of fellow human and nonhuman animals as machines due to accustomisation?
- Attempts at artificial general intelligence (AGI, that is, a general-purpose artificially created exact replica of an intelligent human), with robots entering classroom among human children in order to learn like humans (Goertzel et al 2010). Further to previous speculation, will this reduce the value of humans?
- Neural computing (Silver et al 2016) and brain simulations (like the European Union funded Blue Brain Project, Markram 2006; Markram et al 2015; Costandi 2015) as well as so-called self-aware robots (as the test described in Hooton 2015) helping with the understanding of human brain's functionality, thus potentially preventing problems associated with its development. If such projects of modelling the human brain exist, what if they are successful and implemented in projects like the above AGI experiments allowing the existence of fully mechanical, fully intelligent beings? And since they will benefit by the rapid calculation speeds of the computer, won't they surpass human intelligence quite easily?
- Working androids like Google Atlas (de Waard, Inja, & Visser, 2013). What if these "stupid" robots that are meant to assist workers in lifting heavy weights, become the physical support of human-level AI with access to the entirety of human knowledge on the Web? Who will be in control of such powerful robots and what kind of databases will they employ in order to improve their behaviour?
- Art-producing algorithms (Ghosh 2017). Firstly, does not the mechanisation of producing improbable artistic (say, musical or visual) pose a threat to the very concept of originality in creation? Who is the owner of what an algorithm created, if the artwork is very good – and who will be accountable if the artwork is considered offensive? Secondly, what if these "creative skills" are implemented in the above technologies?
- Military robots, such as bomb-disposal robots and lethal autonomous weapons including military drones and robotic soldiers (Singer 2009), robot nurses, as well as robot caregivers, robotic pets, and sex robots replacing human jobs that are considered to be precarious in various ways (with analytical descriptions of the technological innovations and their ethical implications to be found in Lin, Abney and Bekey 2012). If such technologies are widely adopted, will not humanity adapt to a cruel, dehumanised reality? Will not the military applications of automation result into a climate of fear similar to the one created by the existence of atomic weapons? And what if the aforementioned technologies are enhanced with this set of autonomous applications? And will an increasing amount of robotisation of such jobs become a stepping stone for a total robotisation of more types of jobs?

- AI applications are also used in medical prosthetics for the assistance of amputees but also for the enhancement of soldiers (as in the case of biomechatronics and dermoskeleton projects, Bedard 2011). Will this create an increasing hybridisation and confusion between what it means to be human and what it means to be a machine (as put by O'Connell 2017)? What implications will this have? And who will have access to these very expensive technologies?
- More recently, "generative AI" (or GenAI) applications for text and image generation, based on the employment of large language models (LLMs) and large annotated visual databases, such as GPT-3/ChatGPT and DALL-E (Pavlik 2023).

This list is by no means exhaustive, but is conducted in order to give a general impression as per what are some of the main applications of AI and robotics today and that they have varying degrees of success, affecting indeed a vast array of human activities, from military operations and health to entertainment and education, but at the same time giving rise to risk speculation (more technical and historical details on the evolution of AI as a concept, below). Is the speculation justified? The answer is not only technological, but also historical and societal. Hagendorff & Wezel (2019) have published a list of challenges whose interplay is relatively obscured by the AI hype[6].

Methodological	**Societal**	**Technological**
The data AI systems use do not correspond with reality	AI practitioners need knowledge about technological consequences	Human thinking is very different from intelligent machines
AI based on machine learning only perpetuates the past	Values embedded in AI technologies have to be reflected on	AI applications often lack explainability
	Organisations need to address the lack of diversity in AI research and industry	Many learning algorithms are highly inflexible in their functionality
	Technological capabilities are limited by the scarcity of talented programmers	Labels are a scarce resource, but also a precondition for many AI systems
	The success of AI applications is tied to their acceptability in society	AI systems struggle with the extraordinary
	AI systems do not work without producing hidden costs	Building secure AI applications is nearly impossible
	AI systems rely on instable infrastructures and material prerequisites	

Table 1 Existing challenges in AI extracted by Hagendorff & Wezel (2019), adapted in form of a table by me.

[6] A much more thorough (and cited) outline of opportunities, challenges, and research agenda setting for AI techniques especially in relation to policy can be found in Dwivedi et al (2019), wherein I have contributed with STS perspectives during early stages of analysis of the present research.

This table is indicative of numerous directions contemporary AI research has taken. While the authors do not present this categorisation in the form of a table, to view this typology as such reveals the relative disparity between social research on AI and its technical foundation. AI policy, although not within the empirical scope of this thesis, is an emerging social world which exerts influence on the development of AI in response to questions concerning out-of-control scenarios and undesired outcomes (outlined in chapter 3).

In October 2016 three documents published respectively by the United Kingdom (House of Commons. Science and Technology Committee 2016), the European Union (EU) (European Parliament. Directorate-General for Internal Policies. Policy Department C 2016; following the draft report of European Parliament. Committee on Legal Affairs 2016), and the United States of America (Executive Office for the President. National Science and Technology Council Committee on Technology (2016)) have all called for close inspection of the ethical and legal implications of AI and robotics, with an emphasis on the protection of human rights from becoming manipulated by robots, AI's relation to legal personhood, and more issues relating to AI-out-of-control scenarios[7]. As Cath et al notice, the three reports have probably been prepared independently from each other, and hence – according to their opinion – highlight the effect of an ongoing resurgence in the field of AI during the last years (Cath et al 2017: 2), leaving one, however, with the suspicious feeling of one or more hidden variables (in which case, one might consider the impact of serendipity as exemplified in the above chapter). The differences of approach between the three policy documents showed national-level loci of interest: while the US aimed chiefly at close collaboration between research and private industries, the EU was focusing on the early adoption of regulatory frameworks based on the formation of relevant boards and advisory committees, whereas the UK suggested coordination across policy regulation and industrial development. A surprising common finding in comparing the three documents was the minimal consultation of AI specialists in drafting them, while there was evidential influence by prestigious public commentators with little or no expertise in AI, about the potential existential harms of AI (Galanos 2019).

Currently, new policy documents about AI seem to reach a level of maturity, in that more experts are consulted prior to their publication, from the formation of the EU High-Level Expert Group on Artificial Intelligence (2019) and the recent Artificial Intelligence Act (European Commission 2021) to the UK House of Lords' Select Committee on Artificial Intelligence (2018). This is only indicative of a much broader construction of an AI policy landscape. Indeed, while proponents of the Singularity theory would envisage an exponential growth of AI's humanlike capabilities, what actually grew exponentially was the number of AI policy documents after 2017. Canada was the first country to publish a national AI strategy in 2017, followed by 30 countries by December 2020. National strategies are vital, yet covered only a

[7] See also the very similar suggestions given some months later in a "report written for the Committee on Culture, Science, Education and Media of the Parliamentary Assembly of the Council of Europe" by the Rathenau Instituut, concerned with – as the title of the report reads – *Human Rights in the Robot Age: Challenges Arising from the Use of Robotics, Artificial Intelligence, and Virtual and Augmented Reality* (Van Est, Gerrutsen, & Kool 2017). This report been a follow-up to the European Commission's voting after the demand of Members of the European Parliament (MEPs) for clarification of liability issues with regard to robotic agency in the long run, pointing out the need for regulatory standards in case of damage which took place on February, 16, 2017, (European Parliament 2017), a milestone event for the manifestation of the general attention drawn by AI and robotics. This bears interest as to the history of terminology as well. Whereas researchers outside the core AI development were concerned, roughly until the mid-2010s, with robot ethics and robot rights (e.g. Lin et al 2012, Tzafestas 2016), such discourses are being largely displaced currently by "AI ethics" or AI regulation, closer to the necessary algorithmic components of the technology, rather than its potential physical support.

small fraction of various AI-related policy documents which have been commissioned in the second half of the 2010s and early 2020s (a good overview with links to most documents can be found on the AI Index Report 2021; Zhang et al 2021). This urged the international and intergovernmental initiatives such as the Organisation for Economic Co-operation and Development (OECD) to establish institutional tools such as the AI Policy Observatory (Zhang et al 2021: 165), or, in the case of Algorithm Watch, the AI Ethics Guidelines Global Inventory, being a useful search engine, although excluding final legislations (Algorithm Watch 2021). It should be noted that the European Commission's 2021 Act is, globally, the first draft attempt at a comprehensive, fully-developed legal framework for regulating AI (with subsequent financial penalties for non-compliance) and it appears to be the fruit of a now crystallised perception that AI, in the form of a family of software and techniques, enters the market and has to be regulated; thus, it separates between four categories of risk: unacceptable, high, low, and minimal (European Commission 2021). Nevertheless, things become fuzzier as to when does AI policy begin, if one takes into account the occasional interchangeability between "AI" and "robotics." Indeed, Boden et al (2010; one of the principal refreshers of AI's conceptualisation in the 1980s) have published UK's *Principles of Robotics: Regulating Robots in the Real World*, being, according to co-author Joanna Bryson's website "the first national-level AI ethics policy[8]." More recently, legal scholar Frank Pasquale's book *New Laws of Robotics* blends the two concepts and aims to "warn policymakers away from framing controversies in AI and robotics as part of a blandly general 'technology policy,' and toward deep engagement with domain experts charged with protecting important values in well-established fields" (Pasquale 2020: 15-16[9]).

In 2016, this thesis began with the observation that not too many practical AI experts have been involved in questions asked by humanities scholars and discussions about AI held at a policy level. It appeared as if contemporary AI parlance suffered from the "two cultures" problem expressed by C.P. Snow (1959) on the detrimental effects for solving world challenges stemming from a science/humanities division. The two sides simply do not communicate sufficiently. In the case of AI, when I entered the field, this problem took the following form (with more specific details to be discussed throughout the thesis):

The policymaker's expectations and conceptualisations of AI may not align with the practical, technical, scientific, exploratory, or funding-oriented perceptions of the AI specialist; and few AI specialists understand or care about the needs and rationale of the policymakers. When policymakers (or journalists) use their words for an AI-related technology, they have in mind a specific view of an *either controllable or uncontrollable*, artefact or science of a particular sort; when AI specialists use their flexible, specialisation- and agenda-dependent terminologies about AI-related matters (not technology nor science), they have in mind an approach, a framework, a general area in computing of some sort. And if, as occasionally happens, the two social groups partake remotely in decision-making about funding allocation or regulatory strategies, in processes of advisory boards under bureaucratic slowness and deadline speed, they will not have the time or capacity to place themselves on each other's viewpoint. Thus, the nearly unbreakable glass of alienating spectacle/hype, strengthened by long-term fears and hopes stemming from

[8] https://www.hertie-school.org/en/who-we-are/profile/person/bryson/

[9] This is further telling of a broader rebranding of terminologies to satisfy research trends: questions of information ethics from the 1990s mixed with questions of roboethics in the 2000s and early 2010s, to become the now entrenched blend of "AI ethics." Due to the present thesis' chief focus on "AI" and not "robotics," I have not mentioned in the introduction the influence of science fiction writer Isaac Asimov's infamous "laws of robotics" upon roboticists which has led, during eras of "roboethics" fashion in the 2000s to backlash on behalf or influential roboticists who called already from 2009 for laws of "responsible robotics" (Murphy and Woods 2009). Two decades before that, James Fleck (1984) outlined in detail the impact of Asimov's fiction in the development of robotics and utopian thought in general.

mythology, religion, and science fiction as well as from journalistic mediation and influential non-experts, will ensure that mistakes and confusion will occur from their discussions[10]. The present thesis reports on these early stages of the debate, when expectations and conceptualisations, through their multiple pairings and mismatches established AI as a technology-in-itself, to be governed and funded, instead of a science which envisions and experiments.

In trying to answer questions about the role of expectations in AI research proved to be a complex dynamic of multiple factors. Historical analysis of AI development and empirical engagement with specialists added layers of complexity to the questions and revealed new dilemmas as expressed in the chapters 4 to 6 below: while specialists flag out the potential harms of overpromising and the influence of broader discourse expectations, they also admit to their own overpromising as means to secure funding and how generalised hype acts as a beneficial tool for career opportunities (and opportunistic strategies). Similarly, while they admit the lack of specialist involvement in policymaking regulating their own work, at the same time they show little will to shape such decisions within policy context[11]. This had a crucial implication in the establishment of regulated AI as a technical feature instead of a scientific discipline in the years following my fieldwork, which was held during a period of great uncertainty as to the future but also the very identity of AI. I realised, while interviewing both early career researchers in the field together with "AI veterans" and their in-between generation, that such questions about the vast expectational environment of AI are better understood if situated within the historical timeline which allowed the field to develop. My empirical research captures a relatively recent snapshot of a turning point in AI, what I simply describe here (and further explicate below), as the transformation of AI from science into technology, based on different individual, collective, historical, and regional expectations.

1.4 The Question Concerning Hype and the Fear of AI Winters

Given the social distance[12] between the two aforementioned perspectives (plenty of non-specialist commentators, few AI scientists advocating their knowledge), I focused on the latter option. When proposing this project, I noticed a general lack of empirical work on the topic, and most importantly empirical investigations of AI specialists' views. As expected, however, by the time of submission more empirical investigations appeared, and I have met colleagues at conferences and other academic venues who followed similar approaches. Hence, although this project started initially as an exploratory one about specialist views on futures of AI, it is now narrowed down to questions regarding the interplay between

[10] Legal expert Frank Pasquale identifies a similar problem of sociological distancing between "global technology company bureaucrats" and the "real world consequences of their algorithms" (Pasquale 2020: 116). Pasquale further suggests that what he terms "alien intelligence" (instead of "artificial"), not only alienates one relevant social group from another, but is also a form of machine-based intelligent operation alien to forms of cognition recognizable by humans, thus, often imperceptible in everyday settings. With reference to Jaeggi's "relation of relationlessness," Pasquale is optimistic that with sufficient moderation and education, users should be able to tell "whether tweets and videos represent authentic content or are merely a confected spectacle" (Pasquale 2020: 117-119). While the present thesis focuses on observing and reporting rather than normatively recommending, his reference to alienation and spectacle are acknowledged and indeed interrogated, especially on chapter 7, and the relationship between policymakers and AI practitioners; the post-Marxian concept of "spectacle" (Debord 1988) is employed as an analytical tool in order to explain certain expectation dynamics.

[11] Once again, this statement stems from analysis initially conducted for the purposes of this thesis, however, lack of space does not allow for its elaboration. Open access manuscript under review can be found at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4213120

[12] By social distance I do not intend to make a pun related to the COVID-19 outbreak. Indeed, I have been using this phrase several years before the instalment of social distancing measures.

research practice and the broader arenas of expectations shaping it, as this is profoundly underpinned by questions of AI history and different layers of social rhetoric. This rhetoric, positive or negative, from now on I will denote as hype. Hype is yet another difficult term to define due to its very qualitative meaning. Following media theory and analysis, hype can be seen as "amplification, magnification, exaggeration, and distortion" of existing knowledge, technological or not, usually expressed on mass media (Vasterman 2005: 511). Researchers have been investigating the role of media hype in policy debates; for example, Fijalkow (2013: 54) uses the term unidentified media-hype object (UMO) to describe dyslexia as it appears in public debate, distorting governmental policies of literacy. In the present context, I prefer to speak about *identified* objects of hype[13] of technological interest. Interested in the AI hype, it is often difficult to distinguish between hyped objects such as the internet, ICTs, virtual reality, augmented reality, cyborg technologies, algorithms, big data, quantum computing, and so on. In her recent book about hype, Milne (2020) describes it as a combination of intensive promotion, exaggerated publicity with "bias towards a particular perception," "because of the particular *words* and *narratives* surrounding it" (Milne 2020: x, original emphasis). Milne treats hype as a technical tool, being a "double-edged sword" which can generate excitement towards innovation or technical decision-making, having, however, several limitations, such as the perpetuation of existing social issues being obscured by hype, the shielding of complexity in social processes, chilling effects diminishing action and responsibility, and the "fanatical" takeover of emotional desire and bias against social needs (Milne 2020: 278[14]). To understand what researchers make out of such forms of hype, how they are positively or negatively impacted by the hype, how they assist its generation and transmission will be revealing, in the chapters below, about the integral role of hype in technological development, assessment, definitional negotiation, and governance. Following Van Lente and colleagues' comparison across different hype types, and their conceptualisation of it "as a collectively shared rhetoric about an emerging technology and the underlying innovative activities" (Van Lente et al 2013: 1615), I suggest that hype can be seen and should be viewed as a component of technological trajectories. Through the empirical chapters below, hype's role appears to be double, harmful in some cases, beneficial in some others: on the one hand, the desired outcome might not necessarily follow immediately and there may be loss in trust towards the hype object; on the other, as shown below, in some cases, the desired outcome is reached in the long-term, however, early negative assessment of unfulfilled promises hides the "something came out of it" effect, which, if perceived historically, justifies the role of hype in history.

In previous understandings of hype (such as Van Lente et al 2013), hype cycles tend to complete in moments of disillusionment or tentative stabilisation of knowledge about a technology's capabilities. Moving to AI in particular, a central question triggering the present exploration of contemporary AI hype is based on the assumption that moments of hype are likely to be followed by disillusionment, with the potential for AI, however, to re-emerge as a hype object after a certain period. In the domain of AI, this is known as the interchange between "AI summers" and "AI winters" (Hendler 2008), following the "nuclear winter" metaphor, suggesting stagnation of funding which follows loss in terms of trust towards

[13] Hype objects not to be confused with Morton's hyperobjects (2016).

[14] In the same passages, Milne juxtaposes hype to "facts" and "rational thinking." While I am appreciative of the argument, I am not in support of this phraseology given precisely that hype does shape and has been shaping forms of existing rationality or perceived facticities. Milne also devotes three of her book's chapters to a critical reading of AI-related technological hype, following the approach of hype-reality separation (what hype suggests against what science and technology may achieve). While this is a very useful tool towards responsible approaches to AI, it still leaves unanswered the question concerning hype's generation and its role in research practice.

overpromising research communities. The pattern is indeed impressive. During the course of this doctoral research, I prepared a comparison between two books popularising the concept of machine intelligence, one published in 1955, concerned with intelligent robots, one published in 2019, concerned with AI (Galanos 2022b). The repetition of themes was astoundingly similar. But I treat this as a provocation precisely about the longitudinal re-emergence of AI hype. As I conclude, "I do not want to make another claim about history repeating itself and the 'wow' effect of hype-and-disillusionment cycles – belief in a purely circular history is as reductionist as the belief in the modernist notion of linear progress and innovation" (Galanos 2022b: 83). Moments of AI hype, in this context, will be treated as opportunities to study the complexities of promises, expectations, and motives that lie underneath.

The technical applications outlined above, and their corresponding sets of hopes and fears, are not "technological breakthroughs," as they often presented, but are the outcomes of a long history in computing and adjacent fields. AI's "tumultuous" history (Crevier 1993) is one of identity negotiation built on grand promises and societal expectations, which often leads to disillusionment based on the unrealisability of such expectations. AI has a long history, imbued with conceptual shifts, negotiations, promises, disillusionments, technical achievements, failures, transformation of focus, masking under different terminologies, reinvention, and rounds of renaissance. According to conflicting parts of the literature, at least two "winters" have taken place in AI's history; roughly, one between the early 1970s and early 1980s, and one in the early 1990s until the mid-2000s. Some authors consider an early event marked by machine translation failure and subsequent report (the Automatic Language Processing Advisory Committee/ALPAC report) to be an earlier AI winter in 1966 (Hendler 2008; Grudin 2008; for a thorough outline of this, see chapter 3). For now, it should be mentioned that this lurking possibility for a new AI winter has guided much of this thesis, together with a corresponding concerning expertise: who is credible and responsible for a trustworthy expectation/promissory setting? This thesis also challenges the very notion of the AI winter. I argue that the question concerning whether a new AI winter is coming, or how to avoid it is imbued with deterministic and self-fulfilling undertones ("new" AI winters were expected to come in the past (Waldrop, 1984; Hendler, 2008) as they are now (Shead, 2020)). As I will evidence through the historical outline below, the story about a reoccurring AI winter does not take into account the complex branching of computing technologies (and hypes associated with them) and that while AI promises are not met, other applications spring as relatively unintended outcomes of experimentation, with certain promises to become fulfilled at different settings, at different times. Thus, one of my observations, following my initial exploration of the "AI winter" as a guiding concept, is that its deep association with high risk, makes it a necessary component of the hype itself and has to be studied in close association with it. Moreover, AI winters as a point of departure allowed the revealing of other, institutional processes between formality and informality, relating to funding strategies. The table at the end of this section aims at summarising, with some necessary simplification, what can be viewed as the three rounds of AI hype based on technical applications triggering the hypes, reasons for perceived failure (despite the incremental continuation of research under different brandings) and the ways AI itself was conceived during these eras; a detailed account of those will be discussed in Chapter 3. While my initial intention was to place this table as a summary of Chapter 3's findings, where all terminologies are being explored conceptually and historically, I think it deserves a place in this introductory chapter as a general map and preview of what is to follow. No prior technical/terminological knowledge is required, and the reader is invited to use this table as an orientation tool when reference to specific technical applications is being made in subsequent chapters.

The question concerning AI expectations, expertise, and AI winters is not a theoretical one, although the findings do contribute to theorisation about STS. Technologies that carry sets of intense promises and high expectations (for example, nanotechnology, CO2 emissions, PVC, synthetic biology, and, alas, AI) invite for central regulation by governmental bodies (Collingridge 1980; Konrad and Böhle 2019). The importance and relevance of the topic has practical implications and this is largely exemplified by the post-2015 emergence of an AI policy race (see section 1.3) which generated a further set of linked social science questions such as the following. When is the right time to implement technology policy and regulation? Will regulation obstruct research if imposed too early? Will harmful effects of the technology be impossible to escape if policy is implemented with delay? (more on this dilemma in section 7.1). Who shall be consulted in specialist advisory committees? Do policy experts know enough about technology? Do technologists know enough about policy? Such questions are to be explored in future work (currently in the making), but are indicative of hype's extant performativity potential. The nefarious constellation between expectations, hype, and promises is an integral component of technological innovation and has to be studied carefully. To quote Bakker and Budde:

> "We have shown that technological hypes are potentially powerful phenomena that can trigger actors to engage in an innovation race instead of continuing their waiting game. Hypes can attract actors, funding and favourable regulations (and other institutions) that would otherwise not be attracted. However, hypes are also difficult, if not impossible, to control and expectations are likely to become overly optimistic and subsequent disappointment can cause a standstill once the hype is over." (Bakker and Budde 2012: 550).

To conclude, the question concerning hype will be treated here as the product of formal and informal interplay between expectation and expertise, and shared or contested knowledge about them. This will be argued extensively in the theoretical discussion of section 2.2. To give an illustrative example of hype, I would like to offer the following anecdote, revealing of the multiple layers of actors, promises, and exaggerations involved, and how this influences research practice.

1.4.1 An Anecdote of Hype: The Case of Margiotta

This story was narrated to me by two AI/robotics researchers working at Edinburgh's Heriot-Watt University during independent conversations about the impact of media in the public representation of their fields. The University's Interaction Lab aimed at testing their programming of the humanoid social robot Pepper, a robot that is able to recognise and, to an extent, respond to human facial expressions and emotions. The researchers programmed the robot to be a "shopbot," that is, to assist customers of supermarkets, and hence, they agreed with the local supermarket Margiotta, that the robot would run a pilot performance, so that the researchers receive feedback to advance the robot. After a week, the shop owners were dismayed by the robot's performance that, despite its algorithmic training over thousands of items, was incapable to engage in conversation with the customers due to background noise, further exhibiting minimal mobility thus appearing discomforting to many customers. Although the initial reaction of customers was enthusiastic when they saw the cute anthropomorphic robot, the functional disillusionment caused sufficient discomfort, which led the shop owners to request that the University experimenters collect the robot. For the experimenters, the robot's failure offered numerous lessons about its future optimisation. However, for the media, it was an opportunity to take advantage of, and perpetuate, hype. The initial coverage of the story by a local newspaper was on January 21, 2018 (Herald Scotland, "First robot shop assistant tested at Scottish supermarket"). The article is a balanced report on the robot's

performance, offering reflections by the shop owners and the roboticists. The title mentions nothing about the robot's withdrawal. The next day, January 22, 2018, however, news travelled to further newspapers and a narrative of a robot which was being "fired," "sacked," or "scaring customers" emerged and diffused within few hours (all times shown in 24-hour format, GMT): 07:30: Scottish Business Insider, "Robot 'hired' by Edinburgh supermarket then 'fired' after a week." 10:39 (updated 11:16, January 23, 2018): The Mirror, "Robot working in a supermarket FIRED after a week as it scares human customers." 12:00: The Times, "Store fires robot server Fabio for pushing shoppers' buttons." 12:54: Edinburgh News, "Margiotta shoppers say farewell to first robot worker." 14:00: The Telegraph, "Fabio the robot sacked from supermarket after alarming customers." 14:18: The Guardian Opinion, "If we're spooked by a 'shopbot', we're definitely not ready for NHS droids." By the end of the month, the news story reached the online-based pan-African new outlet on YouTube: January 31, 2018: Africa News, "Robot shop assistant sacked."

The people who have informed me about this (whose identity is protected), referred to the hurdles caused by this event which was followed by negative communication with the robot manufacturing company which complained about the bad publicity. Nevertheless, and my informants agreed with me on this, such a case was enabled at that time because of a broader emerging narrative which saw the emerging robots as capable of having rights and thus be "hired" and "fired." Evidently, this took place three months after the controversial (often characterised as a publicity stunt) gratification of citizenship rights to Sophia the Robot on October 11, 2017, "with critics wondering why a humanoid robot received citizenship while women and foreign workers in the country have less rights, and many humans are practically stateless" (Parviainen and Coeckelbergh 2020: n.p.). Below is a collection of screenshots depicting some of the news coverage from "milder" to more "hyped" versions.

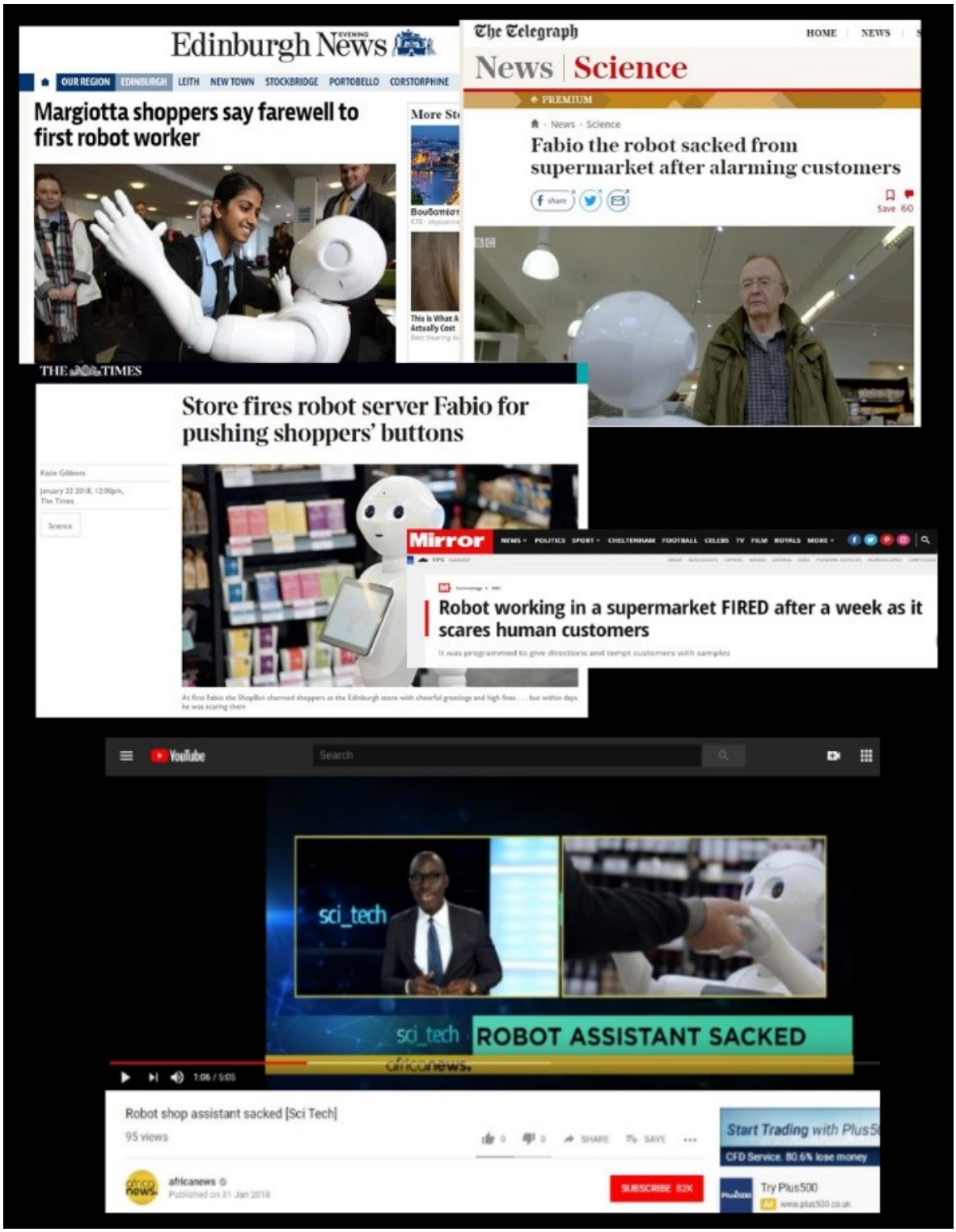


Figure 1 Collage of screenshots depicting headlines about the "Margiotta case."

AI as a moving target	**Round 1: Beginnings (1950-1980)**	**Round 2: International AI race (1980-2010)**	**Round 3: AI of the latter days (2010-present)**
Technical innovations	• Logic theorist (heuristics) • Perceptrons (early connectionism/pattern recognition) Game playing	• Microworlds • Frames • Reinforcement learning • LISP language • Expert systems • Microelectronics • Very large scale infrastructural architecture design • Backpropagation algorithms • Convolutional networks • TCP/IP • Fuzzy sets • Brain-style AI and bioinspired robotics	• Deep learning pattern recognition • Deep neural networks • Large databases/big data • Moving object recognition • Recommender systems • Natural language processing • Generative AI for text, audio, still and moving image
Reasons of "failure"	• Academic discord • Combinatorial explosion in rule-based systems/contextual reasoning • Overpromising/ad hoc evaluations by non-specialists	• Integration failure • "Empty shell" expert systems couldn't be built • "AI effect" – rebranding/spinoffs/branching • Organisational change, user dissatisfaction • Ad hoc evaluations	• AI too vague • Policy regulation • Data lifecycle, ownership/copyright, privacy • Data model biases, transferability limitations, saturation of appications
Perceptions of AI	• AI as imitation of brain processes. • AI as a formalisation of thought. • AI as an attempt at creating robotic intelligent beings. • AI as constant future.	• AI as a patchwork of different disciplines, still negotiating its identity. mirroring its practitioners' motivation. • AI as a set of different methodologies focusing on simulation or programming of intelligence. • GOFAI/connectionist AI • AI as an applications-oriented technical field.	• AI is its applications, an assistive technology. • Separation of AI as a field and as a series of applications. • AI as an agent of social change, risky or beneficial.

1.5 Empirical STS Work on AI

Based on the above, I hereby offer a short literature review of existing STS works on AI that are relevant to and have shaped the present work, and I contribute to by adding further nuances to by the historical and everyday practice explorations. I situate this work as part of a lineage of STS which has been observing AI for several decades. David Ribes (2019) has recently made a case about STS being the observer and contributor to data science (associated to cybernetics and AI), for a long period, to the extent that these domains are hard to be separated. Hence, we cannot deny the role played by STS in the shaping of such AI/data/cybernetic technologies; indeed, STS scholars such as Sherry Turkle or Lucy Suchman have actively shaped perceptions and decisions about AI at a political or design level (Suchman 1984, 2007; Suchman and Trigg 2003; Turkle 1981, 1984; Lachney and Foster 2020). The following STS (or peripheral to STS) contributions which shaped the rationale of this paper are also revealing that social studies of AI involve a continuity of factors and domains, as suggested through the above introduction of hype in AI. Although the following works do not shape exactly the theoretical/methodological framework of the thesis (explicated in chapter 2), I acknowledge them because their suggestions have shaped interpretations of my findings. Works by Fleck (1979; 1982; 1984; 1988; 1992; 1993) have set up the early agenda for the development and establishment of AI as a scientific field, offering valuable empirical, historical, and sociological insights from the field's history in the UK and the US, as well as early analyses of the interplay between science fiction and practical science in AI. Sherry Turkle's early work (1981; 1984, also: Graubard; 1988) shows the psychoanalytical implication of computers as applications AI, suggesting that such technologies can be viewed "as Rorschach," projection screens of their users, developers, or other interested parties' hopes and fears (and therefore shape their definitions of such technologies). Paul Edwards (1996) has outlined and emphasised the military underpinnings of the early foundations of AI in close development with the "cyborg discourse" and enthusiasm, allowing the emergence of a "closed world" of information technologies[15]. More STS contributions to AI include those of: Winner (1977) being an early critic of uncritical belief in automation, and in particular, Marvin Minsky's suggestion that black-boxed AI systems should not be interrogated, Suchman (1984; 2007) who has been an active member of the Computer Professionals for Social Responsibility global network, offering insights from social science perspectives in the 1980s as to the use of AI in military contexts, continuing to research the subject as social-material "configurations" between humans and machines, Collins (1988; 1991; 2018) interested in the social negotiations of expertise left to machines, actively testing machine expertise to showcase limitations and create models of varying types of expertise, and Woolgar (1985; 1987) commenting on the importance of including AI as an established area of research for STS scholarship and, moreover, its treatment as an "actor" within sociotechnical registers. More recently, and occasioned by the novel AI boom, an increasing amount of new STS (and neighbouring) bodies of literature on AI examine the overlapping areas of its history (Agar 2020; Penn 2020), economics (Naudé 2020), research practice (Campion et al 2020), policy (Ulnicane et al 2020; Dexe and Franke, 2020; Stix and Maus, 2021), ethics (Kerr et al 2020), ideology (Katz 2017; Campolo and Crawford 2020), and media representations (Galanos 2019; Brennen et al 2020). Out of these sources, it was Woolgar (1985: 557), who reflected on STS's role

[15] The significant role played by the military in funding AI research will be described in the historical outline below. However, the deep sociological implications of this relationship in the context of contemporary AI renaissance is understudied and mostly needs to be rewritten from the point left in Edwards' book. Another important work looking at the deep implications of military AI is De Landa's 1991 *War in the Age of Intelligent Machines*, also largely outdated.

in to exploring AI sociologically, which I revisit according to the recent hype – surprisingly accurate today, with the exception, perhaps, of the amount of STS scholars investigating AI. Woolgar flagged the co-evolution of growing interest of governments, companies, organisations, and universities for social science contributions to AI, as part of an equally increasing competition between countries (or continents). In that paper's time, Woolgar suggested that this "AI phenomenon" has to be studied more by STS scholars as it is left to philosophers, political scientists, and media scholars, on the side of humanities; while this is not true anymore (as shown above), the current level of maturity in AI/STS debates allows for more synthetic views and dialogue across the different specialisations, focuses, and empirical works of every author.

Such attempts can be found in certain works of Harry Collins (1991, 2018) who was previously encountered as key contributor to the studies of expertise and experience). His 2018 work appears as an extension of the "AI phenomenon" studies, although he speaks of the "AI belief" in his call to STS scholars to maintain disenchantment while carefully assessing the credibility of AI-related assertions:

> "AI belief has a quasi-religious (or counter-religious), ideological element which turns on seeing humans as machines, not just something to be mimicked by machines. […] The AI community has experienced wave after wave of 'hype' when one kind of innovation or another has promised to solve the problem of human intelligence […] Many AI scientists fear that yet another 'AI winter' will follow the latest deep-learning bubble. The large majority of AI scientists want to get on with building devices that work better, and will help humans run their day-to-day lives better, rather than take over the world or prove that humans are merely machines. Indeed, the strongest AI belief seems to come more from philosophers, evolutionary biologists or other outsiders, suffering from the web of enchantment that distance from the frontiers of the technology can weave, and sure that humans can be no more than organic machines designed by the 'blind watchmaker'." (Collins 2018: 26).

Collins further adds that the public receives an image of AI which is distorted from that of the AI lab, calling it "a world of *artifictional* intelligence, available through the newspapers, books and films." (Collins 2018: 13, original emphasis). While the STS literature on AI is growing, the aforementioned sources are indicative of the general pathways the field has been following.

1.6 Chapter Outline

The thesis unfolds in the following sequence:

Chapter 1 has been setting up the stage for the research. Thus far, it explained my motivation to conduct this piece of work, offering an overview of what is meant by "AI" in contemporary parlance. The policy-level relevance of AI, the journalistic hype surrounding it, paired to theological undertones, and the central questions concerning the role of hype and troughs of disillusionment (winters) in AI history have been outlined. The work has been further situated within a growing genealogy of STS works dealing with AI and related disciplines. Some limitations of the present work have been acknowledged, highlighting the relevance of the present work for forthcoming research.

Chapter 2 is dedicated to the tools employed to answer this question, theoretical and methodological. Science and Technology Studies (STS) frameworks related to future-oriented behaviour and sociology are presented, as well as certain vocabularies assisting navigation in the complex interplay of experts and other

interest groups. The theoretical nuances between statements, promises, imaginaries, narratives, and expectations, as well as experts, enactors, selectors, arenas, and exploration-exploitation trade-offs are discussed and outlined here. Certain focus is placed on how this thesis contributes to analysis future-oriented debates through the examination of expertise. This is followed by a discussion on the practicalities of method, justifying the employment of the interviewing method, assisted by critical historical interpretations, and the emerging approach of "scavenging ethnography" (Seaver 2017). Processes of obtaining access, sample criteria, obstacles, considerations on the anonymisation of participants and the present research's limitations are presented here.

Chapter 3 employs the theoretical approaches previously outlined to reflect on the role of expectations, experts, and discourses in the historical shaping of AI. A core argument about understanding AI's expectational environment and the construction of AI hype (which can also be termed "AI craze" or "AI hysteria") is to view it historically, as adjacent to political, social, and technological developments. It proceeds with a close historical examination of the development of AI technologies by looking at how AI's meaning and its associated promises developed and have been negotiated not only by those with technical expertise, but through an increasingly growing (and thus, I suggest, unmanageable) ensemble of actors: scientists, politicians and policymakers, humanities scholars, and journalists. The observed disproportion of AI scientists shaping AI's meaning and purpose lays the foundation for sharpening specific directions as to the impact of vastly hyped environments on basic AI research – supposedly, where it all is expected to stem from. Based on critical literature analysis and document excavation, this chapter aims to offer a clear understanding of the historical context through which contemporary debates emerged. This historical awareness will allow deeper reflections and assessment of the empirical findings and adds the element of historicity to the theoretical approach of expectations/promises assessment. The chapter is divided into two main sections: one focusing on the historical development of AI descriptions; another on the development of AI promises/expectations.

Chapter 4 is the first presentation of empirical data, looking at how AI specialists define (or refuse to define) AI, and thus speaking empirically to Chapter 2's first section. The verification of previously reported lack of consensus about AI's definition and its historical shaping branching into multiple approaches, is assisted by the division of researchers into those who support the need of descriptions and who actively advocate specific terminologies, and those who abstain from definitions for various reasons, such as the impossibility to define AI or the general impracticality of definitions. This allows a useful entry point into the following empirical explorations. AI's official history of rise-and-fall and its guising of the continuous incremental innovation of specific component subfields of AI offers a springboard to assess the promissory environment surrounding its process.

Chapter 5 delves deeper into the role and impact of promises and expectations in AI research. How much do AI specialists know or care about historical developments of AI promises? Do they feel impacted by this history? Do they themselves employ promising in their research? The findings highlight the long-term impact of non-practitioner assessment of AI (chiefly in the form of governmental evaluation reports, but also broader shaping through popular understandings of AI) and projections on behalf of AI communities. Because of previous rounds of disillusionment, AI communities now largely abstain from projecting into the future; indeed, they also abstain from advocating their territory in terms of regulation policy. Given that AI specialists have been alienated from the active shaping of their field's aims, they are left with navigating the expectational landscape by developing funding strategies.

Chapter 6, then, explores the role of promising in contemporary AI funding strategies. An illustrative example examining the roles of serendipity and informal negotiations in shifting national funding policy is presented and theorised in the context of science in the service of the State and technosocial imaginaries. Further views on the topic are explored and AI specialists are found to fall within a spectrum between scientific exploration (promoters of basic research, largely unpopular and thus unfunded) and fund exploitation (opportunist researchers who make use of available resources, tailoring their interests in response to existing funding calls). The open question is posed on whether (a) funds-follow-fashion, that is, do informal and formal promissory competitions on what is considered to be technical state-of-the-art shape national funding schemes; or (b) fashion-following-funds, that is, researchers predominantly adjusting to specific national strategies, implying a hierarchy on which expert is considered best to shape the latter, and thus, the field. While the answer entails a co-construction of the two tendencies, this invites for a closer inspection of the broader AI policy landscape.

Chapter 7 summarises, discusses, and synthesises the above findings. The summary is presented as a response to the research questions; the role of non-practitioner expectations in shaping core technical promises, the interplay between formal and informal promissory negotiations, and the significance of historical awareness in assessing the relevance of expectations and expert legitimacy. AI's broader expectational landscape forms a dynamic of tension with AI scientists' practices. Production of everyday academic AI research is shaped by a long history and negotiations of AI's purpose, in turn shaped by governmental industrial, military, and commercial demands. AI transformed from an intellectual, scientific investigation about understanding intelligence by replicating it, to a practice-based, technological endeavour based on pattern recognition applications and mundane solutions. The role of assessment, either by non-AI specialists in past decades' formal evaluations or by informal prestigious commentators in the post-2010 AI resurgence is persistently harmful, although there is no way to tell who should have the right to be an AI interrogator or not (for example, when advising policy), precisely due to the field's initially intended wide scope. Negotiations of different types of futures are crucial in this mechanism of expertise interplay between scientists, policymakers, and public discourses. An initially unintended finding based on the thesis' findings is a discussion of the properties of hype – a relatively underexplored theme in the sociologies of expectation and promise. This discussion takes the form of three core principles of technological hype. This is followed by an afterword suggesting a future STS-AI alliance.

Past the **bibliography**, the **Appendix** section serves as a collection of technical documents used for conducting interviews (consent form, invitation letter, and interview schedule), including a list of abbreviations.

CHAPTER 2: A HISTORICAL SOCIOLOGY OF EXPECTATIONS AND EXPERTISE - RESEARCH QUESTIONS, THEORY, AND METHOD

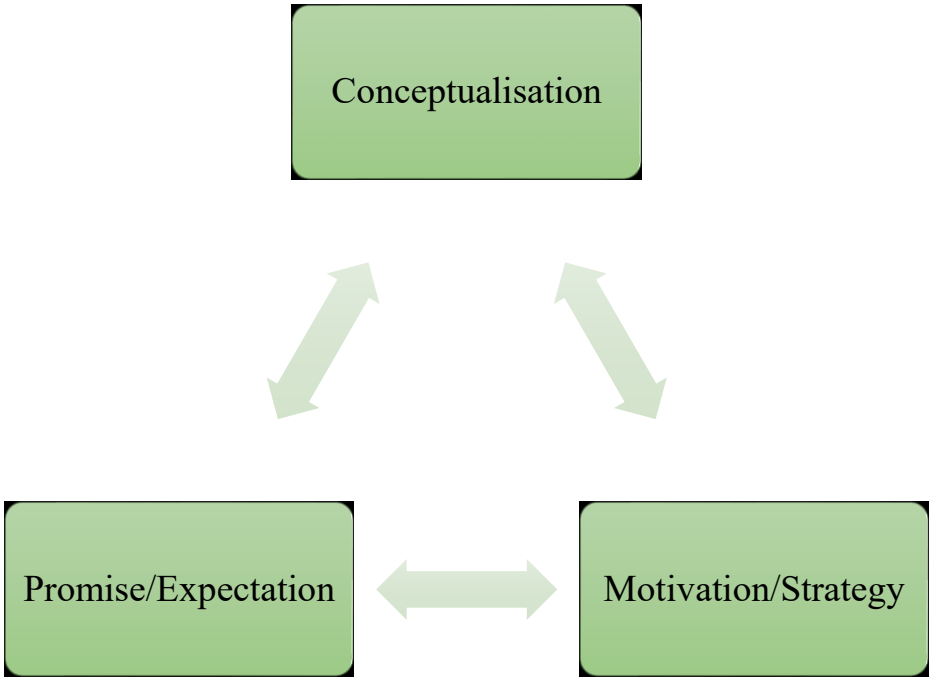
The present chapter distils the specific research questions stemming from the above introductory remarks and describes the theoretical and methodological toolkits employed for carrying out the research, in order to address them. By doing so, I propose a historical sociology of expectations and expertise.

2.1 Research Questions

This thesis' research questions based on the above introductory discussion, can be synopsised as such:

- What is the relationship between AI development and the oscillating promissory environment, including periods of hype and disillusionment (AI winters), in terms of AI conceptualisation, funding, and policy from practitioners' perspective?
- To what extent does informal and non-developer assessment of expectations influence formal articulations of AI communities' practices and strategic foresight? In other words, do non-specialist commentaries impact what scientists are expected to do?
- What can historical examinations of AI's conceptual and promissory settings tell about the current rebranding of AI and how can historical assessment help in better understanding of expectations as an evolving collective process of sociotechnical shaping?

These questions, and the course followed during my empirical journey, splitting time between interviewing and investigating AI history led me to a quadripartite model of understanding the social shaping of AI according to its associated expectations and claims of expertise. That is an omnidirectional shaping between conceptualisation (definitions, descriptions, everyday understandings), expectation (promissory environments, hype, narrative), motivation (funding strategies, curiosity, corporate, academic, or military interests), and regulation (policy landscape). The first draft of this thesis included a chapter devoted to the latter dimension, but due to lack of space and in favour of exploring the promissory dimension in greater detail, it has been omitted and will consist part of future work. For the purposes of the present version of the thesis, I will focus on the interplay between ways in which different conceptions AI are used to create or meet existing expectations, driven by individual and institutional motivations, expectations are shaped by such motivations and depend on previous conceptions, and motivations are shaped by social expectations and accepted definitions which, depending on the variety of motivation, might be accepted or challenged. To visualise this:



The following two sections will present the conceptual frameworks offering vocabularies to describe the empirical findings which, in turn, contribute to such frameworks. This is followed by a description of the methods and choices taken for the present research design.

2.2 Theory: Expectations and Expertise

"[T]he machines are social before being technical. Or rather, there is a human technology which exists before a material technology." (Deleuze 1986: 34)

"There are two futures, the future of desire and the future of fate, and man's [sic] reason has never learnt to separate them. Desire, the strongest thing in the world, is itself all future, and it is not for nothing that in all the religions the motive is always forwards to an endless futurity of bliss or annihilation. Now that religion gives place to science…" (Bernal 1929: 3)

This section will outline the conceptual frameworks which guided this doctoral work. Although there is no single theory examining the complex interplay between hype, practice, and governance, I will highlight the relevance of certain vocabularies that guided my work and justify their usefulness in order to act as core theoretical underpinnings. Most importantly, I will show how my application of future-oriented sociologies onto past settings as well as the study of dynamics between technical experts and broader discourses contributes to theorising a unified subfield of STS which I term sociology of expectations, expertise, and experience (SEEE). It should be mentioned that this theoretical toolkit acts as the basic spine of the analysis. During research, I found that some further conceptual tools were more useful for particular empirical findings, and thus, with SoE and SEE as its main source, the theoretical navigation branches throughout different empirical chapters; to avoid confusion, I refer to these assistive theories in greater detail only within the relevant chapters.

To help understand my rationale behind the interplay between expectations and expertise as a more general contribution to knowledge, I want to begin by a visualisation, the components of which will be unpacked below. While presented at the forefront of theory, it is the result of the empirical

investigations presented in the following chapters. This adheres to an abductive approach, as the theories explicated in the sections 2.2.1 and 2.2.2 have been (among many others) my initial guides during fieldwork – during data collection and upon analysis, however, there has been iterative discarding of other candidate theoretical toolkits, and shaping up of the following expertise-expectations mapping tool which can be useful to social scientists, policymakers, or business/marketing scholars with interest in science and technology. It offers a canvas to situate dynamics across the individual and the collective as well as the micro promise/statement and the macro expectation/grand narrative. In a society where "expertise" becomes a token of a person's endowment with credibility in advising about a field's potential future, it is now plain that the distinction between whether expertise or expectation come first becomes somewhat blurred, or has scarcely been articulated with clarity. The conception of an expert in public talk (or the "influencer") hinders the value of the practical domain expert. After listening to many interviewees and other specialists' frustration with media coverage of AI (e.g. the anecdote in 1.4.1), the continuous impact of technological "gurus" in public debate is becoming a pressing issue which extends the field of AI, and thus I hope this thesis will be of interest to scholars interested in other domains of equal magnitude in terms of loaded expectations and competing expertise such as space industries, medicine, warfare, or education[16]. To quote political theorist Guy Debord: "It is in the interests of those who sell novelty at any price to eradicate the means of measuring it." (Debord 1988: 15). The absence of everyday specialists in public and policy debates indeed eradicates any form of critically assessing everyday consumption of technology-related promises and debates. The influencer is able to set up grand narrative expectations, appealing to those unaccustomed to the hurdles of everyday technical work, and more easily understood due to their conceptual (over)simplifications and the rhetoric of looking at the big picture – further supported by national strategies craving for grand imaginaries and narratives. Science fiction and mythological/religious discourses inspired both strands of, on the one hand, scientific/technological domains which sought to use applied scientific research to explore such questions in a rigorous manner, and on the other, broader discourses attached, ironically, to the mythological/fictitious dimension of AI which is less esoteric than the world of coding, programming, testing, data labelling, algorithm training, and calibrating. Performativity between the collective and the individual, the expert and the expectation is hereby viewed as multidirectional, moving from the confinement of the AI specialists' laboratory to either public (media) or governmental (policy) domains of the non-AI specialists' commentaries or decision-making, and from the development of promises, expectations, and hype to the performance and defence of expertise credentials in formal or informal ways.

Dynamics of expectations and expertise	***Practical domains***	***Grand narratives***
Expectations	Exaggerated promising/abiding by the hype, occasional generation of hype	Large regimes of expectations, national strategies, underlying beliefs
Expertise	Existing credentials, rebranded according to the hype	Prestigious tech/science figures, influential politicians, science fiction and mythology

[16] Following recent entries of the tag "Elon Musk" on influential newspapers such as the *Guardian* or the *Independent* shows how such "influencers" are nearly equated to anything relating to digital, or even space technologies.

This diagram will be revisited throughout the empirical chapters, displaying the variety of its applications, depending on various focal points. Before that, I will outline the main existing theoretical frameworks which shaped its initial formation.

2.2.1 Statements, Promises, Imaginaries, Narratives, Expectations

"The only tradition we can experience is the present moment. And yet we spend most of our lives anxiously hoping we will change – looking forward to things – and doing everything we can to stop this happening. This is why we are only really relaxed, properly at ease, in periods of transition; when we let time join in." (Phillips 1996: 8)

The rather mainstream tendency of studying the future is associated with "futurologists" (often called "futurists[17]"), experts to be consulted about future risks and opportunities, especially after the Cold War (Anderson 2018). Current futurist fashions have stepped away from typical expert consultancy (more on this below, in 2.1.2) and have transformed into a demand for public intellectuals associated with some abstract technical/scientific form of expertise broadcasting views about the future, influencing decisions and public opinion. Max Tegmark (Tegmark 2017) and Elon Musk (Gibbs 2014; Loizos 2017) are good examples of such types of "futurism." The lure about knowing the future based on present states dates at least back to 1814, when, in *A Philosophical Essay on Probabilities*, polymath, mathematician, engineer, and physicist Pierre Simon Laplace hypothesised the possibility of knowing every future or past state based on knowing every present state (Laplace 1814: 4). This deterministic view of the world (sometimes referred to as "Laplace's demon") as a predictable, mechanical whole has branched into both qualitative and quantitative sciences – futurism and AI being both instances of this. This thesis follows the opposite direction, suggesting that our very conceptualisation of a predictable or malleable future has a performative dimension in the ways we negotiate, impact, and predict it – it becomes a form of currency. As succinctly put by Nordmann and Rip:

> "'If-and-then' statements begin by suggesting possible technological developments and then indicate consequences that seem to demand immediate attention. What looks like a merely possible, and definitely speculative future in the first half of the sentence (the 'if'), turns into something inevitable in the second half (the 'then'). As the hypothetical gets displaced by a supposed actual, the imagined future overwhelms the present." (Nordmann and Rip 2009: 273[18]).

In STS, expectations become known from their power to "legitimise, inform and coordinate efforts in research, firms and government" (Van Lente 2012: 779). Their rhetorically vague power renders them performative by creating obligations ("promise-requirement cycles"), since "the only reliable way to validate the claim is to try to achieve it," thus legitimising investments and policy decisions/measures (Van Lente 2012: 772-773; see also, Brown and Michael 2003; Borup et al 2006; Rusconi and Mitchener-Nissen

[17] With no relation to the Italian and Russian art movements inspired by anything which appeared past-avoiding and future-embracing.
[18] It is certainly interesting that widely cited works in contemporary quantum physics also suggests "retrocausality" in nature: quantitative measurements on the possibility of signalling from future to past in the same way the past signals into the future (Leifer and Pusey 2017: 21-23).

2014 for diverse case studies). This field of "sociology of expectations" (SoE) offers an adequate vocabulary to understand how actors from commercial, technical, societal, and overall mixed settings "draw from and add to a repertoire of images, statements and prophecies – and by doing so they contribute to a particular dynamic" (Van Lente 2012: 772). The consistent use by newspapers of stock images from *The Terminator* film whenever an "AI threat" is reported, is an everyday example of this reference to a repertoire of images (on this, see Obozintsev 2018; Royal Society 2018b). Such dynamics can oscillate between broader "collective expectations" and "actor-specific" ones (Konrad 2006). For Van Lente, there is a macro and micro level of expectations broader visions and more specific statements. Longer-term expectations are more abstract, whereas micro expectations tend to be short-term and more specific in content[19] (Van Lente 2012: 773-4), but they affect each other reciprocally. Being effects of coordination and herd behaviour, through their informal construction, as well as their depersonalised function, abstract expectations infiltrate themselves into formal discussions, and tacitly offer directions and reduce uncertainty as the "promising direction is available through the informal expectations circulating amongst technology developers" (Van Lente 2012: 774; cf. Konrad 2006: 431).

Pollock and Williams, place emphasis on "promissory work," but distinguish a spectrum of promissory activity" in the way industry validates available "futures" as a form of currency: "At one end is promissory work that is researched and defended robustly, and which appears to 'matter' to promissory organizations and others who use it. At the other end are kinds of promissory work that seem more like 'provocations' that attempt to capture interest" (Pollock and Williams 2010: 544). Examples in AI-related areas include Moore's Law (its portrayal as a law captures the solidification of the expectation, the singularity argument which despite its fairly unsubstantiated status has mobilised forced against it due to its popularity, and the marketing research charts available in various resources, creating the obligation to catch up or not miss out (See section 6.3.1 for an analytical account). Pollock and Williams import, in their framework of expectations, the notion of performativity of business models, as expressed by MacKenzie (2006). MacKenzie suggests a typology of performativity, in his account of generic economic models, some of which are becoming practically effective, either as successful practical applications of economy as depicted by the initial model formulations, or as counter-performative alternate applied versions of them. While this model of models, of sorts, is very useful for model (and expectations/visions/promises) assessment, it is beyond the scope of this thesis to assess the realisability of promises. Instead, I focus on the complex dynamics of expectation formation and their adoption by relevant actors – hence, only speaking to the initial description of MacKenzie's model, the movement from generic to effective performativity, without assessing the "effects" precisely because part of my argument has to do with the interaction of actors who assess such effects, based on specific individual or institutional agendas, and, to use Van Lente's terms, the employment of micro expectations.

Thus, micro expectations, statements, and promises, when accumulated can construct macro expectations, or, what is occasionally termed a "sociotechnical imaginary": "collectively imagined forms of social life and social order reflected in the design and fulfilment of nation-specific scientific and/or technological projects," extending to "nation-building" (Jasanoff and Kim 2009: 120). A recent application

[19] Much prior to SoE, philosopher Michel Foucault explained how a "family of statements" exists as a "primitive function" within a "correlative space" (as synopsised in Deleuze 1986: 7). That is, as long as a correlative space is defined (for example, AI is a correlative space that several parties may relate to), the "regularity" (Deleuze 1986: 5) of repetition of a statement precedes the agency of author(s) generating it. Similar treatment of statements in STS discourse analysis can be found in Callon (2007: 320).

of sociotechnical imaginaries in the field of robotics for care can be found in Vallès-Peris and Domènech (2020; see also Hyysalo 2006 for explication of "practice-bound imaginaries" for automation in elderly care). It is noteworthy that the term "imaginary" has a long history, at least since the role of social imaginaries has been highlighted in the role of philosophical metaphors generation (Le Doeuff 1989: 3-6) as well as in policy formation (Verran 1998: 238, 243-244). Jasanoff and Kim's analysis of imaginary-based nation building links to the modernist idea of progress. Philosopher/sociologist Theodor Adorno spoke of governmental states using excessive amounts of technological "hurrah-optimism" in their visions to suppress any will for critique and goes as far as seeing optimism as a concept which comes to develop autonomously itself, as opposed to a promising technology that will act as a changer (Adorno 1951: 122). Similarly, Steve Rayner, speaks of the "novelty trap" – the risk confronting effort of competing countries or institutions to "catch up." These risks may be augmented when grandiose claims about emerging technologies are followed by drawbacks, due to the unrealisability of the promised novelties[20] (Rayner 2004: 350-351). This can be tied to older national strategies in the UK, such as the nano-bio-info-cogno (NBIC) strategies (Spinardi and Williams 2005; Williams 2006).

Studying imaginaries in the social construction of the internet, Flichy highlighted the importance of psychological, ideological, and mythological dimensions in the formation of imaginaries which legitimise belief in various "novelties" (Flichy 2007), which might then lead to "compressed foresight" based on "narrative bias" (Williams 2006). Similarly, we can think of AI as a "cultural icon" in the terminology of Nelkin and Lindee (1995) – initial scientific exaggeration paired to age-old fantasies escaping the confinement of scientific communities, becoming registered in the public imagination as science fiction turning fact, thus employed by governments as tokens of technological national strategy building, with or without evidential fulfilment of technical promises: "The danger, then, is not that inflated promises threaten to backfire on the scientific community, but that such promises will long outlive their scientific utility (Nelkin and Lindee 1995: 197). For Arie Rip, however, there might be backlash, when, in his study of nanotechnology expectations, negative public imaginaries might result in "exaggerated interpretation of public concerns" on behalf of policymakers and funders, "seen as an indication of fear, even phobia of the new technology" (Rip 2006: 358). Rip, then, speaks of a potential "nanophobia-phobia—the phobia that there is a public phobia," threatening the scientific community (questions concerning AI policy and regulation will be explicated in chapter 6). While most of the aforementioned imaginary-oriented studies refer to positive/utopian forms of the future, certain technologies, like nanotechnology or AI, lead to dystopian imaginations as well (as in the aforementioned case of the *Terminator*).

Some exemplar cases studying the impact of imaginaries/narratives in AI are the following. Szollosy (2017), from a psychological perspective, suggests that fearful attitudes towards AI express an age-old manifold anxiety of (a) "the mad scholar who seeks knowledge" followed by the "fantasies of being superseded (devoured) by one's progeny" (Szollosy 2017: 434-435), and (b) the projection of human anxieties for an extremely rational and mechanised self, but also the irrational violence of domination and control (Szollosy 2017: 438). The Royal Society in collaboration with the Leverhulme Centre for the Future of Intelligence have published a report on outlining most major AI-related mythological and religious narratives and showed how these have permeated into policymakers' way of talking about AI (Royal Society 2018). Allen discusses how this impacts current media debates by extending Vincent

[20] See chapter 7 on the AI policy race and the historical unfolding of global AI competition triggered by the Japanese "threat" in sections 3.1 and 3.2.

Mosco's concept of the Digital Sublime, in conjunction with Flichy's concept of the internet *imaginaire*, through the examination of "a sample of 55 newspaper articles from [the widely circulated Swedish newspaper] Svenska Dagbladet between 2017 and 2018," indicating the existence of an "AI sublime" through the persistent reoccurrence of four AI myths: "the *intelligent* computer, the *intelligent* robot, the *intelligent* machine and the *intelligent* vehicle." (Allen 2019: 52: original emphasis).

An important aspect in SoE research is the longitudinal study of expectations; this approach drives much of this thesis's rationale. While there are but few available pieces of research, they are also enough to suggest an emerging direction in SoE, which the present thesis enriches by investigating AI. Some examples include: Ruef and Markard's (2010) study of stationary fuels hype from 2001 to 2010; Van Lente, Spitters and Peine's (2013) comparison of voice over internet protocols, gene therapy, and high temperature superconductivity, in their attempt to produce with an analytical theory of hype; Kirkels's (2016) approach to "reconstruction of expectations" in advanced biomass gasification from 1970s to 2015 allowing for critical comparisons and better understandings of disillusionment troughs; Melton, Axsen and Sperling's (2016) case study of alternative, decarbonised fuel transportation from the 1980s onward, monitoring the shifting, reconfigured expectations; Hielscher and Kivimaa's (2019) study of smart meter innovation and implementation in the UK in two rounds (2000-2008 and 2009-2016) offers an additional case study in this direction by showing how sustained expectations surrounding the technology influenced their relevance across a field of shifting energy policy environments; finally, Tarkkala, Helén & Snell's (2019) reflections on a decade of hyped personalised medicine, suggest that practical maintenance is required to sustain sociotechnical imaginaries paired to better, continuous governmental technology assessment. The predominant difference between these studies and the present one is that a field as broad as AI is harder to capture in terms of longitudinal expectations. One would argue that AI as a science is not commensurable with smart meters as a technology. But it is precisely the argument brought about in chapter 3, that the historical examination of AI's promissory environment (a) shows how a scientific field's expectational shaping, especially through governance and hype, transformed it into a technological one, and (b) why SoE and the general study of sociotechnical futures should be applied to study the past. At a deeper level, the present study focuses on the exact interplay science and technology as expressed in the term "technoscience," prominent in STS discourses, with Heideggerian roots suggesting that science is technical arrangements, laboratory settings, and global research enterprises (Shapin 1988: 548, Zwart 2020). AI as a global scientific project is in constant interplay with AI as a series of applied technical settings. The following section focuses on the actors who shape these expectational configurations.

2.2.2 Expertise and Experience, Arenas of Enactors and Selectors, Expectations Governance, Exploration and Exploitation

"When we meet a professional fortune-teller who promises to use his [sic] art to reveal our future, we generally have mixed feelings. On the one hand, the idea appeals to us that someone can look into our future by looking at our hands and relying on a determinism that is inscrutable for us but decipherable by him [sic]. On the other hand, we resist the idea that we are determined, explainable, and predictable beings. We cherish our free will and want to be beyond determinism. But at the same time, we want the doctor to cure our diseases by treating us as structurally determined systems. What does this tell us?" (Maturana and Varela 1992: 122)

"Alone, in our separate kinds of expertise and experience, we know both too much and too little, and so we succumb to despair and hope, and neither is a sensible attitude." (Haraway 2016: 4)

If constructions of the future have an important impact, studying who generates, shapes, circulates, selects, maintains, and dismisses such promises and imaginaries is crucial in order to understand the role of expectations in AI development. With its roots in early foresight studies (Overbury 1969) and Gieryn's (1983) work on the flexibility of scientific boundaries, the need for scientific consultancy in governing science based on potential future outcomes is facing the problem of legitimacy of expertise. This is further reflected on recent calls for responsible research and innovation (RRI; for example, Smith et al 2019; Owen, Mcnaghten & Stilgoe 2012; Stilgoe 2018) where responsibility aims, among others, at the involvement of as many voices as possible in research regulation based on intended and unintended outcomes. Expert studies in STS have initially tried to break with the top-down hierarchy of experts as opposed to "laity," when community and tacit expertise became known to shape and influence scientific advancement; however, multiplicity of voices, misinformation, and vested interests, invited for the reservation of "pockets" of expertise and the demand of exhaustive inspection of one's credentials as to the right to be considered an expert (Collins and Evans 2002). As the two authors have summarised the problem: "Democracy cannot dominate every domain—that would destroy expertise—and expertise cannot dominate every domain—that would destroy democracy" (Collins and Evans 2007: 8). Such ambiguities have been assessed, mapped, and further questioned in the STS subdomain of studies of expertise and experience (SEE), which investigates the roles of experts in the social shaping of technology. In the definition of expertise by SEE, "expertise is socialisation into an expert domain. Society consists of many expert domains of different extent, some small and esoteric, some, like language, large and ubiquitous. Expert domains overlap and are embedded within each other like a fractal" (Collins et al 2020: 63). Collins and Evans' framework and SEE at large have mostly focused on vertical relationships between core scientists and technologists and lay citizens and the question of whether laity may exhibit expertise. An addition to this framework is the concept of the horizontal "expanding expert," the archetype of core scientists who use prestige gained by their own domain to become spokespersons in other domains (Galanos 2019).

As in SoE, systematic longitudinal observations are still limited in studying the evolution of what constitutes expertise within a given field. Collins has written extensively on issues of expertise relating to gravitational physics (Collins and Evans 2002, 2008; Collins 2018). This problem is particularly relevant in AI research, as shown in the historical assessment of Chapter 3 – on the one hand, there have been instances of AI researchers who advocated AI as an all-encompassing field (in the 1990s); yet, recently, they have called for boundary setting, when hype took flight and "too many" people, such as futurists, journalists, or politicians, claimed their share of expertise in AI. While SoE deals with the interplay of collective and individual expectations, the SEE has not yet developed an adequate vocabulary for groupings of different sources of expertise beyond the individual. Therefore, I suggest that the combination the two STS subdisciplines (SoE and SEE) as a means to focus on expertise as a generator of expectations and simultaneously, the formation of (apparently) credible expectations as the verification of certain expertise. Such groupings of shared knowledge, guided by specific constellations of expertise-expectation dynamics can be thought of as "arenas," to borrow phraseology from Bakker et al (2011) who specifically talk about arenas of expectations. Using SoE's reference to allows placing emphasis on the collective temporal transfer of expectations from social group to social group, and thus broaden the scope of both SEE (usually focusing on present-oriented assessment of expertise between individuals) and SoE (usually focusing on present-oriented assessment of future orientation between collectives and individuals). This

enables a more temporal analyses of the horizontal interaction of "multi-regimes of expectations" (and expertise) as well as the vertical interaction between individual and collective expectations (Konrad 2006; Konrad et al 2008; Konrad et al 2017; Budde and Konrad 2019), adding a third dimension of time-sequence. AI has recently opened up the floor for some first explorations of the expertise/expectations interplay, such as Dandurand et al (2020) who have already begun to show how heterogeneous ensembles of actors with variations of expertise which under different circumstances can be guised as "AI" or related to AI can further shape expectations and motivate decisions. The present work adds to such debates (a) by empirically investigating experiences of contemporary AI communities on the intersection between expectations and expertise, and (b) by historically examining these communities and the expectations/expertise interplay as product of long historical negotiation of promises, imaginaries and arena interaction. Ways in which promises and expectations are generated and acted upon as well as expertise is performed and becomes credible is hereby shown by rich descriptions of the emergence of "alliances of the most diverse actors from different political and cultural backgrounds" admitting that "[i]mplicit or explicit, socio-technical futures are inherently political" (Konrad and Böhle 2019: 102), as SoE authors have already noted. One final term borrowed from the SoE literature is the distinction between enactors and selectors, which enables a criterion for describing moments of availability of expertise and expectation variation and the decision points in which certain actors solidify a certain expectation. In Bakker and Budde's work, multiple arenas co-exist in shaping and developing technical options, promises, and expectations, from scientific venues such as conferences and academic journals, to media representations, and policymaking. Often, such arenas overlap and become blurred, but it is precisely in those interactions where promises travel and expectations crystallise. The arenas framework offers efficient vocabulary for the interplay between individual actors' influence towards collective expectations and vice versa:

> "enactors are those actors that develop and simultaneously 'enact' a (radically innovative) technological option. Part of the enactment is the voicing of positive expectations of their option [or visions]. As there are many technological options that are being developed, there are many expectations, and promises. Not all of them become collective and some selection is necessarily made. The selection process, on the basis of different types of assessments, relates to selection in terms of funding allocations by governmental agencies and also in terms of firm-level decisions on viable R&D trajectories. Moreover, at the same time the selection process relates to expectations as well, the so-called selectors assess the different expectations and promises in terms of credibility and their judgments are crucial to the emergence of collective expectations. From their interplay, the actors that voice the expectations and the actors that assess them, collective expectations emerge." (Bakker and Budde 2012: 551-552).

In the case of AI, and in the context of this thesis with AI technical practice as its main research focus, and academic AI practitioners as the main source of empirical interview resource, these arenas can be viewed as such (as unfolded in their exertion of influence in the historical chapter 3): academic AI practitioners, commercial developers, commercial sponsors, military developers, military sponsors, academic non-practitioners (from disciplines such as mathematics, psychology, philosophy, social science, religious studies, art – who have no direct involvement with developing AI techniques), journalists and media discourse, and marketing analysts. The research is thus biased towards representing the views of the first arena, however, this is justified according to the initial premise in that academic AI practitioners were not sufficiently voice during the beginning of the fieldwork. However, there is good reason to focus on the voices of those who have direct experience (or expertise) with the production of a given science or

technology, as their uncertainties might be revealing in the context of exaggerating certainties on behalf by those who have interest in the technology but have little experience and expertise in developing it, as argued by MacKenzie (1998), carefully synopsised in the following qualitative diagram:

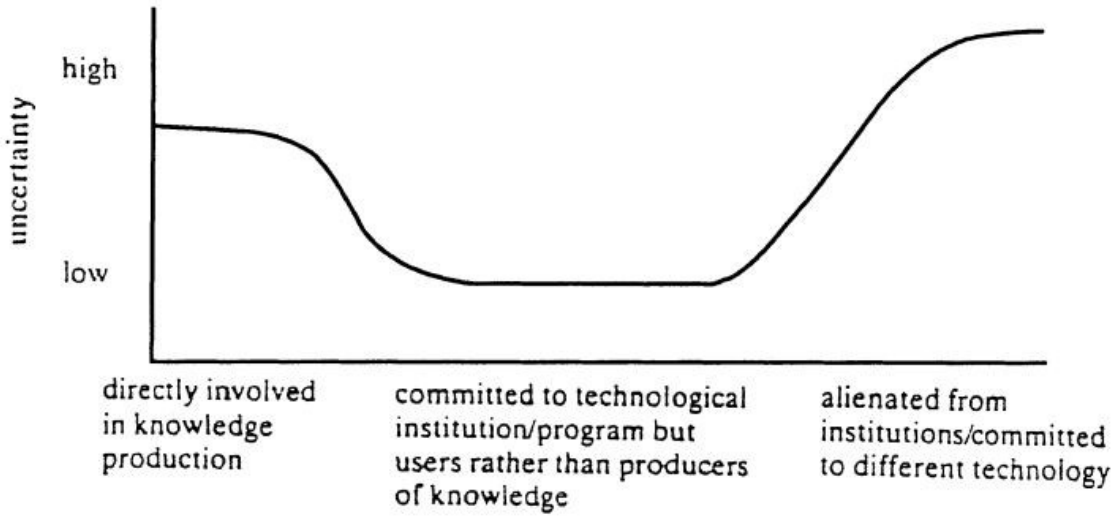


Figure 2 The certainty trough (MacKenzie 1998: 325)

It is useful to note that the certainty trough has been previously employed in SoE studies as in Van Lente (2006: 774-5) and Brown and Michael (2003). Nevertheless, MacKenzie's approach should not be taken at face value – not all AI practitioners are necessarily devoted to novel knowledge production, even if directly involved with it. An additional analytical lens I employ to interrogate findings in chapters 5 and 6 is that of the exploration-exploitation trade-off in individuals' choices to promise as in relation to secure funding. This framework has its roots – ironically – in the post-cybernetic study of adaptation in biological and mechanistic structures. Holland (1975: 181) has suggested that living organisms are "trying to balance exploration (acquisition of new information and capabilities) with exploitation (the efficient use of information and capabilities already available)" in order to ensure survivability and adaptability to novel environments[21]. His framework, however, was mostly celebrated by organisational studies scholars. March (1991) was the first to adopt this framework within the organisational management context as such:

> "Exploration includes things captured by terms such as search, variation, risk taking, experimentation, play, flexibility, discovery, innovation. Exploitation includes such things as refinement, choice, production, efficiency, selection, implementation, execution. Adaptive systems that engage in exploration to the exclusion of exploitation are likely to find that they suffer the costs of experimentation without gaining many of its benefits. They exhibit too many undeveloped new ideas and too little distinctive competence. Conversely, systems that engage in exploitation to the exclusion of exploration are likely to find themselves trapped in suboptimal stable equilibria. As a result, maintaining an appropriate balance between exploration and exploitation is a primary factor in system survival and prosperity." (March 1991: 71)

[21] He suggested in the same book that intelligent machines should aim at a similar purpose, already expanding on artificial algorithms imitating genetic algorithms. Given that the year of his book's publication falls within the register of the first "AI winter," its appreciation by a very different research culture might be partly explained.

I adopt this framework in order to highlight nuances in the motives of AI practitioners chiefly in chapters 5 and 6. While a treatment of AI practitioners as a homogeneous mass with shared characteristics would be simplistic and lacking reflexivity about the sample's vested interests, this framework, adapted accordingly in the mentioned chapters, allowed a more critical comprehension of AI practitioners expertise, as an outcome of its interplay with non-practitioner expectations. With this in mind, focusing on both individual and institutional/collective, temporal and spatial, enacting and selecting aspects of future-oriented aspects of AI, allows a more precise understanding of AI's environment.

A note on the interplay between individual and collective influence in expectation management: often certain collective groupings, or arenas, adopt a specific name to represent their shared knowledge and interests. Sometimes they do not, and it is left up to the researcher/analyst to group them according to a given feature, such as an approach to technical innovation, a political ideology, or a philosophical position. This is often found to be a useful analytical device but also comes with caveats of arbitrariness[22] – where a certain arena begins and ends is arbitrary, as much as the complete agreement between any more than one individuals apparently sharing the same approach. It seems that any assessment of collective enaction or selection of expectations boils down to individuals. But this, the assessment of personal influence in enacting and selecting visions and options, can be also challenged, and indeed has been so in STS, as shown in Mialet's work, deconstructing the work of Stephen Hawking, the individual, showing the immense collective construction of his self via the numerous research assistants, journalists, and machines (Mialet 2012), enabling the credibility of Hawking's expertise and the (apparent) legitimacy of expectations he advocated for. In the following chapters, there will be weight placed on individuals' promissory statements, sometimes as indicative of a generalised sentiment being expressed through them. It is difficult to assess the creation of specific arenas via individualistic statements, however, especially in the historical analysis, this is paired with the parallel formation of terminologies, opening up a further question about the interplay between conceptualisation and expectation in AI. In the interview sections, where anonymised individuals report on their experiences, this becomes indicative of the performative influence of broader sociotechnical regimes on their work and statements, but, hopefully, with less will to influence the expectational environment directly.

2.3 Method

This thesis employs a mixed historical-interview methodology. Chapter 3 offers a social-historical exploration of AI in two parts: a chronological presentation of AI definitions and a chronological presentation of promissory debates and published future-oriented statements. This is used to generate a series of findings in support of the value of treating expectational environments longitudinally through an examination of different arenas of expertise. This is followed by three interview-based chapters, reporting on contemporary AI specialists' views on AI conceptualisation and promise, as well as funding – the latter, being the factor found from both historical and interview-based analysis to be the chief social activity that invites promising on behalf of AI researchers, and thus plays a significant role on how AI communities will negotiate their terminology.

While the historical examination was not part of my initial research design, conducting this first a form of literature review, revealed gaps which could be (partly) answered by practitioners (such as the

[22] Ironically, the same arbitrariness encountered in data labelling for machine learning applications of AI (e.g. Kotliar 2020).

motivational drives for promising or the impact of non-specialist perceptions of AI on practitioner communities). I then realised that this method has been used in previous research with fruitful results (for example, Russell's 2011 study of the history and current state of Black and White students' achievement gap in mathematics achievement). Why history? I follow MacKenzie in that as well: "The 'historical' is the easier part to explain. Those of us who research social processes are seldom able to set up our own experiments. We have to wait for the world to do it for us. The passage of time, and changes it brings in the factors and phenomena that interest us, are our single best source" (MacKenzie 1990: 7). My argument in selecting this is that instead of a dry review of AI definitions and examples of how AI-related social groups and individuals have been promising about it, such conceptualisations and promises, placed in a chronological fashion, allow for a clarification of the (otherwise) messy historical process of which contemporary AI hype is an outcome. Indeed, I have found most historical reviews of AI (outlined below) insufficient in terms of taking into account the vast promissory environment – and similarly, contemporary empirical research that has to do with AI expectations does not take into account the complexity of its historical negotiation of promising and conceptualisation (with very few exceptions). Most AI histories have been written by AI enthusiasts (including practitioners) and thus treat AI's history in a rather deterministic, linear form of development with little, or no, emphasis on the broader social factors shaping it (besides its chief opponents). The interview findings presented later in this thesis are greatly benefitted by the acknowledgment of AI's history; and the contemporary "mess" reported by interviewees is more easily understood based on a historical mess which proves that AI, as either science or technology, is created within and creates a highly promissory environment, dependent on political contexts, practitioner strategies and choices, and assessments by governments, media, or non-AI academic specialists. Returning to MacKenzie: "The present is equally open to examination, just so long as it is studied not in isolation from the past but as a moment in continuing processes" (MacKenzie 1990: 8). In the light of the SoE and SEE frameworks have been covered above, I hereby propose to call this approach a historical sociology of expectations and expertise, bringing together the merits of historical sociology of science and technology as followed by MacKenzie and Russell and those of longitudinal SoE, with the enrichment of SEE's framework.

Some notes on the systematics of the historical method. I display conceptualisations and promissory statements or events negotiating AI expectations employing the chronology approach. In previous work concerned with the influence of Stephen Hawking and Elon Musk's statements in the development of early AI policy drafts (Galanos 2019a) looking at the span of four years (2014-2018), and I hereby extend this timeframe from 1955 to 2021. On the value of chronologies in historical research, I think with Montaña:

> "A chronology is a traditional tool of historians to concisely order a series of events concerning a given subject along a timeline. It can take the form of written register of events in strict temporal order from the oldest to the most recent […] [A] chronology is modelled after the manner in which we, humans, experience the passage of time. It shapes the historical argument through its sequential and rhythmic structuring of time" (Montaña 2017: 14)

It should be noted that I do not claim to have brought to light novel historical findings – as mentioned, this part of the research began as a literature review that was then enriched by systematic investigation, appropriation, rereading, and repurposing of existing histories of AI. I consulted sources and archival records or publications on AI history, that are influenced explicitly or implicitly by their authors' regional contexts, institutional motives or research focus. There are several relatively underexamined and

unclear points in AI history which required additional research. Such references and their corresponding events are cited here, but the terminologies will make sense in the historical outline below. Sources I began with include: the semi-biographical personal inquiries books by McCorduck (1979; 2004; 2019), Rose (1985), and Crevier (1993); Nilsson's thorough technologically focused history (2010); Reichardt's more general work on fact and fiction about robotics (1978); Fleck's works on the establishment of AI as a field (1979) and Agar's work on AI winter (2020, see also below); Feigenbaum and McCorduck's US-filtered narration of the Japanese 1982-1993 Fifth Generation Computer Systems (FGCS) project (1984), counterbalancing its partisanship by defenders of the project (van de Riet, 1993; Shapiro and Warren 1993, Hendler 1994; Garvey 2019); Grudin's (2009) careful examination of the parallel evolution of AI and HCI with emphasis on the international politics and the chain reaction sparked by the FGCS, those being: US's Defense Advanced Research Projects Agency's (DARPA) 1983-1993 Strategic Computing Initiative (SCI; thoroughly covered in Roland and Shiman 2002); UK's 1984-1990 Alvey programme (Oakley 1983; Oakley and Owen 1989; more information about its near-failure because of bulk purchases of unsuitable British computing equipment "just because it is British," according to Sloman, in House of Lords 2018: 161; more about the history of Britishisation of computing in Summer, 2014); and the 1983-1999 European Economic Community's (now European Union) European Strategic Programme for Research and development in Information Technology (ESPRIT) (Van Hove 1991; thorough descriptions of its projects can be found at the digital Archive for European Integration in a series of project synopses reports (the first one: Directorate General XIII 1989). I further consulted alternative historiographies which build upon neighbouring concepts such as self-replication (Taylor and Dorin 2020), artificial life (Steels and Brooks, 1995), machine learning (Plasek 2017), and neural networks (Metz 2021).

The history of AI, as also argued throughout chapter 3, is intertwined with the history of the world wide web, and several historical sources appear online, as "about" tabs or obscure corners of institutional websites, relics of a time when busy computer scientists with a will to preserve their history offered their personal accounts and reminiscences about their research programmes or obituaries to deceased colleagues. Much of the recent history has been covered in part through investigative and technology-focused journalism: newspapers and magazines such as *Wired*, *The New York Times*, *TechRegister*, and many others have offered detailed accounts of AI's recent history. The specific ways in which these sources have informed my research will be better understood as I refer to them in the two chronologies. While they have stood as my primary points of reference, they had to be supplemented by secondary research to cross-reference certain statements or stated reminiscences. I am indebted to my internal PhD examiner Shannon Vallor who sensitised me to look deeper into the recent history of AI for promissory negotiations. While this has been an even more under-researched field, these additions helped fit some important pieces of the puzzle in contemporary understandings of AI. Nonetheless, it is certain that further important pieces of historical information might have been omitted, unintentionally.

Studying in Edinburgh about AI, I soon came to learn about the fire which destroyed the local School of Informatics' AI library, "a collection of AI literature unique in the world, an irreplaceable archive accumulated over the 40 years of Edinburgh's leadership in the field, since its beginning in the 1960s" (School of Informatics 2002), a loss estimated to "5,000 books, 800 journals and 35,000 research papers published by the department" (BBC News 2002). This story sensitised me about the importance of historical preservation as part of STS examinations of technology and made me understand in its utter significance Donna Haraway's reading of Marilyn Strathen: "It matters what stories make worlds, what worlds make stories" (Haraway 2016: 12). I am aware that the historical sources I have found and

comment upon are those who have been documented and disseminated to the extent that I can reach them. My choice in placing emphasis on certain strategic and conceptual points adds however the nuance required for an authoritative reading of AI history which can assist the understanding of contemporary attitudes towards it.

My interpretation of AI's history is highly influenced by the following two historical STS papers; their findings are to be hereby borrowed and extended, in accordance to the central theoretical frameworks of expectations and expertise. Olazaran's (1996) examination of the history of the Perceptron controversy among AI circles (more on this term below) showed that history of technology can operate on an "official history mode" and a "research area mode" (1996). Official history includes what becomes narrated among scientific communities, fabled, printed in canonical textbooks and Wikipedia articles (which did not exist when Olazaran published his work). Depth engagement with AI's history and researchers led me to treat the "hype-and-disillusionment cycle" and the "AI winters" as parts of "official history" opposed to the research area mode which reveals that work in AI never stopped; as we are about to see, papers and conferences continued to get published and organised respectively, research was conducted, and guised under new terminologies, practitioners found ways to attract funds. Jon Agar's (2020) examination of the 1973 Lighthill controversy, concerned with the report commissioned by the UK government which highlighted most of AI's unrealised promises, shed further light into the esoteric politics that support applications-based technology, as opposed to basic research. Agar showed archival evidence that Sir James Lighthill's decision to report that "[i]n no part of the field have the discoveries made so far produced the major impact that was then [early 1960s] promised" (Lighthill 1973: n.p.) was steered by the Science Research Council who commissioned him to write the specific report about AI's state-of-the-art in 1973, assessing the outcomes based on the promises. Agar's work showed the tension between AI (and science at large) as a field of experimental research and practical problem solving; he further stressed the role of underlying personal and institutional motivations in the shaping of AI (thus, it is closely tied to relevant areas of the sociology of expectation and promise, in terms of the impact of research councils as selectors on scientists as enactors). As it is further revealed from official histories of AI and from contemporary conceptualisations of AI in this section, AI *is* as long as there are its applications. AI's revival as a field of applications-based technology, sustained by a broader science fiction or mythological narrative, greatly benefits by its terminological vagueness and/or interpretative flexibility.

Considered jointly, these two papers explicate the way in which certain selectors exert influence not only in the technological options supported by funding, but also in the writing of technology's history, and even the transformation of a field from a "science" into a "technology." In a sense, Agar's emphasis on the under-the-record communications between selectors (research councils) voiced by non-AI experts such as fluid mechanics specialist Lighthill who appears as enactor (or an "expert for hire"; Yearley 2005: 162), shape the official history of AI. Disillusionment of promises is enacted and preferred (selected) by specific circumstances and intentions, rather than in a deterministically projected "cycle." This opened up an additional route for the present thesis to place emphasis on specialists' motivation to compete for funding schemes and the associated politics.

So much for history. The following subsections will focus on the more subtle methodological approach of interviewing.

2.3.1 Sample, Access, Relevance and Structure of Site

This section will deal with issues concerning methods employed to get clearer insights on AI practitioners' perspectives on AI expectations, its conceptualisation, and the promissory environment as this relates to funding and policy. The methodological issues covered are interconnected, however, I have organised them according to the following questions: (1) What was the site examined and why is it important? (2) What kind of method was employed, why, and have been its particularities with regard to the specific research? (3) What challenges and dilemmas were encountered and what decisions have been made to address them? (4) How did fieldwork progress and how will data be presented?

During the course of my research, I have interviewed 25 AI specialists between 2017 and 2020, the majority (19) of which being participants at the aforementioned large collaborative effort in AI and robotics between two major universities in the UK. Out of the six external ones, five have been recommended by interviewees on the basis of respondent referral and one was encountered serendipitously, invited by a mutual personal acquaintance over a friendly meeting. Externals acted as means of comparison with researchers in different universities, based in different cities and countries, or working in different sectors such as public services or industrial development. The following list is an aggregate of specialisations represented in the sample detached from quotes employed in the presentation of findings to ensure as possible non-identification of interviewees. The list is indicative of the variety of specialisations associated with the very broad field of "AI," reflecting the very issue concerning conceptualisations (section 3.1 and chapter 4): mathematical logic and theorem proving for computer language development, systems engineering, computational neuroscience, cognitive science, philosophy of logic, computational linguistics, conversational agents, control mechanisms, neural networks, bioinformatics machine learning applications in biomedicine, genetic algorithms for bio-inspired evolutionary computing, computer vision, object recognition, semantics in AI, AI agent development in virtual reality, video games graphics, game-playing programmes, AI simulation training, aeronautical engineering model development, finite elements analysis for software design, probabilistic programming, commonsense learning, machine learning modelling for construction engineering, biomedical engineering, signals and sensing for autonomous cars and further autonomous robotic vehicles, bio-inspired robotics, walking robots, swarm robotics, assistive medical robotics, design engineering for robotic pets, robotic planning, spatial reasoning, fluid dynamics, marine robotics, intelligent robotics, fluid robotics, cyber-physical systems, microelectronics, IT maintenance.

Given that the sample consists predominantly of academics, it is worthwhile to note their seniority representation as a means to express the balance across different career stages, especially in relation to their experience with promising strategies and lessons from disillusionments. The sample includes nine Professors (out of which 2 being retired), two assistant/associate Professors, nine Lecturers (including Chancellor's Fellows and Senior Lecturers). Six interviewees were outside the institute's sample, but were contacted and interviewed after being recommended by institute insiders. Besides two Professors included in the listing above, the externals included a semi-independent scholar, a robotics design company owner with academic collaborations, an IT specialist with a MSc in AI acquired in the mid-1990s, and a Research Fellow who works for a different University. 16 interviewees admitted to have strong connections to industry, besides the aforementioned one who owns a robotics company, thus, having connections to academia, instead of the opposite. This demographic characteristic is to be taken as an indication of AI's deep relationship with industrial applications, itself a topic that is analysed empirically in chapter 3 as part of the transformation of AI from a theoretical scientific field to an applied technological one.

It has become somewhat commonplace that computer science and AI have been criticised for the structural male dominance. Although gender did not play a huge role in the present analysis, I make reference to it to address the great imbalance I have perceived as well. However, most interviewees did not include preferred pronouns on their websites or email sign-offs, and therefore, while I acknowledge the increasing multiplicity of gender identities, I resort to an indicative binary gender division for the present demographic for the sake of allowing the reader to get a better picture of the interview environment and process. Therefore, four out of 25 interviewees can be identified as female during the course of fieldwork.

While the topic of AI winters will be explicated in greater detail in chapter 3, it is useful to pinpoint here as to how this is represented in the sample's experiences with it – indeed, I have considered experiences of AI winters as a vital consideration in my sample's demographics. Three interviewees have lived through, and shared memories about, this first UK AI winter (roughly, 1973-1980) – one was acquainted to a key actor personally, attending one of his courses. Seven interviewees have shared experiences about the second UK AI winter (roughly, 1992-2005) – one of these interviewees happily shared his academic legacy to Marvin Minsky (one of the four academics who formally coined the term AI in 1955), him being his PhD supervisor's PhD supervisor. Given that rounds of hype and disillusionment have guided much of the research, and the impact AI winters had on UK's AI community, awareness or experience of the AI winter concept/era usually suggested stronger opinions about the topic of possible disillusionment.

The institution I selected (I will explicate in 2.3.3 why it remains anonymous), which incorporates cultural diversity, teaching staff and students from around the world, with prior experience in different countries and laboratories, made a very good candidate to address the objective of the study. Indeed, their annual review for 2015/2016, combined with secondary research I conducted about the members' publications and research projects revealed this centre to be of excellency in the fields of AI and robotics. As I am not looking at a particular branch of the vast AI landscape, the research questions required a site which would allow me to establish some comparisons between various domains, such as laboratories focusing on different aspects of AI, but also different experiences of expectations through time and region. Again, this centre, with its many research areas (as shown by the diversity of expertise outlined above), allowed fruitful comparisons between the construction of various types of expectations: from more to less hyped branches of AI, promissory strategies as expressed by younger Principal Investigators or lessons learned from nearly or recently retired Professors. The centre is a collaboration of two major Universities in the same city in the UK, with a long history and important contributions to these fields. That means, that further comparisons could be drawn between empirical accounts stemming from participants of different Universities, although, past the analysis.

Moreover, the particular collaboration of Universities, because of its simultaneous global leadership in both fields of AI and robotics was useful for the following reason. Although the two terms can mean very different things technically, for several technologists one is not a necessary component of the other. For example, a search engine algorithm does not require a robotic physical support and an industrial robot does not make use of complex AI programming. However, general discourse or imaginaries tend to blend the two concepts – most fictitious robots are artificially intelligent and the sci-fi-like fear of "machines taking over" stems partly from the possibility of convergence between the two. Indeed, as it was shown in the historical outline of sections 3.1 and 3.2, such attempts at creating robots with artificial brains, have been considered to be the "holy grail" for some relevant scientists and technologists. The studied centre's excellence and reputation lies precisely on the fact that its teams tend to

work towards this direction through the combinatorial initiatives between applied AI specialists and roboticists. Hence, the interviewed specialists were able to reflect on topics and promissory environments about the intersection of these technologies.

I further wish to highlight the possibility of the institution's indirect benefit from my research. Often, organisations with several domains of research, multiple labs, and culturally and academically diverse members of staff might have difficulties in understanding their inner social dynamics. By gaining access to such a remarkably productive institution, and interviewing the key players through a social science scope, I suggest that they can be informed about their own advantages from a social perspective; something which tends to be invisible to the scientists within the closure of their laboratories. In addition, I should mention that during the course of my study (and, alas, during various moments of hype peaking), members of the centre have been informed through my research about social dimensions relating to their own field, resulting in interesting conversations, which, in turn, as some informed me, have been incorporated as components of introductory courses to AI they have used with their own students. Such references to the social dimensions of their own work either motivated some to participate in the first place, or was found to be beneficial after the interviews.

The aforementioned advantages associated with the selection of the specific site should not obscure limitations and backdrops concerning this choice. These are reflected on the locality, the domain, the research culture, and the demographic dimension. More specifically: the institute is UK-based – that means, most participants' reflections do not necessarily represent the experience of expectational environments in areas where technical promising paired to availability of venture capital guides perceptions about AI (Silicon Valley's "California ideology" is a well-known case where digital utopian thinking emerges out of technologically deterministic rhetoric; Barbrook and Cameron 1996). This is further associated with the fact that my sample consists of mostly academics, with the exception of one participant who works closely with academics, while much, if not most, of AI's current promissory environment is constructed around industrial settings, private companies, and military projects. Through chapter 3's historical outline, I show how the initial emergence of AI as an academic endeavour was appropriated (and abandoned, during AI winters) by military and industrial operations; academic institutions never lost interest in AI (or its rebranded guises), however, much of the contemporary funding would probably not exist if it were not for commercial and battlefield interests. This limitation on behalf of my sample nevertheless shows an important aspect about AI expectations relating to academic AI practitioners' experience of such an environment, their varied responses, and approaches. This is not to say that academic AI practitioners are completely detached from commercial settings – many of their projects are partly, sometimes fully, funded by industry[24] – for some of them this appears as an opportunity to unify the practical with the theoretical; for others this appears as a constrain against scientific curiosity. This nexus between commercial, potentially military, and academic settings is certainly not captured in its totality through my one-sided report of it; nevertheless, I doubt there can be a view from nowhere. A more heterogeneous set of interviewed actors would be likely to earn in breadth of voices represented, however, the more homogeneous group studied here earns in depth of nuance about the specific research culture, which may find correlates across different university settings. As much as the present study lacks

[24] It should be noted that several of my interviewees admitted no connection to military funding, some have expressed and explicitly anti-military stance. For others, this could relate to obligation towards secrecy about potential military projects, or, as some connections have suggested, to a general informal legacy of this institution's denial to partake in military research.

representation of core voices shaping promissory settings in upstream private industry domains, it equally lacks representation of what has been revealed to be, during the period of this doctoral study, a domain of increasing concern: the downstream data labour domain. That, being the core foundation (the "bread and butter") of AI maintenance and operation in its currently predominant approach, involves data clearance, labelling and clustering, giving rise to increasingly more required, as well as precarious, professions associated with the curation and optimisation of AI product outputs (Sambasivan et al. 2021). Experiential accounts of data workers, as indirectly shown by Sambasivan et al's focus on "everyone [wanting to] do the model work, not the data work," offer a rich account not just of expectational settings, but also concerning pressing issues of expertise. As argued in the thesis' results, expectations and expertise exist in a mutual shaping relationship; therefore, AI expertise as expressed by AI/data maintainers is shaped by, and shapes, AI expectations. This leads to the last, but definitely not least, of the present sample's limitation accounts: the admittance of lack of intersectional representation.

The aforementioned data workers are often represented in worlds regions other than the Western, or "global north" mainstream centres of AI R&D. AI as a sociotechincal enterprise carries the speciesist, classist, ageist, racist, and gendered politics of every technology as a social construct. A very small percentage of my sample identified as female and/or of colour, echoing the legacy of computing science being "a [white] man's world." None identified as nonbinary, trans, or gender non-conformative. A few acknowledged, when asked, the lack of female representation, but none reflected on broader LGBTQI+ communities. (Indeed, little has been written about gender non-confirmist representation in AI practice; an early step is Os Keyes's studies of AI misgendering of trans people, and trans woman and computer scientist Lynn Conway's early advocacy for trans rights in computer science in the 1980s; Keyes 2018; Roland and Shiman 2003). While the sample contained a number of non-White Caucasian members (albeit I retain my scepticism about such categories), none would identify as Black. The role of Black representation in contemporary AI will be explored in sections of chapter 3, however, a great consolidation of pressing issues and opportunities is to be found in Ruha Benjamin's book *Race After Technology* (2019). Given the participants' social status, being receivers of academic salaries, none would easily be identified as "lower class" (some would debate this; I would not). While some of my interviewees could be classed under the label of "elderly," this is not to claim that there are representations of expectational/experiential apprehensions of AI by younger users exposed to algorithmically curated content (Allyn 2021), growing up in environments where the interface between AI/algorithmic mediation and marketing opportunities becomes increasingly relevant to future job seeking. Important dimensions of the intersectionality spectrum not covered within the sample involve representation of nonhuman animal rights in the face of AI development. While, due to anthropomorphic and anthropocentric reasons, it is difficult to write about representations of nonhuman expectations and expertise, works like Bellet's (2019) are also first steps towards involving nonhuman organic creatures in the debate about AI governance. Reflections about nonhuman elements' relationship to AI have only but tangentially offered in the responses of my interviewees; nevertheless, influential work such as Strubell, Ganesh and McCallum's (2020) raise awareness about the environmental costs of AI, keeping up with a legacy of critical scholars studying the environmental impact of computer, digital, and information technologies at large.

Given the delineation of this work's limitations, I further wish to highlight the possibility of the institution's indirect benefit from my research. Often, organisations with several domains of research, multiple labs, and culturally and academically diverse members of staff might have difficulties in understanding their inner social dynamics. By gaining access to such a remarkably productive institution,

and interviewing the key players through a social science scope, I suggest that they can be informed about their own advantages from a social perspective; something which tends to be invisible to the scientists within the closure of their laboratories. In addition, I should mention that during the course of my study (and, alas, during various moments of hype peaking), members of the centre have been informed through my research about social dimensions relating to their own field, resulting in interesting conversations, which, in turn, as some informed me, have been incorporated as components of introductory courses to AI they have used with their own students. Such references to the social dimensions of their own work either motivated some to participate in the first place, or was found to be beneficial after the interviews. A valid criticism of this influence on my behalf of the interviewees is a reminder of the STS interpretations (and Constructive Technology Assessment (CTA), in particular) of the "going native" question in anthropology. Does the researcher have the right to influence the studied actors? Would, or should, that even considered to be a duty? While these dilemmas are typically treated as limitations to the research, I adhere to CTA's embracing of a "soft" influence as part of a co-evolutionary approach to sociotechnical change, not as elements of activist, or confrontational, "push," but as an interactionist call for reflexivity by admitting the insertion of the researcher at different levels of technical debates (from everyday practices, to networks (in my case), up to policy transformation). In Rip and Robinson's words:

> "CTA agents are change agents, but softly, through support and attempts at opening up, rather than pushing. If there is pushing, it is a push for more reflexivity. […] When moving about, it is the CTA analyst (as a social scientist) who inserts herself in these worlds. But in doing so, she leaves traces and thus creates small changes: the CTA analyst is already a CTA agent" (Rip and Robinson 2014: 136)

As the authors further suggest, a "key element in achieving these objectives is making visible what was invisible to actors, not by explaining (although that might occur), but in interaction with actors" (Rip and Robinson 2014: 139). I suggest that my interviewees admitting their incentive to research such topics deeper and become increasingly sensitised about them, past the interviews, confirm this. While I do not claim that my own "insertion" was influential beyond the micro level described above (raising awareness among individual researchers about social dimensions of their specialisation), and while this influence was not planned as part of the research in order to report back about it, it is situated within a broader emergence of critical AI studies, involving multiple strands from the humanities informing AI trajectories. I suspect that my questioning of AI researchers as to their knowledge of ethical debates and social issues of AI has found equivalents in similar research studies of AI specialists carried out simultaneously. Current[25] macro-level regulatory schemes of AI technologies are indicative of multiple micro-level interactions between AI and humanities specialists. Quoting the same authors: "As a knowledgeable visitor, and based on her diagnoses of the situation, the CTA analyst can actively probe views and interactions, so as to find out about the forces at play. […] [T]he insertion can continue over the course of a few years, so that changes can be traced." (Rip and Robinson 2014: 138).

Access to the site was secured via a double route. Firstly, an anonymous academic who was connected to a highly positioned member of the institute made a mutual introduction during an informal setting (although, I am uncertain as to whether a funeral can be considered informal). Around the same time, another important member, and Deputy Director of the centre, agreed via email communication to facilitate access and connections to members of staff as potential respondents. Both gatekeepers have been

[25] 2022 being the time that these final notes are being added to the finalisation of the thesis.

generous in referring to other respondents and around half of the centre's members did participate in my research (more details about demographics below). While the centre alone may stand as a unique locale to conduct fieldwork, both collaborating Universities employ key players in AI and robotics who are not necessarily affiliated with the centre, but teach in the same departments involved, and their contributions to the research was not rejected when a respondent referred me to them; the same was true about researchers beyond the two Universities who, however, worked or studied there. During earlier stages of the research, alternative sites have also been considered, however, there was no need of advancing such plans.

A few remarks need to be made about the epistemological hierarchy of the centre. When looking at the multi-faceted mapping of the site (developed by myself and not available in such a form online), one is struck by the abundance of labels such as "institutes," "groups," "laboratories" (more often denoted as "labs"). For example, what is labelled a "lab" turns out on inspection to involve a senior researcher's personal project turned into a working group which employs younger researchers (such as Fellows and doctoral students) working on a more specific direction. Such labs are usually (not always!) constituents of research groups or institutes working on more general "umbrella" terms of the relevant area, with variations of collaboration being forms as more or less temporary clusters, some of them being more lasting than others. Often, there are material components to such labs, but this is not a prerequisite; likewise, multiple "labs" might employ the same workspace to test robotic applications, and so on. Association of individual respondents to such divisions allowed me to think more easily in terms of researcher seniority when analysing findings. The type and number of "labs" led by a researcher was indicative of their experience in the field, in addition to all other available credentials. Relatively younger researchers' descriptions allowed a better understanding of the AI hype as it is experienced by AI researchers who might have no prior knowledge about AI's long historical negotiation; relatively senior researchers' descriptions allowed connections between their assessment of the promissory environment in order to situate it within a longitudinal percept of AI, as to include younger researchers' apprehension of it too.

2.3.2 The Interview Method and Further Notes on Methodology

Given that this project investigates the complex nexus of expectations and expertise, not only in the present, but through a lens of temporality, interviews allowed the exploration of the expectations/expertise interplay as they were in previous decades and in the present, in addition to already situating the findings below within the historical outlines above. Simply put, respondents can think back and compare what it meant to do AI "back in the days," "before the hype," today, and in the future[26]. Moreover, the lively and direct style of the interview process revealed impacts of imaginaries and narratives as described above, paired to spontaneous remarks which triggered additional theoretical investigation specific to different empirical chapters.

The primary unit of analysis stemmed from interviewing AI researchers, further aided by an extended documental research of published material to conduct the above literature review (technical books and articles with descriptions of AI, policy and funded project reports, histories of AI), which was, to minor extent, guided by specialists' recommendations. My research was enriched by participant observation during almost one hundred AI-related events, with occasional interventions at Q&A sessions

[26] This, of course, depends on the respondent. Some preferred to avoid speculation and this was among the main reasons they agreed to be interviewed for a project on the performative dimensions of expectations.

to enquire about specific terminologies or views about AI promises and expectations. I was also benefited by presenting some of my early work (Galanos 2018) amid technical experts at a conference organised by the International Federation for Information Processing, one of the major computer science organisations since 1960; discussions and observations during this massive conference were also beneficial towards getting a better grasp of the technical expert's conference life. Some events bringing together technical experts and social scientists with interest in the social dimensions of AI were co-organised by myself as member of the Edinburgh-based AI Ethics and Society research group[27]. Hence, some might argue that the present work contains element of ethnography, therefore, before proceeding with the technicalities of the interviewing method, I wish to offer some clarifications regarding ethnographic research. While the term is open to different interpretations, for the context of the present research I discard its classical STS meaning, that is, participant observation within a relatively defined setting (such as a laboratory). Ethnography has been found quite inadequate during my first year's pilot studies in order to research such deep topics, mainly because ethnographic settings are quite open and eponymous, thus, not allowing the confidentiality expressed during one-to-one discussions. However, a rather relaxed understanding of ethnography would suggest that my research drew a lot from ethnographic experience through my physical presence and visits to my interviewees' working spaces (sometimes homes) and numerous AI-related events and conferences I attended in order to familiarise myself with terminologies and gain an overall image of AI practical specialists' routines and scientific attitudes. Following research challenges similar to those outlined by Nick Seaver's investigations of algorithm conceptualisation, I adhere to his approach of "scavenging ethnography" (Seaver, 2017) which suggests that all bits and pieces of participation at events, follow-up formal and informal conversations and spontaneous online or offline communications, and engagement with various material documents (from pamphlets to books) are all part of such an ethnography. Certainly, the world of AI is full of what Seaver calls "terminological anxiety," supplemented by jokes, puns, ironies (all meta-features of showing how human intelligence plays out in machine's (in)ability to showcase intelligence); hence, this slow immersion in the field between the years by attending multiple conferences, presentations, and workshops, revealed the dynamics of defending one specialist group's perception of AI against another's (for example, a machine learning specialist arguing against a symbolic AI researcher) while at different circumstances they might both defend what appears to be a "unified AI" (for example, against a grouping of cognitive psychologists or philosophers who criticise AI's integrity). So much for ethnography; now, to the technicalities of the interviewing method as applied here.

Following Mason's (2002) instructions on building up a qualitative interview, I planned a semi-structured interview (or a "conversation with a purpose" as described by Burgess in Mason 2002: 62). Qualitative interviewing builds upon the ontological position that people's direct forms of expression are instances of the reality to be explored (Mason 2002: 63). First order data were extracted deriving from my interviewees' lively reactions, the situated knowledge of dialogue exchange guided by a topic which interests both them and the interviewer. Added to my initial introducing questions some improvised follow-up, probing, and specifying questions were performed to navigate the topic closer to the desired first-person experience, intentionally letting few seconds of silence to occur sometimes. Such techniques allow the participants to elaborate at ease, often employing interpreting questions, to verify my understanding of their positions (Bryman 2008: 445-447). The participants' passion about the topic ensured detail, vividness, nuance, and richness (the criteria for structuring a good interview by Rubin & Rubin 2005: 129-134). Indeed, there have been cases (three of them quite characteristically) where I

[27] More on the group's activities: https://www.ai-ethics.org/

started with one single introductory question finding myself asking two or three follow-up ones during the course of 60-90 minutes, usually starting with the phrase "you keep responding to my questions without me asking them, but I would be curious to know more about…". Given that I was interviewing specialists, I became aware of the discussion regarding power and dominance when interviewing elites (Kvale 2006). However, this has been counter-argued by Smith's (2006) argument for the non-existence of an elitist/non-elitist barrier in social research interviews. I offer more details about such considerations in the next section, concerning the ethics of interviewing academics.

A note must be made about the choice of terminology as to how I refer to the specialists I have interviewed. Given that "expertise" is a slippery term which is precisely being negotiated in the context of this work (as discussed in the theoretical section above), I have avoided using that word which is also loaded with potential irony (consider the widely used phrase "so-called expert"). I have contacted people whose institutional profiles contained relevant terms, such as AI, machine learning, or robotics (and interrelated branches), hence their publicly displayed credentials indicated expertise. There have been at least five cases in which respondents emailed me back expressing their will to assist, however admitting that they are "not exactly an AI expert." It proved, exactly by interviewing those people, that by denoting their humble non-expertise, they confirmed they knew more about AI/robotics and relevant areas than others did who simply agreed to be interviewed without casting any doubt on the area of expertise. In a sense, this confirmed MacKenzie's certainty trough: the *more* and *less* one knows about a subject, the more aware they are of the uncertainty imbued in it; in-between, people with partial knowledge tend to be more certain about a given technology (MacKenzie 1998; more on this in chapter 5). Their depth of knowledge of AI matters made them quite confident about their uncertainty on whether they are AI experts or not; with some, this unfolded in discussions as to whether *anyone* can claim they are AI experts! For the course of this thesis, I will refer to them as applied AI/robotics specialists, or simply "specialists" with occasional reference to their exact specialisation when relevant for comparisons. An analytical table containing all participant names (pseudonyms), specialisations, and other demographics is provided below in section 2.3.4. I reserve the term "specialists" to further denote their practical experience with basic science and technology, tentatively distinguishing them from less directly involved stakeholders (such as funders, public commentators and influencers, philosophers, and, alas, social scientists), again, thinking with the certainty trough model (MacKenzie 1998).

Considerations have been made as to decision between reference to these specialists as either "participants," "informants," "subjects," or "respondents" (for a detailed, yet, straightforward discussion on the differences between these terms, see Morse 1991). For reasons of practicality, I preferred avoiding entering the debate, mostly employing the terms "interviewees" denoting the certainty of an interview that took place, and "informants" in that indeed the specialists have informed my view on the topics. Some further notes on my degrees of interference: I aimed at listening to what they have to say about the topics of interest. Sometimes, prior to responding to my questions, specialists were enquiring about my personal prior knowledge of these subjects. My honest response in several cases, given that I had conducted documental research on technical manuals and historical accounts of AI was that "I do have certain degrees of knowledge, but for the purpose of this interview, I pretend to be a blank slate." Thus, such AI/robotics practitioners informed with their practical knowledge of their domains my analysis which is rooted in social science, as they would for any non-technical, yet interested audience.

Interviewees were informed of the nature of my research project. When initiating contact, either after finding their institutional email addresses on the centre's website, or via respondent referral, I offered

a brief abstract of my research goals and questions, allowing them time to process and ask for any clarifications on my behalf, always eager to provide with further details prior to the interview and after. A shorter version was contained on the email body and a longer (one-page) was attached. Candidates who responded positively filled in consent forms prior to the session where they (a) agreed on the confidentiality which is kept, (b) were informed about the password-protected storage of the recording sessions on my personal computer[28] and my intention to send them transcriptions of the material which they had the right to assess, comment upon, and correct possible errors or fill in inaudible gaps, and (c) were asked to sign for their general agreement in participating, and their power to withdraw at any time, acknowledging the fact that they might be quoted anonymously. One might argue that the interviewees' right to correct/alter their responses upon receiving transcripts, might be unfair to their original spontaneous formulations which might capture a more "honest" reality. While I admit that this might be possible for other types of research, my decision was justified by the purpose of this project which leans towards exactitude of opinion and not originality of expression. In other words, it matters little to me whether the informants have expressed something they find to be incorrect or irrelevant – the project is in need of well-formulated opinions that will shed light on the topic – I am not studying the persons, but the field. The Appendix contains copies of both documents candidate participants received (project information/objectives and consent form) as well as the interview schedule I held during the process.

Before proceeding to the ethical challenges of the methodology, I wish to make one final point with regard to the relationship between the interviewing method and the aforementioned "scavenging ethnography." Offering a "safety net" of control to interviewees over their data, access to transcripts in case they wanted to test for errors, their content was protected and greater exactitude with regard to their responses was achieved for the benefit of the research. Interestingly, this secondary communication once I had prepared the transcripts resulted, at the very least, in more complete interviews with fewer parts inaccessible due to inaudibility, and at the most, in interesting follow-up email conversations. One retired interviewee invited me to his home to spend an evening for correcting a few errors. This was followed by an excursion to his personal library, collecting AI/robotics-related material from the 1950s until today; this evening was of particular importance for the development of the research, as I had the opportunity to take pictures of documents which I later researched at my own pace, expressing the AI/robotics hype of previous generations.

2.3.3 Ethics of Interviewing Academics: Anonymity, Mind Games, and Challenges

Beyond the formal procedures of establishing rapport and filling concise forms of consent (Thorne 1980), I had to confirm the confidential nature of what individuals shared. For example, when I asked participants to give examples of recent projects they worked on, sometimes they mentioned explicitly that they can give a very tentative outline as the rest is confidential on behalf of the funding body. Often, the findings of such projects are protected by agreement with funders or for fear of plagiarism, prior to official circulation[29]. Involvement within specialists revealed academic alliances or hostilities. The extent to which

[28] For the recordings, I have used a Zoom H1n Handy Recorder, a high quality dictaphone, saving the archives in MP3 format of 320KBPS frequency. Once recorded on the dictaphone's digital memory card, data have been transferred and stored securely on two separate personal external hard disks.
[29] This itself speaks to the expectations literature, as such confidential material, if analysed, might reveal the valuable differences between promise/expectation and actual progress/state-of-the-art/outcomes.

respondent referral created a tentative bias on my research towards a specific tendency within the centre remains an unknowable, however, the occasional external interviewees confirmed that my findings were sufficiently indifferent to particular departmental interests. Respondents were reminded that this project did not study the content of the technology, but the social aspects of it; happily, this enthused many of them who claimed to have never thought about these dimensions.

During the planning stage of the research, a question of power arose. How does a doctoral student study academics and researchers? Interviews are often considered ethically challenging a method, because of the assumption (often a reality) that the question of vulnerability in social science has to do with protecting "vulnerable" groups (such as victim*ised*, patient*ised*, in general vulnerabl*ised* people) from the "invulnerable" prestigious researcher. Such a point of view is not relevant when researching researchers, sometimes thought of as "elites" (I will explicate below why I contest this term too in my study). One of the mostly cited pieces of research about methodological and ethical considerations when "studying up" (military specialists in particular) offers an overly dramatic image which nonetheless captures a number of relevant issues in a caricatured manner:

> "To become too closely identified with one party can close down access to other groups and individuals. The process of distancing is, however, far from straightforward with the powerful. In the environment of secrecy, where access is limited and precarious, contacts and informants have to be cultivated. Trust and confidence is vital in such circumstances. To cultivate trust some degree of empathy, even sympathy, is needed. To be guardedly neutral is not enough. There is a thin dividing line between building up trust and becoming too closely identified with a particular contact." (Williams 1989: 253).

Such uncertainties have been reflected in recent pieces of literature and have been epitomised in Lancaster's suggestion "that there is a need to reconceptualise notions of authority, sensitivity, vulnerability, and power not as fixed qualities inherent to the researcher or the participant, but rather as fluid and relational" (Lancaster 2017: 95). While fluidity and relationality may appear as appealing in contemporary contexts, their practical reality is problematising ethics.

The difficulty of stating a precise ethics of "studying up" lies in that is a can of worms that once opened a number of "what if" questions emerged and intermingled. This uncanny can of worms relates to a very realistic paradox about the meaning of confidentiality. As stressed by Yu, while "confidentiality is about keeping data revealed by the participants to the researcher him/ herself, further implying that no one besides the researcher would be able to access the data," this is not really the case as certain quotes, or interpretations of statements are used in research, ending up in publicly available Theses, papers, books, and other formats of publication (Yu 2008: 163). Hence, to start engaging with the problematic of confidentiality, it is useful to state that "results of this research will be used for academic purposes only, which is quite different from the common promise given" (Yu 2008: 163). This was done and agreed upon with the interviewees. Then, one must examine more closely the previous works dealing with the nature of the problem, in order to discard certain options. Hence, I hereby present issues identified in the literature concerned with studying "elites," "researchers," through the filters of "anonymity" and "confidentiality." Generally, most authors admit fluidity and case-specificity with regard to anonymity; however, when it comes to cases similar to the one examined, anonymity appears to be preferable.

A typical concern is that the interviewee's specialised position tends to dictate the discussion about what is relevant to the examined issues creating a barrier between the informant's authority as an

"expert" and the junior researcher's "inexperience." According to Lancaster, who studied policymakers as a PhD student, such dynamics tend to be very fluid and dependable upon the informants' prior knowledge of whether their names, hence, knowledgeability, will appear in wider audiences (Lancaster 2017: 97). To this issue, my own experience was much milder. All interviewees have been excited by the fact that someone interviewed them and although sometimes they have tried to go deeper into technical issues to explain certain points, I would not claim that this dictated the discussion, but allowed them to give a more complete answer in the end – usually, a quotable one[30]. Once again, in the case of AI where, as I argue in Chapters 3 and 4, "what is relevant" is also dependent as well as it shapes "what is AI" – therefore, the more AI researchers tried to convince me about how their specialisation in one aspect of AI or another is relevant, was actually contributing to understandings of the social shaping of concepts, promises, funding strategies, and governance.

Keeping interviewees anonymous or not is considered to be a trade-off that shapes what they report during an interview (Farquharson 2005: 351 in Lancaster 2017: 102). The response to this is that despite the potential will of some informants to be advertised through the research, a full anonymisation will ensure the richness and sincerity of their responses. To make the real names of my interviewees and their respective institutions would definitely add prestige and realism to my research. However, prestige is often paired with gossip and layering of honesty: if one knows that what they are saying is attached to their names, one might wish to be more careful and self-constrained. At the same time, several interviewees expressed their will not to remain anonymous if possible without giving justification. Some have given interviews for the press, and they might have considered this as something to be added on their list of works with non-technical outreach. One interviewee (and indeed a retired "explorer" of science, as argued in Chapter 6) was explicit to be recorded saying that if there is a chance to bypass formalities, he would like the interview to be publicly available with or without his name associated with it. On the other side of the anonymisation spectrum, an interviewee pinpointed specifically towards the recorder and prior to sharing a very important piece of information, he started with the phrase "and this will be for the anonymisation part, ok?." As Vainio has succinctly put it in more theoretical terms, "[f]irst, ontologically, anonymity is a way of turning into 'data' what someone has said or written. Second, anonymization as 'analysis' turns the participants into examples of specific theoretical categories, and as such is a part of the data analysis" (Vainio 2013: 685). Anonymity was particularly useful when my interviewees expressed gladness being given room to express their honest opinions about the research world without feeling that their status is somewhat threatened. Anonymity, after all can be seen as professionalism on behalf of the researcher (see Yu 2008: 167), and a way to ensure another type of scientific rigour, not through specificity of names, but through politeness of attitude, with respect to individuals' privacy.

Certain sections of the interviews that were found to be particularly sensitive, either by myself as a researcher, considering that some references might identify individuals, or by interviewees asking for such protection themselves, were treated with care. For example, there have been cases where critical statements about other scholars, or relevant influencers or policymakers were made, references to interdepartmental rivalries and (un)intended criticisms towards other members of staff. In such cases, the dilemma on the researcher's behalf lies on whether one should let informant decide/have control over the data or being

[30] Some of the findings about incremental innovation in chapter 4 are good examples. A senior researcher insisted that prior to conducting my research, I should first try and code some software as a way to understand the chores of not getting the required result because of simple misspelling – something which, if followed by most speculators, would decrease the amount of overpromising.

"protective" over them and anonymise them for their own good (yet, thus echoing a superior/invulnerable researcher status). According to Wiles et al, these two orientations "can more usefully be seen as extremes on a continuum with researchers located at different points according to their research approach, the context of their research and the specific issues they faced" (Wiles et al 2008: 425[31]). Again, following Lancaster, such sections are "treated with care in analysis" and in the case of interviewees requesting to see how data will be used, this was carefully negotiated "to ensure that participants [do] not have 'control' over what could be reported (analysis and interpretations) while also respecting participants' concerns so as to maintain access and participants' ongoing consent" (Lancaster 2017: 100). Wiles et al propose a particular methodological schema on what it means to treat such sections with care when stating they had to:

> "[…] be mindful of and to attempt to distinguish between: 1) what was public knowledge in terms of our participants' expressed views in their presentations and their research; 2) what was data generated in our study for public consumption but which must be anonymized; and 3) what was private knowledge that we had gained from our research that we did not have individuals' consent to use, or knowledge gained from our personal contact with an individual." (Wiles et al 2008: 291).

Such decisions were made when interviewees spoke of their relationship to funding bodies and policymakers, as well as particular research projects. It remains unknown whether, due to high confidentiality of a given project, despite the anonymity (or because of it), interviewees might have veiled the existence of projects relating to "great" expectations; however, I am quite confident that their experience on such topics has been distilled in other answers they have given. It would be unfruitful to completely disembody the interviewees' names from their specialisations, because what they state might be indicative of their academic seniority, age, gender, or other variables. Therefore, the pseudonyms chosen are playful slight indications of their specialisations, and when quoted as part of the findings, I may refer to relevant elements of their demographics, when this relates to the analysis – in most cases, it does not. As one interviewee told me when I asked him to describe his specialisation, "just by saying this, it identifies me." A question relating to this had to do about how much value one should put in descriptions of researchers' personal lives. For example, a handful of researchers, when asked about reasons that motivated them to apply to the department they were based at, explained that a life decision (such as family-making) dictated their professional career; something which incentivised the inclusion of serendipity as a factor in assessing research cultures (as opposed to linear trajectories of progress). Hence, I disagree with Alvesalo-Kuusi and Whyte's straightforward and aphoristic response to this: "Those of us who research the powerful are generally not interested in their personal lives" (Alvesalo-Kuusi and Whyte 2018: 147). Nevertheless, these details are never shared or mentioned throughout the thesis. Reed et al have suggested a more opportunistic approach to departmental conflicts, which is transforming the identification of such discrepancies into a tool, namely, conflict management, that is, a way to speak about a phenomenon through the interdepartmental and other types of rivalries (Reed et al 2009). Therefore, both personal lives but also inter- or cross-departmental conflicts can be indicative of expectations setting.

A final challenge involved with full anonymity, as Wiles et al. flag, is the possibility of people propelled to making guesses about small samples (Wiles et al 2008: 425). As per idiom, "it is a small world"; researchers or other people associated with the topic may be interested in the identification of the

[31] It is worth stressing that Wiles et al studied couples and families; hence, their context, differed a lot from researching researchers – their finding, however, is relevant for the present research and applied differently than Wiles et al's research.

centre studied and given that while there are enough such sites to generate an ambiguity, there are not so many to ensure full anonymity after a dedicated investigator's research. This is acknowledged in the consent form by stating my attempt to "ensure full anonymity as much as possible." Which was further approved by the University of Edinburgh's School of Social and Political Science's formal procedure in obtaining ethical clearance for the project.

2.3.4 Analysis and Presentation of Data

I have transcribed on my own using a combination of personal skills and the aid of a trusted speech to text software. This resulted to over 500 pages of interview transcripts, which, given the *semi*-structured nature of the interview and my intended will to let interviewees expand on topics with minimal guidance to ensure authenticity, made quantified coding rather difficult, if not pointless. Initial stages of analysis involved coding through the software NVivo 11. NVivo allows users to mark content of the interviews and create tags of coding ("nodes") which helps navigate and compare pieces of text. While this was proven to be useful during preliminary analysis, the rather unique approach of every researcher's style of expression, as well as the intentionally open-ended style of the questions asked, did not allow for a strict coding. I would suggest that it might be a peculiarity of qualitatively interviewing academics, that a mixture of their enthusiasm, passion, personal agendas and speculation about/interpretation of certain words, often leads into conversations which, if transcribed, tend to resist the constraints of rigid, closed coding. Nevertheless, an experimental feature of NVivo (at the stage when it was used), allowed for pattern-based coding, that is, the software recommended relationships between pieces of text based on previous coding. While this feature was interesting to make some initial thoughts, the complexity of the subject as described in theory (for example, the subtle differences between future-oriented behaviour and the way these might be expressed) would make consistent use (and training) of NVivo more of a hurdle than a solution. Moreover, the possibility of existing and recommended coding creating an "NVivo bias" made me rely purely on my own analytical skills. Rodik and Primorac (2015) is an excellent outline of arguments against the great expectations (in accordance to the present work's subject!) associated with Computer-Assisted Qualitative Data Analysis Software (CAQDAS) in social sciences.

The semi-structured interview schedule (available in the Appendix) was prepared according to the tripartite scheme suggested by Legard et al (2003) for in-depth qualitative interviews. This scheme divides questions into themes of:

(a) Contextual information and background. I used this set of questions to gain information about the respondent's specialisation, motivation to work in the field (which often included hints about expectations during the time of the researchers' early studying years), the institution they are working in and their working routines, involving recent projects and funding strategies.

(b) Ground mapping questions. This section enquired on definitions of AI, current state-of-the-art, obstacles to innovation, and the possibility of AI's transformative powers (dimension mapping questions); this was followed questions regarding AI's history or controversies, previous and current rounds of promises and expectations, and how this is felt and performed on behalf of the researcher (perspective widening questions).

(c) Content mining. This section was less strict in its unfolding and was highly dependent on the previous course of every conversation. In several cases, interviewees responded on questions contained here as parts of previous responses, when it came to issues of ethics and other social challenges, governance, policy, public portrayals, and science fiction.

As mentioned, the results were very rich and the empirical findings presented in this thesis are but a fraction of gathered material. For example, I have not included very important observations interviewees have offered about their own relationship to science fiction, their views on public outreach, or their technical views on what would make an AI breakthrough[32] (I employed that term in a playful manner, acknowledging the highly expectational nature of it). I have used thematic analysis of emerging patterns following a mix of approaches, mostly following the open coding approach (Robson 2002), that is, interpretative analysis as the interview material accumulated, followed by additional and concurrent literature review conducted along the way and the iterative emergence of themes. In presenting data, I adhere to the displaying of open coded data, through the introduction of interview vignettes. Simply put, I treat transcripts as rich text sources which show the different ways interviewees speak about the theoretical concepts, issues, and challenges addressed above.

Thematic analysis (Nowell et al 2017) poses a number of benefits and challenges that space does not allow for depth consideration. Most importantly, thematic analysis benefits by the organic, real-life emergence of themes that are to be problematised and mainly suffers by the question concerning credibility and establishing trustworthiness in the presentation of data. I am particularly sensitive to the problem of confirmability, that is, the confirmation that the researcher's "interpretations and findings are clearly derived from the data" (Nowell et al 2017). Through the display of descriptive data, I suggest that such interview vignettes can stand as evidence and confirmation of interpreted material. Such vignettes can be imagined as the equivalent of interview footage during an educative documentary, preserving the respondent's original expressions, enthusiasm, irony, humour, and allowing the reader to judge whether the conclusions derived are conclusion based, if not on "real life," at least on the lives of sampled interviewees. This style of interpretative data display was also employed by Shoshana Zuboff's 1988 *In the Age of the Smart Machine: The Future of Work and Power*, which I found very effective in presenting the qualitative livelihood of workers in areas traditionally considered as of very quantitative nature (e.g. Zuboff 1988: 27-95). Zuboff's work is probably the earliest precursor to theses like the present one, presenting empirical data on similar areas in a thorough manner; and hence, data display in this way alludes to a tradition of interview-based AI developer studies.

Only one out of a further set of 25 invitees responded negatively, expressing his will to assist, yet his unavailability due to workload. There can be imagined several interpretations for the 24 who did not respond to invitations (my strategy involved sending each person an email up to a maximum of three times). Workload, inbox overload, and possibility of spam, and general disinterest on the topic can be few of them, however, this is just speculation. I did not observe any notable pattern dividing those who did respond from those who did not, as both sets involved approximately equal representations of academic seniority, age, gender, and technical specialisations.

2.4 Closing Remarks

[32] It is my intention to return to this material and prepare relevant future publications on these three topics.

Given the qualitative nature of the research and the availability of specific individuals who shaped my empirical findings, it is certain that the empirical results could have been very different should a different sample have been studied or a different theoretical toolkit. It is my conviction, however, that the following empirical findings are of sufficient generalisability given the historical support in tackling the question concerning expectations and expertise, and can be found useful not only to STS scholars and social scientists, but also members of the AI community with interest in the social dimensions of AI and robotics, as well as policymakers and further members of groups of interest who wish to gain a closer look into research cultures when in discussion with their members. With this summary of theory and method employed being complete, it is now time to proceed with the presentation of historical and empirical findings.

CHAPTER 3: A HISTORY OF AI DEFINITIONS, PROMISES, AND WINTERS

This chapter consists of an extended critical literature review of AI's history, rereading it through the lens of its conceptualisation (3.1) and the promissory negotiations from its initial conception until the writing of this thesis (3.2). This historical review contains original interpretations of the AI's history and offers archival findings that have not been situated within STS/historical accounts of AI before, mainly due to the lack of historical awareness in contemporary attempts to map expectations and promises of AI. While initial writing of this chapter was aimed at becoming part of a literature review, it was proven that additional research had to be conducted in order to clarify certain connections in AI history, and these, in turn, revealed the value in applying SoE and SEE vocabularies in the past. Therefore, section 3.3 will summarise the findings from sections 3.1 and 3.2 and offer reflections which contribute to a longitudinal assessment of AI expectations and expertise.

To my knowledge thus far, there has been no attempt to comprehensively trace the social-historical-political shaping of AI's meaning, and hence the present chapter will be of particular relevance to any scholar, policymaker, or AI-related worker who is being requested to define AI for drafting a document and stumbles upon problems of lack of consensus. Tantalisingly, besides some small collections of definitions and taxonomies (reviewed below), no other piece of work appears to exist comparing AI definitions and descriptions, so this adds further historical value to the present thesis, while it sets up the scene for arguments better explained through empirical statements by researchers offering their views on AI definitions – themselves partly impacted by AI's social-historical construction. Hence, this section aims at reporting AI's interpretative flexibility (Pinch and Bijker 1984) as a product of its history and the moulding by different arenas of expertise during periods of hype and disillusionment. It critiques the hype cycle notion as reductionist, diverting one's view from silent, yet important incremental innovation, taking place during the waiting games without hype. It also finds that periods of so-called disillusionment, if examined carefully and with less focus on funding rounds, reveal that AI gains in popularity through non-AI scientists during those times.

3.1 A Timeline of AI Descriptions and Definitions, Events, and Development of Applications

The present section takes a closer look at AI's historical development, its landmark events and technical applications developed attached to the evolution of AI as a field, closely inspected in relation to failures in AI's canonical history, periods known as AI winters. This timeline is sprinkled with definitions and descriptions of AI throughout its development showing shifts in expectations and perceptions about the field's purpose and essence as shaped by its technical developers, further shaped by broader discourse, eventually – relatively – crystallised in policy domains by regulators. The timeline's division is unavoidably arbitrary, partly based on canonical histories of AI, but partly based on my own interpretation of themes in its development and focus on winters – it should be taken with a pinch of salt. A first central argument advanced here is that a critical rereading of AI history reveals that the fable of the AI winter is a

convenient reductionist tale which veils AI's incremental innovation as a field consisting of multiple branches, an ensemble of different specialists who, in different epochs, might make more or less frequent use of the term AI. A second argument is that AI's initial conception as a broader scientific field, because of the AI winter's performative aspect and the increasing demand for practical applications on behalf of funders, makes AI increasingly being perceived as a technological field. This bears implications for the examination of research cultures as examined in the following chapters, as acknowledgement of an applications-oriented funding and policy landscape minimises motivation for visionary basic ("blue sky") research.

The documental analysis below, in conjunction with the empirical statements about contemporary AI specialists' understandings of AI (in Chapter 4) will offer a novel subversive reading of AI's history, as it is being shaped by an interplay of expectations about what AI means and should be and what circumstances, institutions, and individuals co-determine it. An interesting observation, looking at AI history through a timeline of its descriptions, is that AI appears to gain popularity in general discourse during eras conventionally thought of as those of disillusionment; therefore, I suggest that disillusionment chunks of time can be viewed as "waiting games," to borrow a term from Bakker and Budde (2012), and hence my division into five historical sections, three of hype (and increased mobility of arenas) and two of waiting games (where AI is sustained by non-specialist discourse about AI).

a. Early foundations of AI, promises and early disillusionment, 1955-1975.

Like with many fields, it is difficult to locate a specific beginning to AI. As mentioned earlier, the fascination with replicating intelligence or body functions is a trait which can be found in ancient mythology, religion, philosophy, and alchemy. This means such mixed discourses and narratives preceded any modernist scientific treatment of AI. Cybernetics was the main modernist scientific precursor of AI, given that (a) it was the field which first studied systematically control and communication processes in animal and machine systems, and (b) three out of four researchers who proposed the term were participants at the Macy Conferences where cybernetics was developed. In September 1955, young assistant Professor in Mathematics at Dartmouth College, also working at IBM, with the assistance of Marvin Minsky, Claude Shannon from the Bell Laboratories, and Nathanael Rochester from IBM, submitted a funding proposal to the Rockefeller Foundation for a summer workshop which would aim, precisely, at synthesising and clarifying the simultaneous advances in what was rather informally called "thinking machines" or more formally "automata studies" and "complex information processing," as a unified field termed "artificial intelligence." The proposal did not provide with a specific definition, although, after the term was proposed, it read: "The study is to proceed on the basis of the conjecture 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 1955: 14). AI began under the hypothesis that to the extent that intelligence *exists*, there should be a way to break it into parts (processes and mechanisms), describe each and every one of them, and find ways to simulate it. The summer workshop brought together competing researchers in the field which was already in the making with several alternative name proposals. These early discussions on AI resulted in approaches which involved digital representations of knowledge and semantic networks connecting them which would allow the development of game-playing programmes based on heuristic search: according a specific mental representation of knowledge, a mechanism "seeks" for the optimal solution to solve a task. Newell and Simon (1960), also participants at the Dartmouth

workshop, have suggested and for some years defended the greater preciseness of the term "complex information processing" instead of AI.

One year after the initial formulation of the AI proposal, a second approach to electronic computers was proposed, that of the 'perceptron model.' While no reference to AI was given, it is useful to read the definition provided as this has been further encompassed as an alternative route to AI methodologies by Minsky himself and his colleague Seymour A. Papert: "The proposed system depends on probabilistic rather than deterministic principles for its operation, and gains its reliability from the properties of statistical measurements obtained from large populations of elements. A system which operates according to these principles will be called a perceptron" (Rosenblatt 1957: 2). This paved the way for early pattern recognition, such as letters and basic shapes, for example in Rosenblatt's machine Mark 1 Perceptron (and thus was the most important predecessors to contemporary data-driven recommendations based on pattern recognition within large datasets). Conceptually, the two approaches signify that the assistance of machine replication of intelligent processes can shed light on the nature of intelligence, one side of which being rule-based, hierarchical, reasoning through sequences of deductive syllogisms, and the other being that of experiential learning based on trial-and-error according to interpretation of givens ("*data*" in the Latin sense of the word). For Dreyfus and Dreyfus (1995), the core difference between the two approaches has to do with "creating a mind" (that is, creating machines who think based on symbol manipulation in McCarthy et al's version of AI) versus "modelling the brain" (that is, creating machines based on the assumption that brain neural networks can be imitated by machines). As Olazaran (1996) has shown, however, this sharp binary is quite superficial in AI's historical assessment. Two years after Rosenblatt's formulation of the perceptron, the Simulmatics Corporation was established. Simulmatics, a political and research methods offspring of the think tank RAND (Research and Development) Corporation, with close ties to DARPA, applied data collection and statistical science (the pre-qualitative versions of social science) in order to identify and manipulate voting behaviour (Lepore 2020). Jill Lepore's recent excavational study shows how huge hype surrounded the Simulmatics Corporation due to their deep association with John F. Kennedy's 1961 US presidential inauguration and the similarity between their methods and those employed by Cambridge Analytica to assist the Donald Trump's 2016 presidential election (see additional details on the importance of this event as to AI's expectational environment in the following section). Very much like in the case of Cambridge Analytica, the workers at Simulmatics:

> "[…] ran a simulation through the computer, analyzing the effect of further discussion of religion on each of 480 voter types concerning '(1) its past voting record; (2) its turnout at the polls; and (3) its attitude toward a Catholic candidate.' As a result of this analysis—the computer simulation of an election that had not yet happened—Simulmatics recommended that Kennedy confront the religion issue head-on, not to avert criticism but to incite it: 'The simulation shows that Kennedy today has lost the bulk of the votes he would lose if the election campaign were to be embittered by the issue of anti-Catholicism." (Lepore 2020: 119).

Kennedy took on board the advice and won. Interestingly, at that time, the Simulmatics methodology was not considered to be "AI" and it is unknown whether Rosenblatt would have been in contact with any members of RAND or Simulmatics, although RAND had at the same time employed Simon and Newell to advance heuristic AI; later on (in 1964) hiring philosopher Hubert Dreyfus to conduct the first assessment (and indeed, a polemic) of Newell and Simon's AI, titled *Alchemy and Artificial Intelligence* (1965; see also Dreyfus 2006; 2012). At that time, what is now termed "data-driven AI" or "machine learning AI"

was a separate technical field, awaiting its convergence. I will return to Simulmatics promissory game and the disillusionment surrounding the hype in section 3.2, concerning promises of that era.

In the meantime, DARPA allocated money to the MIT on Project MAC in 1963, for developing an early attempt at real-time access by several users to a single computer (thus a precursor to contemporary internet), while John McCarthy started the heavily funded Stanford Project at the University of Stanford. In the same year, AI centres (employing the term AI) were found in Carnegie Melon and Edinburgh Universities. By that time, and as AI gained traction, competing propositions were made. For example, cognitive psychology pioneer Ulric Neisser, who helped computer scientist Oliver Selfridge (also a participant of the Dartmouth workshop) in developing an alternative perceptron-like model, contested the then predominant AI approach: "The very concept of 'artificial intelligence' suggests the rationalist's ancient assumption that man's [sic] intelligence is a faculty independent of the rest of human life. Happily, it is not" (Neisser, 1963: 197). The mechanistic universe in which AI of the time seemed to belong is indicated in this passage.

In the late 1960s and early 1970s limitations have started to appear. Minsky and others have proposed the concept of "microworlds" in AI, the study of domain-specific intelligent behaviour often overlooked by those who wish to study intelligence at large (Minsky 1968, 1975). Terry Winograd's programme SHRDLU and Carnegie Melon's robot Shakey, were early examples of task-specific applications of the microworlds approach, however, the researchers admitted the difficulty in extrapolating the findings and connect them to broader systems of general purpose intelligent machines (Nilsson 2010). Nevertheless, this came in accordance with the US's Mansfield Amendment, an act which restricted scientific and technological funding to works with direct military application in 1969. Despite obstacles caused by some early distrust such as the ALPAC report on the failures of machine translation in 1966 (more details on this and the Mansfield Amendment in section 3.2), Minsky continued gathering and developing approaches to AI. The following description which is often attributed to as a definition of AI by Minsky, is often cited via secondary or even tertiary sources – it took me a sufficient amount of time to actually get access to the original collection of papers to which the following passage acts as a preface opening paragraph, whereas the "definition" cited consists of only the last half part of the sentence, usually beginning quoting from "making…": "How can one make machines understand things? This body is a collection of studies in artificial intelligence, the science of making machines do things that would require intelligence if done by men [sic]." (Minsky, 1968: v). At least a couple of interpretations may follow by the fact that this definition stayed (indeed, some of my senior interviewees referred to this without referring to Minsky). Chiefly, it is evident that Minsky promoted this field as a scientific one; hence, it suited AI researchers who wished to belong within an established scientific environment. Secondly, one cannot but notice that "intelligence" in "AI" referred to a technical rendition of *human* intelligence as a prototype (and possibly male if one does not take "men" with a grain of linguistic salt).

It should be stressed that none of the above descriptions have been presented in their original formulation as strict working definitions of AI or the perceptron as a subject, and I could not find any document containing such a definition, although definitions tended to be popular in similar technical fields (see for example, the use of definitions in W. Ross Ashby's precursory *Design for a Brain* (1954) or throughout the early cybernetics conferences transactions in which, not only the field is constantly defined and redefined, but sections are dedicated to specific nomenclature, e.g. MacKay 1951 in Pias 2016: 222). To my knowledge, the first definition of AI, or the "ABC of the subject" came from a non-AI specialist (yet, a well-known specialist in fluid dynamics) and AI opponent: Lighthill. Its purpose, I interpret, was to

offer something which appeared to be scientific (and thus thoroughly concise) enough in order for the rest of the criticism to stand as worthwhile by someone who understands the subject so well, so that they can define it, to the extent that can even make alphabetical puns about it:

> "The ABC of the subject. There is a general consensus about which main areas of research are to be grouped within the broad field of AI. […] letter A stands for Advanced Automation: the clear objective of this category of work being to replace human beings by machines for specific purposes, which may be industrial or military on the one hand, and mathematical or scientific on the other. […] letter C stands for Computer-based CNS [central nervous system] research. […] Category C is concerned, then, with theoretical investigations related to neurobiology and to psychology. […] letter B stands not only for Bridge activity, but also for the basic component of that activity: Building Robots. The whole concept of Building Robots is, indeed, seen as an essential Bridge Activity justified primarily by what it can feed into the work of categories A and C, and by the links that it creates between them." (Lighthill, 1973: 3).

It is unknown whether Lighthill's negative assessment of AI, suggesting AI's impractical output is directly or indirectly related to the impact of the Mansfield Amendment. Nevertheless, both assessments, with ALPAC predating them, aimed at short-term practical impact, and, I argue, has further impacts on AI's conceptualisation. So far, Minsky and McCarthy's AI was only peripherally associated with intelligence's physical support (in AI's case, the robot) and automation, as indeed McCarthy, by proposing the term AI wished to step away from "automata studies," i.e. the study of automated robots. Nevertheless, in the public mind, machine intelligence and electronic brains were associated with the popular robotic images and the experimental humanoid robots presented at conferences (for example, see Sluckin's 1960 Pelican introduction to "Minds and Machines," or Strehl's 1952 "The Robots are Coming"). Lighthill's report to the UK government *enacted* (à la Budde and Bakker 2012) a specific view of AI, which, until then, was not crystallised by AI scientists and to a certain extent this might be said about the US understanding of AI too. But until then, the main carry home message was that intelligence was conceived as something that can be described and AI is a means to describe it by offering a way to model brain processes. In a sense, after a long time for AI to become established, its thought of definition was almost equated with its thought of collapse.

b. Waiting game 1: The silent years of applications-oriented AI, and simultaneous popularisation, late 1970s.

Under the context of the above assessments and political repurposing of AI, several funded projects kept emerging, although they were not nominally presented as "AI." DARPA's Speech Understanding Program (1971-1976) and the DENDRAL (Dendritic Algorithm) project are notable examples from that era, which were based on the direction of specificity required by the funders. Researchers like Edward Feigenbaum sought at mimicking specific expert skills by coding certain sets of knowledge representation. This gave rise to the early "knowledge-based information systems" or "knowledge-based expert systems," often referred to simply as "expert systems." Research in this area was incremental and was not characterised by the exaggeratingly promissory environment of the earlier stage. Slow, yet steady advances in computer vision and mobile robotics were made. It is interesting that, as I intend to show here, although research in "AI" was not presented as such, AI remained popular among non-AI specialists who published books partly in response to an existing public excitement about AI at the time. During the post-Lighthill and post-Mansfield AI winter, AI as a concept remained relevant in its non-technical representation within arenas

fascinated by the prospect of AI, but with little or no knowledge about its technicalities. In such books, usually written by scholars who were in partial contact with AI experts, consulting them for their writing needs, AI becomes more closely associated with robotics, while it appears as an established field. Jasia Reichardt's book *Robots: Fact, Fiction, Prediction* offers the following definition: "Artificial intelligence means artificial behaviour, or simulation of artificial behaviour, by computer programs and robots." (Reichardt, 1978: 158). Reichardt, an art curator famous for her exhibition *Cybernetic Serendipity*, the first artistic exhibition focusing on the interplays between control and chance as well as robotic/cybernetic devices and art, published that book for the well-known art publisher Thames & Hudson as an accessible entry to prevalent fictitious and scientific views on robotics and listed numerous AI and robotics experts who were consulted in writing it. It is thus shown that for both public and scientific mind of that time, robots and AI have become mostly synonymous. Philosopher and mathematician Aaron Sloman, in his influential book *The Computer Revolution in Philosophy*, moved away from Minsky's 1969 definition and aimed at crystallising the purposes of AI by adding the philosophical dimensions of AI in its agenda, focusing on AI as a device which would assist the understanding of intelligence proper:

> "AI is not just the attempt to make machines do things which when done by people are called "intelligent". It is much broader and deeper than this. For it includes the scientific and philosophical aims of understanding as well as the engineering aim of making.
>
> The aims of Artificial Intelligence
>
> 1. Theoretical analysis of possible effective explanations of intelligent behaviour.
>
> 2. Explaining human abilities.
>
> 3. Construction of intelligent artefacts." (Sloman, 1978: 17).

A year later, another influential work, cognitive scientist Douglas Hofstadter's Pulitzer prize-winning *Gödel, Escher, Bach: An Eternal Golden Braid*, further subtitled "a metaphorical fugue on minds and machines," almost following Sloman's suggestion (without quotation, however), took AI as a point of departure in order to make arguments about the nature of metaphors in information processing. Hofstadter focused on prospects of AI (e.g., Hofstadter 1979: 129) to show what machines can tell humans about the formation of metaphor. He thus defined AI's purpose purely as a methodological device that aims at connecting:

> "the seemingly unbreathable gulf between the formal and the informal, the animate and the inanimate, the flexible and the inflexible. This is what Artificial Intelligence (AI) research is all about. And the strange flavor of AI work is that people try to put together long sets of rules in strict formalisms which tell inflexible machines how to be flexible." (Hofstadter 1979: 26[35]).

Towards the end of his book, Hofstadter introduces what he calls Tesler's theorem: "AI is whatever hasn't been done yet" (Hofstadter 1979: 601). Hofstadter misquotes personal communication with his associate Larry Tesler, a well-known computer scientist who, in the future careered in companies such as XEROX, Yahoo!, and Amazon. Although Tesler corrected the misquoting several years later ("What I actually said

[35] The large influence of Hofstadter's book is captured in Mitchell's preface to her recent introduction to AI, where she describes how, as part of an advisor to Google's AI programme, she, alongside many other contemporary AI figureheads, was excited to meet Hofstadter as a guest of honour to this meeting, who expressed his concerns about the future of AI, as being "terrified" – something which Mitchell aims to disprove throughout the book (Mitchell 2019: xiii-xxvi, 351).

was: 'Intelligence is whatever machines haven't done yet'," Tesler 2019), this humorous "theorem" gained popularity among AI scholars (including Minsky), associated with the "AI effect" (AI Effect 2019), the idea that whenever a practical accomplishment is made by AI research, it stops being considered AI, dismissed as mere computation instead of genuine intelligence.

Thus, it becomes clear that the period known in AI history parlance as the first major AI winter, was a "winter" only in terms of AI funding. It saw, however, AI gaining popularity in broader discourses (including science fiction) through publications by non-specialists who were expressing interest in the topic, while AI researchers continued to conduct work in areas such as early expert systems which, under other circumstances, might be identified as AI and indeed set up the stage for next rounds of AI hype. The themes emerging about AI are now more stabilised and include: (1) AI as a scientific attempt at imitating brain processes, (2) AI as a technique aiming at formalisation of thought, (3) AI as an attempt at creating robotic intelligent beings, and (4) a more reflexive and humorous understanding of AI as a constant future. Such definitions have their promissory value implicated in ways they have been expressed, thus becoming an active shaping of future agendas.

c. International AI race, second round of disillusionment, and further incremental innovation, 1980-1998

Descriptions of AI offered in this sub-section came during the period of the first international AI race, following the 1980s resurgence for AI with new arenas emerging from previous and new research. 1980 saw the founding of the American Association for Artificial Intelligence (now, Association for the Advancement of Artificial Intelligence, AAAI), together with numerous trade shows and further symposia around the world. In 1982, the Fifth Generation Computer Systems (FGSC) programme was announced by the Japanese Ministry of International Trade and Industry (MITI) to be held between 1982 and 1993. In response, DARPA funded the Strategic Computing Initiative (SCI, 1983-1993) programme to create a large-scale computer-based intelligent system, which would develop further several AI-related technologies, so far developed in isolation, and then integrate these component technologies into a very large system, equivalent to the telephone (this was parallel and closely associated to the development of early internet technologies, such as the Transmission Control Protocol/Internet Protocol (TCP/IP); Roland and Shiman 2002). The UK's response to the Japanese was the Alvey programme (1984-1990), a collaborative effort between government, academia, and industry to compete with Japan. The European Economic Community (predecessor to the EU) responded to FGCS with the European Strategic Program of Research in Information Technology (ESPRIT, 1983-1998), which contained lots of AI-related aspects (Van Hove 1991, Directorate General XIII 1989). On the technical side, important advances have been in preparation during this period on what was known as the connectionist front of AI, but in these decades avoiding identification with the term AI. Notably, Hinton's work on backpropagation algorithms was published in 1986 (Rumelhart, Hinton, and Williams (1986), LeCun's work on bio-inspired convolutional networks was published in 1989 (LeCun et al 1989), Bengio's contributions to neural networks applications to speech recognition and his collaborations with LeCun (Bengio 1993, LeCun and Bengio 1994), all laid the foundations for what was later to be appreciated as deep machine learning AI, but again, not labelled as AI in that time[36]. In 1986, the Conference and Workshop on Neural Information Processing

[36] It is interesting to note that when in 2019 the three researchers received collectively the ACM A. M. Turing Award on March 27 2019, the press (*The Telegraph*, *The Verge*, *TechRegister*, *TechTimes*, and *Forbes*) pronounced them as "godfathers of AI."

Systems (NeurIPS) was also established, first as a workshop attracting the rather few scholars interested in the neural-like computational processes in the mid- and late 1980s, growing to become one of the largest venues attracting machine learning AI researchers – and their promissory games which will be touched upon in section 3.2 (Else 2018). While these figures, techniques, conferences, and research programmes will be examined in greater detail below in terms of their promissory components, it is useful to show how AI was perceived during this time.

This period, in addition to the new research configurations, was further marked by proliferation in publication of books about AI which were aiming to satisfy semi-popular demand, often written by non-AI specialists who contributed to the notion of AI based on their own fields' perspectives. Clinical psychologist Neil Frude, in his popular science book *The Intimate Machine* describes AI as the area "concerned with producing programs which emulate certain human or animal functions, particularly those which we normally refer to as 'intelligent', like complex problem solving" (Frude, 1983: 40). To my knowledge, this is the first time, at least in popular science, that an AI description moves away from *human* intelligence and incorporates animal intelligence as well; Frude was speculating about the emotional bonding between humans and nonhumans – it is thus quite likely that his experience with human-animal bonding led him expand AI's notion to animal intelligence simulation as well. Two years later, the very influential for future understandings of AI book *Artificial Intelligence: The Very Idea* by philosopher John Haugeland was published. Haugeland, critical of his time's "commercial ventures" (1985: 5) of expert systems, dismissed the idea that simulation of intelligence would count as intelligence; hence, if a GPS system simulates human instruction by voicing "in 100 metres turn right," this should not count as intelligent behaviour. Taking a deep philosophical look at AI, examining the meaning of reasoning as part of intelligence, he defended early AI specialists who supported the idea of symbolic manipulation as a form of intelligence and separated his time's approach to AI from Good Old Fashioned Artificial Intelligence (GOFAI), a term which became standard among several AI scholars, as a contrast with alternative approaches including the neural network types. For Haugeland, "the claims essential to all GOFAI theories" were on the one hand "our ability to deal with things intelligently is due to our capacity to think about them reasonably (including subconscious thinking)" and on the other "our capacity to think about things reasonably amounts to a faculty for internal 'automatic' symbol manipulation" (Haugeland, 1985: 112-113). Cognitive scientist, philosopher, psychologist, and computer scientist Margaret Boden offered a more expansive view on AI, to embrace all machine-related research which allows better understanding of knowledge processes in her book *Artificial Intelligence and Natural Man* [sic]:

> "By 'artificial intelligence' I therefore mean the use of computer programs and programming techniques to cast light on the principles of intelligence in general and human thought in particular. In other words, I use the expression as a generic term to cover all machine research that is somehow relevant to human knowledge and psychology, irrespective of the declared motivation of the particular programmer concerned." (Boden, 1987: 5).

In their edited volume *The Foundations of Artificial Intelligence: A Sourcebook*, Patridge and Wilks (1990) gathered notable AI researchers from different research institutions of excellence around the world to offer their views on AI's state-of-the-art and future prospects. Influential AI researchers from Yale, Chicago, and Texas Universities as well as entrepreneur Robert Schank's contribution is key to understand an insider's view on AI's discipline in relation to other disciplines and the impact of AI's publicity in its conceptualisation. Interestingly, Schank problematises the concept even for insiders and turns the

argument around, but suggesting that the concept's confusion with other disciplines is actually reason to view AI as *relevant* to all other disciplines:

> "Artificial intelligence is a subject that, due to the massive, often quite unintelligible, publicity that it gets, is nearly completely misunderstood by people outside the field. Even AI's practitioners are somewhat confused with respect to what AI is really about […] Is AI mathematics? […] Is AI software engineering? […] Is AI linguistics? […] Is AI psychology? […] AI should, in principle, be a contribution to a great many fields of study. AI has already made contributions to psychology, linguistics, and philosophy as well as other fields. In reality, AI is, potentially, the algorithmic study of processes in every field of inquiry. As such, the future should produce Al/anthropologists, Al/doctors, Al/political scientists and so on. There might also be some Al/computer scientists, but on the whole, I believe, AI has less to say, in principle, to computer science than to any other discipline. […] In some sense, all subjects are really AI." (Schank in Partridge and Wilks, 1990: 3-4, 13).

A slightly opposite (yet somewhat complementary) view from the same volume comes from Edinburgh University's Alan Bundy who, instead of explaining AI as relevant to many fields, approaches AI as an outcome of various motivations shaping its context, thus, being closer to a social shaping approach, in line with the lessons learned in section (a) about the post-Lighthill and post-Mansfield applications-based AI, creating a tripartite distinction, outlining descriptions as seen above:

> "The different kinds of AI correspond to different motivations for doing AI. The first kind, which has become very popular in the past five to ten years, is applied AI, where we use existing AI for commercial techniques, military or industrial applications, i.e. to build products. Another kind of AI is to model human or animal intelligence using AI techniques. This is called cognitive science, or computational psychology. Those two kinds of AI have often been identified in the past, but there is a third kind on which I want to concentrate most of my attention. I call it basic AI. The aim of basic AI is to explore computational techniques which have the potential for simulating intelligent behaviour." (Bundy in Partridge and Wilks 1990: 216).

Such views can be seen as responses to concurrent criticisms of AI's practical output. The following polemic passage by software engineer David Lorge Parnas in his abstract to an article provocatively entitled *Why Engineers Should Not Use Artificial Intelligence* reads:

> "(a) the terminology used in many AI discussions is poor, (b) that many techniques widely touted as revolutionary are ad hoc, 'cut and try,' methods that will not lead to trustworthy products, (c) that many claims about AI and expert systems are exaggerated, and (d) that the fundamental research is more philosophical than practical." (Lorge Parnas 1988: 234).

This passage is indicative of the early perception of a new AI winter; the second one in an "official history mode." Roland and Shiman's (2003) assessment of the SCI, Crevier's (1993) history of AI, synopsising tables which can be found in Grudin (2009; also cited in Wikipedia), would agree that the failure of component integration into intelligent expert systems as well as the failure in creating "empty shell" expert systems, resulted to funding drawbacks in Japan and the US, paired to the unrealisability of the vision that LISP (list processing language) expert systems would dominate the market (more on these promises in section 3.2 below), resulted into a second AI winter. Nevertheless, research continued. The first attempt at a taxonomy of AI definitions was made in Stuart Russell and Peter Norvig's *Artificial Intelligence: A Modern Approach*, the book that has become the standard textbook for AI students and researchers since its first edition in 1995, followed by another two updated impressions, in 2003 and 2009, with nearly

40000 citations according to Google Scholar, by the time writing. Russell and Norvig do not offer their own definitions, but create a quadripartite taxonomy of types of AI research, in the form of a grid (table adapted from Russell and Norvig, 1995: 5, without the examples of definitions, some of which have been outlined above):

1. Systems that think like humans.	2. Systems that act like humans.
3. Systems that think rationally.	4. Systems that act rationally.

The publication of such an influential handbook such as Russell & Norvig's does not correspond with the this above reading of AI history; it is unlikely that if an "AI winter" was such a devastating event, one of the most influential handbooks would be prepared and published during that period. Hence, such accounts appear to be quite reductionist, failing to take into account the multitude of parallel developments in AI's many fields, as outlined in the beginning of the present subsection. What is crucial for the further development of AI hype in the post-2010 era is that in 1986, the three-page long article by Rumelhart, Hinton, and Williams (1986) appeared on *Nature*, which introduced the method of backpropagation (an algorithmic method to reduce errors in statistical reasoning in neural networks; again, more on this in 3.2), meant to mark the technical foundation of recent promissory environments. This article does not mention "AI" throughout its three pages; however, two out of its four references are given to Minsky and Papert's *Perceptrons* book and Rosenblatt's original article on the perceptron. Of particular importance in official modes of AI history assessment is that the recent popular-level book *Genius Makers* by *New York Times* and *Wired* journalist Cade Metz retells AI's history with its timeline beginning with Rosenblatt, and an entire chapter devoted to Hinton's evolution and subsequent adoption of his method in tech-giant services such as Facebook and Google. Some interesting themes found from this section include the following, partly contradictory accounts, telling of the field's expansion and increasing interpretative flexibility (including awareness of its interpretative flexibility!): AI is viewed as a patchwork of different disciplines, still negotiating its identity, mirroring its practitioners' motivation. It consists of a set of different methodologies focusing on simulation or programming of intelligence. Views of AI practitioners, depending on their research agendas or affiliations, differ – some treat it as an established science studying the imitation and coding of intelligence and some treat it as an applications-oriented technical field. Some see AI as different across generations, and therefore, for the first time, at this stage, we see reflective descriptions from AI practitioners, who offer views informed by historical awareness of the field and acknowledging multiplicity of approaches. This multiplicity, nevertheless, is shaped by numerous expectations, dependent on a social context in which AI becomes recognised by military and commercial interests as a field of investment. Lack of definitions does merely signal a disagreement in technical approaches, but a will to repurpose and refresh AI as a concept, either towards a curiosity-driven scientific endeavour (e.g. Bundy, Schank, , an applied technical domain of "systems" (Russell and Norvig), or something in-between (Boden).

d. Waiting game 2: behind the internet scenes, 1999-2008

In my view, this period marks the most shadowy era of AI in terms of terminological publicity. An explaining factor can be the popularisation of the term information and communication technologies (ICTs) and the advancement of Web 2.0 technologies (user-generated internet content advancing a bottom-up structure, different to the early hierarchical modes of internet; O'Reilly 2007) – these created sufficient

hype for large-scale expectational environments and public imagination (for example, Flichy 2007; see also the discourse on information superhighways and multimedia technologies, Emmott 1995; Williams 1997). Many of my interviewees, when asked about the source of contemporary interest in AI, they referred to the proliferation of smartphones, enabling user-generated data to be uploaded and easily accessible for algorithms to get trained (with all privacy dilemmas being encountered), essentially built on the Web 2.0 vision of users "taking over" the internet. Behind the scenes of internet excitement and hype, AI technical applications and approaches continued to advance: new hierarchical models were introduced which, together with dramatic improvement in moving object recognition, more types of driverless cars and other vehicles, virtual autonomous agents and multi-agent systems (for example in videogames), and natural language processing, could create complex "grammars" of layers of synthesised visual, acoustic, and rule-based associations (Nilsson 2010). I find it particularly interesting that, with its now 45 years of history, in a time of relative obscurity, most descriptions I encountered (examples below) were presented as definitions, something which did not happen earlier with only few exceptions.

Dependent on background or research focus different scholars offer different definitions. Neuroscientist Anthony J. Bell, with no reference to any previous similar suggestion, claims that "AI's ultimate purpose is to build a robot that lives in the world with a computer for a brain. It therefore assumes that the essence of the living and/or thinking process can be captured in digital computation" (Bell 1999: 2914). This was the same era that transhumanists, scholars who suggested technological enhancement of humans is necessary in order for humanity to survive a forthcoming surpassing by intelligent machines, gained popularity (e.g. Warwick 1998, published by Penguin; Warwick 2000). Hence, there was a rise in circles of specialists whose work relates indirectly to AI, who, taking advantage of that time's relatively inactive state of earlier AI scientists, associated their names with their version of AI as a field which could achieve the creation of humanlike thinking robots, being a potential threat for humanity. Besides Warwick, other prominent figures in this arena include Kurzweil (Kurzweil 2005; Kurzweil and Kopor 2009), and Chalmers (2010); O'Connell (2017) traces Bostrom's intellectual trajectory as an extropian in the 1990s (a type of early transhumanism) into the passionate herald of "superintelligence" in the 2010s).

On the other hand, there were insiders whose work in the field was never halted. Rodney Brooks, pioneer of the Artificial Life (or ALife) movement in AI and robotics, which is based on mimicking smaller-than-human intelligent structures and follows an evolutionary path in achieving human-level or other forms of advanced AI, makes a distinction between situatedness and embodiment in artificial systems, in which, variations of degree between the two exist across the two extremes: "Under these definitions an airline reservation system is situated but it is not embodied. A robot that mindlessly goes through the same spray-painting pattern minute after minute is embodied but not situated" (Brooks, 2002: 51-52). Earlier on, Brooks justifies his expansion of AI's concept as to include a broader variety of types of intelligence: "Judging by the projects chosen in the early days of AI, intelligence was thought to be best characterized as the things that highly educated male scientists found challenging. […] The things that children of four or five years could do effortlessly […] were not thought of as activities requiring intelligence" (Brooks 2002: 36). A *Beginner's Guide* to AI by specialist Blay Whitby crystallises that AI is "the study of intelligent behaviour (in humans, animals, and machines) and the attempt to find ways in which such behaviour could be engineered in any type of artefact" (Whitby, 2003: 1), thus offering a symmetrical approach to all three categories' types of intelligence. Aaron Sloman, whom we encountered two subsections ago, offered an alternative humorous definition, to signal the thoroughly established by then impact of science fiction on AI research and a constantly future-oriented vision: "My colleague

Russell Beale once suggested a useful introductory definition of Artificial Intelligence (AI) for people who know nothing about it: "AI can be defined as the attempt to get real machines to behave like the ones in the movies" (Sloman 2003: n.p.). By 2009, AI received a quadripartite entry in the Webster dictionary which, in turn, was cited in the influential United Nations Educational, Scientific, and Cultural Organization's Encyclopedia of Life Support Systems' entry on AI. That time's Webster entry read as such:

> "1. An area of study in the field of computer science. Artificial intelligence is concerned with the development of computers able to engage in human-like thought processes such as learning, reasoning, and self-correction. 2. The concept that machines can be improved to assume some capabilities normally thought to be like human intelligence such as learning, adapting, self-correction, etc. 3. The extension of human intelligence through the use of computers, as in times past physical power was extended through the use of mechanical tools. 4. In a restricted sense, the study of techniques to use computers more effectively by improved programming etc." (Webster dictionary cited in Kok et al 2009).

From this, it can be inferred that although technical AI research has escaped the boundaries of human intelligence-only imitation, "official" definitions were, on the one hand referring to classical understandings of AI, and on the other, have been influenced by the "extension" or "augmentation" approach, possibly after transhumanist influence. We also get a first glimpse of AI as a "restricted" set of techniques, more adjacent to contemporary understandings[37]. Nils J. Nilsson offered what appears to be the most comprehensive, although as he admits "generous," definition of AI, written by someone who has dedicated an entire academic career to the field, to encompass the various research domains which would be recognised as AI in different circumstances. This definition appears on the first page of Nilsson's *The Quest for Artificial Intelligence: A History of Ideas and Achievements*, being thus far among the most comprehensive historical accounts of AI, albeit with a rather technical and mostly US-centric focus:

> "Artificial intelligence (AI) may lack an agreed-upon definition, but someone writing about its history must have some kind of definition in mind. For me, artificial intelligence is that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment. According to that definition, lots of things – humans, animals, and some machines – are intelligent. Machines, such as "smart cameras," and many animals are at the primitive end of the extended continuum along which entities with various degrees of intelligence are arrayed. […] For these reasons, I take a rather generous view of what constitutes AI. That means that my history of the subject will, at times, include some control engineering, some electrical engineering, some statistics, some linguistics, some logic, and some computer science." (Nilsson 2010: xiii).

Nilsson's description is further telling of another problem several AI researchers would acknowledge when asked to offer AI definitions – the difficulty to define intelligence, paired to the changing conceptions of what intelligence is about. This sheds some additional light into the problem of AI conceptualisation. From AI descriptions of this stage, there is an apparent fixation in two themes: Firstly, AI appears to be an established field with no negotiation of its identity; although its present identity, due to alternative

[37] The current version of (now) Merriam-Webster's definition is much shorter and seems to acknowledge encompassing of nonhuman intelligence traits and removed reference to the augmentation approach: "1: a branch of computer science dealing with the simulation of intelligent behavior in computers 2: the capability of a machine to imitate intelligent human behavior" (https://www.merriam-webster.com/dictionary/artificial%20intelligence). An exploration of the entries' authors and sources of knowledge exceed the scope of this chapter.

approaches such as ALife, has expanded as to include the study of intelligence in machines and nonhuman animals. Secondly, the relationship with science fiction robots or grandiose visions of AI becoming synonymous with intelligent, humanlike robots, became established. In Geraci's thorough review (2007), the late 1990s and early 2000s mark the period where religious studies become increasingly more interested about AI, but this is initiated first through AI depictions in science fiction – it was in 1998 that a theologian from MIT, Anne Foerst, suggested a theological interpretation of Rodney Brooks's Cog robot (still from MIT), reflecting on technosalvation and technowrath (Geraci 2007: 962-963). Such debates did not receive much publicity beyond the then esoteric circles of robotics, AI, and religious or literature studies. During that time, AI, overshadowed by the internet hype, stayed behind the scenes as a field to receive big funding and its researchers entered a waiting game, throughout which they continued to conduct incremental research. But it was precisely the massive popularisation of the internet which would allow the technical foundations for contemporary AI approaches to be built.

e. AI of the latter days, 2009-present.

Nilsson's book is the latest attempt at a fully comprehensive account of AI history. Since then, and especially after the recent hype, contributions to AI history research took the form of journal papers (Plasek 2016; Garvey 2019), personal accounts/memoirs (McCorduck 2019) or Wikipedia articles (History of artificial intelligence 2021). What most authors agree upon, is that the abundance of data availability through popularisation and wide access to mobile smart devices, combined with increasing computational power, wireless networks, and large databases, enabled massive parallel pattern recognition. Approaches such as that of the perceptron, the backpropagation algorithm, and reinforcement learning, which became overshadowed by AI's official history mode (and supporters of GOFAI), showed signs of very promising success. An application of the backpropagation algorithm, an AI method advanced in 1986 (see subsection c, above), with its roots in the 1960s perceptron development, was Krizhevsky, Sutskever and Hinton's (2012) employment of deep neural networks to extract pattern recognition from the large database of annotated images ImageNet triggered a series of promises. ImageNet was first presented in 2009 as an attempt to classify image databases semantically, following the same approach to WordNet, a 1980s database of words and models of semantic association[38] and in accordance with the early 2000s demand for reliable descriptions of data (metadata; Doctorow 2001). The main investigator was computer scientist Fei Fei Li[39] (Li et al 2009; Deng et al 2009; nowhere in these first two publications the terms AI, machine learning, or neural networks are to be found). Learning from large databases fuelled by a ubiquitous internet of things and its mobile devices and personal assistants, enabled this strand of AI application to perform increasingly better in speech and visual object recognition. In turn, this generated an environment of high expectations for applications such as online recommendation systems, medical prediction software, business and trading automations, facial recognition and detection applications, as well as search and rescue robots (themselves also increasingly hyped via online "viral" videos showcasing short successful performances; de Waard, Inja, & Visser, 2013; for a criticism of such videos' credibility, see Aylett and Vargas 2021; for the high hopes associated with the field in general, see Mishra 2021; for AI hype assessment in medicine, see Matheny et al 2019).

[38] https://image-net.org/about.php
[39] Fei Fei Li's significance in the promissory environment of AI will be analysed in section 3.2.

By the middle of 2010s, although AI as a term gained traction among wider discourse (such as journalists), ethicists, and policymakers, were mostly focused on fields such as ICTs and robotics in their discussions on information ethics or roboethics (e.g. Capurro and Nagenborg 2009, Lin et al 2012, Tzafestas 2016; moreover, tracing headline trends on articles published on the journal *Ethics and Information Technology* is quite revealing on the various types of "ethics" attracting publication hype). Thus, AI as a field became dispersed across different computer-related domains which can be viewed under certain circumstances as extensions of its own early development. Definitions of AI circulated at that time were not significantly different from those of previous generations; however, the interesting step came when AI's successes proliferated, the term became part of everyday parlance, and policy actors intervened in order to regulate AI. AI specialists such as Mitchell (2019) agreed that definitions are rather hard to obtain for AI, and her updated introduction to the field (a Pelican Book, 60 years after Sluckin's 1960 *Mind and Machines*, being published by the same Penguin subsidiary), is a rather descriptive and informative take on applications, subfields, and concerns, avoiding essentialist definitions. Admittedly, this was a period of generalised confusion about AI, when the massive hype and promissory environment was paired to dystopian views about AI's existential threat in public statements by influential commentators which triggered a further regulatory environment (Galanos 2019). While chapters 5 and 6 look closer into the interplay between scientists, policy/funding actors, and the impact of public hype, it is useful to trace how current understandings of AI are shaping and shaped by policy documents. Of the many (see chapter 5) AI policy documents, I have selected three highly influential ones, based on the magnitude of their producing institutions (and thus high citation rates) and impact in terms of establishing AI as a fixed entity in the contexts of scholarship, legislation, and public outreaching; such documents have now become the standard references and recommendation guidelines that AI developers, researchers, and intermediaries are expected to adhere to (thus often citing them) when, in the dissemination of their research and products, are required to ensure their abiding with ethical principles. While such documents are themselves producing expectations, I suggest that they are also shaped by expectations. The intergovernmental regulatory Organisation for Economic Co-operation and Development (OECD) viewed AI from a deterministic perspective, having transformative dimensions, considering it a "technology" instead of a field, which is general-purpose, in contrast to most of AI sceptics:

> "Artificial Intelligence (AI) is a general-purpose technology that has the potential to improve the welfare and well-being of people, to contribute to positive sustainable global economic activity, to increase innovation and productivity, and to help respond to key global challenges. It is deployed in many sectors ranging from production, finance and transport to healthcare and security." (OECD 2019).

In the same year, The European Commission's appointed advisory group for an AI strategy called High-Level Expert Group on Artificial Intelligence (HLEGAI) offered a balanced distinction between AI as a field of technical application and as a scientific discipline, however, with almost zero reference to early AI foundations of symbolic reasoning – hence, a definitional product of its own time:

> "Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a

> numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.
>
> As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)." (High-Level Expert Group on Artificial Intelligence 2019: 8).

This group's discussions evolved into the first comprehensive legal framework for regulating AI industry, business, research, placement, and usage according to a scale of potential risk. Three years later, the proposed "single future-proof definition of AI" (European Commission 2021: 3) was in response to consulted "stakeholders" who "mostly requested a narrow, clear and precise definition for AI" (European Commission 2021: 8) and moves away from HLEGAI's broad and highly encompassing definition. For the Commission's 2021 understanding of AI, an:

> "'artificial intelligence system' (AI system) means software that is developed with one or more of the techniques and approaches listed [below in the document] 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." (European Commission 2021: 39).

Robotics and other types of hardware have vanished[40], while AI as a scientific field is not part of any of the discussions within the 100-page long document (this will be important in chapter 5, when discussing specialist views on regulating AI, whether this is possible, and its types of impact). Elsewhere in the document, AI is presented as "a fast evolving family of technologies that can bring a wide array of economic and societal benefits across the entire spectrum of industries and social activities" (European Commission 2021: 1, similar variations in pages 18, 34). Almost in line with OECD's deterministic definition, such optimistic descriptions of AI's transformative capabilities counterbalance a large document that covers the different types of AI-related risks. In the context of AI becoming increasingly more the topic of socio-political debate beyond policy, with several reasons for that to be explicated in the following section, media scholar employed at Microsoft Research Kate Crawford's 2021 investigation of AI's "atlas" begins with a negation of the term's components[41] in order to proceed with a constructivist definition:

---

[40] By 2022, while amending corrections to the thesis, the EU-funded European Institute of Innovation & Technology's (EIT) Artificial Intelligence Community published a landscape map titled "Creation of a Taxonomy for the European AI Ecosystem," where the eight main areas of AI are defined as "Computer Vision, Computer Audition, Computer Linguistics, Robotics, Forecasting, Discovery, Planning, Creation" (Artificial Intelligence Community 2022), with robotics being returned to the map of AI landscape. The report suggests compatibility to HLEGAI's principles, and therefore can be taken as a more updated perception of AI at an EU level. As human-computer interaction specialist with recent interest in human-centered AI Ben Shneiderman (2022b) notes, the interesting dimension of this taxonomy is that it "combines technologies (the first four) and applications (the last four)." He proceeds in the same note to create his own taxonomy which includes "(1) Physical Devices with Mobility, (2) Active Appliances with Fixed Locations, (3) Voice User Interfaces, (4) Personal and Group Supertools, (5) Recommenders, Decision Aids, and Prediction Supertools, and (6) Knowledge Work and Creativity Supertools" (Shneiderman 2022). Nevertheless, AI as science is missing completely from both accounts.

[41] See also my 2018 paper "Artificial Intelligence Does Not Exist: Lessons from Shared Cognition and the Opposition to the Nature/Nurture Divide" presented at the 2018 edition of the International Federation for

> "I argue that AI is neither *artificial* nor *intelligent*. Rather, artificial intelligence is both embodied and material, made from natural resources, fuel, human labor, infrastructures, logistic, histories, and classifications. AI systems are not autonomous, rational, or able to discern anything without extensive, computationally intensive training with large datasets or predefined rules and rewards. In fact, artificial intelligence as we know it depends entirely on a much wider set of political and social structures. And due to the capital required to build AI at scale and the ways of seeing that it optimizes AI systems are ultimately designed to serve existing dominant interests. In this sense, artificial intelligence is a registry of power. […] I use AI to talk about the massive industrial formation that includes politics, labor, culture, and capital. When I refer to machine learning, I'm speaking of a range of technical approaches (which are, in fact, social and infrastructural as well, although rarely spoken about as such)." (Crawford 2021: 8-9[42])

Crawford's definition might not be used frequently in technical papers – however, her book and work has become widely recognised amid AI research communities. Her definition of AI as a product of forces of power is finally indicative of the latest years' reflexivity within (at least some) AI communities about AI's challenging dimensions. The ways in which this was achieved, however, will be understood in the next section concerned with the promissory environment that shaped, and is being shaped by, these understandings of AI. As Crawford also suggests, despite the occasional loss of trust to AI as a term, "the nomenclature of AI is often embraced during funding application season, when venture capitalists come bearing checkbooks, or when researchers are seeking press attention for a new scientific result" (Crawford 2021: 9). The following, conflicting in part, themes can be extracted from this final stage of AI's conceptualisation. AI has now become a "thing," and mostly refers to tangible systems; it is nearly equated to its applications, although some wish to draw a line of separation between AI as a research field and as a series of applications. The interesting dimension, which will be better understood sociologically in the following section through the examination of promises, is that AI is presented as a deterministic agent of social change, risky or beneficial. What I will argue in the next section concerned with a history of promises, is that a series of sensationalist and industrial expectations highlighting contemporary AI's financial potential paired to technical and social critical negotiation of its limitations, further created an established view of AI's risky or beneficial potential which was adopted by regulators.

It becomes clear from the above quinquepartite development of AI conceptualisations that AI specialists had little to do with their field's officially accepted meaning. While McCarthy wished to move away from automata (robotics) in his first description, Lighthill added them back. When novel AI approaches such as ALife appeared, as to include nonhuman intelligence and physical systems, AI policymakers established that AI is software applications. Experts associated with themes relevant to AI but not AI *per se*, performing on the boundary between specialist arenas and non-specialist arenas, generate regulatory environments; and in turn regulatory environments crystallise their own perceptions of AI. Such actors, are credible enough to exert influence on what is AI for policy or popular science, but not

---

Information Processing's conference. While Crawford negates the two terms by placing emphasis on the "natural" components making up AI's infrastructure and the association of intelligence with power instead of rationality (and I agree under given levels of communicational abstraction), I suggest that the very dichotomy between natural and artificial is to be negated, also in favour of a shared, systems-based perception of intelligence as something "in-between" intensities (commonly perceived as "entities") rather that "within" brains, chips, or actions.

[42] It is always interesting to note how constructivist vocabularies about social, economic, and infrastructural shaping of technology or anti-rationalism are sometimes mixed with the chief lexicological identification of scientific determinism and rationality: "in fact."

sufficiently credible to be respected as AI scientists within the AI community. Going back to Olazaran and Agar's papers and considering them jointly, they explicate how non-technical actors exert influence not only in the technological options supported by funding, but also in the writing of technology's history, and even the transformation of a field from a "science" into a "technology." By enacting certain views of AI based on broader discourse, they led to a selection of a view of AI that *can* be regulated because it *should* be regulated, and thus, it has to be confined in the form of *system*, rather than a *science*. In a sense, Agar's emphasis on the under-the-record communications between research councils voiced by non-AI experts such as fluid mechanics specialist Lighthill who appears as an "expert for hire" (Yearley 2005: 162), shape the official history of AI. To use a quote from one of the few sources casting doubt on the AI hype cycle story: "Rather than framing controversies within a rise-and-fall narrative, we might therefore interrogate if and to what extent they were a functional and integral component to the construction of the AI myth" (Natale & Balatore 2017: n.p.). The following poetic quote found in T.S. Eliot's *Murder in the Cathedral* describes sharply the machinations of AI historical development: "However certain our expectation / The moment foreseen may be unexpected / when it arrives. It comes when we are / engrossed with matters of other urgency" (Eliot 1936: 56). As the internet became a matter of urgency, with all relevant literature on its social implications and technical futures emerging after the mid-1990s, the AI project of machine translation became better fulfilled during that time; similarly, the application of backpropagation algorithm was effectuated in times of the internet hype in 2009 – in other words, it during waiting games that AI expectations justify their existence, not through the hype seasons. The following section of this historical component to the literature review about AI's development is concerned with mapping the promises of AI. This will take a closer look at the competing motives and social events which shaped the historical understandings of AI and will thus act as a stepping stone to understand interviewed specialists understanding of AI, its promises, and motivations.

3.2 Historical Examinations of AI's Promissory Environment

> *"[O]ver the Christmas break, Allen Newell and I invented a thinking machine." (Herbert Simon, 1956, talking to the graduate students of his course Mathematical Models in the Social Sciences, cited in Feigenbaum 1992: 194)*

I have so far exemplified the development of AI through its conceptualisation. However, for the present introduction to the relevance of AI in being studied from an expectational perspective, it is important to review the above timeline based on the evolution of its promises and expectations. Technologies that mirror and simulate human or animal intelligence and actions fascinate and worry both practitioner and non-practitioner groups associated with AI. This section will look into the development of AI's promissory environment in a more historical fashion. Although plenty of the expectations associated with AI are captured in its conceptualisations over history, outlined in the previous section, it is important to situate the empirical research in the chapters to follow within a long historical context where AI communities, AI affiliates (people with interest in AI as a theme, but without sufficient technical expertise), and AI critics, have spoken about AI in a future-tense from the 1950s until now. I have collected such phrasings from a variety of heterogeneous documents including, but not limited to, AI histories, scientific articles, policy

reports, and popular readings[43]. These are presented in three sub-sections, based on the tentative timeline explicated above: early AI beginnings, post-FGCS programme international responses, and contemporary hype and critical approaches from within and beyond the AI community.

a. Early Promises and Predictions

Prior to promises following Dartmouth, it is historically important to refer to slightly earlier predictions from scientists whose work has been associated with the early foundations of AI before the coinage of the term. In his influential publication *Computing Machinery and Intelligence* (1950), Alan Turing introduced his imitation game (widely known as the 'Turing Test'), which invites readers to consider the ethical implications of robots reaching equal levels of intelligence with humans, concerning the possible indistinguishability between human and machine intelligence. His imitation game suggests that if a human interlocutor communicating with a human and a machine reaches a point where she cannot distinguish the difference between the participants, the machine has passed the test. Turing's estimation was that around the end of the 20th century such tests will be trivial, and "the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted" (Turing, 1950, p. 442[44]). Turing was member of the Ratio Club, the UK mini-equivalent, according to some, of the US cybernetics group. In the same year, cybernetician Norbert Wiener published his book *The Human Use of Human Beings: Cybernetics and Society*, in which he offers numerous technical predictions about the future of human society mingled with various technical applications of cybernetics, as well as general predictions about human endeavours based on findings from cybernetics. Advances in computing, for Wiener, were part of a new industrial revolution, and would have a great impact in the workplace. Wiener suggested that the overall goal would lead to the positive outcome of humans being liberated from repetitive tasks, entering a phase of generalised leisure. However, "the new industrial revolution is a two-edged sword. It may be used for the benefit of humanity, but only if humanity survives long enough to enter a period in which such a benefit is possible" (Wiener 1950: 162[45]). Both Turing and Wiener have blurred the boundaries between esoteric science and public discourse – they were producing both pioneering technical work and dedicated time to public outreach (Wiener admittedly more than Turing). Therefore, much of the pre-AI intelligent machines discourse was shaped by the engineering and mathematics community.

Two years after the Dartmouth workshop, two of its participants, Newell and Simon, initial supporters of the "complex information processing" term, on paper presenting their findings on heuristic problem solving, made the following predictions, largely quoted by investigators of causes for AI winters:

> "On the basis of these developments, and the speed with which research in this field is progressing, I am willing to make the following predictions, to be realized within the next ten years:

[43] Collection of such documents was based on a mixed approach of researching specific keywords, guided by senior interviewees, and researching University libraries' sections on AI as well as many second-hand bookstores where "outdated" treatises on technical subjects tend to concentrate and be sold for very low prices.
[44] In section 7.3.1 where I discuss the performative impact of the Singularity hype, I discuss Turing's prediction from the year following this article, about the possibility of intelligent machines outstripping humanity from an evolutionary perspective.
[45] Consider the semblance with one of Stephen Hawking's predictions cited below.

> 1. That within ten years a digital computer will be the world's chess champion, unless the rules bar it from competition.
> 2. That within ten years a digital computer will discover and prove an important new mathematical theorem.
> 3. That within ten years a digital computer will write music that will be accepted by critics as possessing considerable aesthetic value.
> 4. That within ten years most theories in psychology will take the form of computer programs, or of qualitative statements about the characteristics of computer programs.
>
> It is not my aim to surprise or shock you if indeed that were possible in an age of nuclear fission and prospective interplanetary travel. But the simplest way I can summarize the situation is to say that there are now in the world machines that think, that learn, and that create. Moreover, their ability to do these things is going to increase rapidly until in a visible future-the range of problems they can handle will be coextensive with the range to which the human mind has been applied." (Simon and Newell 1958: 7-8).

Contemporary assessments of AI history written by AI practitioners (Nilsson 2010) or people affiliated with the latter (McCorduck 2019) would suggest that such predictions, although unrealised by the specified timeframe, were realised with delay; thus, the critique against them has been unreasonably harsh. Fleck (1982) suggests such grandiose promising act as strategic moves towards the establishment of a field.

Between Dartmouth and Simon and Newell's predictions, Rosenblatt's perceptron approach, precursor to contemporary machine learning emerged. As shown in the previous chapter, Minsky and McCarthy were not dismissive of this approach. On the contrary, they embraced it and highlighted limitations based exactly on the absence of sufficient data to feed in the machine in 1969 and in consecutive reprints of their book on perceptrons, until the late 1980s. It was the broader AI research community, however, which interpreted this as rivalry and, to a certain extent, Rosenblatt's over-optimism that stirred some disillusionment (Olazaran 1996). Indeed, in 1960, Minsky included pattern recognition in his summary of the goals of AI which were: search, pattern-recognition learning, planning, and induction (Minsky 1960). In the same work, he even foresaw contemporary challenges faced in the use of deep ML, which gave rise to "explainability" issues (that is, showing in an understandable fashion how an algorithmic process gave a certain result; for example, Hagendorff and Wezel 2019). Minsky, describing limitations of such perceptron-based pattern recognition systems, speculated about the potential incomprehensibility of the programmes. The following passage shows how, as an AI insider, he not only embraced the idea of the perceptron (contrary to what mainstream AI history suggests), but even defended his computer programmer colleagues as to their ethical agency: "[I]t does not follow that the programmer therefore has full knowledge of (and therefore full responsibility and credit for) what will ensue. For certainly the programmer may set up an evolutionary system whose limitations are for him [sic] unclear and possibly incomprehensible" (Minsky 1960: n.p.). Almost identical to contemporary debates about the ban of lethal robots and ethical discussions about explainability, early political STS casted doubt on Minsky's approach through Langdon Winner's critical lens:

> "Minsky may believe that the products of such incomprehensible programs will be uniformly positive, that is, that programs that behave in ways its makers could not predict will achieve marvels of performance in, for example, games of checkers. But what is to prevent possible detrimental effects following from such work? What is to stop an incomprehensible program from developing a lethal edge? Certainly not the programmer." (Winner 1977: 304)

Rosenblatt and Minsky's treatment of the perceptrons took place in parallel with the aforementioned development of statistical social science at the Simulmatics Corporation, which, as suggested above, Ithiel de Sola Pool, one of the central figures of Simulmatics commented in an article about their company's success:

> "The outcome of the present study gives reason to hope that computer simulation may indeed open up the possibility of using survey data in ways far more complex than has been customary in the past. The political 'pros' who commissioned this abstruse study were daring men to gamble on the use of a new and untried technique in the heat of a campaign. The researchers who undertook this job faced a rigorous test, for they undertook to do both basic and applied research at once. The study relied upon social science theories and data to represent the complexity of actual human behavior to a degree that would permit the explicit presentation of the consequences of policy alternatives.
>
> This kind of research could not have been conducted ten years ago. Three new elements have entered the picture to make it possible: first, a body of sociological and psychological theories about voting and other decisions; second, a vast mine of empirical survey data now for the first time available in an archive; third, the existence of high-speed computers with large memories. […] It is our belief that this is now possible which was put to a test by the campaign research reported here." (de Sola Pool and Anderson 1961: 183).

If Simulmatics was such a pioneer in behaviour prediction, to the extent that its methods are considered to be responsible for Kennedy's victory, why is it that its story is so well buried, waiting for Lepore's 2020 excavation of it? While Lepore does not offer and explicit answer to this, from her work I infer this might have to do again with the perception of a disillusionment trough. Like AI of the same time, Simulmatics' early success on one case, allowed optimistic speculation and promise. Lepore's study touches briefly on the disillusionment the company entered soon after the great optimism:

> "When Simulmatics' sales pitches failed, they failed for two reasons: the high cost of the Simulmatics model (many clients 'seemed shocked by our price scale') and the insufficiency of data on which to train the model ('calls have all bogged down on the "past data" question'). Even as the company readied for its public stock offering, Pool raised the question that Simulmatics would never really answer: 'What is the data we would need for this model?'" (Lepore 2020: 139; disillusionment was also thought to be brought about due to another core Simulmatics contributor's heavy alcoholism).

Although Lepore does not draw explicit parallels with today's data-driven AI, it is clear that contemporary concerns about AI's forthcoming success rate are variations of the same set of questions: the high environmental and hidden labelling labour cost of AI, and, if not insufficiency, then inadequacy of datasets, further paired to a better understanding about the social world's resistance to be purely quantifiable (Hagendorff and Wezel 2019; Cantwell-Smith 2019). I suggest that what is to be remembered from the Simulmatics optimism based on the company's development of the techniques, is not another hype-and-disillusionment story, but awareness of its methodology's simultaneous development with early GOFAI as well as the perceptron model, which had to wait until the next century to be successfully converged with them and become part of AI's contemporary rebranding. Simultaneous beginnings in automation, cybernetics, statistical/behavioural theory, and AI took a long time branching until they met each other on the way.

As demonstrated above, some early cyberneticians took responsibility of both developing the science and technology of proto-AI, but also becoming spokespersons of it through publications accessible

to wider publics. Minsky appeared to follow a similar path and this is shown in the following note by Hubert Dreyfus, the longstanding critic of AI's grandiose claims: "Marvin Minsky, head of MIT's Artificial Intelligence Laboratory, declared in a 1968 press release for Stanley Kubrick's movie, *2001: A Space Odyssey*, that 'in 30 years we should have machines whose intelligence is comparable to man's [sic]'" (as cited in Dreyfus 2006: 44). This passage does not only show that Minsky's expectational vision was similar to Turing's, but also his role as a technical specialist who influences broader discourse relevant to AI; *2001: A Space Odyssey*'s AI character HAL is often used as a typical example of science fiction narratives which have distorted the meaning of AI's actual technical capabilities (consider, for example, the subheading "Why the Problem with Artificial Intelligence is H.A.L. (Humanity At Large), not HAL" in an article by philosopher Luciano Floridi (Floridi 2015).

While discourse was fuelled by expectations which eventually reached Kubrick degrees of outreach, early disillusionment about narrower technical promising appeared. What was considered by some the cause for a first AI winter, was the 1965 Automatic Language Processing Advisory Committee ALPAC report, an assessment of machine translation (an application of the aforementioned AI approach of microworlds), commissioned by the US government. The report resulted in loss of trust in machine translation for a decade which was considered by some to be the "silent years" of machine translation (Nilsson 2010: 181; I have abstained from mentioning ALPAC earlier mainly due to lack of reference to AI terminologies). In the context of cold war stratagems between the US and Russia, it was considered possible that the relatively finite content of Russian and English languages (vocabulary, grammar, and syntax) could be precisely matched and thus translated, at least in the cases of political reports and scientific publications which employed relatively neutral language, without the peculiarities of idioms and everyday natural language[46]. This echoed the early cybernetic visions of systems theory of control, wonderfully captured in the following passage from a 1947 letter to Norbert Wiener by Warren: "When I look at an article in Russian, I say: 'This is really written in English but it has been coded in some strange symbols. I will proceed to decode it'" (Weaver 1949: 18). While ALPAC is considered by some (e.g. Grudin 2009; Nilsson 2010) to be the first form of an AI winter, due to its application-specificity, it contradicts the simultaneous increasing interest in AI events; notable examples are the first International Joint Conference on Artificial Intelligence (IJCAI) in 1969 with over 600 participants and sponsorship from 16 different technical societies from the US, Europe, and Japan, held biannually since then, and the invitation-only "Machine Intelligence" workshops held at the University of Edinburgh, organised since 1965 by Donald Michie. It was at the 1968 edition of these workshops that Edward Feigenbaum of Stanford announced the first results of his research group's early accomplishments with what was eventually to be known as knowledge-based expert systems or simply expert systems. Such systems, with DENDRAL and MYCIN as the most prominent early applications in medical diagnosis, proceeded on the premise that "success depended upon the amount and the quality of the expert knowledge that the program captured" (Roland and Shiman 2003: 191). Hence, if human intelligence could not be replicated in total, a partial, well-defined, and thorough set of knowledge-based intelligent task might be more successful. All this, together with the Kubrick-level AI discourse, denotes the high degree of expectations sustained at that time. Simultaneously, however, as the US was entering war with Vietnam, the US Congress passed the Mansfield Amendment to the 1970 Defense Procurement Authorization Act, restricting funding to

[46] Although the machine translation specialists were impacted, AI as a broader field continued to expand. As it has been shown above, AI branched into several fields and machine translation has been one of them, closely attached to information theory and science. See the statistics-based promissory comeback of machine translation in the next sub-section, concerning the oft-quoted statement by Frederick Jelinek.

scientific and technological research projects "with a direct and apparent relationship to a specific military function or operation" (Public Law 91-121, cited in Nilsson 2010: 203).

The second AI winter (or, for UK researchers, the first one; for others the only one) was marked by the Lighthill report, which was encountered in the previous chapter. It should be noted that although AI was researched in several UK institutions, Edinburgh was particularly affected because it was an internationally known centre of excellence in AI. The following passages synopsise the overall sentiment of the report, mainly focusing on the problem of combinatorial explosion – simply put (and in a fashion very similar to contemporary criticisms), too much heterogeneous data about the worlds very rich informational environment would lead to system failures:

> "Most workers in AI research and in related fields confess to a pronounced feeling of disappointment in what has been achieved in the past 25 years. Workers entered the field around 1950, and even around 1960, with high hopes that are very far from having been realized in 1972. In no part of the field have the discoveries made so far produced the major impact that was then promised […] [O]ne rather general cause for the disappointments that have been experienced: failure to recognize the implications of the 'combinatorial explosion'. This is a general obstacle to the construction of a […] system on a large knowledge base which results from the explosive growth of any combinatorial expression, representing numbers of possible ways of grouping elements of the knowledge base according to particular rules, as the base's size increases." (Lighthill 1973: n.p.).

While in the case of ALPAC, it seems that external funders cut machine translation funds solely based on unrealisability of promises and in favour of competing projects (such as HCI; Grudin 2009), in the case of Lighthill, internal disputes within the research community acted as criterion to governmental strategy, complicating assessments between forms of enaction and selection – technical options are not the sole criterion of selection; selection takes place as an assessment of social behaviour that becomes an indicator for the integrity of a technical option worth sponsorship (namely, AI). According to Edinburgh-based AI pioneer Jim Howe's reminiscence, "[u]nfortunately, the high level of discord between the senior members of the School had become known to the its main sponsors, the Science Research Council. Its reaction was to invite Sir James Lighthill to review the field" (Howe 2007: n.p.). Donald Michie, also from Edinburgh, commented almost ten years after the event:

> "Work of excellence by talented young people was stigmatised as bad science and the experiment killed in mid-trajectory. This destruction of a co-operative human mechanism and of the careful craft of many hands is elsewhere described as a mishap. But to speak plainly, it was an outrage. In some later time, when the values and methods of science have further expanded, and those of adversary politics have contracted, it will be seen as such." (Michie 1982: 220).

Richard Gregory, pioneer in the psychology of visual perception, optical illusions, and in early AI developments of visual recognition, collected most of his important publications in 1974, a year after the Lighthill report's effectuation, when he was still Edinburgh-based. While the book mostly covers material on human dimensions of visual perception, it is interesting that Gregory decided to finish the book with a perspective on AI.

> "[A]s we develop the power to predict so there is less novelty in the events predicted. Scientific theories destroy the appearance of intelligence in things as prediction becomes possible. This is true for biology as it is for physics. This does however lead to a paradox for sociology. Sociologists are concerned to predict the effect of changes on future society. But is prediction *in principle* possible

> when intelligence is involved? If intelligence is the production of novelty, prediction might seem to be strictly impossible. However this may be, it seems that the present trouble about social prediction is simply that there are no adequate theoretical models of societies. […] We find ourselves in just this position in trying to assess the implications of future intelligence. […] In these circumstances the best we can do is to write fiction from our past; and hope that the story we like best turns out to be true. […] The vital point about intelligent machines is that once they are trusted they will take decisions, and these decisions will directly affect us." (Gregory 1974: 637).

This was probably the first time that an interdisciplinary alliance of benefit between the social sciences and AI was proposed as part of an expectational game. It is evident that Gregory either foresaw (if this passage was actually written in 1965) or took on board Lighthill's combinatorial explosion argument. However, his response as an AI scientist with a keen eye for crossing boundaries was that the main problem lies in societies' inadequate models which would produce inadequate results when fed into AI. Once again, another problem noted today as "algorithmic bias" based on ill-modelled data (for example, Selbst et al 2019).

b. International AI race and generalised sensationalism

I want to advance here an argument based on previous suggestions in the introductory chapter whereby I treated AI winters as convenient reductionist fables. Focusing too much on the UK's landscape makes the notion of the AI winter easily believable. However, the new directions in AI opened up across the Atlantic suggest the opposite view – the AI winter existed in one region; it was not worldwide. The following historical note by Crevier is revealing: "[Lighthill's] 1973 report called for a virtual halt to all AI research in Britain. This recommendation led to the quasi-dismantling of top-flight research groups, such as that at the University of Edinburgh, and to the emigration of eminent British AI workers to the United States" (Crevier 1993: 117). If one wishes to trace the evolution of AI through its promises and expectations, one needs to take into account that AI as a field developed through transition of situatedness. Indeed, Hinton, core figure in the development of the backpropagation algorithm in neural networks, and thus for contemporary AI renaissance (whose work was encountered above and will be revisited below), began his PhD in AI at Edinburgh University exactly on the year that Lighthill was commissioned to write his report. He was awarded the title in the aftermath of the report, thus, remaining jobless, working on a connectionist model which was considered to be eccentric amid AI communities, at a time of the AI winter. He thus moved to the US and then Canada, in an attempt to avoid the Ronald Reagan's regime, in order to continue his work which would culminate in the mid-1980s and recognised widely in the early 2010s (Metz 2021: 34-45). As Sherry Turkle asserts, AI's practical applications in data analysis across the military-commerce complex rendered AI's sustainment broader social/political issue:

> "By the mid-1970s AI was no longer marginal. It had its own academic programs, its own journals, its own conferences. It was well funded because of its value in the marketplace and to the military. Expert systems were used to analyse stock prices, data from oil well drillings, materials from chemical samples. Companies competed to hire AI graduates to start in-house departments. The future of the field became part of a heated discussion about Japanese-American industrial rivalry." (Turkle in Graubard 1988: 254).

And on the technical level, Roland and Shiman note: "just when disappointing results in one branch of AI seem to have discredited the entire field, promising developments in another branch revived expectations and set off another scramble for the promised land. […] In the late 1970s, expert systems sallied forth to take up the fallen banner of AI" (Roland and Shiman 2003: 190). I will show below, in the UK and Europe, a similar phenomenon took place during the period of what is usually described as the second AI winter; researchers carefully dismissed the term "AI" from their applications and publications, focusing on different terminologies, yet directly related to technical AI research.

During the period between 1975 and 1980, seeds were planted towards contemporary successes in AI, added to the connectionist approach of perceptrons and the already emerging expert systems: the frame approach and the reinforcement learning approach, both foundational in contemporary machine learning in setting contexts on the one hand, and learning from previous examples based on algorithmic training on the other. Minsky continued to be a key player in AI research and it was he who developed the frame approach:

> "Here is the essence of the theory: When one encounters a new situation (or makes a substantial change in one's view of the present problem) one selects from memory a substantial structure called a frame. This is a remembered framework to be adapted to fit reality by changing details as necessary […] Once a frame is proposed to represent a situation, a matching process tries to assign values to the terminals [the detailed features] of each frame, consistent with the markers at each place […] Most of the phenomenological[47] power of the theory hinges on the inclusion of expectations and other kinds of presumptions. A frame's terminals are normally already filled with default assignments. Thus, a frame may contain a great many details whose supposition is not specifically warranted by the situation. These have many uses in representing general information, most-likely cases, techniques for 'by-passing logic' and ways to make useful generalizations." (Minsky 1975: 212-213).

The promise seems like an extension of the initial premise of AI; intelligence can be precisely described, and although different contexts have different expressions of intelligence, such contexts can be modelled in detail as well. Richard Sutton, Canadian AI researcher, towards the end of the 1970s, was a MSc student in computer science at Stanford University (in which John McCarthy was already a Professor). In 1979, he developed the idea of reinforcement learning, one of the key techniques which allow contemporary algorithms to adjust their results ("learn") based on previous behaviour (such as "likes"). It is interesting to read his memory of describing the beginning of the theory which, instead of the contemporary, more neutral, term reinforcement learning, was named after the much more anthropomorphic (and thus, as proven to be in the future, more public discourse-friendly) hedonistic behaviour:

> "[I]n 1979 we came to realize that perhaps the simplest of the ideas, which had long been taken for granted, had received surprisingly little attention from a computational perspective. This was simply the idea of a learning system that wants something, that adapts its behavior in order to maximize a

[47] For readers interested in phenomenology, Minsky's appreciation of the term was, in my view, quite unique. In most of his papers referring to phenomenology he did not provide with references as to the phenomenological flavour he refers to. Dreyfus and Dreyfus (1995), however, are probably correct in asserting that Minsky's phenomenology was of rather Husserlian flavour, especially given his will to quantify expectations, as per the cited passage (indeed, prior to SoE, expectations have been studied by Husserl as cognitive objects, with Merleau-Ponty extending the Husserlian diagrams as to the multiple impacts imagined expectations might have to a given subject; Merleau-Ponty 1945: 445).

> special signal from its environment. This was the idea of a 'hedonistic' learning system, or, as we would say now, the idea of reinforcement learning." (Sutton and Barto 2015: viii).

From this passage I extrapolate the simultaneous continuation not only of AI research (and evidence that reinforcement learning is part of a long trajectory), but also the role of general discourse in selecting terminologies which bear anthropomorphic traits. Likewise, a thought similar to Turing's prediction that robotic children might be mocked by human ones towards the end of the 20th century, was further expanded in the views of UK-based philosopher and computer scientist Aaron Sloman in 1978, who argued against future robot exploitation and slavery or even "racial discrimination against them" (Sloman 1978: 272-273). His book *The Computer Revolution in Philosophy* is among the first ones (if not the first) targeted at rather close audiences of philosophy and computer or cognitive science scholars, thus, broadening the scope of AI as to include pure philosophical debate; at once challenging the technical boundary of AI in two ways: showing AI's relevance to philosophers and incorporating lessons from philosophy for future AI[48]. Therefore, and as shown in 3.1, AI research was never permanently halted; its scope and conceptualisation changed. The following passage from an article on Business Week from 1982 captures the excitement of the time and if it is taken out of chronological context, it could very easily pass a "Turing test" for a post-2010 quote:

> "The world stands on the threshold of a second computer age. New technology now moving out of the laboratory is starting to change the computer from fantastically fast calculating machine to a device that mimics human thought processes–giving machines the capability to reason, make judgements, and even learn […] Experts are now convinced that it is now only a matter of time before these 'thinking' computers open up awesome new applications in offices, factories, and homes." (Resnick et al 1982: 66).

In the aftermath of an early 1980s AAAI conference at Stanford, Scott Fahlman surveyed Carnegie Mellon trends of alternative solutions to a noticeable halt in the development of LISP language, which was core to the development of AI since its very beginnings (Fahlman 1981). The survey captures the enthusiasm of transition between a purely LISP-based and personal computer-based application of intelligent systems, considering different options and possible disadvantages, adhering towards a hybridity of the two. According to Crevier, at that time "a cloud hung over the upstart AI industry. It was the result of a poor business decision by the academic founders of the first AI companies, blinded as they were by their unquestioning faith in the specialized LISP language" (Crevier 1993: 209). This was based on a tripartite circumstance: (a) cost; LISP machines did not benefit from economies of scale, as there was no agreed upon hardware type (such as laptops) which would make use of LISP, (b) lack of network connectivity between data banks companies (something which enables today's machine learning training through internet connectivity), and (c) the LISP language was unknown beyond the AI communities, hence programmers were scarce and thus expensive (Crevier 1993; Fahlman 1981).

Nevertheless, the Japanese FGCS programme mobilised international responses, through the promissory language employed in its official reports. According to one such report, these systems were "expected to have advanced capabilities of judgement based on inference and knowledge-based functions, and capabilities of flexible interaction through an intelligent interface function" (as cited in Nilsson 2010: 277). This was the same period that Japanese industry was dominating multiple markets globally, from

[48] Sloman's role as an advisor in protecting UK AI research from national vision traps will be revisited below in this sub-section's footnotes.

Sony music listening equipment and Nikon photo-cameras to Toyota cars; therefore, such an announcement caused sufficient alarm around the globe. Roland and Shiman (2003) and Nilsson (2010) both assert that, at the time, it seemed quite likely that Japan might succeed in AI applications. Edward Feigenbaum, PhD student of Herbert Simon and then Professor at Stanford University, considered to be "father of expert systems" in AI and married to a Japanese woman (and thus frequenting to Japan and its computer conferences), co-authored and published in 1983 with Pamela McCorduck, considered to be the first historian of AI, their book *The Fifth Generation: Artificial Intelligence and Japan's Computer Challenge to the World*. The two authors, and Feigenbaum's wife Penny Nii, had visited Japan and Feigenbaum was among the participants of MITI's initial conference which announced the FGCS. In the first pages of his book with McCorduck, they mention:

> "The Japanese plan is bold and dramatically forward-looking. It is unlikely to be completely successful in the ten-year period. But to view it therefore as "a lot of smoke," as some American industry leaders have done, is a serious mistake. Even partially realized concepts that are superbly engineered can have great economic value, preempt the market, and give the Japanese the dominant position they seek. We now regret our complacency in other technologies. Who in the 1960s took seriously the Japanese initiative in small cars? Who in 1970 took seriously the Japanese national goal to become number one in consumer electronics in ten years? (Have you seen an American VCR that isn't Japanese on the inside?) […] We are writing this book because we are worried. But we are also basically optimistic. Americans invented this technology! If only we could focus our efforts, we should have little trouble dominating the second computer age as we dominated the first. […] America needs a national plan of action, a kind of space shuttle program for the knowledge systems of the future." (Feigenbaum and McCorduck 1983: 2-3).

The book read as an "explainer" of these new technologies, yet imbued with national vision rhetoric. While Feigenbaum would be considered an insider with technical expertise, McCorduck helped make the book accessible enough to influence broader readerships. According to McCorduck's memoir, the book's outreach was immeasurable, given that several unauthorised copies have been made (acting as further indication of popularity):

> "The book became a best-seller in Japan and the United States and eventually, with authorized and unauthorized copies, sold about half a million copies around the world. Ed Feigenbaum and I enjoyed the giddy experience of being nine-day wonders. Yes, it's fun to walk along Madison Avenue and see your own book in bookshop windows. Translations abounded, the phone rang constantly, and the publishers and we were happy at least. Congress held hearings: was Japan's Fifth Generation a threat to national security? Should the National Science Foundation or DARPA invest more dollars in computer research? Snarky researchers claimed that Feigenbaum and I had written the book only to beef up his research budget." (McCorduck 2019: 297).

However, Feigenbaum's role was double. The first sentences of the above passage from the two authors' book ("The Japanese plan is bold … the dominant position they seek") are sharply identical to the testimony Feigenbaum gave to the US Congress during the hearings of the *Japanese Technological Advances and Possible United States Responses Using Research Joint Ventures* session. The following part of the testimony is revealing about his deep engagement with the US policy as an influential insider:

> "As often happens in science and technology, the most important part of the creative act is asking the right question or placing the right long-term bet. I believe that the timing of the fifth generation project

> is exquisite, and the bet is a very good one. Indeed, in my view, the only gamble is one of timing, not orientation. The era of reasoning machines is inevitable. It is the 'manifest destiny' of computing.
>
> Now what options do we have for an American response? […] Option zero: Nothing special, do our normal thing, hurry up later, hope for the best and it might work out. Option 1 is some kind of orchestrated national response. […] The nearest thing we have to a MITI for computer technology is the Defense Advanced Research Projects Agency (DARPA). […] Though it has a defense mission, it has always acted more broadly in the national interest. Under its funding and guidance, artificial intelligence research has been nurtured for 20 years. […] Arpanet is a good bet for a national response. Option 2: A national center for the next generation computer technology, as a manifestation of the national will to maintain our No. 1 position in the computing world. This would be an ICOT-like center[49], staffed by scientists and engineers from universities, the computer industry, and Government laboratories." (Feigenbaum in United States. Congress. House. Committee on Science and Technology. Subcommittee on Investigations and Oversight 1983: 119-120).

According to Roland and Shiman's history of DARPA's SCI, "[t]he computer community derided Feigenbaum as 'chicken little,' but Congress embraced him as a seer prophesying doom. In testimony before the House Committee on Science, Space, and Technology, he infected the legislators with his sense of alarm" (Roland and Shiman 2002: 92). According to the same authors, Switches expert Robert Kahn's three reasons for seeking generous funding for DARPA's Information Processing Techniques Office (IPTO[50]):

> (1) "faith of a true believer in the importance of computers and communications" following his associates and predecessors (such as J. R. Licklider) who believed computers "had the potential to revolutionize society."
>
> (2) the belief that "several new technologies were poised on the verge of major advances" – these technologies were "microelectronics, multi-processor computer architecture, and artificial intelligence" – his vision was to connect them: "[o]ne produced switches, one arranged switches in space, and one sequenced the switches. All connected" – Kahn would finally attempt to employ the LISP language in within a single architecture.
>
> (3) announcement of Japan's FGCS[51] was the third reason which inspired Kahn to seek funding (Roland and Shiman 2003: 33, 36-37). Robert Cooper (the second Robert in leading the SCI),

[49] ICOT was the Institute for New Generation Computer Technology, funded by MITI, responsible for carrying out the FGCS project.

[50] Kahn had recently become the director of IPTO, seeking to refresh its goals through a new project, following the vision of its first director, J. R. Licklider, who was involved in the early cybernetics conferences (Roland and Shiman 2003; Pias 2010).

[51] In the context of promising, it is important to understand what the "fifth" in FGSC means, and why this bears semblance to contemporary discussions about quantum computing. As comprehensively synopsised by Roland and Shiman: "The first generation machines had been vacuum tube devices, of the kind that John von Neumann worked with in the period during and after World War II. With the invention of the transistor in 1948, a second generation of computers, extant between 1959 and 1964, was made possible. The electronic components in these machines—transistors, diodes, and so forth—were physically placed on printed circuit wiring boards and connected with copper wires. The third generation, such as the IBM 360 system of the 1960s, consisted of integrated circuits (ICs) in which the components were fabricated on single silicon wafers. Each of these ICs or chips was then connected by wires to the other chips to make a circuit of small scale integration (SSI). By the late 1970s the United States was well into very large scale integration (VLSI), in which an entire processor, less memory and peripherals, could be fabricated on a single chip. The Japanese were now promising to make a quantum leap to ultra large scale integration (ULSI) and produce machines powerful enough to achieve AI"

> thought that because the Japanese mainly invested in linguistic operations of AI systems, FGCS's impact would not be of the same magnitude as the Japanese dominance of stereo, automobile, and other electronic and electrical markets which were not specific to Japan's unique linguistic structure; this path did "however, provide a rationale and potential political lever for attracting new money to US computer research" (Roland and Shiman 2003: 50, 141-142).

To get a closer look at SCI's finalised promise and premise, as expressed in the programme's written documents, one should consider the following dispensation of AI's proposed pyramidal development scheme that Kahn proposed in his 28 October 1983 *Strategic Computing: New-Generation Computing Technology: A Strategic Plan for its Development and Application to Critical Problems of Defense*, which "is proposal to fund the computer research community, but it reads like a manifesto to transform the military" (according to Roland and Shiman 2003: 72):

- "Applications drive requirements of intelligent functions.
- Intelligent functions drive requirements of system architectures.
- System architectures drive requirements of microelectronics and infrastructure." (Roland and Shiman 2003: 72)

Hence, as a side-note to section 3.1, the notion of what constitutes intelligence, and therefore its artificial version and the technical context in which this could be built in (the system architecture and infrastructure) was dependent on, and made to fit, the military applications which secured allocation of funds. The promising of this stage involved reliance to "breakthroughs" according to an earlier document called *Strategic Computing and Survivability*. This plan:

> "had admitted that success would depend on 'technological breakthroughs.' By definition, a breakthrough is a leap into the unknown, a step up to a plateau of technological capability heretofore impossible. They may be anticipated, perhaps, but they cannot be predicted or foreseen; otherwise they would not be breakthroughs. Unknown and unknowable, they left applications managers little to work with. A capability might come, but researchers could not say when it would appear or what it would look like." (Roland and Shiman 2003: 80-81).

Robert Kahn took advantage of the time's expert systems enthusiasm and enveloped within the goals of SCI their advancement for military purposes. One of the main technical obstacles during the early 1980s was the difficulty in building domain-specific expert systems from scratch: expert knowledge had to be gathered, coded in the system, and new induction rules had to be developed for every different domain – this service was offered by expert systems firms. Kahn, using military as the main goal application, promised the possibility of creating generic, "empty shell," adjustable expert systems which would be easily adapted to different context quickly; this would be more easily accomplished by the time SCI would end, given that computing speed would increase dramatically within the decade - work in this area was among SCI's tasks (Roland and Shiman 2003: 194-195). According to a memorandum directed to the company Teknowledge which was commissioned almost $2 million by DARPA to work for the SCI in advancing Kahn's proposal, the expectation set was to "design and develop a robust software architecture suitable for building expert systems," which would have to be "modular, broad extensible, suitable for large-scale applications, distributable, and transportable," including AI features such as "reasoning with

(2003: 37). Although not clearly mentioned here, this is closely related to the "microchip revolution," another term that came to replace "AI" – consider the Edinburgh-based publication in Michie (1979).

uncertainty, knowledge acquisition, and cooperative systems" (Request for a New ARPA Order, memorandum cited in Roland and Shiman 2003: 201).

SCI was approved by DARPA and $1 billion was spent in it between 1983 and 1993. Through this, it is shown that a double round of enacting took place in order to convince the government to allocate funds. Feigenbaum managed to strategically influence AI communities, governmental funding bodies, and, through his book with McCorduck, the general public, paving the way for Kahn to draft the final proposals and give shape to Feigenbaum's testimony through promising in the context of national vision. It is important to consider this promissory game through the applications-oriented lens. SCI Programme Manager Larry Roberts, on a 1988 paper recalling memories from the post-Mansfield researcher attitudes, notes:

> "The Mansfield Amendment […] forced us to generate considerable paperwork and to have to defend things on a different basis. It made us have more development work compared to the research work in order to get a mix such that we could defend it. […] The formal submissions to Congress for AI were written so that the possible impact was emphasized, not the theoretical considerations." (Roberts 1988: 229-230).

This was in sheer opposition to early AI researchers who were also funded by DARPA. According to Newell, one of the Dartmouth workshop members, "[t]hey [ARPA] didn't have any control over the money they'd given us; that was our money now to go do what we wanted with. We didn't tell people in ARPA in the '60s what we were going to do with the money. We told them in sort of general terms. Once we got the money, we did what we thought was right with it" (Alan Newell, 1991 interview cited in Roland and Shiman 2003: 23).

McCorduck, defending the philosophical nature of AI research (1979, 2004, 2019), and reflecting on the history of AI as it progressively became the product of governmental, military, and industrial funding, suggested that the measure of AI's success was in executable programmes, no matter the depth of philosophical contribution; a dramatic imbalance for philosophers whose written paragraphs sufficed to count as contributions: "AI had a problem that philosophers had never faced: its researchers needed to write programs that demonstrably *worked*" (McCorduck 2019: 320). This was paired to a further problem unique in the field of AI, as noted by Shapiro: "[A]s soon as a task is conquered, it no longer falls within the domain of AI. Thus, AI is left with only its failures; its successes become other areas of computer science" (Shapiro 2000: n.p.). The SCI, and its counterpart Strategic Defense Initiative (SDI) initiated in the same year as SCI, mobilised a group of computer scientists who worked at the California-based computer company Xerox/PARC and at Stanford University, who, by 1983 formed the Computer Professionals for Social Responsibility (CPSR), a national non-profit organisation who openly opposed the military uses of computer technology[52]. This in itself generated an established notion of a separate arena of politically conscious computer scientists willing to defend their anti-militarist values, open to discussion with ethicists and social scientists with interest in computer science. Notably, STS scholar Lucy Suchman, one of the co-founders of CPSR authored the only document thoroughly criticising SCI (Suchman 1984[53]).

[52] http://cpsr.org/about/history/

[53] Elsewhere on CPSR's website, one can find Suchman's early report concerned with femininity in computer science. While this being a very important topic, the framing which extends beyond AI will not allow me to analyse it thoroughly here. Yet, an additional direction for future work is the comparison between Suchman's work on defending femininity from an anti-military stance with Lynn Conway's biographical accounts. Conway

The expectations set by DARPA were not met and considered to be disappointing at the level of component integration. In order to be sufficiently marketable and distributable, numbers of rules had to increase dramatically, like computing speed's orders of magnitude, inference sets had to be broadened, requirements for creating and extending knowledge bases had to be extended and simplified – and all this had to connect. It did not (Roland and Shiman 2003: 208-210). Likewise, expectations in computer vision were also unmet, mostly due to the high optimism following early 1980s publications which motivated DARPA to invest largely in it. There were, however, areas where SCI's AI was successful; most notable speech recognition and natural language processing. Paired to that, although no generic expert system was made, the advances made in the area of expert systems through the ABE ("A Better Environment") software architecture from Teknowledge were sufficient enough for securing a practical role in industry (Hayes-Roth and Jacobstein 1994; Hayes-Roth et al 1991). To the extent that the high-level expectation of general purpose expert systems is considered, AI historian Daniel Crevier's assessment appears apposite in that "the expert systems flaunted in the early and mid-1980s could not operate as well as the experts who supplied them with knowledge. To true human experts, they amounted to little more than sophisticated reminding lists" (Crevier 1993: 209; cf. Collins 1987, 1990). As long as low-level, applications-based practical outcomes are concerned, McCorduck's assessment seems more apt, considering SCI as a response to FGCS programme: "But the Fifth Generation's major accomplishment was not negligible: it trained a generation of young scientists in the field. […] The field seemed to have passed from revolutionary to normal science[54]" (McCorduck 2019: 304, 307). This can be evident through the August 1994 US Commerce Department's "Critical Technology Assessment of the US Artificial Intelligence Sector." According to the report, "[i]n 1993 the global market for AI systems was estimated at about $900 million, 60% of which was made up by the US market alone. […] In terms of sales, the most successful and dominant AI tools are knowledge-based systems, neural networks, fuzzy logic systems, and natural language systems" (Charles 1995: 70). Such calculations certainly challenge the strict perception of a second AI winter in the early 1990s. This is further supported by the way AI scientists of the late 1980s spoke about the opportunities opened up by the "AI business" which, however, have had to be seized with caution and interest in producing novel theoretical bases:

> "Due to the increase and success of AI 'business', the number of people actively working in AI is undoubtedly growing rapidly. But the financial rewards of most current practical applications of artificial intelligence is resulting in a tendency for many to move away from theoretical, long-term research. The future of commercial AI products rest on the pace of technical advances because without significant theoretical developments, vendors will be selling 'more of the same' to consumers on cheaper and more powerful hardware rather than supplying new and improved kinds of products like second generation expert systems. The tone and quality of the papers included in this book however

---

was one of the leading scientists at SCI, being one of the pioneers of the very large systems integration (VLSI) model, which Kahn would hope to employ in conjunction with his packet switching methodologies to meet SCI's final integration goals. Conway was also celebrated among women and feminist scientists, according to her own accounts, as one of the female engineers breaking the barriers of masculine dominance in computer science and engineering. In 2000, however, she revealed on her website and elsewhere ("came out," according to her account and inverted commas) about her gender transition in 1968 which resulted to the dissolution of a marriage and herself being fired from IBM where she was currently been employed. After 2000's revelations, she became a transgender activist. The open question in comparing Suchman and Conway's stance towards SCI might lead to interesting empirical understandings about the complex interface between technology, military, and gender. (Conway's homepage: http://ai.eecs.umich.edu/people/conway/conway.html)

[54] Probably a reference to T.S. Kuhn's (1963) normal science paradigms.

indicate that healthy fundamental research is continuing in AI and give cause for optimism about the future." (Hallam and Mellish 1987: n.p.).

While certain histories report the existence of an "AI winter" in the early 1990s, others claim success. What is the reason behind this conflict of reports? One argument is point of focus. As demonstrated, the Japanese FGCS project's chain reaction did not stop at US, but further mobilised the European ESPRIT programme and the UK Alvey programme. It is particularly interesting that the promissory game played in these regions employed less "AI" terminologies and worked based on different sets of computer science-related expectations. This shift of focus might have allowed them to be considered "successful," or at least not failed, paired to an evolving environment of different methods of evaluation (more on contemporary approaches to research assessment in chapter 6). While FGSC and SCI failed to integrate the various component technologies they have advanced, Alvey and ESPRIT managed to avoid entering the arena of AI promising, with projects relating to expert systems and other types of ICTs. There was sufficient lack of reference to the term in relevant summaries and documents (Oakley, 1983; Oakley and Owen, 1989[55]). This will be further verified via interview material from a key player in the field in the empirical chapters below. Nevertheless, it is worth noting that Lighthill was contacted to offer advice as per the usefulness of the Alvey-related AI techniques. He was far more optimistic this time, however, because of the detachment of AI promising from the 'millennial' expectation of AI dominating the 1960s, and its applications-based orientation: "the Alvey recommendations for major R & D and educational initiatives in the field of Intelligent Knowledge Based Systems seem to me to be exactly right. They are combined, as they should be, with major proposals in three other, equally vital, technological areas. They are directly linked to industry and its needs" (Lighthill, cited in Agar 2020: 306). Private and confidential sources beyond my interview sample have informed me about further internal disputes within the AI community of the time, based on personal affairs paired to ability of a key actor in the project to "cover up" the failure by writing down an alternative assessment. Looked at from a distance, it seems as if AI researchers decided to refer to a given set of technologies as AI when AI needs to be credited; whereas, they will offer a new terminology in need of avoidance of the term (for example, in times of contemporary hype, scientists who have expertise in AI systems from a time of post-Lighthill AI innovation may now refer to their field as AI, whereas they would not do so back then). According to the director of the programme Brian Oakley, and with allusion to British poet John Milton, "[i]f the Lighthill Report of the early 1970s was paradise lost for the AI community, the Alvey Report of the early 1980s was paradise regained" (Oakley 1990: n.p.).

Another argument about the paradox concerning the historical acceptance of AI winters, is their treatment as a self-fulfilling prophecy. If AI specialists perceive the threat of funding stagnation as real, they might live within that fear. The very concept of "AI winters" was for the first time printed during the era of this stage of second AI hype, in the proceedings of the AAAI conference panel from 1984, entitled *The Dark Ages of AI* which commented on the inflated expectations following the post-FGCS responses. In particular, computer scientist with AI expertise Drew McDermott observed the "commercial hustle and bustle" of the time around AI and suggested a hypothetical scenario in which all Japanese and US AI projects fail to meet their expectations creating:

[55] Alvey programme is said to have been nearly unsuccessful due to a different reason: stubborn support of locally made computer machines; an aspect of national strategy which was supported bulk purchases of unsuitable British computing equipment "just because it is British," according to Aaron Sloman, as evidenced in a recent report from the House of Lords (2018: 161) which, among others, considered witnesses in developing a counter-AI winter strategy. More about the history of Britishisation of computing in Summer (2014).

> "a big backlash so that you can't get money for anything connected with AI. Everybody hurriedly changes the names of their research projects to something else. This condition, called the "AI Winter" by some, prompted someone to ask me if "nuclear winter" were the situation where funding is cut off for nuclear weapons. So that's the worst case scenario. I don't think this scenario is very likely to happen, nor even a milder version of it. But there is nervousness, and I think it is important that, we take steps to make sure the "AI Winter" doesn't happen-by disciplining ourselves and educating the public." (McDermott et al 1985: 122).

The identity of the people ("by some," "someone") who were carriers of this "nervousness" remains unknown. This continuous process of shaping AI's character for different purposes has been at the same time beneficial for setting scientific boundaries for AI, yet at the same time detrimental as to its amorphous conceptualisation, making it vulnerable to non-specialist scrutiny. If specialists cannot define an established view of their field, non-specialists have the right to challenge its rigour. The "nervousness" of the time sparked a period in which several AI researchers aimed at redefining AI (see, for example Schank or Bundy's 1990 defence of AI in the previous section), while others aimed at carefully demarcating between hype and reality.

By 1984, and parallel to the need of guising AI under different names, the deterministic notion of "information revolution" gained traction (Galanos 2014). Occasionally, AI, computers, and information technologies have been linked or used interchangeably. For example, Winner whom we encountered in his earlier criticisms of Minsky, ended his article entitled "Mythinformation" by stating: "Some observers forecast that 'the computer revolution' will eventually be guided by new wonders in artificial intelligence. Its present course is influenced by something much more familiar: the absent mind" (Winner 1984: 596). In 1985, Tom Forester's collected volume The Information Technology Revolution gathered articles by numerous authors, some of them placing focus on AI – and the de-sensationalising of it. Margaret Boden, whom we encountered earlier as the author of the influential book *Artificial Intelligence and Natural Man* (1987), remarked in her chapter 'The Social Impact of Thinking Machines' that:

> "Sensationalism feeds on ignorance, and many descriptions of artificial intelligence in the media, and in popular books about the subject, are sensationalist in nature. Whether proclaiming the 'wonders' or the 'dangers' of AI, they are not only uninformative but highly misleading–and socially dangerous to boot. They suggest that things can be done, or will be done, tomorrow which in fact will be feasible (if ever) after decades of research. Unfortunately, these sensational reports are sometimes encouraged by ill-judged remarks from the AI community itself." (Boden in Forester 1985: 102-103).

It is interesting to notice how authors from this volume appear to follow McDermott's suggestion for an inward disciplining of AI communities paired to outward education of the public. In the same volume, we encounter the ambivalence of Joseph Weizenbaum's views on AI: on the one hand considered a pioneer in the field, on the other a fierce critic of its potential societal outcomes, aimed at disambiguating between imagined and practical definitions of "thinking machines." In his chapter 'The Myths of Artificial Intelligence' he refers to John Von Neumann's (one of AI's predecessors associated with the first cybernetics movement) understanding of what is meant to describe thinking with precision – thus, suggesting that all contemporary AI researchers should be aware of the limits in describing mental processes:

> "As a computer scientist, I agree with Ionesco, who wrote, 'Not everything is unsayable in words, only the living truth' […] Actually, von Neumann had a standard answer for anyone who asked him whether

> computers could think, or be intelligent, and so on. He argued that, if his questioner were to present him with a *precise* description of what he wanted the computer to do, someone could program the computer to behave in the required manner. Whether he thought there were some things in the human experience that could not satisfy his criterion, I simply don't know. The position that every aspect of nature, most importantly of human existence, must be precisely describable, and its corollary that all human knowledge is sayable in words, is central to the credo to which all true believers in the limitless scope of artificial intelligence must hold." (Weizenbaum[56] in Forester 1985: 93).

Another example stems from philosopher of mind and language John Searle, notable for his division between "weak" and "strong" AI[57] as well as his development of the 'Chinese room argument' aimed at disproving the possibility of genuinely intelligent machines[58]. Searle, interviewed during an early ethnographic project of AI specialists at the time suggests:

> "[…] there's a lot of nonsense that comes out about AI, like the idea that computers are a deep threat to human beings and that computer achievement will destroy our sense of human dignity. That's crap! I have a pocket calculator that can beat any mathematician in the world, but that's no threat to anybody's dignity." (Searle, cited in Rose 1985: 1965)

Terry Winograd, developer of the 1968 SHRDLU computer programme (a pioneering application employing natural language processing for the moving and manipulation of digital objects; precursor to contemporary conversational AI) with engineer and politician Fernando Flores, in their 1988 thorough assessment of AI *Understanding Computers and Cognition* highlighted not only the problems of calling such systems AI, but also the expectational challenges of the "expert systems" terminology:

> "Calling a program an 'expert' is misleading in exactly the same way as calling it 'intelligent' or saying it 'understands.' The misrepresentation may be useful for those who are trying to get research funding or sell such programs, but it can lead to inappropriate expectations by those who attempt to use them." (Winograd and Flores 1988: 132).

Challenging the way in which attractive terminology is used for the allocation of funds, they further predict the soon-to-come failure of systems integration (as happened in Japan and the US) and the spinoffs of successful AI-related applications which were – as per the AI effect – detached from AI and associated with the information revolution:

> "The grandiose goals, then, will not be met, but there will be useful spinoffs. In the long run, the ambitions for truly intelligent computer systems, as reflected in this project and others like it around the world, will not be a major factor in technological development. They are too rooted in the rationalistic

[56] It should be noted that Weizenbaum's popular book *Computer Power and Human Reason* (1976) presents AI as technique instead of science; thus may count as a point past Lighthill and Mansfield assessments in the late 1960s-early 1970s, and early enabler of contemporary policy assessments of AI as technology; see above chapters. See also McCarthy criticism of Weizenbaum's book as "unreasonable" and difficult to summarise ("IT'S HARD TO FIGURE OUT WHAT HE REALLY BELIEVES…" – capitals in original; McCarthy in Kuipers, McCarthy and Weizenbaum 1976: 5).

[57] A term similar to "general purpose" AI, ultra- or superintelligence, hence, not examined in detail above. Searle has been also notable for his Professor Emeritus position being revoked after multiple allegations for sexual harassment (Weinberg 2019).

[58] The Chinese Room argument is a thought experiment about a person inside a room with no prior knowledge of a given language (for example, Chinese), who receives queries in that language and gets trained into responding "programmatically" based on patterns of association rather than genuine understanding of the language and its meanings (Searle 1980). Machines who perform such tasks count, for Searle, as weak AI, whereas, nearly impossible to create, machines reaching such genuine understandings would count as strong.

> tradition and too dependent on its assumptions about intelligence, language, and formalization." (Winograd and Flores 1988: 139).

Saymour Papert commenting on the controversy stirred by his own and Minsky's earlier criticism of the perceptron (which was proven to be falsely perceived *as* controversy in Olazaran 1996), aims at counter-hyping the late 1980s overpromising in a chapter interestingly titled 'One AI or Many?'. Papert employs the metaphor of Snow White being the long-sleeping field of AI, with new algorithms and large datasets acting as Prince Charming who is expected to awake her. Papert does not dismiss the possibility of AI's wide adoption, however he suggests that availability of computers and people's cultural familiarisation with them will rather enable any form of societal change based on technology (indeed, the passage below could very much be relevant in 2022, time writing):

> "A purely technical account of Snow White's awakening goes something like this: In the olden days of Minsky and Papert, neural networking models were hopelessly limited by the puniness of the computers available at the time and by the lack of ideas about how to make any but the simplest networks learn. Now things have changed. Powerful, massively parallel computers can implement very large nets, and new charming algorithms can make them learn. No romantic Prince Charming is needed for the story.
>
> I don't believe it. The influential recent demonstrations of new networks all run on small computers and could have been done in 1970 with ease. […]
>
> A more sociological explanation is needed. Massively parallel supercomputers do play an important role in the connectionist revival. But I see it as a cultural rather than a technical role, another example of sustaining myth. Connectionism does not use the new computers as physical machines; it derives strength from the 'computer in the mind,' from its public's largely nontechnical awareness of supercomputers." (Papert 1988: 13-14).

Papert's remarks might refer, although tacitly, to what was considered a war between AI engineers of the connectionist front who made use of statistical understandings of the natural world against the qualitative human science interpretations of the world. A core event from the same year was the slogan-isation of the phrase "Whenever I fire a linguist our system performance improves" stated by engineer Frederick Jelinek at a natural language processing workshop in 1988; as Hajič and Hajičová (2007) mention, this caused the establishment of an age-old unspoken competition across statisticians and linguists[59].

Technical applications and approaches developed during that time, at more experimental levels during the early 1980s and further tried and tested through the mid-1990s, involve methods more closely associated with contemporary machine learning and neural and Bayesian networks. Notably, the backpropagation algorithm was formulated in that time, playing a pivotal role in current recommender systems (Rumelhart, Hinton and Williams 1986) with the addition of bio-inspired convolutional networks in 1989 (LeCun et al 1989) as related to unsupervised and reinforcement learning techniques[60]. Further advancements include the rapid development of microelectronics, and multi-agent systems, leading to

[59] Interestingly, Jelinek himself aimed at ending the war by claiming in a subsequent publication in 2004 that "some of my best friends are linguists" – in turn, linguists Hajič and Hajičová turned back the favour by titling their presentation "some of our best friends are statisticians" (2007: 2-3).

[60] The names of Geoffrey Hinton and Yann LeCun will be encountered again in the following section c, as the two researchers being technical developers of the component technologies supporting contemporary AI hype, have resurfaced under the roles of enactors.

successful military applications such as early autonomous vehicles and remote robot agents with the Deep Space 1 spacecraft being probably the most notable achievement for intelligent robotics and chess playing programme Deep Blue's "victory" over world chess champion Garry Kasparov in AI reasoning toward the end of the 1990s (Nilsson 2010). All this was in parallel with the early popularisation of the World-Wide Web, which, although detached nowadays from common conceptions of AI, can be considered as part of AI's development to the extent that memory storage and heuristics (search) were among the foundational prospects of AI according to Minsky (1960), and, very importantly, that the TCP/IP protocols which enabled internet infrastructures were developed by Robert Kahn in 1969, who aimed at using them as integration components of AI at his SCI programme (Roland and Shiman 2003). This paved the path for hypes replacing AI nominally: the internet imaginaries, excitement with multimedia technologies, as well as the information superhighway rhetoric (Flichy, 2007; Emmott 1995) can be said to have taken away some of the fascination with AI during the 1990s and 2000s. According to Grudin (2009: 55), it was the late 1990s internet boom which enabled some of the popular speculations about ultraintelligence and the singularity (e.g. Kurzweil and Kapor 2001, Kurzweil 2005, Warwick 1998).

Therefore, while in the mid-1980s, AI communities aimed at keeping the promissory environment constrained through sufficient internal criticism, protecting such debates from non-specialist interpretations, non-technical discourses sustained the AI myth. A quote by Czech philosopher of communication Vilém Flusser: "It will ultimately be possible to build machines that will replace human work in all fields, and people will be 'free.' Machines will be the slaves of the future, and all human beings will become subjects of history, as they are freed from alienated labor" (Flusser 1991: 15). As mentioned in the previous section, it is interesting that although basic research in "AI" was not consistently presented as such, AI remained popular among non-AI specialists who published books partly in response to an existing public excitement about AI at the time. During the post-Lighthill and post-Mansfield AI winter, AI as a term was mostly encountered within non- scientific/technological domains; while non-specialist assessment caused the AI winter in part, other types of specialisation with interest in AI as a theme (artistic, philosophical, literary), assisted the conservation of the term. During the post-FGCS winter, and after a period of AI communities aiming at taming AI expectations, such expectations were fuelled again by non-specialist discourse. AI had to wait until the early 2010s to be hyped again.

c. Contemporary AI hype

While authors have commented upon the various performative impacts of hype and promises on AI's contemporary resurgence (Matheny et al 2019, Galanos 2019, Kerr, Barry, and Kelleher 2020), little light has been shed on the multiple arenas of promissory games. In this section, I will trace significant examples that constructed the post-2010 expectational environment. This is important in order to comprehend the relevance of two further aspects analysed below in this thesis: the impact of hype on AI research communities, contemporary hype as partially a historical product of the previous rounds outlined above, and the broader societal reasons which led to the establishment of the post-2020 AI policy landscape.

While contemporary AI history is still in the making and no peer reviewed publications of it have been made, perhaps with the exception of Metz (2021), there is general agreement (from Wikipedia to my interviewees suggestions) that the recent AI hype was triggered by and associated with applications of deep learning algorithms. A crucial moment was related to the publication of the results about extracting

patterns from the large database of annotated images ImageNet was an application of the backpropagation algorithm (Krizhevsky, Sutskever and Hinton, 2012), a landmark event which legitimised expectations around AI. As shown earlier, Hinton's algorithm was presented in 1986, however, with no reference to "AI," yet citing Minsky and Rosenblatt, evidencing the technological continuity from 1957 perceptrons and Minsky's appreciation of connectionism, to the second round of AI excitement in the 1980s, all the way to contemporary advances. As mentioned above, the 2009 papers by Fei-Fei Li, Jia Deng and their colleagues never mentioned AI-related terminologies either. Instead, their work is situated within the data and metadata literature. From the abstract:

> "The explosion of image data on the Internet has the potential to foster more sophisticated and robust models and algorithms to index, retrieve, organize and interact with images and multimedia data. […] ImageNet aims to populate the majority of the 80,000 synsets of WordNet with an average of 500-1000 clean and full resolution images. This will result in tens of millions of annotated images organized by the semantic hierarchy of WordNet. […] We hope that the scale, accuracy, diversity and hierarchical structure of ImageNet can offer unparalleled opportunities to researchers in the computer vision community and beyond." (Deng et al 2009)

What was considered "breakthrough" in several assessments of machine learning in the following years was but a series of incremental innovations – and by the time mainstream internet scholarship became disillusioned by online environments' surveillance structures (Mayer-Schönberger 2011), seeking a "right to be forgotten," a new hype had to be created. In 2011, a combination of machine learning techniques for natural language processing generated hype with IBM's *Watson*, a computer system which won a live game of the quiz show *Jeopardy!*. Dave Ferrucci, principal investigator of Watson, stated during a talk at the Computer History Museum about the event: "I had to get funding […] I told the executives I could do this in 3-5 years. I kind of just guessed. […] We want Watson to enable better judgement by humans in decision-making, whether it be in medicine, law, finance or services" (IBM Research Editorial Staff 2011). The article dedicated to the game's outcome as presented on the IBM website, finishes with the following lines: "Watson came away with the win, but left the auditorium with tremendous enthusiasm for this computer and its impact on the future of technology" (IBM Research Editorial Staff 2011[61]). Around the same time, hype was surrounding the notion of total human brain simulations, such as the EU-funded Blue Brain Project (Markram 2006). The hype was peaked when fragment of an infantile rat was presented as fully simulated, enabling speculation about human-level mapping of the brain, based on Moore's Law-type of future inference (Markram et al 2015; Costandi 2015).

This environment generated further resurgence of AI as an existential threat with influential figures associated with fields of science, technology, and entrepreneurship offering public statements about the hazardous potential of AI developing beyond human levels of intelligence and getting out of control; such figures include Stephen Hawking, Elon Musk, Bill Gates, and Max Tegmark (Galanos 2019). Hawking, in particular was quoted to say: "The rise of powerful AI will be either the best, or the worst thing, ever to happen to humanity" (Treblin 2016). His name was often cited in AI and robotics regulation proposals from the same year (consider the semblance with Wiener's prediction cited in subsection a above; Wiener 1950: 162 – it is interesting to observe what appears to be repetition of history, on the one hand, and yet, the passage from early AI/cybernetics communities warning about their own creations'

[61] For the sake of an argument which will be made in the empirical examination of promissory environments, I have to mention that in three instances of the article, the term "revolutionary" or derivatives is employed to characterise either Ferrucci or *Watson*.

undesired outcomes to outsider scientists commenting on fields external to their own). Such negative statements were following Nick Bostrom's book publication (eventually, a best-seller), thus entrenching a specific narrative of AI as an existential threat. However, these negative views acted as part of general awareness about technical availability. Borrowing a notion from the field of marketing science, Berger, Sorensen and Rasmussen (2010) suggested that "negative publicity can increase purchase likelihood by increasing product awareness," especially on relatively unknown products, contrary to previous studies suggesting the overall negative effects of negative advertisement. Although this notion is not particularly important for previous or further theoretical settings, I refer to it here because it allows the better understanding of the non-dualist dynamic effects which led to AI's multi-level establishment through the following series of positive/negative mixtures.

Kevin Kelly, editor of the popular technology magazine *Wired*,[63], wrote on the magazine's blog in 2014 an article which took into account the historical advancements of AI from the 1960s onward, comparing them to his recent visits to IBM. Although he dismissed singularity-type of arguments[64], he foresaw the purely applied version of today's AI (if not foresaw, admittedly shaped, given the magazine's influential status):

> "The AI on the horizon looks more like Amazon Web Services – cheap, reliable, industrial-grade digital smartness running behind everything, and almost invisible except when it blinks off. […] Like all utilities, AI will be supremely boring, even as it transforms the Internet, the global economy, and civilization. […] [T]he business plans of the next 10,000 startups are easy to forecast: Take X and add AI." (Kelly 2014).

The same closing sentence was used in his popular science book *The Inevitable: Understanding the 12 Technological Forces that will Shape Our Future* (2017[65]). AI, thus, appeared to offer a fertile ground for large investments according to major industry advisors, who began to include AI as an important investment target. Such advisors include the IDC's Worldwide Semiannual Cognitive Artificial Intelligence Systems Spending Guide (International Data Corporation 2017), the McKinsey Global Institute's discussion paper on AI's impact on businesses, industries, governments, and developers (Bughin 2017), and Grand View Research's AI market analysis (2017). IDC's predictions were very promising in 2017, as the service:

> "[…] forecasts worldwide revenues for cognitive and artificial intelligence (AI) systems will reach $12.5 billion in 2017, an increase of 59.3% over 2016. Global spending on cognitive and AI solutions will continue to see significant corporate investment over the next several years, achieving a compound annual growth rate (CAGR) of 54.4% through 2020 when revenues will be more than $46 billion." (International Data Corporation 2017)

[63] *Wired* has been the latest, and most popular, instance of Kelly's relationship to digital technologies. He was also co-editing the counter/cyber-culture magazine *Mondo 2000*, one of the early radical venues which brought together technical knowledge from the Silicon Valley, the underground world of hackers, and a blend of alternative modes of thought (from oriental philosophy to drug experimentation; psychedelics meeting cybernetics). His experience in 2014, thus, was embedding knowledge from many rounds of AI hype.
[64] 2014 saw the publication of Nick Bostrom's *Superintelligence*. A thorough analysis of this strand of AI is offered in section 7.3.1.
[65] Also, considering terminologies, Kelly's article is certainly among the earlier ones to make use of "AIs" in the plural. Cf. the interview quote by Ravi Autonomaskar in chapter 4, about AI being uncountable, as in "physics."

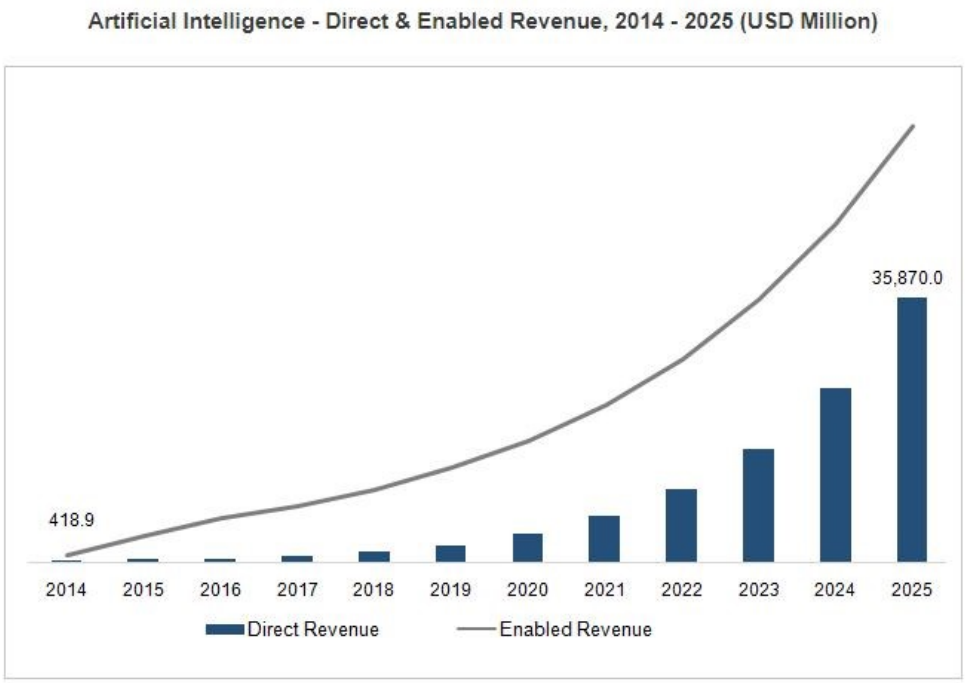
As the analysis suggests, the sectors involved in these investments are not restricted to companies who adopt such technologies to match their products to their clients' desires, but also extend to manufacturing, retail, and healthcare. These sectors are expected to make significant expenditures in AI and cognitive systems in the forecasted period. The 80-page analytical report by the analysts of the McKinsey Global Institute verified from the very beginning that "[e]xpectations are high" as:

> "[c]ompanies at the digital frontier—online firms and digital natives such as Google and Baidu—are betting vast amounts of money on AI. We estimate between $20 billion and $30 billion in 2016, including significant M&A activity. Private investors are jumping in, too. We estimate that venture capitalists invested $4 billion to $5 billion in AI in 2016, and private equity firms invested $1 billion to $3 billion. That is more than three times as much as in 2013. An additional $1 billion of investment came from grants and seed funding." (Bughin et al 2017: 6).

Their estimates about future investments took into account both cautious and promissory forecasts, generated from the Factiva, Tractica, and Transparency Market Research groups, reflecting the mixed signals by public defenders and critics of AI:

> "[…] analysts remain divided as to the potential of AI: some have formed a rosy consensus about AI's potential while others remain cautious about its true economic benefit. This lack of agreement is visible in the large variance of current market forecasts, which range from $644 million to $126 billion by 2025." (Bughin et al 2017: 6-7).

Grand View Research (2017) has provided with an "Artificial Intelligence Market Analysis By Solution (Hardware, Software, Services), By Technology (Deep Learning, Machine Learning, Natural Language Processing, Machine Vision), By End-use, By Region, and Segment Forecasts," analysing data from 2014 to 2017, offering predictions until 2025. Their estimations can be synopsised in the following paragraph and accompanying chart:



> "The global artificial intelligence market size was valued at USD 641.9 million in 2016 on the basis of its direct revenue sources and at USD 5,970.0 million in 2016 on the basis on enabled revenue and AI based gross value addition (GVA) prognoses. The market is projected to reach USD 35,870.0 million by 2025 by its direct revenue sources, growing at a CAGR of 57.2% from 2017 to 2025, whereas it is expected to garner around USD 58,975.4 million by 2025 from its enabled revenue arenas." (Grand View Research 2017).

It is interesting to notice the semblance between such graphs of exponential revenue and future projections of exponential growth of computer intelligence, surpassing human capabilities (e.g. Bostrom 2014). In his popular science book *Life 3.0*, physicist and cosmologist Max Tegmark, cited economist Erik Brynjolfsson's views on AI acting as a liberating actor, leading humanity towards a world of leisure, based on the idea that an AI-driven society will look like ancient Athens, with machines replacing human slaves (cf. Flusser's similar view cited above; yet so politically divergent).

> "The reason that the Athenian citizens of antiquity had lives of leisure where they could enjoy democracy, art and games was mainly that they had slaves to do much of the work. But why not replace the slaves with AI-powered robots, creating a digital utopia that everyone can enjoy? Erik's AI-driven economy would not only eliminate stress and drudgery and produce an abundance of everything we want today, but it would also supply a bounty of wonderful new products and services that today's consumers haven't yet realized that they want." (Tegmark 2017: n.p.).

Interestingly, Tegmark, as well as Bill Gates have publicly supported Hawking and Musk's views on the dangerous potential of AI in 2014 (Galanos 2019). Moreover, Tegmark became president of the Future of Life Institute, while Brynjolfsson is director of the Digital Economy Lab at the Stanford Institute for Human-Centered AI (HAI); both, instances of contemporary "ethical AI" initiatives built on rhetoric about the uncertainty of AI's future developments. The glass half full in terms of AI employment replacement was seen as half empty following the negative "job losses" narrative[66]. A very decisive moment was circulation of a preprint in 2013 of Frey and Osborne's study "The future of employment: how susceptible are jobs to computerisation." Google Scholar metrics already indexed citations to the article, thus increasing its visibility after relevant search queries[67], however, it was not until 2017 that the paper was published with *Technological Forecasting and Social Change*. By the time writing, the paper has nearly 9000 citations, and it has been used as evidence in justifying policy measures in regulating AI due to its very alarming tables of jobs lost to AI automations. The abstract reads:

> "In this paper, we address the question: how susceptible are jobs to computerisation? Doing so, we build on the existing literature in two ways. First, drawing upon recent advances in Machine Learning (ML) and Mobile Robotics (MR), we develop a novel methodology to categorise occupations according to their susceptibility to computerisation." (Frey and Osborne 2017: 254[68]).

[66] It would exceed the scope of the present work to extend references to the very old narrative of jobs lost to automation from the early industrial revolution and the Luddite movement, throughout the 20th century.
[67] This is mostly a personal observation. I first encountered the article in 2015, citing it in a paper which was eventually published in 2017 as a preprint (Galanos 2017), only days prior to Frey and Osborne's final publication.
[68] Although AI communities have been relatively silent in public discourse during this period, with only but a few exceptions, it is pertinent to cite Toby Walsh's humorous comment on Frey and Osborne's paper: "The authors used machine learning to predict precisely which of over 700 different jobs could be automated. It is of course ironic that a report about the automation of work was itself largely automated" (Walsh 2018: 100).

Tegmark's (and Brynjolfsson's) utopian visions of an economy liberated from labour's "drudgery" was also promoted through the lens of communist ideology in support of AI developments in China. Professor of Law at Tsinghua University Feng Xiang's article 'AI will spell the end of capitalism' on the *Washington Post* reiterated a very similar argument, however, in response to the high privatisation of AI tools; in Xiang's view, the massive pervasiveness of AI systems, combined with communist attitude, should first turn AI in a public good, and as a corollary, liberate humans from work:

> "China's socialist market economy could provide a solution to this. […] More than anything else, the inevitability of mass unemployment and the demand for universal welfare will drive the idea of socializing or nationalizing AI. Marx's dictum, 'From each according to their abilities, to each according to their needs,' needs an update for the 21st century: 'From the inability of an AI economy to provide jobs and a living wage for all, to each according to their needs.' […] It is the very pervasiveness of AI that will spell the end of market dominance. […] The dream of communism is the elimination of wage labor. If AI is bound to serve society instead of private capitalists, it promises to do so by freeing an overwhelming majority from such drudgery while creating wealth to sustain all. […] The communism of the future ought to adopt a new slogan: 'Robots of the world, unite!'" (Xiang 2018[69]).

Be it socialised or nationalised, AI still appears to be determinant force which will transform humanity, speaking directly to Jasanoff and Kim's (2009) notion of the nation-building sociotechnical imaginary. Optimism and pessimism in the context of ideological visions make extended use of AI's promissory potential. These polarised messages about the inevitability of a future brought by AI automation was not only imbued with religious determinism, but sparked "AI religions" as well (adding a layer of literal expression to the theologisation of AI discussed in section 1.2): Inventor and engineer Anthony Levandowski, co-developer of the self-driving car software Waymo, later sold to Uber, established in 2017 *The Way of the Future*, informally known as "church of AI" (Harris 2017a; 2017b – interestingly these articles which brought wider attention to the church were published on *Wired*), an event highlighting the esoteric and religious dimension this AI narrative, as suggested in the introduction. This church promotes "a peaceful and respectful transition" towards humanity's step into the singularity (Way of the Future 2017[70]).

Within this wide field of AI being re-established as a pole of attraction, carrying technical, financial, religious, and existential expectations. In the same year that Frey and Osborne's article got published, *The Way of the Future* was established, and ethical AI initiatives began flourishing, more critical perspectives challenging the AI promise in commerce and politics emerged. Since Google bought

[69] One cannot but observe the dialectic nature of Xiang's argument, abiding by communist principles: thesis (AI development as an automation tool), antithesis (AI exploited by massive corporations for private benefit and causing unemployment), synthesis (AI's pervasiveness leading to its transformation into a public good, liberating from work but also creating new forms of employment).

[70] The relationships between AI and religiosity do not end here. Floridi suggested that AI is divided into two churches: "AItheists," disbelievers of the possibility of AI achieving genuine intelligence, dismissing the entire project, and "Singularitarians" who act as future-tellers, referring to "futures that are conveniently close-enough-to-worry-about but far-enough-not-to-be-around-to-be-proved-wrong" (2015: 8). Others have previously presented the singularity as Cargo Cult Science; a metaphor employed by Richard Feynman (1974) to denote how an unknowable cognitive object might evoke religious feeling (Fernaeus et al 2009). Levandowski later became known for his 18-month prison sentence after being charged with "33 counts of theft and attempted theft of trade secrets while working at Google" (Korosec and Harris 2020). The founding of a religion by someone who admitted to have stolen trade secrets bears semblance to the aforementioned cases of "Ethical AI" proponents associated with harassment allegations.

the company DeepMind, its applications of neural networks and reinforcement learning allowed short-term human memory to be mimicked; the company generated hype when *AlphaGo*, an AI Go-playing programme, was victorious over the world Go champion Lee SeDol (Hern 2017, Oh et al 2017). This form of hype fostered a peculiar set of impact. On the one hand, AI expectations were inflated because, although chess-playing machines beating human champions were dismissed as threatening because of chess following formal logic, Go was considered to be context-based enough for any machine to ever become competent enough in winning against human champions. On the other hand, China censored broadcasting of the game between Lee SeDol and AlphaGo, based on Go being considered one of China's national games; thus, a machine outperforming a human, would mean loss of China's national pride (Hern 2017[71]). It was in the same period with DeepMind's hype about game playing neural networks, that the company exerted a similar pair of enthusiasm and controversy about patient data being used in the company's deal with the Royal Free London NHS Trust. This was an attempt towards aligning purposes, the NHS (the UK's National Health Service) aiming to invest in data-driven technologies to assist kidney disease prediction and DeepMind aiming to improve its AI services by training its models on 1.6 million patients' data. Revelations by *The New Scientist*, however, suggested that data acquired by DeepMind's app Streams from patients of three major London hospitals extended far beyond what was legally acceptable as relevant to the specific disease's prediction, while their further use by DeepMind and Google remained nefarious – resulting in legal requirement to delete all related data according to EU and UK regulations (while the case is still open in court). Public statements on behalf of DeepMind suggest an interesting direction in the promissory games around AI private companies in that time: "In our determination to achieve quick impact when this work started in 2015, we underestimated the complexity of the NHS and of the rules around patient data, as well as the potential fears about a well-known tech company working in health" (Revell 2017). Public distrust towards data-driven companies' handling of personal data makes now privacy one of the key factors for potential AI disillusionment.

The period between 2015 and 2019 marks the revival of critical and activist voices within the AI community, echoing CPSR's aims towards responsible, anti-militarist, and inclusive scientific practice. A number of events involving members of the AI and broader computer science community raised awareness about the intersectional harms associated with data-driven AI, leading from a series of observations to a systematic arena of critical AI scholars within the field, shaping its expectational trajectories and potential solutions. It was also following a period when the promissory framing of the internet's potential as a source of big data transformed the internet imaginary (Flichy 2007) into, first, a place of potential information overload, requiring data deletion to be recognised as a virtue (Mayer-Schönberger 2011), to, second, such gatherings of "big data" becoming a source for useful pattern extraction, labelling, sorting, and recommending (Mayer-Schönberger and Cukier 2013[72]). One of the critical opening points was software programmer's Jacky Alciné's viral tweet reporting on an early version of automated machine learning-based classification of images resulted in eighty images of one of his Black friends stored in Alciné's computer to be classified under the label "gorilla" (Alciné 2015). Academic policy discussions about the racist and sexist tendencies infiltrated, perpetuated, and augmented in AI outputs have began to emerge (Garcia 2016) together with more systematic and broadly reaching treatments of the topic such as

[71] I find it nevertheless ironic that past this decision, China developed one of the most advanced AI strategies (Allen 2019), appearing to some as the new oriental equivalent to Japan's FGCS (Walsh 2018: 243-245, McCorduck 2019).

[72] I leave for future research the interrogation of the two popular books written by the same author within a span of three years, offering two very opposing views on the topic, yet, under the same deterministic tone.

the popular book *Weapons of Math Destruction* by mathematician Cathy O'Neil (2016). In the spirit of "big data" as the key deterministic factor of critical research trends, O'Neil only situates the problems of social discrimination in the context of AI and machine learning in only few instances of the book, possibly indicating that the massive hype about AI as a term was yet to take-off as O'Neil was preparing the book – these references exist in the two chapters debating targeted online advertising. This was all paired to the transition from Barack Obama's presidency to the rise of Donald Trump as president of the US, and the role of algorithmically mediated dissemination of "fake news" on social media (Allcott and Gentzkow 2017), as well as the role of behavioural microtargeting and psychographic applications of machine learning algorithms by company Cambridge Analytica which continued to stir debate as to the extent by which has been of assistance to Trump's victory by the identification of voter types (Laterza 2021; Kanakia et al 2019 – consider the continuity with statistical reasoning in the previous section's references to Simulmatics; O'Neil hints to the initial revelations about Cambridge Analytica's involvement in collecting voter's data in chapter 10; O'Neil 2016). Detection of mis- and disinformation online was also paired to increasing awareness of visual disinformation in visual "deep fakes" following the 2014 publication of Ian Goodfellow and colleagues' work on generative adversarial networks (GANs, currently also known as "generative AI"). Following Hinton's success with pattern recognition in large visual datasets, Goodfellow et al (2014) showed how a computational training process of a model for a given image dataset, can produce variations of the same image according to prescribed parameters (currently known as "prompts"). This methodology's realistic output in offering alternate versions of digital content (visual, audio, text) paired to its increasing efficiency (as recently outlined by Goodfellow et al 2020; in collaboration now with "godfather" Bengio), raised further concern about the combinatorial effects of such algorithmic techniques with the internet's networking information dissemination speed. As Phillip Isola, another influential convolutional network scientist stated in the New York Times:

> "It took a few years until people started to make systems that could do the opposite: not take an image and recognize that it was a cat, but take the label "cat" and synthesize an image that looks like a cat — the inverse problem. You could make photos of really low-resolution faces. Very rapidly after that, people were able to use these things for face-swapping and deepfakes and all of that. The technology advanced so quickly right around those years. It went from 'O.K., this is a really interesting academic problem, but you can't possibly use this to make fake news. It's just going to produce a little blurry object' to 'Oh, you can actually make photo-realistic faces'." (Isola, quoted in de Luca and Beltran 2019).

When Facebook's Mark Zuckerberg was called to testimony to Congress about his company's involvement with the Cambridge Analytica scandal in 2020, among the over 600 question he was asked, many had to do with fake news. Interestingly, Zuckerberg suggested that what other saw as a challenge caused by deep learning, he saw as the solution, referring to deep learning detection of terrorist content which was successfully employed on Facebook's services[73].

[73] Facebook's chief technology officer Mike Schroepfer announced in 2020 a contest in which he challenged AI practitioners to develop AI software to detect AI-generated deepfakes: "I was pretty frustrated by the amount of time the ML industry spent making deepfakes better compared to the time spent combating the harm they could create. So we used a competition plus open science to spur more focus."
https://twitter.com/schrep/status/1271470864593121280
Consider also the following statement by one of my interviewees, reminiscent of a 1980s conference: "And then this issue came up. Should we allow the use of robots controlled by AI in warfare? And I thought the discussion

Computer vision company Clarifai took advantage of these challenges to instigate a research programme on content moderation to be able to distinguish inappropriate content within given contexts[74], such as pornographic from non-pornographic material. An intern working on that project in 2017 was Inioluwa Deborah Raji, a woman of colour, who quickly realised (a) that the images consisting the company's training datasets appeared to be predominantly White, out of which, the vast majority appeared to be male, and (b) that the dataset used to train the company's system to recognise pornographic material included the vast majority of all datasets' Black people. According to her personal accounts, this was a revelatory moment, after which she became increasingly attentive to such patterns of Whiteness and masculinity in social settings involving computer studies, such as conferences. The principle example being the 2017 NeurIPS conference – which, by that time, had become the main venue for machine learning practitioners to exhibit their work – whereby, according to Raji, there "were about 8,000 people […] and maybe less than 100 black people and not many women at all, so it was very overwhelming" (Gorey 2020[75]). A very similar account was offered by another female Black AI scientist, Timnit Gebru about the previous year's edition of the same conference:

> "What really just made it accelerate was [in 2016] when I went to NIPS and someone was saying there were an estimated 8,500 people. I counted six black people. I was literally panicking. […] At the same time, I also saw a lot of rhetoric about diversity and how a lot of companies think it's important. And I saw a mismatch between the rhetoric and action. Because six black people out of 8,500—that's a ridiculous number, right? That is almost zero percent." (Gebru, quoted in Snow 2018)

In a consecutive Facebook post following this event, Gebru pointed at a shift in the expectational environment of AI, at the time dominated by existential threat hypes:

> "I'm not worried about machines taking over the world. I'm worried about groupthink, insularity and arrogance in the AI community. […] AI is working for a certain very small segment of the world population. And the people creating it are from a very minuscule segment of the world population. […] And the people creating the technology are a big part of the system. If many are actively excluded from its creation, this technology will benefit a few while harming a great many." (Gebru, quoted in Metz 2021: 232-233)

---

was getting bogged down, so I offered a proposal: not only should we allow it but we should arrange for all the wars are between groups of robots shooting each other, but nobody liked that."

[74] Clarifai's founder, Matt Zeilier, was the 2013 receiver of the ImageNet prize, received by Hinton and colleagues in the previous year.

[75] The conference's original acronym, NIPS, sparked further critical debate in the same year, leading to two signed petitions aiming at changing the name as to avoid sexist connotations; while the first petition was dismissed given the observed bias at the conference's audience, the second one directed to a more diverse and global audience of machine learning practitioners generated the critical mass required: "The letter highlighted "disappointing behavior" at the 2017 event and said that the 'acronym of the conference is prone to unwelcome puns'. It gave examples from previous years such as an unofficial pre-conference event named TITS" (Else 2018). It could be assumed, however, that the workshops' first editions focusing on both natural and synthetic neural information systems employed the pun probably due to the gender neutrality in the existence of nipples in most mammals, in the spirit of investigation across mechanical and biological information processing (consider the 1990 edition poster: https://media.neurips.cc/Conferences/1990/Poster/NIPS-1990-Poster.pdf). Another conference following a similar approach in echoing the social demand for avoiding discrimination in the selection of humorous acronyms was ACM's conference on Fairness, Accountability and Transparency, being abbreviated as FAT* from 2018 until 2020, when the conference changed the acronym to FAccT for the 2021 edition to avoid discriminatory stances on the basis of the social stigma of obesity (https://twitter.com/facctconference/status/1222916502392909825).

At the 2017 edition of NeurIPS, Gebru, together with her colleague Rediet Abebe, organised a workshop titled Black in AI where Raji was invited to be a part of. Black in AI grew to be a much larger non-profit organisation raising awareness about racial inequality in computer science[76]. A similar organisation, focusing on intersectional elements of inequality, with gender and race as points of departure and a "full spectrum inclusion" to be stated afterwards, was founded the year before by Gebru's younger colleague Joy Adowaa Buolamwini, named the Algorithmic Justice League[77]. Buolamwini was currently working on her second MSc dissertation to be submitted in 2017, titled "Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers," one of the first (if not the first) computational approaches to challenge intersectional criteria's calculable reflections in the face of AI. Buolamwini was later revealed to be inspired by Frantz Fanon's book *Black Skin, White Masks* (Fanon 1952[78], Benjamin 2019: 124; Metz 2021: 235), a metaphor about the social guising of Black people to fit the White norm; Buolamwini reversed the social metaphor turning it into technical literal application and tested a white mask against computer vision/facial detection systems which functioned poorly on darker skin tones, especially female. Her thesis included dedications to Gebru, Raji and O'Neil. Her research on such "gender shades" turned into a research programme at MIT, leading to a series of publications and collaborations highlighting numerous variations of AI's functional fallacies (Buolamwini 2017; Gebru, Hoffman and Li 2017; Raji and Buolamwini 2019; Raji et al 2022), extending to corporate influence on tackling issues of discrimination.

In the same year that Black scholars identified the predominant Whiteness at AI communities, Margaret Mitchell, an ex-Microsoft Researcher applying Hinton's methodology in linguistic models who had moved to Google, exposed her views on a *Bloomberg* interview, after estimating that within five years she had been collaborating with hundreds of men and possibly ten women: "I call it a sea of dudes […], I do absolutely believe that gender has an effect on the types of questions that we ask" (Mitchell, quoted in Clark 2016). Mitchell belonged to, and partly shaped, an arena of AI scholars who, having credentials as AI scientists, wished to depart from mainstream (and very male) long-term expectational notions of AI as an existential threat, however, without submitting to an uncritical view of AI technology, thus steering within Google a group focusing on short/mid-term immediate issues on "Ethical AI" (Metz 2021: 237). Becoming aware of Gebru and Raji's work, she invited them to co-author a paper for FAT* on "actionable auditing" of data models, inviting companies with large datasets to acknowledge within "model cards" published together with a model's release (Mitchell et al 2019; a similar call for transparency on datasets was proposed recently in Gebru et al's 2021 "datasheets for datasets" paper). These papers became

[76] https://blackinai.github.io/#/about
[77] https://www.ajl.org/
[78] Black Marxist psychoanalyst Frantz Fanon's book *Black Skin, White Masks* was more than appropriate to use in order to reverse a metaphor and use it in a literal sense in Gebru's methodology to test AI systems. Reading Fanon's work in light of AI technology reveals his seemingly prophetic account on racial discrimination based on datafied segmentation in parts: "'Dirty n***r"' Or simply, 'Look, a Negro!' I came into the world imbued with the will to find a meaning in things, my spirit filled with the desire to attain to the source of the world, and then I found that I was an object in the midst of other objects. Sealed into that crushing objecthood, I turned beseechingly to others. […] I demanded an explanation. Nothing happened. I burst apart. Now the fragments have been put together by another self" (Fanon 1952: 109). Being a Marxist, he was also fully aware of the role of technological tools in the perpetuation of colonial practices and that such socially constructed technical arrangements are also responsible for the invention of discriminatory terminologies which he rejected in the spirit of Lacan's negation of *the* woman: "I, the man of color, want only this: That the tool never possess the man. That the enslavement of man by man cease forever. […] The Negro is not. Any more than the white man" (Fanon 1952: 231; for his Marxist and Lacanian heritage, cf. 161-164 footnotes, and 233 epigram).

influential to the extent that companies such as Google have adopted them[79]. At the the overall shift in AI's trajectory from a social justice perspective led to the formation of Partnership on AI, a "non-profit partnership of academic, civil society, industry, and media organizations creating solutions so that AI advances positive outcomes for people and society[80]." Its founding members in 2016 have been Google, Facebook, Microsoft, IBM, DeepMind, and Amazon but currently listing most major AI-related firms globally, partly showcasing how the promissory environment between overlapping arenas of corporate interest and activist science forged an agenda of AI terminologies associated with social change, as argued in section 3.1.

The corporate interest in "ethical" or "human-centred" practices did not come without consequences for advocates of social justice. Gebru had by that time received two invitations to join Google, once by Samy Bengio (head of Google Brain, Google's main AI department and brother of Yoshua Bengio) at the 2019 edition of NeurIPS after the Black in AI workshop, and by Mitchell in 2020, to join the Ethical AI team. Simonite (2021) offers a detailed account of Gebru's trajectory as a migrant throughout her career at IBM and Apple, until her involvement with Google. The latter choice proved the corporate interest's frailty in managing the will to absorb social critique, when, on the verge of publication of one of Gebru's papers, sponsored by Google, and highlighting the limitations and potential harms associated with GPT-3's GAN-based language model Google was imitating in some of its search engine function, after a rapid series of complex negotiations resulted in the ousting of both Gebru and Mitchell from the Google group. This sensitised a number of researchers in both academia and industry to reconsider the purpose of "ethics" as part of large corporate branding. According to Mitchell's later reflections: "It was like people really appreciated what I was saying, and then nothing happened" (Simonite 2021). At the 2020 edition of NeurIPS, taking place only days after Gebru's ousting from Google, Microsoft machine learning researcher Hanna Wallach delivered a keynote on a track devoted to broader social implications of AI, publicly addressing these reflections about the expectational motivations of AI researchers:

> "indeed these activities [addressing societal implications, stating assumptions, receiving feedback during public outreach, science communication] are dis-incentivised by our current research practices which reward the fast-pace submission of as many papers as possible. Moreover, in contrast to most other disciplines, we somehow picked up a culture of immodesty in which overselling is the norm and talking about limitations and negative results are discouraged. This means that even when we do see potentially concerning broader impacts and societal implications we are uncomfortable calling them out. But to set appropriate expectations for other groups of actors in the research to practice pipeline researchers need to communicate effectively and honestly about their work without overselling […] [U]niversities are losing people to industry or simply not hiring them in the first place and none of this [equal opportunities to multidisciplinary AI practice] seems realistic. […] Even if researchers do deeply understand the sociotechnical nature of machine learning and they have the skills to communicate it carefully and effectively about broader societal implications they might not be empowered to do so, or to do so publicly and honestly, especially if they are pre-tenure or work in industry. Take last week's events involving Timnit Gebru for example." (Wallach 2020)

[79] https://modelcards.withgoogle.com/about - consider such documentation strategies' incorporation in Shneiderman's recent textbook guide to human-centred AI as part of verification and validation testing techniques (Shneiderman 2022: 160).
[80] https://partnershiponai.org/about/

Wallach also refers to the "homegrown" culture of AI practitioners, being in part a legacy of computer culture of anarcho-capitalist individualism (similar to the "California Ideology"; Barbrook and Cameron 1996). An additional nuance to these approaches is the mix of anti-corporate interest and "homegrown" approaches with profit based on the idea of open data. One of Hinton's co-authors of the 2012 ImageNet paper, Ilya Sutskever, who worked at Google Brain between 2013 and 2015, "envisioned a lab that was entirely free of corporate pressures, a not-for-profit that would give away all its research, so that anyone could compete with the Googles and the Facebooks" – this thought resulted into the company OpenAI which received the financial backing of Elon Musk amounting to more than a billion US dollars (Metz 2021: 163-165[81]). OpenAI was responsible for the development of GPT-3, the large language model whose criticism led to Gebru and Mitchell's ousting from Google.

Such reflections about ethical banners being guises of capitalist interest, and capitalist interest being further guised by narratives of openness were further paired to increasing awareness on the pressing issues concerning "hidden labour" behind big data-driven AI practice, concerned with the data labelling, filtering, and collecting processes, something that the large corporate companies did not address as part of their efforts towards inclusive policies. The following passage from the Harvard Business Review captured the mismatch between the promissory value of AI and the hidden labour costs underneath, standing in stark contrast to Tegmark's neo-utopian AI vision about a labour-free AI-driven society:

> "Facebook created a PR firestorm last summer when reporters discovered a human "editorial team" – rather than just unbiased algorithms – selecting stories for its trending topics section. The revelation highlighted an elephant in the room of our tech world: companies selling the magical speed, omnipotence, and neutrality of artificial intelligence (AI) often can't make good on their promises without keeping people in the loop, often working invisibly in the background." (Gray & Suri 2017).

This has led a faction of researchers to speak about corporate "ethics washing" (the appropriation of ethical guidelines by high-level stakeholders without evidential implementation of them), further transformed into "ethics bashing" (the trivialisation of this phenomenon to the extent that it harms any potential of ethical assessment; Bietti 2020). Such critical stances towards the enterprise of Ethical AI have been further supported by criticisms concerned with the character of such enterprises' sponsors. In his article 'The Invention of 'Ethical AI': How Big Tech Manipulates Academia to Avoid Regulation,' Ochigane (2019) speaks about Joichi Ito, former head of the MIT Media Lab and regular financer of the Lab's $27 million 'Ethics and Governance of AI Fund' initiative whose financial connections with financer and sex offender Jeffrey Epstein were disclosed only after Epstein's criminal acts have been publicly revealed, yet previously known to Ito[82]. In the same spirit, Katz wrote the following assessment of AI as a manufactured

[81] Musk, deeply associated with AI's existential threat narratives had already prepare an excuse to invest in AI back in 2014 when during an interview claimed his investments come "not from the standpoint of actually trying to make any investment return […] I like to just keep an eye on what's going on with artificial intelligence. I think there is potentially a dangerous outcome there" (Galanos 2019: 424).

[82] The following passage by Ochigane deserves attention: "A former Media Lab colleague recalls that Marvin Minsky, the deceased AI pioneer at MIT, used to say that 'an ethicist is someone who has a problem with whatever you have in your mind.' (In recently unsealed court filings, victim Virginia Roberts Giuffre testified that Epstein directed her to have sex with Minsky.) Why, then, did AI researchers suddenly start talking about ethics?" (Ochigne 2019). Other researchers have come to criticise such institutionalisation, operationalisation, and bureaucratisation of "ethics" as ritualization (Åm 2019). While ethics is one of the widest of all theoretical (and practical) disciplines, and while AI ethics has grown into a distinct field reflecting manifold typologies of expectations and social concerns (e.g. code of conduct, design principles, privacy, manipulation, transparency, autonomy, automation, bias, morality, safety, well-being, and existential threat; Dignum 2019; Müller 2020), it is important to note that such specific criticisms refer to the idea of a reductionist

product which serves the benefits of large corporations. As a follow-up from the previous chapter, it is hereby shown how terminologies which appear as fixed entities in science and technology, can carry expectations and guise vested interests:

> "The label 'AI' has in fact recently undergone a rebranding. Corporations have helped manufacture an "AI revolution" in which AI stands for a confused mix of terms—such as "big data," "machine learning," or "deep learning"—whose common denominator is the use of expensive computing power to analyze massive centralized data. AI has essentially become a convenient redressing of a stale vision long promoted by Silicon Valley entrepreneurs. It's a vision in which truth emerges from big data, where more metrics always need to be imposed upon human endeavors, and where inexorable progress in technology can "solve" humanity's problems. Powerful companies have played a crucial role in the rebranding by hiring academics working on statistical analysis of big data (a term now interchangeable with AI), intervening more aggressively in academic research, and dominating mainstream discourse on AI." (Katz 2017: 2).

As he mentions below, "[i]t seems the term "AI" can be made to fit nearly any cutting-edge computation offered by computer scientists" (Katz 2017: 3). Pamela McCorduck, (see earlier discussion on her impact assessment of Japan's FGSC in the 1980s in section b, this section) reflects on the difference between early AI romantics of the Silicon Valley and their contemporary successors:

> "I sometimes wonder how the AI pioneers would regard present-day Silicon Valley. They'd sure be very pleased that AI is so prominent, highly honored, and pursued. [….] They might be less enchanted by a culture that revolves so single-mindedly around making money. Each of AI's four founding fathers lived modestly, in houses they'd acquired when they were new associate professors, houses where they'd brought their children up, where they ended their days. Science, not the acquisition of capital, drove them." (McCorduck 2019: 344).

Albeit the latter two references seem to be unrelated political and emotional commentaries, they share a common trait in that they indicate the lack of scientific involvement in contemporary settings of AI's promissory and expectational environment. STS scholar Harry Collins, who entered the discussion about AI's feasibility during the expert systems round of hype (Collins 1987; 1990), observes, on the one hand the impact of critical non-specialists on the AI community, and on the other the obligation for AI practitioners' involvement in the process of AI governance:

> "The more ambitious members of the AI community often seem engaged in trying to win a game against an opposing team – the critics – rather than searching for the truth (with the critics often falling into the same pattern). Yes, the goalposts are always being moved by the critics, but the real problem is that it is the critics who are moving them. […] To repeat, when AI was an orphan discipline starved of funds there was an excuse – even if not a very good one – to put the best possible gloss on what was being accomplished; nowadays, those at the frontiers of the AI community are pretty well the most powerful body of research scientists in the world, and without their input the policing and evaluation of AI's accomplishments could not be as thorough as it should be. They should, like the physicists, become an object lesson for the proper conduct of science. I know this is possible because I've seen it with the physicists." (Collins 2018: 177).

---

understanding of both ethics and AI, whereby AI can be "ethical" as much as it can be a salvation or a catastrophic agent.

The general lack of practitioners entering public debates about AI was noticed in the absence of AI/robotics regulation policy documents published in 2016 (Galanos 2019), which has led a few, yet influential, AI researchers to offer responses to prestigious figures such as Musk. The case of MIT's Rodney Brooks, a constant defendant of public outreach in AI is among the first ones to speak openly about AI policy goals; Brooks criticised Musk's suggestions to regulate AI because of existential threat fears as nonsensical due to lack of foreseeability of such a scenario (Loizos 2017). What Collins, being in the UK, probably did not see in his 2018 reflection, is the interesting mingle of both technical and social scientific expertise boiling during that period when elder generations of "critics" were again concerned with the re-emergence of ultraintelligent/singularity debates. It is important to take into account the US political climate. At a symposium titled *AI Now* hosted by the White House and New York University's Information Law Institute on July 7, 2016 during Barack Obama's final months of presidency, Microsoft researcher and media studies scholar Kate Crawford and Google researcher, technology entrepreneur, and rhetorician Merdith Whittaker co-founded the AI Now Institute, being "the first of its kind, bringing together experts and researchers across computer science, economics, law, academia and other sectors to explore the socially responsible creation and use of AI technologies" (Ryder 2017). AI Now aimed at diversifying the sources of data for AI models, as to protect mis- or under-represented groups and create partnerships between researchers, the industry, and international organisations such as ACM, IEEE, and AAAI for the standardisation of AI codes of ethical conduct (Crawford et al 2017). AI Now continues to publish annual reports on topics they consider pressing concerned with AI's design and implementation, receiving, among many others, the support of Lucy Suchman, being the active STS tie between CPSR and contemporary critical discourses in AI. While Suchman preserved the connection to the 1980s AI resurgence, younger STS and media studies scholars also joined these alliances, such as Ruha Benjamin, being an active supporter of Black in AI (Benjamin 2019[83]) or Alex Campolo being a regular contributor to the institute's reports.

Ironically, this is contradicted by other specialists, that leave little space for the researcher but to consider them as members of very distinct arenas. One such specialist is Stuart Russell, who has supported Hawking's statements and the establishment of ethical AI institutions (2019[84]). Collins' remark on AI community's goals being defined heteronomously by non-specialist commentators was particularly visible in the responses to politician, diplomat, and ex-US Secretary of State Henry Kissinger's warnings about the future of AI. In 2018, Kissinger published an article on *The Atlantic*, titled 'How the Enlightenment Ends: Philosophically, intellectually – in every way – human society is unprepared for the rise of artificial intelligence' in which he suggested, among other deterministic claims that "AI may change human thought processes and human values" (Kissinger 2018). The article received sufficient attention (perhaps not as wide as statements by Hawking or Musk), however, AI communities took the chance of defending AI against false claims. Gong (2018) collected a number of tweets and further online responses by AI

[83] And indeed, featuring on the group's Wikipedia page by the time writing, albeit not an official member; moreover, highlighting the significance of Fanon's work in Buolamwini's Gender Shades.
[84] Hawking was the first among the four authors (Russell, Tegmark, and Wilczek being the rest) who co-authored for *The Independent* the article bearing the title containing Hawking's name: 'Stephen Hawking: 'Transcendence Looks at the Implications of Artificial Intelligence-But Are We Taking AI Seriously Enough?' (Hawking et al 2014). Interestingly, Russell was also keynoting at the NeurIPS 2020 panel when Wallach expressed her concerns about the future of ethical AI from within the industrial sphere. We have also encountered him in 3.1 as the co-author with Norvig who has produced with the first AI definitions taxonomy.

researchers. One of them, Yann LeCun (whom we encountered previously as the developer of convolutional networks in 1989) stated:

> "I think this "old and overblown narrative" is a consequence of two things: 1- serious scientists not venturing into speculative futuristic predictions to preserve their credibility as serious scientists (nothing wrong with that). 2- non scientists, who are used to holding the levers of power in society, trying to maintain their diminishing control over things by telling scientists that they know nothing about the "real world". They are the ones worrying about "robots taking over" because it's their job to "take over". But as we all know, this desire to take over is somewhat inversely correlated with intelligence." (LeCun, cited in Gong 2018).

Other statements include Oren Etzioni's ("Hey, Dr. Kissinger--do you want to hear my thoughts about the Vietnam War?"), Mic Wright's ("AI is a long way from being so dangerous that it plans secret bombing campaigns against Cambodia, murdering more than 4,000 people"), and Suresh Ventakasubramanian ("How about an article by an AI on the threat of Henry Kissinger? Should be easy to generate"; all cited in Gong 2018). Gong proceeds in showing the international impact the original article had however, in opposition to the close circles of AI community tweets, including its consideration by famous politicians (such as US Defense Secretary Jim Mattis) and other notable figures:

> "Alarmingly, however, a search on Twitter for Kissinger's message had a high proportion of international engagement, with messages sharing it in French, Spanish, and German among others. The message was indeed broadcast and shared worldwide. In fact, the First Minister of Scotland, Nicola Sturgeon, even shared the article as something to 'get your intellectual juices flowing'." (Gong 2018).

Kissinger repeated the same arguments on a follow-up article in *The Atlantic*, this time joining forces with business magnate, software engineer, and ex-Google CEO and executive chairperson Eric Schmidt, and computer scientist and Amazon co-director Daniel Huttenlocher (Kissinger, Schmidt and Huttenlocher 2019[85]) which seems to repeat the pattern of alliance between a famous non-specialist commenting on AI with certain supporters from the technical community, as in the case of Stuart Russell's alliance to Stephen Hawking.

On the far end of AI specialists who embrace the views on AI's unwanted consequences, there are those who go as far as heralding the singularity. One such case is an article published on the popular online magazine *Futurism* titled 'The "Father of Artificial Intelligence" Says Singularity Is 30 Years Away.' Jürgen Schmidhuber, "the Co-Founder and Chief Scientist at AI company NNAISENSE, Director of the Swiss AI lab IDSIA" suggested that the singularity "is just 30 years away, if the trend doesn't break, and there will be rather cheap computational devices that have as many connections as your brain but are much faster" (Creighton 2018). Interestingly, Schmidhuber is labelled as 'father of AI,' although, a quick search reveals that he was born in 1963, eight years after the terminological birth of AI and the historical advances outlined above; thus showing the way in which media reflect historical unawareness of the field and how distorting effects might be enabled. It took some time for AI researchers to get closer to what Collins proposed and it is not yet possible to say whether the current stance of AI research communities is a form of response to the wider social enabling through sensationalist writing. An early defence against more general forms of hype (and a warning about AI winters) came, again, from LeCun, however, published on the *IEEE Spectrum*, mostly read by the AI communities without general outreach. Among

[85] The three authors have also co-authored a book (Kissinger, Schmidt and Huttenlocher 2021).

other topics, some quite technical, discussed during an interview, LeCun takes opportunity to speak about the hype, after being asked to expand on misunderstandings surrounding neural networks:

> "My least favorite description is, 'It works just like the brain.' I don't like people saying this because, while Deep Learning gets an inspiration from biology, it's very, very far from what the brain actually does. And describing it like the brain gives a bit of the aura of magic to it, which is dangerous. It leads to hype; people claim things that are not true. AI has gone through a number of AI winters because people claimed things they couldn't deliver. […] It [hype] sets expectations for funding agencies, the public, potential customers, start-ups and investors, such that they believe that we are on the cusp of building systems that are as powerful as the brain, when in fact we are very far from that. This could easily lead to another 'winter cycle'." (LeCun, cited in Gomes 2015[86]).

LeCun has nevertheless shown further interest in public outreach in the forthcoming years. An interesting example of public outreach is his 2017 public debate at the New York University's Center for Mind, Brain, and Consciousness, where he aimed to outline reasons to consider the current AI hype as probably harmful and outline limitations. The debate was against Gary Marcus, a researcher on the intersection between psychology, biology, and cognitive science, whose early work with his PhD supervisor Steven Pinker used expert systems and connectionist methods as an example of failure to model human intelligence, as part of their studies on children acquisition of knowledge (Marcus et al 1992). Marcus has since 2012 shifted his attention more closely to AI, tracing the concept's realtions in psychology and behavioural science (Marcus 2018: 7), and has become an advocate of "critical appraisal" to the field, as exhibited in his 27-page ArXiv preprint publication, where he reflects on past history of deep learning, its current successes, but also highlights ten factors which may lead current AI into a "wall":

> (1) "Deep learning thus far is data hungry"
>
> (2) "Deep learning thus far is shallow and has limited capacity for transfer"
>
> (3) "Deep learning thus far has no natural way to deal with hierarchical structure"
>
> (4) "Deep learning thus far has struggled with open-ended inference"
>
> (5) "Deep learning thus far is not sufficiently transparent"
>
> (6) "Deep learning thus far has not been well integrated with prior knowledge"
>
> (7) "Deep learning thus far cannot inherently distinguish causation from correlation"
>
> (8) "Deep learning presumes a largely stable world, in ways that may be problematic"
>
> (9) "Deep learning thus far works well as an approximation, but its answers often cannot be fully trusted"
>
> (10) "Deep learning thus far is difficult to engineer with" (Marcus 2018)

To further sustain my argument about the transferability of expectations from internet/ICTs debates to machine learning/AI ones, consider the following 2001 critique of the metadata (data describing data) hype by digital rights activist Cory Doctorow where he aimed at criticising the "meta-utopia" suggesting that the Web 2.0 could become an ideal space for every knowledge-seeking human:

> "2.1 People lie

[86] Interestingly, during the 2015 edition of NIPS, LeCun had also warned Sutskever about his potential fail with the OpenAI organisation due to its non-for-profit structure (Metz 2021: 166).

2.2 People are lazy

2.3 People are stupid

2.4 Mission: Impossible -- know thyself

2.5 Schemas aren't neutral

2.6 Metrics influence results

2.7 There's more than one way to describe something" (Doctorow 2001)

What is indicative of this comparison is that AI's far outreach as a concept, influenced by the science fiction narratives of previous decades, requires critique which take it (in the form of deep learning) to be the subject of failure – whereas in a field as niche as metadata, without mythological and literary connotations, it is much easier to shift the blame to human societies instead of using deterministic language, even for a non-deterministic cause. What becomes increasingly relevant at this stage of promissory debate, as in the case of LeCun debating Marcus, is that the debate shifts from whether AI can become sentient or not, but to a more nuanced conversation of varied balancing between the ability to defend a field and the will to critique and highlight limitations. While Marcus's critiques are not disagreed upon by LeCun, it is his prior closer expertise in AI motivating him to enter a "debate." When Ben Poolio, researcher at Google Brain tweeted about the ten challenges in Marcus's preprint, and expressed disappointment about the lack of technical recommendations, LeCun's response was the following: "The number of valuable recommendations ever made by Gary Marcus is exactly zero. Criticisms abound, though" (Marcus, responding to Poole 2018).

Certain AI scientists aim at taming the hype by acknowledging findings from qualitative/social sciences or commonsense thinking. Computer science/philosophy specialist Brian Cantwell Smith suggests an initial alliance between AI and the qualitative sciences which could later lead to incorporation of the latter's findings into novel AI research agendas[87]. For him, AI should take into account the world as a "plenum of surpassingly rich differentiation," and he further suggests:

> "AI needs to take on board one of the deepest intellectual realizations of the last 50 years, joining fields as diverse as social construction, quantum mechanics, and psychological and anthropological studies of cultural diversity: that taking the world to consist of discrete intelligible mesoscale objects is an achievement of intelligence, not a premise on top of which intelligence runs. AI needs to explain objects, properties, and relations, and the ability of creatures to find the world intelligible in terms of them; it cannot assume them." (Cantwell Smith 2019: 35).

Similar proposals for the need to investigate the alignment of human values and AI systems from a phenomenological perspective which takes into account the deeply contextualised nature of the world have been expressed (Han et al 2020). Nevertheless, AI communities become increasingly ambitious in their goal-setting. A report from MIT's DSpace laboratory researchers Holmes and Winston outlines their attempt at stories-recognition software; that being a very difficult task in AI due to the requirement of vast contextual reasoning for a given entity to perceive series of words as "stories." Their report is rich in expectations setting:

> "We believe that tomorrow's AI will focus on an understanding of our uniquely human intelligence emerging from discoveries on par with the discoveries of Copernicus about our universe, Darwin about

[87] Cf. Collins (2018) who, although not a computer scientist offers his expertise as an expert on knowledge to counter AI hype through very similar arguments on commonsensical thinking.

> our evolution, and Watson and Crick about biology. These cognitive mechanisms will take to another level applications aimed at reasoning, planning, control, and cooperation. Tomorrow's AI applications will astonish the world because they will think and explain themselves, just as we humans think and explain." (Holmes and Winston 2018).

While so far I have explored contemporary AI hype as the product of technological advancement, financial projections, public figures and commentators, as well as specialists' promises, it is important to refer to DARPA's renewed interest in AI given its pivotal role in funding AI research during both previous rounds of AI hype and as evidence of continuity of the military shaping of AI (as exemplified in Edwards 1996 and DARPA insiders such as Fouse, Cross and Lapin 2020). Between August and the fall of 2017, DARPA representatives visited Google and Clarifai to seek assistance in developing advanced military drone technologies which could identify targets. During a discussion with Clarifai's engineers, the military personnel enquired whether Clarifai could promise identification of specific buildings (like mosques, "often converted into military headquarters by terrorists and insurgents," according to military visitors) or distinguishing between men and women, particularly in the cases of men wearing dresses – according to the military representatives, human soldiers have learned to identify such men through gaps between the legs – but they required the promised accuracy of machine learning to be able to do that more quickly (Metz 2021: 239-240). Clarifai, the same company where Raji interned during the same year, secured a contract with DARPA and continues to support what was later revealed to be called DARPA's Project Maven: "Every member of Clarifai's Project Maven team agreed to work on the project, and the two people who chose not to participate were assigned to different efforts across the company," (anonymous Clarifai spokesperson, quoted in Conger and Metz 2018; for Clarifai's continuous support to DARPA, see Brewster 2021 and the company's official statement[88]). While only one Clarifai employee did quit after the meeting when killing was formally discussed with Clarifai, this was not the case with Google.

Google's services had already been used for military purposes: "Some of that collateral damage is now landing on the heads of British soldiers, as insurgents are reportedly using images gleaned from Google Earth to pinpoint mortar and rocket attacks against the most vulnerable targets inside military bases" (McNamara 2007). It was during Obama's administration that Eric Schmidt, then chairperson of Google also became chairperson of the newly created civilian organisation Defense Innovation Board, "that aimed to accelerate the movement of new technologies from Silicon Valley into the Pentagon" (Metz 2021: 242). After a series of negotiations between August and September 2017, in which Google's chief executives conversed with the US Department of Defense's representatives, and have received advice by Google's AI department that incorporating ethical principles by design within military AI systems was far from realisable, they proceeded with signing the contract with Project Maven (Fang 2018). A series of email communications between Google's sales team who wondered whether the achievement should be publicised or not was leaked with Fei Fei Li, the scientist who initiated the ImageNet project. Li, by then an Assistant Professor at the University of Illinois Urbana-Champaign moving to Stanford, becoming director of the Stanford Artificial Intelligence Lab (SAIL; founded between 1962 and 1965 when McCarthy and Feigenbaum received ARPA sponsorship to work on AI[89]) had already become chief executive of Google Cloud's AI department. In the same year, she was also acting as supervisor of Timnit Gebru's thesis, and a public advocate of the human-centred AI theme. Li's response to Maven was very

[88] https://www.clarifai.com/blog/why-were-part-of-project-maven
[89] https://ai.stanford.edu/about/
https://amturing.acm.org/award_winners/mccarthy_1118322.cfm

positive but she insisted that a diplomatic stance should be kept, in order to preserve the ethical image built around Google:

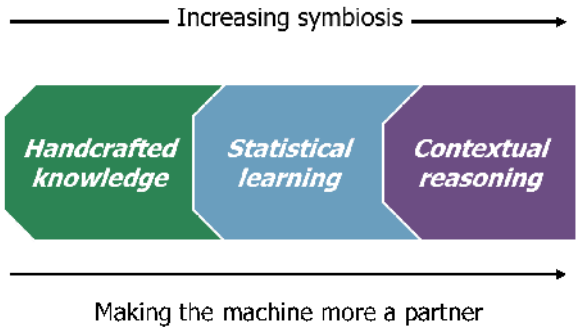
> "It's so exciting that we're close to getting MAVEN! That would be a great win. I think we should do a good PR on the story of DoD collaborating with GCP from a vanilla cloud technology angle (storage, network, security, etc.), but avoid at ALL COSTS any mention or implication of AI. Google is already battling with privacy issues when it comes to AI and data; I don't know what would happen if the media starts picking up a theme that Google is secretly building AI weapons or AI technologies to enable weapons for the Defense industry. […] Weaponized AI is probably one of the most sensitized topics of AI — if not THE most. This is red meat to the media to find all ways to damage Google. You probably heard Elon Musk and his comment about AI causing WW3. […] Google Cloud has been building our theme on Democratizing AI in 2017, and Diane and I have been talking about Humanistic AI for enterprise. I'd be super careful to protect these very positive images." (Fang 2018; Metz 2021: 245)

The company agreed to avoid advertising; however, within a few months, Google employees have discussed internally the ethics of Google being in support of military practices. By February 2018, Meredith Whittaker of the AI Now Institute began petition addressed to Google executives and Li that demanded the detachment of Google from DARPA. Within two months, 3.100 signatures were gathered, including Bengio and Hinton's, leading to Google's eventual dismantlement of the contract with Project Maven (but not its further collaboration with DARPA; Brewster 2021). What is crucial for the development of AI's promissory environment, is the position of researchers between arenas. As Metz notes: "Li was caught between what her bosses wanted in industry and what her peers wanted in academia" (Metz 2021: 249; he further notes that Li reported receiving death threats on Chinese message boards, possibly in light of her Chinese origin and China's growing interest in military AI investment). It was during the period of internal Google unrest that Gebru was warned to avoid joining Google because of that reason. After Gebru was ousted by Google, she stated "We need more support for external work so that the choice is not 'Do I get paid by the DOD or by Google?'" (Simonite 2021). In September 2018 announced *AI Next Campaign*, US's Department of Defence's novel effort in supporting AI. It is noticeable that no mention is given to DARPA's cut of funds to AI in previous rounds of AI hype; DARPA now presents itself as a constant supporter of AI research:

> "Today, DARPA continues to lead innovation in AI research as it funds a broad portfolio of R&D programs, ranging from basic research to advanced technology development. DARPA believes this future, where systems are capable of acquiring new knowledge through generative contextual and explanatory models, will be realized upon the development and application of "Third Wave" AI technologies. DARPA announced in September 2018 a multi-year investment of more than $2 billion in new and existing programs called the "AI Next" campaign. […] AI Next builds on DARPA's five decades of AI technology creation to define and to shape the future, always with the Department's hardest problems in mind. Accordingly, DARPA will create powerful capabilities for the DoD by attending specifically to the following areas." (AI Next Campaign 2018).

With this, AI comes full cycle in its deep association with DARPA's rounds of investments in AI. Interestingly, articles published in *AI Magazine* aim at showing continuity in investment ("sustained investments," Highnam 2020: 83; "DARPA has always been interested in AI frameworks," Fouse, Cross and Lapin 2020: 3). Despite what have been considered to be landmarks of AI winters through "DARPA's frustration with the Speech Understanding Research program at Carnegie Mellon University" and

"DARPA's cutbacks to academic AI research in general" between 1971 and 1975, as well as the "cancellation of new spending on AI by the Strategic Computing Initiative" (Wikipedia 2021b; cf. Crevier 1993, Roland and Shiman 2003) in both official and informal historiographies, military actors have the ability to rewrite history, showing continuity across their strand of AI research. New capabilities based on contextual reasoning have attached various new modifier to AI: robust AI, adversarial AI, high performance AI, and next generation AI are all new versions of AI's promissory environment with the question concerning its official "success" or "failure" to remain open. The following table, taken from DARPA's website, summarises the promise of contextual reasoning, presenting AI as a series of three waves, from "GOFAI" to connectionism and contemporary attempts at context-understanding machines who will work as partners to humans in light of US's national security.



By the end of 2021, Schmidt's co-authored book with Kissinger was published.

3.3 Discussion: The Longitudinal Assessment of Expectations and Expertise in AI

Viewed together, the two divisions of this chapter, historical tracings of AI's conceptualisation and promising reveal a number of lessons relevant to the SoE and SEE. The collective "herd behaviour" property of expectations formation is longitudinal as a waveform exchange between grand narrative fantasies and individual, actor-specific political/economic decisions, interests and beliefs. Early AI researchers have been trying to influence both scientific and public environments through various promissory games. While much of contemporary misunderstanding of AI may have stemmed from science fiction depictions of AI such as *2001: A Space Odyssey*, such depictions have been partly shaped by AI researchers in their attempt to gain visibility as scientists within a nascent field. Much of contemporary AI imaginary assessment might take the form of "there is difference between actual AI and science fiction AI," "science inspires science fiction," or "science fiction inspires science." However, through AI's longitudinal analysis, we can see that there is a finely nuanced spectrum of influence, which, at earlier stages, allowed AI to expand as a broad field of interest – Minsky would probably consider it a success for something like HAL to be created; his final book on creating artificial emotion and commonsensical thinking can stand as evidence (Minsky 2007). Narrower, arenas of AI expertise were visionary enough, in their overall struggle to establish recognition of a new field and by doing so generated an entire follow-up culture fascinated not only by science fiction, but of the very possibility that there are actual scientists working on what can be found in science fiction books and films. But precisely due to this nearly intentional promissory blurrification between scientific research and promissory imaginaries, escalating claims about AI by mixed arenas of more to less practical expertise have resulted into the mandate for AI regulation. It is not that Hawking, Musk, or Tegmark have been completely ignorant about AI scientific discourses and that they draw their knowledge from purely scientific claims; their own statements came in

the aftermath of publication of books like Minsky's, or roboticist David Hanson's extensive promotion of his vision to create intelligent robots, with Sophia being one of the most prominent examples in 2013 (Hanson, Bar-Cohen and Marom 2009; Parviainen and Coeckelbergh 2020).

The second round of AI hype was the outcome of national strategies and international competition; a similar story is observed in the contemporary "AI race." Much of the excitement about AI was earned as researchers made use of the term in order to attract governmental funds. Use of terminology was not consistent and different views are being reported: on the one hand, defenders of AI aimed at lowering the expectations about AI, yet sustain the term; on the other, those who feared a new "AI winter" preferred to create alternative terminologies. Longitudinal observation of shifts in expectations shows that the latter create shifts in conceptualisation, and vice versa. All this will be further reflected in the ways contemporary AI specialists speak about their field in the interviews examined below.

While non-technical disciplines with interest in AI sustained notions about AI's future, AI technologies advanced through adjacent fields and a migration of expectations (Konrad 2006: 439) took place, together with the literal, material migration of researchers from the UK to the US during times of AI funding stagnation. This did not stop researchers from using the term in contexts different than publications or grant proposals. AI proceeded as a series of technological and cultural spinoffs; certain AI applications became parts of everyday computer technologies during the steady popularisation of computers and the internet from the 1990s onward, with ICTs and internet technologies being the most prominent cases, along with developments in minor branches, such as object recognition. In the 2010s, excitement about pattern extraction from large datasets, paired to publicity of certain events (AlphaGo victory over Lee SeDol, progress in brain simulation) led influential figures to make provocative public statements about AI's hazardous potential; this was further reflected in a series of policy documents which enquired about AI regulation. At the same time, investment and management consulting companies presented AI as an area of growing interest. AI specialists were late in publicly defending AI's state of the art against critics; possibly due to their own focus on the positive side of the hype wave. Some exceptions of insiders supporting AI in public did appear, however, paired to alternative voices by specialists who did not abstain from considering narratives of the existential threat character. Moreover, what constitutes an "expert" on AI becomes challenged due to the high popularity of public figures commenting on AI, ranging from politicians to business magnates and astrophysicists. "Boring" statements by specialists do not sell as much as the exaggerations on non-specialists; with some specialists being eager to collaborate with the latter in partaking in their glory. Due to lack of knowledge about each actor's vested interest, it is difficult to make concise interpretations. What can be said, is that the complex ensemble of polarised positive and negative expectations, often evoking religious undertones, rebranded the "AI phenomenon" or "AI belief" in the context of massive corporation interests, the military, and certain strands of academics who interpret AI flexibly, associating their names with AI expertise, showcasing their ability to keep AI under control in an ethical and human-centred manner.

History of AI appears as highly interpretative depending on degrees of hype: while some would consider DARPA or Lighthill markers of AI winters, the "winter" dissolves once DARPA wishes to fund AI again, presenting itself a continuous supporter of AI development, or when social groups such as market analysts project graphs of exponentially increasing investment returns. At this stage, AI's unique form of hype blurs the boundaries between insiders and outsiders: AI specialists appear to make use of the highly promising AI environment not out of technical or scientific curiosity, but in order to remain relevant, either by responding to critics, or by creating alliances with influencing non-specialist actors which will offer them publicity, visibility, and funding. AI specialists are now able to use their expertise as currency to protect the audience of those who fund them from an imaginary enemy and created threat. What is left for this thesis after the present historical outset, is to show what interviewed AI specialists commented on the promissory environment in relation to AI as a concept, its attached promises, its funding strategies, and the role of policy. Another lesson from the historical analysis of AI conceptual/promissory

trajectories is the importance of taking interdepartmental discord, and informal communications more seriously into account when assessing the presumed neutrality of technoscientific progress and speculation; this invites for a problematisation of the concept of enactors/selectors in the SoE/SEE context.

Thomas Kuhn, in his description of the social structure of scientific revolutions placed emphasis on the role of generational shift (1963) – paradigms change effectively once members of the previous paradigm generation cease to cast influence on contemporary debates. The early symbolic/GOFAI pioneers are all retired or dead by now. Figures of contemporary influence include those who in late 1970s-1980s period advanced the connectionist methods in the form of convolutional neural networks and pattern recognition. While the two technological options were developed almost simultaneously by enactors (Rosenblatt's perceptron being the first instance of connectionism, only two years after the Dartmouth proposal), various types of interdepartmental (in the case of Edinburgh) or "official history" (the "perceptron controversy") discords, and lack of vast datasets and computer power, did not allow early association of connectionist methods with AI hype. But the contemporary dominance of AI debates (and its public advocacy) by people who developed alternative approaches to classic AI, such as Hinton, LeCun, or Brooks, shows, on the one hand a perceived paradigm shift which is, unfortunately, detached from its historical context. It is probably left to the alliances between the technical and the social to maintain the responsible curiosity-driven research which detaches itself from the large corporate and military interests. Suchman's involvement in both CPSR and AI Now groups acts as symbolic bridge across decades that preserves potential for a socially and historically aware AI technoscientific field. While Li's position as both herald of a humane AI theme within large capitalist corporations and being split between academic exploration and military exploitation, acts as a symbol of contemporary society's simultaneous fragmented and yet mutually shaping structures. Symbolic AI forgetfulness results in confusion: the general/mythological AI imaginary of human-level brain imitation feeds into the regulation of connectionist methods: the dead paradigm's fear governs the current paradigm – thus inviting for contemporary researchers' demarcation of their position in terms of scientific development and concern foresight.

In the table below, I have tried to summarise most key actors, institutions, and events that formed arenas of promise and formulations of expectation enaction in a grid which aims to trace the longitudinal assessment of multi-regime dynamics (Konrad et al 2008) based on different variations of expectation-expertise instantiations. It should be noted that numbered arenas in one phase do note necessarily correspond to arenas of another phase, although there is definitive evolutionary overlap, as much as schisms, alliances, and introduction of new arenas. At this stage, I am hesitant to make claims about "selectors" (in Bakker and Budde 2012's terms); while selection has taken place within given contexts (e.g. Lighthill, campaign against Project Maven, EU AI Act), to speak about a longitudinal selection would be monolithic, at the very least, without taking into account the precise argument advanced, concerning fluctuations and transferability (migration, to echo Konrad 2006) of hype. I hope that the following table can be useful to social science assessment of broad technoscientific expectational environments of large-scale fields with long historical negotiation, especially when temporality of expectational and promissory shifts is considered. Thus, I recommend that the expertise assessment, drawing loosely from Collins and Evans (2002), should proceed by considering not just the occasional credentials of a given scientist or statesperson. AI's metamorphic ability (to transform itself due to its social shaping, not itself transforming society) invites for holistic and careful assessments of AI representatives which should reflect on their understanding of the technical and social dimensions of the field. This actor mapping should then consider lessons from the SoE assessment of individual/collective, formal/informal, micro/macro expectations development, and thus place actors, groups, and events of significance within arenas of promise, motivation, and positionality. This, although highly dependent on available data and the will of researchers to conduct the tedious detective and scavenging work, can allow delineations between nation-building imaginaries (Jasanoff and Kim 2009) and social/psychological imaginaries mixing sensationalist hype and existential/religious concerns and fascinations (Flichy 2007).

Phase 1: 1955-1973

Promises and definitions as motives to establish or criticise a new field

- Arena 1: Academic researchers trying to establish AI as a scientific field - different approaches emerge (symbolic AI and perceptrons, McCarthy et al, Rosenblatt, Newell and Simon)
- Arena 2: Academic researchers opposing the effectiveness of AI influenced by practice-oriented governmental policies
- Arena 3: Military funding for AI after academic convincing (DARPA)
- Arena 4: Parallel evolution of statistical simulation and connectionist methods used for political purposes (Simulmatics)
- Arena 5: Science fiction initially influenced by AI research then preserving AI narrative (2001: A Space Odyssey)

Phase 2: 1974-1980

AI debates redirected to non-technical fields, symbolic AI becomes rebranded to create technical expectations

- Arena 1: Symbolic AI researchers conduct theoretical research as AI, defending against critics (e.g. Minsky, McCarthy)
- Arena 2: AI researchers develop commerical expert system technologies and reward/punishment algorithms usually without referring to AI (e.g. Feigenbaum, Barto) - while computer scientists depart from the AI front in order to criticise the vision of human reason replaced by automation (e.g. Weizenbaum)
- Arena 3: Social scientists criticising black-box-ness of AI (e.g. Winner)
- Arena 4: AI debates transferred to philosophy, psychology, literature, and art (e.g. Sloman, Hofstadter, Reichardt)

Phase 3: 1981-1999

AI's identity is negotiated in light of its industrial and military orientation, neural network promises are developed not as AI

- Arena 1: Symbolic AI researchers strategise as to prefer or avoid the use of AI as a term - reevaluate its merits and weaknesses, or expand its scope (e.g. McDermott, Schank)
- Arena 2: Market of a consumer-oriented AI business minimises theoretical AI research (e.g. LISP machines, expert systems)
- Arena 3: Academic researchers developing neural network methods without calling them AI (e.g. LeCun, Hinton, Bengio)
- Arena 4: Military funds AI in light of international competition with researchers responsible for the internet's development (SCI, FGCS)
- Arena 5: Social scientists advocate for social responsibility as part of AI communities (e.g. Suchman and CPSR)
- Arena 6: Governments invest in general computerisation (e.g. ESPRIT, Alvey, Information Superhighways)

Phase 4: 2000-2008

AI debates redirected to sensationalist discourse and non-technical fields, applied AI develops not as AI

- Arena 1: Symbolic AI applications are fragmented into different fields, robotics developments concentrate more attention than AI (the "AI effect"; e.g. object recognition, ALife)
- Arena 2: Computer scientists develop data annotation and classification techniques after the internet Web 2.0 boom (e.g. ImageNet, IBM)
- Arena 3: Singularity/transhumanist debates become more popular (e.g. Bostrom, Kurzweil, Warwick)
- Arena 4: Humanities scholars interested in social/ethical implications of ICTs and robotics (e.g. Floridi, Mayer-Schonberger)

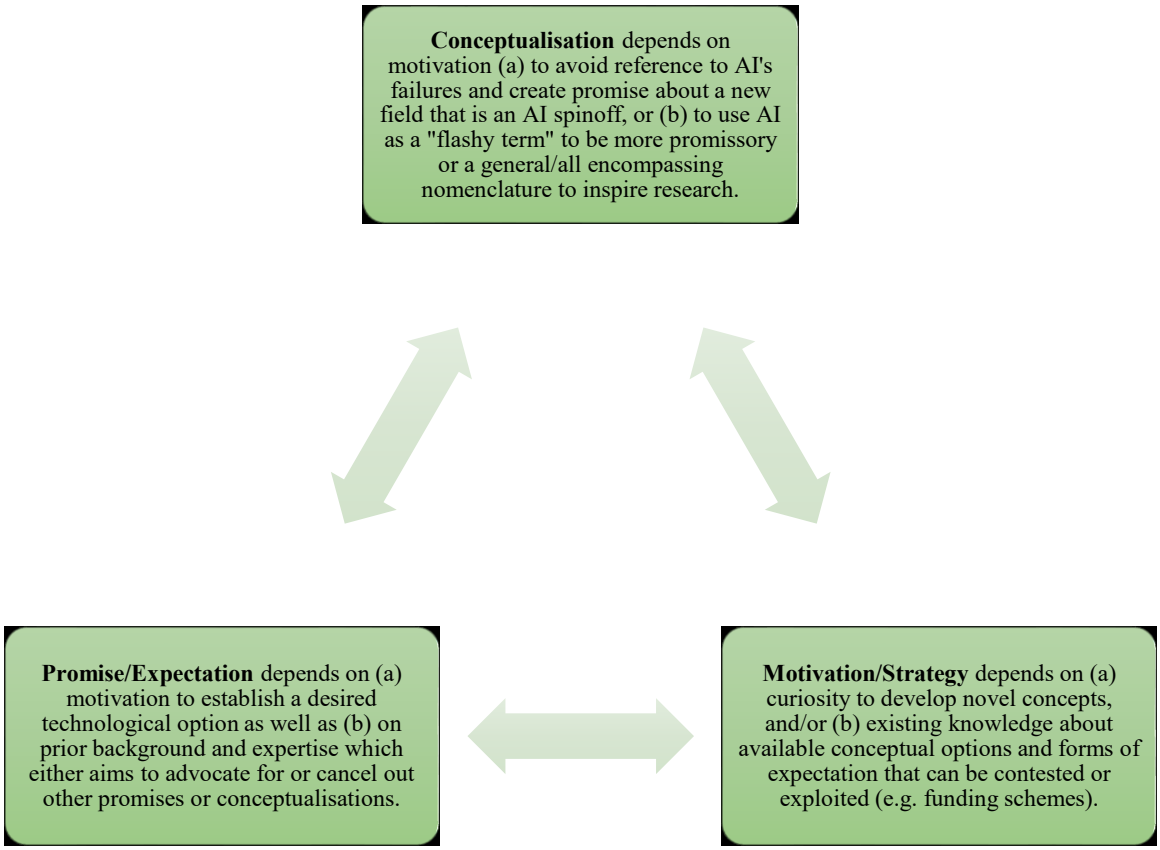
Phase 5: 2009-2021

Intensification of competing types of expertise and expectations, neural networks prevail but expectations remain mixed

- Arena 1: Connectionist neural network researchers rebrand their approach as AI (e.g. Hinton, Sutskever)
- Arena 2: Market analysts and investors project a very promissory future for AI (e.g. McKinsey)
- Arena 3: AI researchers between academia and industry highlighting limitations with intention to improve systems and against military uses of AI; join forces with the humanities (e.g. Gebru, Suchman, Crawford, Whittaker, AI Now Institute, Black in AI)
- Arena 4: AI researchers between academia and industry in support of military-industry collaborations (e.g. Li, Clarifai)
- Arena 5: Military sponsors AI aiming to collaborate with industry (e.g. Project Maven)
- Arena 6: Singularity narrative fuelled by non-specialists, however, attracting specialists who use this as veiling of other challenges (e.g. Bostrom, Hawking, Kissinger, Schmidt, Russell, Levandowski, various future-oriented institutes)
- Arena 7: Policymaking and regulatory institutions adapt to various waves of hype and critical AI voices (e.g. 2016 policy documents compared to 2021 AI Act)

To synopsise the present reading of AI's historical sociology of expectations and expertise with emphasis on the AI technical community: AI winters happen only in nomenclature's official history mode, not in research history mode – a waiting game is a game of rebranding with occasional tensions between the ability to maintain and defend the field from critics, and at the same time remain humble without overpromising. At the same time, hype-producing nomenclature is transmitted to and preserved within non-technical fields, waiting to reassign itself within AI communities, once the technical infrastructures seem to offer fertile promissory ground.

Lastly, I want to revisit the tripartite scheme proposed in section 2.1, based on the discussion advanced in chapter 3. The nexus concerned with the definition or understanding of a technology, the ways in which actors decide to promise about it, and the motivations and strategies built to perform the expectational game and suggest conceptualisations, can be viewed as mutual. The next chapters, following the historical construction of conceptualisation and promise (which highlighted a variety of motivations) will focus exactly on these three domains.



A hype-historical afterward: Hooton's report on *The Independent* (2015) is an interesting case which received plenty of media attention at the time, when roboticists from the Rensselaer Polytechnic Institute claimed their robot "passed the classic King's Wise Men puzzle which serves as a test of the awareness of the self." This news piece was released only a short period after Hawking and Musk's statements. As I was conducting research, it seemed that media attention slowly stepped away from self-aware AI robots as to express different concerns due to alternative reasons to worry about AI and its regulations. Only days prior to submitting this thesis, however, "Ilya Sutskever, head scientist at the Elon Musk cofounded research group OpenAI," tweeted on February 09, 2022 that "it may be that today's large neural networks are

slightly conscious" (Al-Sibai 2022). Keeping the above historical examinations in mind, one can highlight three things: (a) the return of AI hype, despite a seeming disillusionment, (b) that, given the short span of seven years difference, 2022 witnesses instant reactions from the AI community suggesting that such statements damage AI research, serving marketing practices (Al-Sibai 2022), and (c), the following body of empirical work, although speaking to the 2018-2020 period of uncertainty is evidencing its relevance anew.

CHAPTER 4: HOW DO AI PRACTITIONERS CONCEPTUALISE AI?

4.1 Introduction, Relevance, Research Questions

It is customary to begin various types of scholarly work by defining one's terms. However, as thoroughly discussed in section 3.1, it has become somewhat commonplace throughout the decades that there is no agreed upon definition of AI (for example, Murphy, 1985: 33; Nilsson, 2010: 13; Truby et al 2020: 111), chiefly because there is no agreement upon what constitutes intelligence, but also, as argued throughout section 3.2, because of the multitude of competing arenas of expectations and expertise, across academic disciplines, corporations, and governmental or military institutions. STS has been particularly attentive to the profoundly social background of scientific and technical definitions; although definitions appear to be fixed descriptions of apparently static cognitive objects, their social uses are shaping and being shaped by particular individuals, groups, institutions, and customs (an example is the treatment of words such as "synthetic" or "natural" in Calvert 2010). This became known in the STS approach of Social Construction of Technology (SCOT) as "interpretative flexibility" (Pinch and Bijker 1984) – the relative liberty of various ensembles of actors (mostly human users of human language) to interpret meanings and uses of various technologies and terminologies, in a way which shapes the construction of technoscience towards its closure. "Closure" has been contested as a concept in that even technologies and sciences which appear to be stabilised in their form and function with great certainty, continue to be interpreted and new ways of using them or mixing them with other technologies are invented (Rosen 1993). However, AI, either as scientific or as technological trajectory, seems to have halted at the level of interpretative flexibility without any sign of reaching closure (albeit, as I have already shown in 1.2.2, this can be contested given the influence of policy pressure). This brings in opportunities of novel approaches attached to AI, but also challenges of unmanageable lack of terminological consensus with implications in decision-making which requires tentative clarity of terminologies: from funding applications and interdepartmental collaborations to policy documents for AI governance and national strategies.

In this chapter, I contrast the descriptions of explicated in chapter 3 (section 3.1) with responses by interviewed AI specialists, showcasing the long-term effects of AI's history of terminological guises. How much of this history is present in their sayings? What is the impact of this change on their own understanding of their work? Such views will be presented in the form of quotes categorised in themes, revealing specialist attitudes towards AI, which will be further enriched by the following empirical chapters. Contemporary academic AI specialists exhibit a wide and nuanced variety of ways describing their field. This captures the "tumultuous" historical character (Crevier 1993) of AI's evolution, as a mix between struggle for identity establishment (Fleck 1982) and terminological vagueness which allows high degrees of interpretative flexibility. This chapter's discussion will synthesise the historical and interview-based conceptualisations into a critical discussion about what it means to conduct AI research today in the context of a strong interplay between expertise and expectations. I conclude by connecting the findings into an argument to be utilised in discussions about research cultures (enactors) and policy practices (selectors) in AI, which branches further into the empirical chapters below; different apprehensions of AI by researchers themselves are correlated with different approaches in conducting research, different types of promising (chapter 4), and different motives (chapter 5). From my own experience, speaking with

policymakers with interest in AI, there is a tentative belief that AI is a unified concept, and that bringing an expert for audit is indicative of the entire field. Through history and empirics, I wish to challenge that view, and problematise the situation, especially in the context of hype formation and its impacts.

This historical unfolding of AI's concept being too wide to be contained in utilisable definition enables interpretative flexibility to become an instrument of hype and expectations generator which impacts research cultures in a variety of ways; either in the case of AI being exploited by researchers in order to push forward agendas compliant with funding fashions (chapter 4), or in the case of AI being (mis)interpreted by policymakers as an item for regulation in the emerging AI policy landscape, opening up the question of expertise in AI governance. The following research questions specific to the chapter guide its development:

- How can AI's history be seen as a shaper of its current conceptualisations?
- How does this influence contemporary researchers associated with the field?

4.2 Contemporary descriptions of AI by interviewed specialists

Following the presentation of the historical malleability of AI from 1955 to 2021 (section 1.2.2), this section presents 12 (out of 25) interview vignettes by AI specialists exemplifying contemporary conceptual attitudes towards AI by practitioners. During the interview process, I have asked all participants to describe their disciplines, background, and specialisation, and how this relates to AI. This was usually followed by questions of definitions: how/do you define AI, robotics, or machine learning? I followed a "blank slate" approach, in that I did not disclose how much I already knew about AI terminologies (sometimes I have been asked how much I know already, and I explicitly claimed that "I have conducted personal research but for the purpose of this interview, I pretend to be completely ignorant"). This had multiple benefits; placing myself in this humble position, respondents felt the responsibility of offering honest views, as truthful to their knowledge; it allowed them to reflect live and admit, for example, partial ignorance or difficulty in responding; in the later context of the interviews, this question had occasional impact when discussing the role of experts, for example, in policy or in demarcating science from fiction. Some respondents (mostly younger ones) never thought of a definition before and justified that in numerous ways. Some others (seniors) smiled nostalgically, claiming they have not been asked to think of a definition for decades. I will hereby bring reminders of each participant's specialisation, showcasing the variety of domains and application orientation as well as hints about generational traits, prior to commenting on the themes in relation to the above historical findings. I would not claim that such descriptions are influenced by a single factor; it becomes probable that in most cases degrees of interdisciplinarity or specialism are interplaying with seniority and research practices, as well as personal traits such as opportunism or humour. I have divided presentation the responses into two main themes. On the one end, there are specialists who are (and want to show) that are historically conscious, aware of changes in meaning of AI. They do not necessarily dismiss contemporary AI approaches, but admit change and express occasional nostalgic remarks. This type might be the outcome either of age (people who lived through the change) or personal studying of the field's history. On the other end, there are "agnostics." They also come in different forms; sometimes, because of their knowledge, they admit definitions are difficult to formulate or impractical and constraining; in other cases, they view them as expressions of different interests. In between (although I have not dedicated a theme on those), there are cases like the

following, senior researcher with expertise in autonomous vehicles, who first attempt to define AI extrapolating from their own domain, offering broader descriptions only upon reflection:

> "Well usually when people talk about artificial intelligence I think they mean demonstrating human like capabilities. So that, yeah well that would be some artificial intel... So humans act primarily on their senses but obviously humans can't think abstractly without any senses and that would still be artificial intelligence. But a lot of human reasoning is about sensory input. Okay. So I guess the sensory input itself is not intelligence. The reasoning about the sensory input is. But then some people will say a chess playing computer is demonstrating artificial intelligence. And I guess that has no sensory input, it's entirely abstract. So artificial intelligence doesn't require sensory input but sensory input drives most of the reasoning we do." (Otto Sensious).

Implications of this phenomenon will be more relevant in chapter 4, when interrogating research practices via funding schemes. Committees advising research councils about funding plans are shaped by the terminologies used by the experts who consist them. Hence, if this researcher advised a committee about AI, it might be likely that more funding would go on sensing. But this will be explored in detail later. For now, it is pertinent to view the impact of AI's historical shaping on AI researchers' understanding of it.

a. Definable AI: Historically Conscious, Nostalgic Remarks and Contemporary Applications

Among the first people I interviewed was a pioneering AI language developer, Theodore Prover, who has worked in the early AI domains of theorem proving, and has lived through the Lighthill report era. He did not offer a definition, but let his understanding become apparent through descriptions. For him, AI's goal is about understanding intelligence and consciousness and finds it ironic that the 1980s approaches to AI (which, as said, were not presented as AI) are now presented as such.

> "I think I've also always had a fascination for artificial intelligence, I think that understanding intelligence and understanding things like consciousness are some of the major intellectual challenges that, you know, face science. And this seemed like the best approach to get a handle on that. […] now, you know, machine learning is, has become AI in many people's eyes and other more traditional areas of AI, the more symbolic approaches have somehow got side-lined, it's kind of ironic because in the 80s the machine learning people wanted to distance themselves from AI and see themselves as a different discipline and now they're kind of taking over AI. So it's a bit strange but I've been very much on the more symbolic wing of the field rather than the statistical, sometimes called subsymbolic, field, so, so yes, my approach to AI is somewhat side-lined […] So it's not trying to build symbolic models of the world, it is just confronting the machine learning system with huge numbers of examples and it's just trying to generalise from that that." (Theodore Prover).

For Fabio Informatetti, bioinformatics machine learning specialist, who, although experienced in the field since the early 2000s, has personal interest in AI history, AI hype distorts understanding of AI history and meaning. He admits then that contemporary AI means patterns extracted out existing data, that is, no reference to early AI:

> "It's very difficult to give definitions. You know, these terms are laden with some history in the case of AI and there are also the boundaries of what is statistics and machine learning, for example, have become blurred between the 90s and 2000s, and now they've become more clear again. So it's a bit of a discipline influx. The kind of generic definition of machine learning which is probably what people think of the core of AI – I mean, AI has got a very long history of logical AI and inductive logic and

> things like that that are not in my opinion really what is the AI hype at the moment – the general definition of machine learning will probably be some sort of algorithmic methods for extracting mathematical representations out of data that are predictive typically that will be the semi-technical representation or extracting patterns from data." (Fabio Informatetti).

Ravi Autonomaskar, another senior machine learning expert, specialised in autonomous vehicles, responsible for his institute's public engagement and with vast intercontinental experience (and academic roots to Minsky), suggests a return to the origins of AI as a scientific field and strongly opposes contemporary conceptualisations of AI as a (countable) technological application or artefact:

> "I can give a definition, that's not to say it would be widely accepted. So the way I understand it, artificial intelligence is to mean the study of phenomena around intelligence and ways in which you can computationally treat them. So sometimes people think about this as 'OK, there is an AI.' I've never really talked like that. To me, AI is more like biology, in that biology is the study of living things there are many different living things and many principles that…, and so on. And so like that AI is to me the study of intelligent things." (Ravi Autonomaskar).

Elsewhere, he defined his own domain, machine learning, as a subcategory of AI, based on the broader suggestion that learning is only one aspect of intelligence. Maurice Constructeur, construction engineer with expertise in machine learning applications in architecture was not an AI specialist by training, but felt the increasing demand for machine learning applications to enter his domain as a reviewer for a top construction engineering journal. The impact of AI in his work was so profound, he now has to teach about machine learning to his students and based his response on his personal understanding of the topic. His experiential conception of AI then, distinguishes between AI as a broad field which includes visionary goals, with machine learning being the actual everyday method to complement human decision with machine precision:

> "I mean the only thing I was teaching mostly in this year was my opinion but I suspect that might not actually be very well agreed, but my view is more AI is a bit of a vision to create machines that are in fact intelligent. […] AI is just an overall field with a major vision. And then within this, there has been effort in essentially using machine learning techniques first, which are, let's say, highly supervised with, you know, where we tell the machines what to look for explicitly in the data to be able to recognise things, and deep learning is been looking more and say OK, let's try to give as little clue as possible and let the machine more, learn more on its own to figure out what is actually relevant because maybe as humans we are fools and actually tell the machine to do the wrong thing, the machine might find out that there's something actually, a better way to sort this out. So, that's how I generally present it to the students." (Maurice Constructeur).

A similar account is offered by Paolo Oceanio-Marinetti, who applies machine learning in oceanography and marine robotics and has basic training in fluid mechanics (and thus recognised Lighthill's name because he studied his scientific work but had no knowledge about his impact on AI!). For him, AI is sustained because of a broader public grandiose vision, but the more one gains a specialist's view, the more the process becomes demystified and revealed as simple mathematical calculations and extrapolations:

> "So, I would approach the conversation in this way, you know, usually there's this thing that artificial intelligence is a very resounding name. It sounds so cool, right? The fact that you can have something that behaves almost like a brain, you know, if you want. But if you really break it down, it's not. I mean

> you're just having, if you really want to, say, take all the fascination for it, all the charm, what you're left with is an algorithm that tries to optimise something essentially. And if you want to even break it down even more, you have a function and you have several parameters that you want to try to get your system to approximate these function as much as you want and what the artificial intelligence does is, you know, tweak all those parameters until, you know, you basically are able to approximate this function." (Paolo Oceanio-Marinetti).

Before closing this subsection, it would be useful to refer to some empirical evidence about AI's guise as "expert systems" in the time of the Alvey programme in the UK. Aaron Auticous, with expertise in aeronautical engineering, software design, and robotic planning, who received funding from the Alvey Programme, shows in the following statement that, although for an insider a certain field counts as AI, historical circumstances such as the Lighthill report, invite for new terminologies:

> "I've never seen so much research funding in my life again. It was kind of curious. […] I moved into this project and this project was essentially about trying to use AI techniques, knowledge-based system techniques to support designers. We call this an intelligent knowledge base design support system. Because back in those days in the Alvey programme we weren't allowed to call it AI because this was still post AI winter stuff and then the, what was the report of, I've forgot the name of the guy…" (Aaron Auticous).

Interestingly, Colin Garvey notices something similar happening in the Japanese context following the failure of the FGCS programme: "Tokyo AI researcher Yutaka Matsuo has observed that from 1997 to 2002, AI was so taboo that to even use the Japanese word for it (jinkō chinō) drew condemnation" (Garvey 2019, 657). This first type of historically aware, or practice-based defenders of AI pointed towards a definition or description, that agrees with the historical timeline of AI. For them, AI development continues with or without use of the label "AI" – new guises can be invented. AI's original meaning is mostly being forgotten and replaced by the machine learning/neural network/connectionist strand. AI is a field of broad vision, and is sustained by that, but its current success breaks down to statistical calculation for the technical expert whose arduous work testifies the difference between organic intelligence and AI applications: AI is a means to assist human and machine collaboration in the workplace, not a competitor to human expertise. These fit closely with the historical exploration in chapter 3. Several researchers understand AI as machine learning, mainly because their background was not in early AI approaches and they were called to adapt to these methods because of increasing demand. Those who have lived through the change, or have studied it, acknowledge the difference.

b. Indefinable AI: Intelligence Unknowability, Definition Deniers, and Social Opportunists

The other side of interviewees have been particularly cautious in providing any type of specific definition, using a variety of – complementary, in my view, – reasons. Dalia Virtualia, a senior researcher on the intersection between virtual reality and robotics, is vehemently opposing the possibility of a good AI definition mainly because of the sociopolitical shaping of intelligence as a concept (in a sense, echoing Brooks's assertion about intelligence quoted above), and secondly because of the different distorted meanings attached to AI by public commentators:

> "Well, the definition of AI is somewhat bogged down by the fact that nobody knows what we mean by intelligence in the first place. So there is no single, I mean, many of us wouldn't accept that intelligence is a single thing anyway. […] Many of us don't believe in IQ tests, ok? What would be artificial

> intelligence in that case? OK so, the definition I've always liked best is the engineering one which is an intelligent system is a system that successfully does something which if a human did it we'd call intelligent. So it finesses the argument about... now the problem is that what we call intelligent is a moving target, so had you been around 500 years ago and able to do long division, you would have been one of the world's elites because the algorithm for doing long division was known to only very few people […] So it's no longer regarded as being really, really intelligent and a computer can do long division millions of times a second without any difficulty. Is that intelligent, yeah? So this is a social label. Intelligence is social label, not a scientific label and it means different things to different people at different times. So this of course bedevils discussion of what artificial intelligence is. Some people think they mean by artificial intelligence something which would be indistinguishable from a human and other people claim to think that artificial intelligence is something that would be superior, whatever that means, to a human, 'superintelligent,' I'm not sure they know what they mean either and I'm damn sure we're not going to get at... yeah? So that's artificial intelligence. It's a morass." (Dalia Virtualia).

Besides theoretical reasons suggesting AI's indefinability, there are practical justifications too. Wolfgang Swarmroboter, senior researcher on the intersection between computational neuroscience and swarm robotics offers a viewpoint between utility and tentative cynicism. Practical output is what counts; if one begins with a certain definition, this might constrain innovation, while further branding (in the form of a definition) may follow after: "We don't work with definitions, so, it doesn't matter if I do something that it's more towards AI or more towards robotics. Either way, it's fine, it should be interesting and it should produce some output, that's the only requirement and everything else is a matter which can be considered after the event" (Wolfgang Swarmroboter). A similar approach is held by Marta Objectividez, early career researcher in object recognition and computer vision with background experience between large companies and academia. She suggests that definitions are not important when addressing fellow members of the AI community who already know the context of a domain; when addressing broader communities, a domain expert can roughly define their domain, but the broader AI field is rather difficult to define:

> "When I write a paper or something I don't have to define these things because they are already intended for an audience who already knows what I mean. So, I choose my language according to the audience. If I had to talk to a layperson then, and I was going to describe computer vision, I would say it's the problem of extracting useful information from visual data where visual data can be images or videos. So that's how I would define computer vision. AI, honestly, I don't really know how to define it." (Marta Objectividez).

Social pragmatic approaches to definitions can also be the outcome of historical knowledge. Pioneering computer language developer with operation systems background Gerald Compulangus, with vast experience in advising the ESPRIT committees, also a student of Lighthill, when asked whether he uses definitions he responded: "No, I don't, actually." After a short pause, he spent a more than five minutes explaining how the term carries plenty of history dating way before the Dartmouth workshop, to Alan Turing and the mysticism of Golems. Therefore, to define AI according to a restricted view, is unfair to its history and fellow colleagues. Lloyd Fluidic, younger researcher in fluid robotics specialist with further expertise in public outreach and strong ties to industry and government, offers a more cynical and opportunist view. Being aware of the way different technological hypes work through branding, he sees terminologies as trends, expressing temporary "pack leaders":

> "So, all the different terms or the different disciplines, all these sort of things are really just a way of people, I mean, it's tribalism, it really is the answer. Academia is a sort of conglomerate of tribes who

> are all sort of existing in the same physical space and all have sort of chieftains which sort of are in charge of that one tribe and in order to succeed in academia, then you have to become the sort of tribal leader[90].” (Lloyd Fluidic).

Prior to closing this empirical section, I wish to quote design engineer Joseph Petrobotter, who specialises in companion/pet robot design, owning a successful company which produces such artefacts, while collaborating with academics in writing articles about his findings. His humorous comment, based on personal experiences with international meetings, exemplifies the problematic dimensions lacking a common language:

> “Well I'm all for buzzwords providing people understand what the buzzwords mean, because I was in a meeting once with forty different European roboticists and they spent half an hour discussing the meaning of the word sentience. And of course sentience is an important aspect of robotics and, you know, what does it actually mean, how do you define it. You know, words are for communication, they are a currency of ideas. And so if one person is using a word and is getting it to represent one idea and the other person is using what he [sic] sees it as representing a different idea, you get a miscommunication which is why the European Union needs a common language not a common currency. But that's a different story [laughs].” (Joseph Petrobotter).

To sumarise this subsection’s sentiments, a number of historical lessons have been confirmed: AI *cannot* be defined because intelligence or sentience cannot be defined. At the same time, AI *should not* be defined because this could constrain research. In partial confirmation of more contemporary (policy in particular) understandings, AI is its applications. But in contrast to singular policy understandings of AI, such applications are contextually defined, to the extent that AI specialists will not need to use the term AI when they talk to peers – they will use their niche specialisations’ terminology. AI, then, cannot be defined because it is too broad, its conception is shaped by non-specialists, and acts as a label, mostly empty of content, used to satisfy certain vested interests. This latter finding justifies the further focus on promises and motivations. AI definitions are of little value without examining the motivations behind constructing them. This combination of agnostic, opportunist, or cynical views on AI’s meaning can be further tied to an increasingly political shaping of AI as a product of its industrial success and the meanings attached to it by its policy selectors. Researcher enactors show little or no interest in establishing or shaping AI; instead, AI becomes a currency they choose to enact *upon* in order to proceed with their own agendas.

4.3 Discussion: *Quo Vadis*, AI?

> *“If you’re a skeptic I want to make you a believer – and if you’re a believer, I want to make you a skeptic.” (Patrick Winston, 1984 head of the AI laboratory at the Massachusetts Institute of Technology, about AI’s potential, cited in Waldrop 1984: 804)*

The present chapter reviewed practitioners’ conflicting understandings of AI. Considering these after journeying in chapter 3 from AI’s early visions and until the current “new AI” which saw the transformation of brain modelling and the imitation of humans by machines[91] into AI as an agent of change

[90] This quote will be revisited later on, as the same interviewee expands on this statement as to address the impact of such tribal formulations of terminologies in developing funding strategies and convince resource allocators.
[91] This will be termed, in chapter 5, “high epistemology” AI.

and a means to achieve statistical correlations in practical applications[92], one is greatly surprised to see that nearly nothing from the early fascination about AI's prospects has remained in everyday practitioners' descriptions of their craft. Exempt are the senior scholars who preserve reminiscences of AI's debates at least as these have been outlined in the last three decades of the 20th century. It has been shown that, for example in the early 1990s, AI practitioners viewed AI as an all-encompassing field. Although this was visionary on their behalf, and an attempt to sustain AI as science proper, this allowed experts from different fields to be attached to, and shape it, bringing in their own traditions; from psychologists to philosophers. Going back to Olazaran's (1996) distinction between official history mode and research history mode allowed chapter 3 to think of periods of so-called disillusionment as periods of "waiting games" (Bakker and Budde 2012) in which specialists and non-specialists research and write about the field, waiting for a hype trigger to turn them into enactors. The same happens within today's hype construct: different AI specialists within a spectrum of awareness about, and emergence from, AI's history, become enactors of different technical options, not within the limited scope of what their technology can or cannot do, but on the broader field of what their technology (or science) *is*.

During the historical outline, we saw the transformation of AI as a field which struggled to be established through its practitioners' exaggeration in the 1950s-1970s, into a field expected to produce practical applications to satisfy military and international political races, shaped by its funders and science fiction in the 1980s-1990s, into a purely practical domain, shaped by policymakers and sustained by broader narratives. AI's expectation for practical outcomes are side-effects of its interpretative flexibility and terminological vagueness; if anyone can be an AI expert, it is easy for funders to hire experts who transform AI into a narrow technical application instead of a science. Lighthill and DARPA won when it comes to AI adoption and contemporary pressing demands for "impact," while imaginaries and mythology prevail when it comes to fuelling of expectations.

Is this necessarily problematic? It depends on the interests of a given arena – or individual. One interpretation is that AI science is lost to AI technology's manipulation by big companies. "The vagueness of AI helps refuel existing patent arms races, and some investment sites have been keeping score. Microsoft is leading the current AI patent spree, according to one investment site, having filed more than 200 'AI-related' patents since 2009" (Katz 2017: 14). This is a very important observation, highlighting the problematic aspect of interpretative flexibility. The post-2010 AI-related patent could be an ICT-related or multimedia-related patent if treated with hyped terminologies of previous decades. Generation of AI-related numbers by those who benefit by the numbers and set the trends for what counts as AI-relatedness, paired with a loss of the thinking subject in a broader numerical mechanism of data harnessing, leads Katz to suggest that "AI is a vehicle for promoting this thinking and imposing governance by the numbers" (Katz 2017: 15). One empirical question stemming from this chapter for future work is: how much do AI practitioners care about their field's appropriation by the big companies?

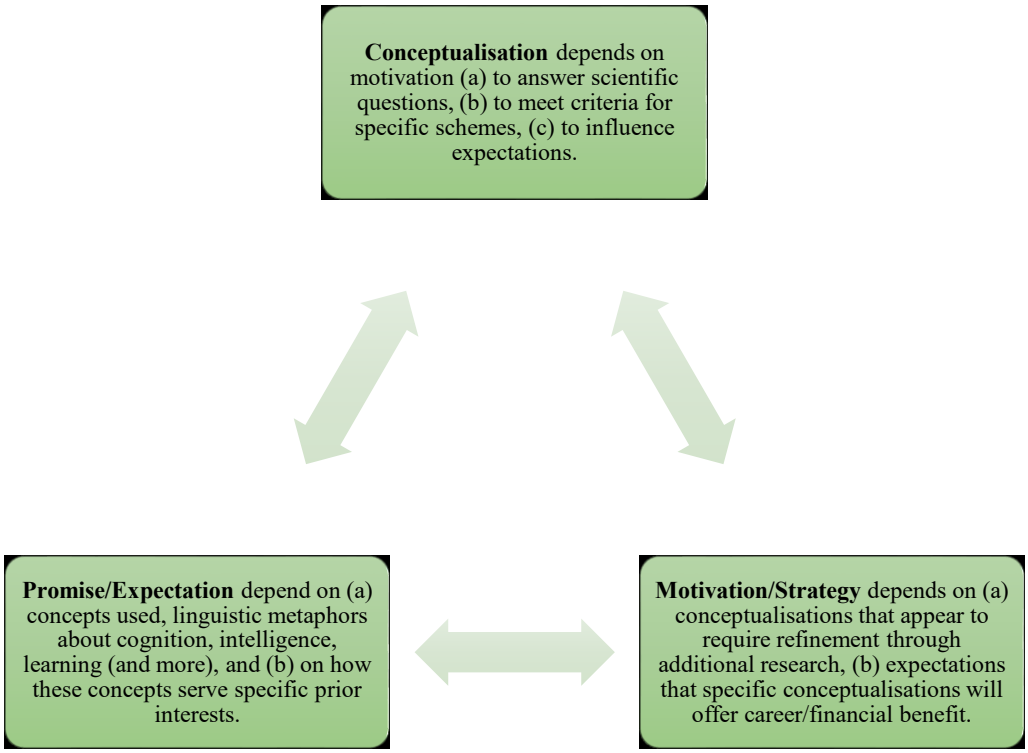
A synthesis of the two types of responses examined leads to the following reflections. AI can be seen as a field of changing nature and occasional guising and reappearance, on the one hand, and on the other, as something indefinable used as a label serving certain purposes. Researchers, being aware of the difficulties in defining AI, show less interest in actively shaping the concept; even refuse to attempt shape or interrogate it, accepting the contemporary machine learning and applications-oriented policy and industry landscape as it is; in other words, they willingly remain enactors at best, but never aim at

[92] This will be termed, in chapter 6, "low epistemology" AI.

becoming selectors. Edmond Intelligendson, Professor with over 40 years of experience in the fields of intelligent robotics and AI game-playing programmes, while reflecting on the "big questions" about AI, philosophical and conceptual, expressed his dismay over contemporary researchers' lack of interest in those. "We had those discussions back then. Now we don't. What we have is a bunch of people using their own tools and adopting almost an engineering approach to the development of the subject, you know, 'we built it and it goes forward'" While talking, he started bringing down books about AI from his office's bookcase, including some of the ones cited in the introduction (for example, Sloman 1979, Boden 1987). He continued:

> "I mean, what were the dates? I used to write on them. […] These books were written in the early 80s, late 70s, and they were asking some pretty big questions about where we were going with AI. Those questions don't seem to be asked nowadays. I think we do need to keep asking. And, as I say, my main concern is, I think, a lot of the current generation of people working in AI, they don't even know those questions exist." (Edmond Intelligendson).

What are the implications of such great interpretative flexibility (or vagueness) and lack of AI research cultures' interest to play active roles for AI's research practice? In the following scheme, I revisit the model presented in 2.1 in its third iteration, with more weighting on conceptualisation and its influence in promising and motivation. Based on the present historical and empirical account for AI being specific enough to be considered a field open to exploitation and regulation, yet broad enough to be shaped by a multitude of actors, the next three chapters will explore how this conceptualisation is correlated with promising strategies, motives for research behind AI practitioners, and their relationship to regulation of their field.



CHAPTER 5: PRACTITIONERS' VIEWS ON AI'S PROMISSORY ENVIRONMENT

5.1 Introduction – Relevance, Research Questions

"We think, in our company, that we can read anything that is printed, and we can even read some things that are written. The only catch is, 'how many bucks do you have to spend'" (Rabinow 1962; included in the proceedings of a 1962 conference reporting on developments of optical character recognition)

"[King Krool]: 'For twelve years now any constructor who fails to meet my demands, who promises more than he is able to deliver, indeed receives his reward, but is hurled, reward and all, into yon deep well – unless he be game enough (excuse the pun) to serve as the quarry himself' […] [A]nd the two constructors immediately began to consider the various possibilities, drawing on their knowledge of the deepest and darkest secrets of the arcane art of cybernetic generation. […] Not knowing where or how to place the controls – that is, the brain – so that they would be safe, the constructors had simply made everything brain, enabling the beast to think with its leg, or tail, or jaws (equipped with wisdom teeth only). But that was just the beginning. The real problem had two aspects, algorithmic and psychoanalytic" (from the chapter The Second Sally, or The Offer of King Krool contained in The Cyberiad by Stanislaw Lem; Lem 1975: 62, 65, 82)

The present chapter offers a deeper exploration of the contemporary promissory environment of AI, corresponding to the historical outline of promises in section 3.2, in a homologous correspondence of the previous chapter's empirical continuation of AI's history of conceptualisations. In academic discussions about AI promises, there are at least two dominating tendencies on behalf of those who are aware of AI's history, as I have observed them in various conferences and relevant events: firstly, there are optimist arenas suggesting a "this time, we can do it" sentiment; secondly, there are sceptic arenas suggesting a "history repeats" sentiment. In other words, the history of AI as a cycle of hype and disillusionment, enables attitudes constrained within the circular metaphor: AI has to either break the cycle or follow its predetermined path. However, given that the very understanding of AI changed within research and policy environments, it is worthwhile to consider more carefully – and thus, empirically – what is the evolution of the promissory landscape and to what extent can we still speak of circular paths. If AI has changed, it might be the case that circular metaphors are irrelevant to AI's analysis in a way similar to analysing chemistry based on alchemy's failures. Nonetheless, and as it is important to consider the slow historical transformation of alchemy into chemistry, AI's processual transformation into a contemporary applied field, is pivotal to understand its potential pitfalls too. The main assumption which aims at being hereby tested, based on the above analysis, is that non-specialists sustain the more romantic future view of AI, with profoundly transformative aspects scaling from transformations in everyday life to the creation of general purpose intelligent robots, whereas specialists retain less farfetched promises, however, fuelled by the media to influence decisions. In other words, I am in partial agreement with Mishra's recent suggestion that

> "The AI hype is analogous to a virus that finds a new 'host' every so often, and that has kept it from dying. Initially, it was the scientific community and the research labs that hosted this virus. Next, it was

the defense industry and Hollywood. Lately, businesses and the general public have joined in to keep the virus afloat." (Mishra 2021).

However, in this chapter, I want to show that arenas of expectations and expertise, are not as clear-cut as they might appear. We already saw in section 3.2, during the slow historical unfolding of AI, how alliances and discrepancies between the academic, technical, business, and policy sectors form and dissolve, challenging notions of expertise and perceptions of technology as fixed entities; from McCarthy influencing Kubrick and Feigenbaum influencing the US government to Lighthill "killing" AI and the AI romanticism being sustained within adjacent disciplines. As shown, all this has been done through an interplay of expectations and expertise, which qualifies the question of contemporary promising. By the end of the 2010s, grandiose promises were generated more frequently by non-AI specialists; however, it is crucial to monitor specialists' responses to the new hype as well. While there have been occasional specialists responding to the hype publicly (and thus cited in 3.2), section 5.2 takes up the task of reporting on the empirical findings about researchers' comments on the expectational environment, which leads into the chapter's conclusions. By enriching empirically the AI promissory environment's historical evolution a double question about AI's uniqueness is being asked:

- What is unique about AI generating an eternal occurrence of its promissory settings? What can we understand about AI through its expectations and what can we understand about expectations through AI?
- How do large-scale expectational environments impact AI communities? Do AI communities contribute to the promissory game?
- What is unique about this round of promissory environment, suggesting that it should be studied in detail? What can we learn after nearly a decade of new AI hype?

The first and second questions will be answered, to a large extent, throughout the three following sections. The third question will be enriched by additional clues, and the purpose of this chapter is to build the ground for responding to this question in the course of following chapters involving AI research funding practices and policy. Nevertheless, the significance of these findings will be further highlighted in the concluding chapter 7.

5.2 Practitioners' Views of Promissory Environments

In the present section, I offer material from 14 interviews, concerning the role of promises in research, the impact of AI winters, and what enabled contemporary excitement with AI. As mentioned in the methods section, when asking questions, I have initially assumed a "blank slate" role, pretending to know as little as possible about all of the above information. This often triggered interviewees to ask me whether I know, or not, about previous rounds of hype, the Lighthill report, or fascination with certain strands of AI, such as expert systems. Once I nodded positively, they would elaborate in detail as to the impact of such rounds of hype. It should be noted that most of the interviewees that have not been selected to contribute to the present section, belong to the younger generation of interviewed researchers. Their views were, nonetheless, valuable in that they all admitted the existence of hype, the occasional media distortion of technical capabilities, and that certain promises are necessary to be made in order to succeed in grants applications. However, the responses by those who had at least 20 years of experience in the field were the ones I considered mostly valuable for the development of the present section, due to their ability to

historicise the evolution of hype, to critique or justify variations of hype through experience, and make recommendations or forecasts about potential future disillusionments.

Given that all interviewees have experience with the UK context, it is therefore subsequent that there are more references to Lighthill as associated with the notion of the AI winter; most interviewees abstained from commenting on DARPA or other regional contexts of hype. The arguments presented here take the form of a three-step process, based on the challenges outlined above as to the various perceptions surrounding AI's historical trajectories and its promissory environments:

- Firstly, I reflect on the fixity of the very term "AI winter": what can we learn from the winter fable about contemporary promissory landscapes?
- Secondly, what do practitioners think about promising? Do they adhere to promising/overpromising? Do they see any dangers in it?
- Thirdly, are there alternative views to viewing AI's history as a series of rise-and-fall stories?

Insights based on these questions will allow situating present research communities within the historical context of AI's development, leading into remarks about the uniqueness of this round of AI hype and broader remarks about hype in technology.

5.2.1 Do AI Winters Exist?

I had the chance to speak to a well-known researcher, now retired, who was appointed a visiting fellow at the Edinburgh University on the year Lighthill's report was prepared (1972) and later member of the audience at the BBC Controversies show featuring McCarthy, Michie, and Gregory's debate against Lighthill (BBC TV 1973). In his response about the effect of Lighthill's report, Isaac Cognispatious, who specialises in the development of cognitive abilities within given spatial contexts, refers to the broader academic environment and the inter-departmental jealousy paired to Lighthill's engineering mindset (which did not allow him to clearly understand the complexity of AI's project) that were responsible for the effects of the report, thus less than the criticisms themselves. Due to the significance of the interviewee, I quote in length:

> "Well, the effect of Lighthill is somewhat exaggerated I think. I think the problem there was that Lighthill was an applied mathematician and he thought that anything good that was already coming out of AI, that was what he called advanced automation, was already happening completely independently of AI in mechanical engineering and computer-assisted mechanical engineering and control development of control systems, and so on; […] in the Lighthill framework the researcher would think about how to think about the problem and then would programme the computer to do that, and that it would control the landing of an airplane or something else, all the power switching in the power station, and so on; and that would overlap with AI but it would be more engineering and less trying to understand in some general scientific way how intelligence works, how vision works, how learning works, how memory works, and then some of us would learn how emotions and motivations and preferences and goals and things of that sort of work. So he, I think, hadn't really understood enough about what was going on in AI, Lighthill. And of course, there were some people who were jealous of the amount of money that was getting into AI, because a lot had gone into Edinburgh's, partly through Donald Michie being a very good empire builder, and so on. But also, and this didn't help, the AI people were divided into factions who didn't agree with each other. So often if you wanted to find something wrong with AI, you would go and find another person who didn't agree with it. And I think

> Lighthill cleverly made use of that kind of thing. There were four sort of departments in Edinburgh at that time, I got involved with all of them. […] I thought they were all interesting, […] but to some extent there were rivals and that was unhealthy. And I think Lighthill's report, some of the bad things about it may have come out of that rivalry, internal rivalry rather than external criticism. Anyway. That's all gossip, [laughs] may or may not be relevant to your objectives." (Isaac Cognispatious).

One should be reminded, that Lighthill was probably the first scholar to offer a concise definition of AI as such; indicative of his externality to the field which was still seeking its identity. This opens up room to discuss the boundaries of non-specialist assessments of science and technology. Another researcher suggests that Lighthill's assessment had practical implications on certain branches of AI research, namely the connectionist strand (later revived through different terminologies, with backpropagation). Professor Edmond Intelligendson captures all at once the AI effect of AI technologies being detached from AI once applied (in the case of expert systems) and the overall distrust towards connectionist approaches:

> "So, back in the 60s we had high hopes that that technique they were using would generalise out to cover everything; and it didn't. And then we looked in on ourselves and we said okay, doing things in the general doesn't seem to work too well, so we created expert systems which were a very, very focused way of becoming an expert in a particular area but not trying to do anything with that particular domain. And we had some great successes with those as well, expert systems are now everywhere but no one realises that anymore because we just take them for granted. And then there was a period of dormancy, I suppose, during which the connectionist ideas started to gain traction. I suppose they'd always been there, actually, but they were later killed off by [short pause] killed off by something called the Lighthill report, I don't know if anyone's told you about that." (Edmond Intelligendson).

Intelligendson's view is nonetheless very UK-centric; his own argument would be extended via Olazaran's (1996) study of the connectionist/perceptron dismissal because of what was presented officially (according to written historical sources) as a controversy, however was simply an alternative or complementary approach to other AI strands. Theodore Prover's explanation is somewhat different in that he sees expert systems as guises of previous AI terminologies because of Lighthill's assessment in light of the Japanese FGCS programme. The most interesting input from the following quote is that Prover notices the movement from large-scale integration to very large-scale integration as a terminological transition between knowledge-based systems to intelligent knowledge-based (expert) systems; in a sense, following the SCI programme.

> "So then in the 80s of course was a reversal again, so the Japanese decided that they would have this Fifth Generation Computer Programme and that was taken extremely seriously because, you know, they had already taken over the market for things like hi-fi systems and motorbikes and cars and so on, and so that would seem to happen next, they sort of suddenly coming to an area and dominated and so people like the Americans who have been dominating the computer markets suddenly thought 'oh, these Japanese are going to be making computers in the future, we're gonna be probably out of business,' so there were then investments in across all the different countries to sort of rival that, so here it was the Alvey programme, and the Alvey programme sank a lot of money into AI, it didn't actually called it AI, which was interesting because of the Lighthill report there was a feeling that a lot of AI is still in the doldrums still. So they called it intelligent knowledge-based systems which people in AI hated, but we still took the money, it was that it was a pretty, pretty bad name. But it was based on the analogy with the large-scale integration and very large-scale integration, and so on, the cheat

> names, you know, so it was 'OK, we've been working with knowledge-based systems so the next thing would be intelligent knowledge-based systems'." (Theodore Prover).

As mentioned, the Alvey programme, precisely because it abstained from employing AI terminologies was not considered to have failed in its promises as it did not attract excessive amounts of hype. Through a mutual acquaintance, I met Matt McMeister, an IT specialist who works at a public institution and obtained a Master's degree in AI from a department which was focusing on AI applications in the early 1990s. McMeister learned about the AI winters phenomenon through our serendipitous discussion and then agreed to be interviewed and offer his reminiscences about the AI field according to his experiences. As shown, he remarks that due to his department's specific focus on AI, the second winter did not seem to have been felt, despite the timeline's expectation; even despite that this was post-Alvey, which meant that, as shown above, was dissociated to AI terminologies:

> "That [the winter] wasn't something that was particularly noticeable from my brief enter into AI, when I did the master's degree and then as a research assistant. And that, as it turned out, I was right in the middle of that sort of second AI winter, I suppose. Because I think the computing science department was entirely focused on AI, it wasn't like a sideline of the department. […] there may have been a difficult year in funding but I wasn't aware of that certainly as a student, as a research assistant, that wasn't something that I was aware of. There was certainly a degree of realism and pragmatism in the way that AI was approached. It was applying AI." (Matt McMeister).

Perceptions around AI winters differ according to regional circumstances and situated contexts. Another external interviewee, Evan Replicautious, who would perceive a form of AI winter during the 2000s, prior to the contemporary hype. Replicautious, who is interested in evolutionary approaches to self-replicating algorithms, obtained his PhD in AI in the late 1990s and speaks about the failed business attempts he has made with colleagues during the following years in selling AI applications packages to finance companies:

> "Absolutely, yeah. No, actually, I mean, in 2007-8 it was still very difficult to sell the idea of an AI-powered investment fund. And we were talking to some of the richest family investment houses in Europe and some very powerful investors and fund managers in California. But even talking to these people, it was very hard to sell this idea of an automated black box investment fund where you could get some idea about why decisions were being made but only to a certain extent. And yeah, you're absolutely right. I mean now, I'm not sure *now* now, but suddenly within the last two or three years, it seems like anyone with AI in their business plan could get funding at the drop of a hat. [laughs] Overstating it somewhat, but I think, yeah, certainly, I'm sure it would have been a lot easier if we were doing that now than 10 years ago." (Evan Replicautious, original verbal emphasis).

From the above snapshots it is shown that fixity in terminologies, be that "AI" or "AI winter," alienate observers from complex, multilinear, and unexpected trajectories of AI's evolution. As much as there is hype about AI's capabilities, official histories are written which hype (or at least, overemphasise) the disillusionment, extrapolating from regional contexts. Indeed, non-specialist assessments and personal interests may impact funding of terminologies – but terminological masking ensures the evolution of technology, at the cost, however, of excessive new rounds of hype once a full generational cycle of previous employers of a term is complete. But how does promising impact this complex phenomenon?

5.2.2 Role of promises and a new winter

In this subsection I focus on responses concerning the role of promises in research. Do researchers embrace promise? Does overpromising generate harms? Can there be a strategy made out of it? Would it lead to another winter? Quotes below will assist a better understanding of the multifaceted promissory environment which hosts very ambivalent views on the topic; often quite self-sarcastic on behalf of the researchers. Prior to this, I want to share a quote by Wolfgang Swarmroboter, robotics/AI specialist, who offered his explanation about the waiting game prior to contemporary AI hype, based on a different set of expectations; not about how much AI can achieve, but generating expectations about the most accurate waiting time in order to achieve good results with less efforts. If Swarmroboter's assertion is somewhat correct, this might explain the silent development of AI during "winters" as a waiting strategy, founded on the basis of Moore's law:

> "One question was that until, I don't know, five years ago we had Moore's Law, meaning that you just wait essentially and computers become more and more powerful. So you don't have to do much about it. It's just like a natural law that we get all the computational power if we just wait, yeah? So people even had this kind of discussion, if you have to do a computation which takes maybe years, right? So if you have really something like the 42[93], yeah? [laughs] […] So if I start later I can finish earlier. That's good. [laughs] Because, I mean, just waiting is a linear process but if computation is increasing exponentially then it can gain by delaying things. […] But it was really when Moore's Law was in full power, they said if it takes longer than two or three years, you just wait for the next generation of computers, they are able to do it maybe, two or four times as fast and so you saved the time would you have waited [laughs] to get the result at least at the same time. […] And so people started to think of it as a computational paradigm and especially parallelism of processes that you have huge numbers of processes instead of faster and faster processes […] So having a million computers, this doesn't immediately make things a million times faster. It's like this, impregnating nine women to get one baby in one month, you know. [laughs] So it's a problem that you need to have some kind of parallelisation strategy." (Wolfgang Swarmroboter).

With this remark, Swarmroboter shows, in a rather humorous way, how contemporary AI hype was an instance of "compressed foresight" (Williams 2006); an expectation built upon an expectation, or an intermixed bundle of computer-related expectations: AI will be ready to flourish as a hyped product, after sufficient application/verification of Moore's law. Delaying until that moment possibly explains the silent attitude of AI communities. With effect: AI hype now replaces previous "internet" or "ICT" hypes, generating a new buzzword to replace not the development of a new exciting technological venture per se, but the precise need for hype and excitement. But what are the impacts of hype on research? Maurice Constructeur is a construction engineering specialist who applies machine learning techniques to his domain and who has viewed the increasing amount of AI-related papers submitted to a prestigious journal in his field. According to him, researchers feel the urge to remain relevant and may prioritise creating projects that speak to the hype, or tailoring their work based on hype:

> "[AI is] a buzzword at the moment and it's just sexy to say, you know, 'I'm going to do something with AI' part in your project I believe, especially in an area like the construction industry there is a lot of

[93] This is the interviewee's reference to Douglas Adams's novel *The Hitchhiker's Guide to the Galaxy* in which a supercomputer named Deep Thought needs 7.5 million years to calculate the exact "answer to the ultimate question of life, the universe, and everything" (Adams 1979). It is worth mentioning that the majority of my interviewees often made science fiction references, often making tacit or explicit associations between the latter's mutual development with science. Length limitations do not allow for a separate chapter on that interesting topic, however.

> effort to digitise the construction industry and as it's digitised, people start talking about big data and therefore they say, 'ok, it's a good area of application of AI,' 'a new area of application of AI with great promises,' etc, so... […] if you submit a proposal these days and you say you're going to do traditional machine learning people will just reject you because you say you're not going to do AI." (Maurice Constructeur).

Some researchers have been aware of the hype cycle notion and made instant associations between it and AI winters. Douglas Medicliff, using neural network techniques to extract patterns about Parkinson's disease, sees the disillusionment trough as unfair to the beneficial effects such technologies have, further suggesting that the brain-related metaphor of the "neural" component causes much distortion in the perception of actual capabilities, raising higher expectations:

> "Yeah, that's true of some areas of AI. Neural networks are a classic example of the hype cycle. So there's been at least three peaks of hype to be followed by the doldrums, and then some recovery later on. I think that's the case of older types of AI. Neural networks seem to be particularly affected, perhaps because they drive the imagination more. People see this relationship with their brains and expect them to be as intelligent as animals. [short pause] But then they turned out not to be the best solution to certain problems and people get disillusioned and lose interest. Which is probably unfair." (Douglas Medicliff).

Dalia Virtualia, working across robotics, AI, and virtual environments comments on the lack of unity between AI communities' stances towards AI promising. Reflecting on Collins' remark earlier, that AI communities should collaborate in defending and setting their field's goals, as well as on Van Lente's highlighting on the relevance of institutional/collective vision, the AI effect of the field branching into different disciplines detached from AI, strikes back to the field once hype revives AI. Parallel to this process of multiple co-evolving technical strands of AI, hype about AI takeovers further complicates the situation:

> "It's just that AI sort of splintered into different channels with different expectations and different interests. And I don't think most of us have a vision for AI. We hope to do something useful and incremental. As I said I want things to do the right thing in increasingly demanding environments. And that's quite an ambitious objective. That's quite ambitious enough, thank you very much [laughs] – without having visions of robots taking over the world and the rest of it. Yeah, so I'd like a robot which is a successful domestic helper. Yes, well that's not on the horizon, let me tell you, it's a very, very difficult task. OK? Very difficult task. I don't see that happening in the immediate future at all." (Dalia Virtualia).

Later in the interview, Virtualia shared an anecdote about the near impossibility of avoiding overpromising. Through the story she shared about a colleague of hers, it is shown that overpromising, although criticised now *within* AI communities (who have some historical awareness of its harms), it is entrenched within general funding environments who will assess proposals with little or no understanding of technical feasibility. Meeting the hype's terminology becomes a prerequisite no matter if evidence is brought forth about the unfeasibility of promises:

> "It's very difficult. I mean, researchers, as I said, have to overpromise or they don't get funded. Yeah? We know this: the funding process encourages people to assert what they should not be asserting. I can give an example of this. My colleague who is a psychologist wants to go for some actually not technology funding but social sciences funding. And he wants to do projects in schools with a desktop

> robot which doesn't move around but has a head and things and does a little bit of movement. And he has some good psychological aims in doing this with children with autism and things like this, and social interaction problems. And these days you have to go through an internal process before you are allowed to put it in a proposal at all. So you have to go through the internal process and we talked through the technology feasibility of what he was proposing to do. So we said to him, 'look don't promise you're going to do speech recognition on children because we know it doesn't work at the moment, it might work at some point but we know it doesn't work right now and this isn't a project about speech recognition.' So if he promised to do speech recognition it wouldn't work with children because their voices, they are in the wrong bandwidth for most speech rec systems. And what happened, he goes and does this competition with six other people and the reviewers want him to promise speech interaction, because if he won't promise proper speech interaction, they don't let him go through. Now, they are social scientists, they don't know what the technology feasibility is at all. But, you know, they were essentially asking him to overpromise. He wasn't ambitious enough. It's like 'yeah, you know, be ambitious. Suggest doing something which we know isn't going to work.' [laughs] Very ambitious, even with a technology project focusing on that problem which of course is a different issue. So yeah we tend to get asked to overpromise, we can't help it. Most of us try to restrain this." (Dalia Virtualia).

Isaac Cognispatious verifies the problematic effects of AI community fragmentation, and how this was partly the cause (not the effect) of the Lighthill report. He justifies, however, the promissory attitude of early AI researchers, which was based, according to him, on honesty and miscalculation of timelines for achieving results. He sees expectations as part of a learning process, while fragmentation results into smaller, incremental successes within compartmentalised branches of AI:

> "And they [Minsky and Papert] proved some limitations [of perceptrons] but they did not say that all connectionist architectures have exactly those limitations, they were talking about a particular class. And that highlighted some of the problems for the whole approach of having, collecting statistics and trying to learn from statistical correlations as opposed to, for example, adding logic and other kinds of symbolic structures, grammars and grammars plus semantics. Anyway. So that led to feuds within AI and some of that also fed into Lighthill, I think. But it also fed into other things. Actually, let me get this right. There had been very high expectations, many of the people, not out of dishonesty, but out of excitement were prophesying things that would be achieved. There's a famous claim that Marvin Minsky who I regard as very intelligent, I learned a lot from him, […] at one point thought the problems of vision might be solved in a summer and at Edinburgh they had this thing called the summer vision project and what happened, as in many AI projects, that by actually trying to model computationally this thing that they thought they understood they actually discovered that understanding was very shallow and the problems are much harder. And so what was going on was that all over the place people were discovering that the problems are harder than had been thought. And that meant that predictions that had been made in grant applications were not being met, were not being fulfilled. So it wasn't just Lighthill or any particular group of individuals batting one another; it was that the learning process happening in the community was a bit slower than the growth of expectations, and the predictions and the promises and the things that went into grant proposals. […] But that was mainly because the problems are so hard. But it didn't help that instead of all trying to work together, the people who favoured different subgroups thought of themselves as competing, and they were in a

> sense competing for funds. So we got a winter of nobody winning for a short time. But actually every now and again someone would get a grant, do something interesting[94].” (Isaac Cognispatious).

Aaron Auticous, AI researcher during the Alvey programme, further supports the tendency of AI specialists to build promises, on the basis of branding them as “motivational statements” (thus in agreement with Fleck’s view that early overpromising enabled the field’s establishment; Fleck 1982). Auticous, however, sees also a cross-generational difference between types of hype. While early overpromising served a different purpose, that of offering motivation, contemporary overpromising serves a rather financial and political one; being thus susceptible to further political and public scrutiny. Auticous further verifies Medicliff’s earlier remark about the emotional imaginaries attached to AI which allow the emergence of greater future projections:

> “What I realise is that a lot of what people say, people like Rod[ney Brooks], or Newell and Simon, and McCarthy, the original founders of AI and various others, you know, during the 80s and 90s, when they say, they make statements about what's going to happen, but these aren't predictions and it's unfair to take them as predictions. They might sound like predictions but what they are is, I think, much more motivational statements for the speakers themselves and others to hear and say, ‘oh wow, if that's going to happen I want to be part of it.’ And I like this because it's quite hard to make effective motivational statements in a way it doesn't appear to be predictions. If you kind of soften it or qualify it, it just doesn't have the impact, and therefore it doesn't serve the purpose. […] It took a while to get there but, so, those predictions have come off but again the original statements I think are more motivational than prediction statements. And I think this is going on now with a lot of this deep learning stuff. And I think it could be that some of these deep learning guys are going to get caught out again because the politicians mistaking motivational statements for predictions will start to worry, well actually you haven't done all, we're getting into trouble with these things, and you didn't tell us about this, and, you know, biases and stuff like this. I think the history of AI is that it's been slightly miscast in some ways. Now, of course it's not unique, maybe it's attracted more of this because of artificial intelligence and the somewhat emotional connotations that inevitably go along with anything like that as opposed to, say, cognitive science which has tended to keep his head down and maybe make better progress.” (Aaron Auticous).

Edmond Intelligendson commented on the possibility of a new AI winter based on contemporary overpromising. The high pervasiveness of AI expectations across different domains comes with inflated promises of solutionism; AI appears as the magic recipe which will solve problems at areas like medicine, education, psychology, construction, mobility (just to name a few of the areas interviewees specialise in), and is very likely to be followed by disillusionment. Intelligendson connects this to the increasing need for hype and gives examples of possible replacements:

> “There is a danger of that. Um. [pause] I say that because there is a certain amount of hype around at the moment about how much we can achieve with AI. And it's almost as if any topic you take, you will find within that area, that domain, that sector of industry or the economy, you will find people who are expecting AI to come along and solve their problems for them. And expecting it, not just hoping for it, actually expecting it, almost betting on it. And that kind of optimism is really very optimistic. [laughs] So when those people discover that AI has not come to the rescue that they thought and potentially that

[94] After this passage, Cognispatious returns to Minsky and Papert’s arguments to highlight the pitfalls of contemporary statistical reasoning in machine learning strands of AI and the inflated expectations surrounding contemporary environments.

> they banked on, they will be very disillusioned. And that could lead to another winter, if you like, as people say, well, actually it did some good stuff but it's, that's it, you know, we now need to move on to something else. We must invest more in quantum computing. We must invest in nuclear fusion. [laughs]" (Edmond Intelligendson).

The possibility of a novel AI winter was a topic of interest for those interviewees who were aware of previous rounds. After being interviewed, Aaron Auticous continued the discussion over email exchange and agreed on being quoted on some further remarks. Comparing his own views with one of his ex-colleagues who claimed during a talk that AI is here to stay, he aimed at tracing a middle path of technical understanding of actual applications and the limitations presented by the grandiose expectations. A new winter, if it takes place, will have to do with processual decrease in excitement (and thus funding), but will come out of practical experience, instead of individual reports, as in the case of Lighthill or ALPAC:

> "This time, I think the winter will be different. I see the beginnings of some deflation in the claims we'll soon have AI everywhere, lead [sic] by the spreading realisation that automated driving of road vehicles is not going to arrive as soon as many in the industry and press made out. This, I think, is being backed up by some better reporting, and thus wider appreciation, that many of the (so called) deep learning based systems have some important difficult to cure, and not always easy to detect, disabilities – biases and fragilities and opaqueness. These things will, I think, result more in a slowing down, but not a sharp cold winter." (Aaron Auticous – email communication).

Finally, Ravi Autonomaskar, specialising in autonomous vehicles and with great interest in public outreach, expresses his distrust towards overpromising, partly due to the arrogance of the promising researcher, which in turn feeds into alienation from the question concerning unwanted effects. For him, the question currently remains open-ended, yet needs to be handled with caution: certainly, high expectations will not be met; however, it is unknown whether incremental successes will overshadow the strong promissory foundation which enabled the flourishing of successful developments:

> "I think they're being too arrogant and there are many, many groups that are arrogant like this and they're surely not solvent because they're not asking many important parts of that question. And so the question that I don't know the answer to is what will be the consequence of it; maybe nothing. I mean, maybe the technical tools and engineering that's produced will be enough for people to ignore the scientific promises but I don't know. I mean, I, as a personality, as a scientist, I would prefer it if we didn't make wild promises when not necessary." (Ravi Autonomaskar).

So far, there is general agreement between interviewed researchers on the various types of harms caused by promissory hype, although there are variations and combinations of types of promise, which, under certain conditions, can be proved beneficial to the AI ecosystem. It has been shown that hype itself is the constant while technologies, or terminologies, are variables. While in previous rounds of AI enthusiasm, predictions, and promissory agendas were developed in the service of the field's establishment and can still act as "motivational statements," contemporary environments exist in a contradictory fashion for those who develop technologies. On the one hand, knowledge that non-specialist assessments can be harmful because of misconceptions aligned to AI's emotional connotations is a lurking fear; on the other hand, the environment which enables such processes of assessments (either as funding schemes or as posterior reports) demands overpromising. This internal process of promising intended to enact upon funding schemes will be explored in the following chapter. What is left for the present chapter's scope is a closer view at technological trajectories as unfolded behind the highly expectational façade.

5.2.3 Incremental Innovation Behind the Scenes of Rebranding

The present subsection will explore two remaining themes stemming from the empirical reflection of contemporary promissory environments: the existence of incremental innovation often obscured by the breakthrough appearance of "novelty traps" (Rayner 2004) or disillusionment troughs. This, I argue, has implications and is implicated by the fashionable terminologies of different times; two examples brought by in the end of this section will elucidate this argument. As seen in section 1.2.3, Dave Ferrucci's *Watson* was characterised as "revolutionary," preserving the age-old notion of the technoscientific breakthroughs as a "heroic rescuer" (Williams 2006: 341). Historically aware interviewees such as Theodore Prover aimed at unmasking the revolutionary heroism of hype, showcasing the slow technical development and the role of massive funding available to private corporations enabling wide collaborations, as opposed to public academic institutions:

> "But each of those techniques, you know, the technologies that it was based on, none of them is revolutionary, and it's just like collecting them all together. So there's nothing revolutionary in terms of technology. It's just that they invested a lot of money and got a lot of collaborators and built this massive system that academically we couldn't build, we didn't have the resources to do that. So I think the applications of it haven't really blown the world away, you know what I mean. You know, there were big ambitions, but, you know, I'm not, I've not read anything actually about, you know, this application of *Watson* that's been extraordinary successful and that it revolutionised insurance or health or whatever." (Theodore Prover).

Similarly, Maurice Constructeur, adds on the incrementalism of deep machine learning development that "hit-and-miss" is part of the process, including the promising surrounding it: "I think these things, these techniques, I mean they will, the scientists behind these methodologies they are continuously improving their approaches. I mean, they are doing a bit of like a hit-and-miss, really because they don't quite fully understand how these things work really well" (Maurice Constructeur). In a sense, Constructeur is verifying MacKenzie's (1998) certainty trough in that scientific understanding of what is going on inside the deep neural networks of today is opaque even to the specialists programming them. This becomes a double-edged sword in that it allows impressive results to be showcased, yet (a) creates heroic images of scientists and their creations to be generated while (b) they drive human imagination excessively due to the large unknowability of the technique's complexity. The following long quotes by Edmond Intelligendson were parts of his response to questions concerning the cause of recent hype, what changed in the last ten years, and how does promise relate to the hype. Intelligendson's partly nostalgic remarks – always with humorous undertones, observable by his laughing – begin with the vulgarisation of AI-as-science turned into AI-as-product, followed by explanations about what enabled contemporary AI adoption, soberer than the breakthrough headlines:

> "So yes, I saw that happen. I witnessed it. [laughs] I can remember ten years ago. You would say to people that you're working on artificial intelligence, they would say "oh, that sounds cool," but that would be the end of the discussion. But sometime in the intervening ten years, everyone now knows about artificial intelligence. They start talking about *an* artificial intelligence which is never a way that I would... To me artificial intelligence is more a descriptive thing rather than an object, but even I now talk about *an* artificial intelligence [laughs] so the public discourse has influenced the way I speak as

> well. But why did it happen? It happened primarily, I think, because, well, robots were moving out more and more into people's work lives, not their domestic lives or recreational lives, but in their work environments. They who are working in manufacturing industry, they would see more and more robots coming through into their working environments. They were seeing these robots become more and more capable, so they could see that there was something going on there. […] But I think possibly the biggest thing that happened was the development of the deep learning systems. The machine learning systems which are armed with a massive amount of computation power, and we didn't have that 10 years ago. […] Huge amount of computation power and huge amounts of data. […] So those systems, I think, really hit the imagination of people because they could demonstrate some really clever things. You could show people algorithms which were learning to recognise individual human beings from huge numbers of human beings by their face. You could show systems that were able to control the navigation of a car based on cameras that were mounted on the car. […] Up until then it had all been sci-fi. Now they could see it was starting to become science reality. And, of course, in sci-fi there is no limit. [laughs] All sorts of things are possible. So as soon as you start to shine a light on a bit of a sci-fi, say 'look that's now real,' then people's imagination is 'well, maybe all the rest of it can become real as well!' And that's why everyone's so keen on it, so interested in it. The possibilities are really quite unimaginable. (Edmond Intelligendson)

Intelligendson then continues with reflections on the unknowability about the inner processes of the deep learning systems and how this generates a "brute" approach to a technique which was more associated with exactitude than randomisation; possibly a parallel generation to the vulgarisation of countable AIs. This suggests challenges in terms of responsibility on behalf of the researcher who, however, is tempted by the fascinating results:

> "Up until then there were lots of us working with neural networks, making incremental progress, scrutinising our algorithms in great detail, tweaking them very, very precisely, and then deep learning came along. And what deep learning said was, well, just take a multilayer system, don't worry about counting the layers or the nodes in the layers... this used to be our bread and butter! [laughs] How many layers can I get away with, how many nodes can I get away with...? No, no, just give as many as you want, and then just throw as much data as you need in order to make it work! There's no thought probably out there; but look at the results, the results were fantastic! [laughs] So yeah. So we kind of ended up with a kind of brute force approach to machine learning. I use the word 'brute force' to describe it, because you don't have to think much about it. And that's worrying as well, of course, because we don't know what these systems are actually learning. And if you've not thought about what you're doing with the machine learning system and you don't know what it's learned, then, are you really justified in using it on any real-world problem at all?" (Edmond Intelligendson).

This last sentence concerning the real-world problem applicability ties to Prover's commentary on *Watson*'s delivery of promised results and, viewed through this scope, invites for a complex assessment of AI promise as problematic based on both responsibility towards unknowability and uncertainty regarding delivery. Paolo Oceanio-Marinetti, reflecting on the various waves of hype brought an example from Claude Shannon's 1952 demonstrations of a mechanical mouse which could solve mazes based on an early machine learning technique. The disillusionment following instant fascination and extrapolation about possibilities in mimicking intelligence would not differ much today than in Shannon's time:

> "But the reality is that, I mean, from my own experience when you start jumping on a problem, you have an initial moment where you seem like you are breaking all the big hard bits and the reality is that it's easy to reach a plateau, so you basically solve the major problems, it seems like you're doing

amazing success and then you get to a point where this process sort of stabilises and then sometimes it's hard to get away from the thing. And I think this applies to many fields really." (Paolo Oceanio-Marinetti).

Reaching such plateaus of productivity (à la Gartner hype cycle; Linden and Fenn 2003) is possible if the technology concerned is narrow enough to be monitored and specified; for example, navigation systems are sufficiently self-contained in order to be included in a relatively linear trajectory of high hope, tentative disillusionment, and productivity aftermath. With AI at large, things become more complex and I want to advance the argument that merging of disciplines and terminological rebranding is a form of renewing interest in technologies which appeared to have fallen into a cliff of disillusionment despite continuation of developments which take place incrementally. Connie Converse, with her origins in Germany, received a Master's Degree in 2000, specialising in NLP, currently developing conversational AI applications. She shares an anecdote which helps reflect on the historicising aspect of AI/informatics terminologies and how these blend with new sets of promises based on temporal excitement about specific successes. Converse acknowledges the existence of AI winters (and summers) and foresees the possibility of a new winter coming not so much because of technical undeliverability of promises, but because of obstruction to employment of data which do not interfere with user privacy. She is also in agreement with Intelligendson in the brute approach to machine learning and the uncertainty attached to its potential consequences:

"I remember when I did my Master's at [protected] we were able to choose a title in the end, you know, independent from what we actually did, so I did language engineering, but then you could become a 'Masters of something,' and they offered us Masters of Artificial Intelligence, Masters of Informatics, and I can't remember what the other one was; but in my mind I would never have chosen artificial intelligence because back then it sounded so old fashioned. [shot pause] And I chose informatics because I thought, oh you know, thinking back to Germany, and in Germany we call it Informatik. So yeah. But over the last couple of years suddenly everybody wants to do AI and not even knowing what that is. […] And I think what sort of started this whole AI trend, so you know, there are always AI winters and AI summers and all of that, so currently we've got a big, there's big public enthusiasm, actually not, maybe not public enthusiasm, but people are very excited about what you can do with AI because they discovered deep machine learning; where the promise is basically where you can learn from large non-...well, they are annotated, but datasets which basically don't have annotated features. So in the good old days, machine learning was actually quite a labour-intensive way of, you know, building a model. So you would actually annotate features, extract features semi-automatically, and then, you know, feeding it into an algorithm and then you could predict things like, for example, 'is this tweet positive or negative sentiment,' things like that. Now, basically. you don't have to extract these features anymore but you can learn from raw data and that has led to tremendous progress in fields like vision, for example, but also some fields of NLP made very astonishing progress. So I think that's why people are very, you know, at least in Academia, people seem to be more positive about where the whole field would be going, and a lot of people now suddenly come into the field and want to study AI. So, it's great for us because, you know, we get lots of students and lots of interest. […] and that's always going up and down and people have too high expectations and then worry about where that goes. So I think, you know, the next thing people will realise that there isn't such a thing as a magic AI ingredient you can put into your software and it will suddenly solve everything which you don't know how to solve. I also do think one related problem is that, you know, in order to create these machine learning algorithms, it's not only the algorithms, it's the data you feed into the algorithm and, you know, there's obviously a privacy problem with these datasets, but also there's a problem, what we call, you know, bias in the data." (Connie Converse).

Similarly, the following quote by Evan Replicautious is revealing in terms of how rebranding is not only associated with different trends in technology (such as the displacement of prior techniques in favour of novel ones) but also how this is has to do with generational shifts[95] and the intentional strategy of gathering proponents of specific approaches to form arenas of interest. From Replicautious' remarks, added to those by Converse, rebranding and re-promising is not only a sign of the times, but also a matter of individual/institutional decision-making:

> "But just at the end of the PhD in 1999, that department merged with computer science and cognitive science to become Informatics. And as far as I'm concerned the whole nature of the department changed around that time. Quite a few people who were in the AI department either retired or moved elsewhere at that time, particularly the kind of people who were most closely associated with my kind of work. […] So the whole character changed, and not just the people changing but also, it was kind of, as I see it, the takeover of the Bayesians. All these people, the Informatics started to have a very strong focus on statistical machine learning as around the turn of the millennium. And I mean, I don't begrudge that because obviously that's a central pillar of current success in AI and machine learning and there's a lot of amazing work going on there. But I think it lost some of the diversity of approaches and AI to me isn't just about statistics and mathematics and machine learning. But I guess these biological-inspired approaches, evolutionary approaches were certainly frowned upon, I think, in those days. And things like a lot of my colleagues are doing, interesting work in morphological computing and understanding how the body and physical structure of a robot or an agent helps it to achieve its behaviours and goals. […] It's not, to my mind, all about computation and Bayesian statistics, that's obviously a very interesting and important part, but there's other stuff too. And I think that diversity was kind of lost when AI got subsumed into Informatics. Also [laughs] another point, I was quite already told by [protected name] who was the head of the [protected AI research cluster] at the time, just after I finished my PhD, 'there were some lectureships coming up,' he said, 'oh, you can apply for them but you probably won't get it, because we're looking for people from outside to come in, looking for fresh blood.' So it's probably not what he was supposed to say but he kind of put me off applying." (Evan Replicautious).

This explicates the way in which academic arenas simultaneously adopt and generate promises based on occasional rebranding of various sets of technologies with impact on certain branches of AI research. The need to remain relevant in the age of AI hype causes a double harm to research community: on the one hand, those who are specialised in a branch of AI which does not receive the promissory environment's support, are likely to witness a winter in their branch; on the other hand, those who quickly embrace fashionable AI-ish techniques are able to partake in the AI feast of machine learning. One final empirical item I want to share prior to this chapter's conclusions stems from the industry sector. Joseph Petrobotter who owns a company of AI-enabled pet robots reflects on the practical, everyday role of expectations and how customer experience can help find alternative routes via design to move away from the dangers of disillusionment:

> "I think people expect robots to do a lot more than they can do. People have high expectations from autonomous machinery and, you know, they get, it looks obvious to me that designing something that

[95] Let us be reminded of Max Planck's principle on scientific progress and alternative paradigms: "A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die and a new generation grows up that is familiar with it. […] What does happen is that its opponents gradually die out, and that the growing generation is familiarized with the ideas from the beginning: another instance of the fact that the future lies with the youth" (Planck 1950: 33, 97)

> looks like a humanoid and can't behave like a humanoid is going to be a disappointing experience. So what we started off by doing is saying, oh let's design robotic tables or robotic chairs and take things which people are not used to be, that people have low expectations of. I don't expect the table to be robotic. So if I have a table that exhibits robotic behaviour, it's completely magic! Whereas if I put that same sort of behaviour in a humanoid, it's very disappointing. You maybe can make a cup of tea, oh is that all it can do? Whereas if you do it to a table, it's like, wow, god, that's amazing." (Joseph Petrobotter).

Thus, from Petrobotter's perspective, one should take human extrapolation from available successes based on emotional, anthropomorphic connotations, and thus, instead of falling in rebranding traps, repurpose the appearance of the final artefact. This, of course, is relevant mostly (if not only) to practical applications, which support the establishment of a practice-based AI, as opposed to AI's initial goals as a science.

This section made visible through empirical reflection multiple complexities which go against prevalent views of AI hype based on "breakthrough" images, showcasing the various ways in which promises form a vast environment of enaction across multiple arenas. To try and sum up:

- Promises serve different purposes at different contexts, temporal, regional, institutional, or individual. They may be instances of enthusiasm or "motivational statements," sometimes paired to miscalculations and exaggerations about deliverability of outcomes. They can also attract visibility, in moments of excessive hype.
- Hype, at least in the past three decades, appears to have formed a novel technological environment, seeking for technological terminologies and promises to feed upon (thus confirming the historical analysis of recent events). This new environment, paired to corporational interests and national strategies requires researchers to take active part in overpromising in order to secure funds.
- Incrementalism, the slow development and optimisation of techniques, as opposed to solutionism, the sudden breakthrough view of heroic technologies, becomes obscured within this environment. Several enactors of previous generations become irrelevant in the times of new paradigms selected by those whose promises have earned sufficient power.

In later chapters, the ways in which such promises gain power is explored. For now, a synthesis of the historical and empirical components of this chapter follows in its closing remarks.

5.3 Discussion

"Treason doth neuer prosper, what's the reason?

For if it prosper, none dare call it Treason."

(Harington 1618: n.p)

"It appears to me that this mystery is considered insoluble, for the very reason which should cause it to be regarded as easy of solution - I mean for the outré character of its features. […] In investigations such as we are now pursuing, it should not be so much asked 'what has occurred,' as 'what has occurred that has never occurred before.'" (Edgar Allan Poe, The Murders in the Rue Morgue, cited in Debord 1988: 63)

By replacing "Treason" with "AI" in the above rhyming aphorism by Harington, one can obtain a clear view about the riddle of the "AI phenomenon" (Woolgar 1985) and the "AI effect" (AI Effect 2019). This chapter reported on contemporary AI practitioners' views on promising, showcasing how the promissory and large expectational environment of AI enables grandiose hype about something now called "AI," yet

not being exactly "AI," according to several technical specialists, as shown in the previous chapter; whereas several applications of traditional AI are not considered to be AI anymore. The present empirical findings helped verifying a number of historical challenges in promising and assist in problematizing them further. The following table summarises a typology of expertise and expectations based on the predominant polarised utopian/dystopian expectational environment of AI, followed by explanations and further discussion on the implications of such a quadripartite polarisation of promises based on practice, however, historically situated within the evolution of AI's promising, building further on the table presented in section 3.3.

A typology of AI expectations arenas	**Techno-optimism**	**Techno-pessimism**
(a) Non-specialist assessment – arenas of broader discourse	Large funding, demand for overpromise, national AI strategies, occasional support by cultural mediation	Negative reporting (à la Lighthill), dystopian visions (*à la* Hawking or Kissinger) and media representations
(b) Specialist projection – arenas of developers' expectations	Early enthusiasm and first-step fallacies, opportunist advantage from the hype, alliance with funders	Historical awareness of winters and will to change terminologies, admittance to potential harms and alliance with broader discourse doomsayers

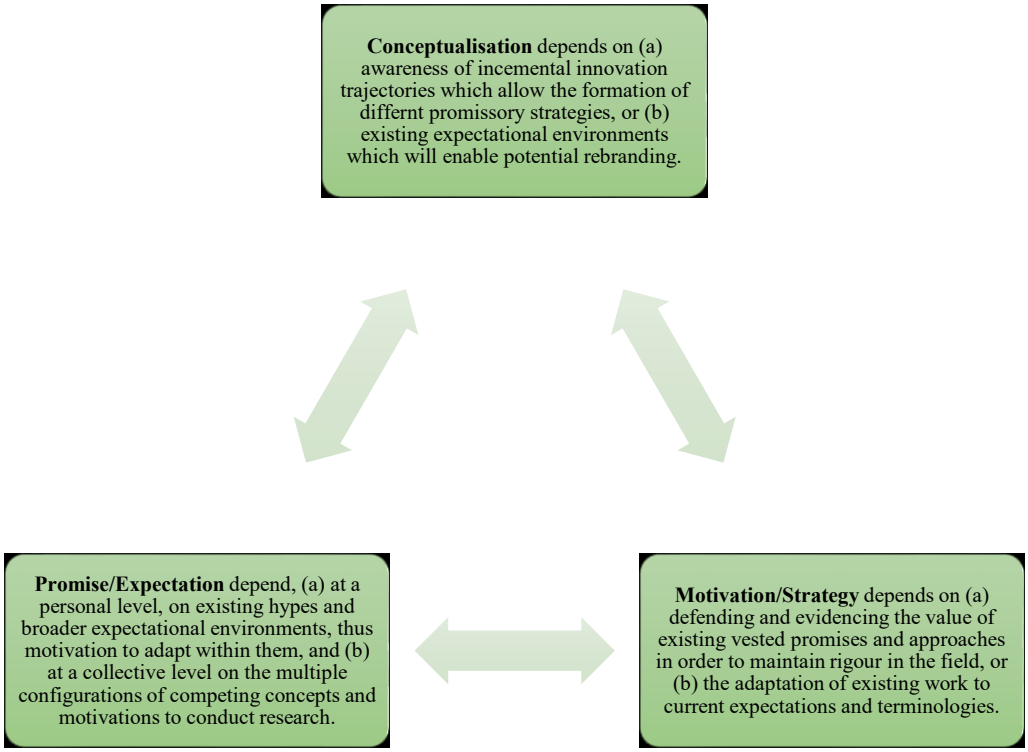
On the question concerning the uniqueness of AI's promissory environment, the role of emotional connotations and the power of AI to feed human imagination has created a very peculiar interplay between AI research communities and broader public discourse. Science and technology do not advance as singular trajectories. Various ensembles form and dissolve and in the case of AI, this has been witnessed as a mix of different dynamics which includes imaginaries of techno-optimism and techno-pessimism which can be extended: (a) outwardly towards selectors of governmental policy and large corporation interests as well as arenas of media representations and cultural products such as popular science books or science fiction references, and (b) inwardly towards enactors of AI research community arenas who become increasingly dependent on the broader environment. While early AI research was more powerful in establishing its aims, the vast diffusion of its conceptualisation through discourse, and the sustainment of interest in AI through public representations, led contemporary AI communities to abide with a large environment which supports hype. This is telling of an interesting social/historical change which would render partly unfair the comparison between "good ol' days" of AI and current debates. The closer one gets to the times of McCarthy and Minsky, even Kahn, the more one perceived hype as created out of curiosity and miscalculations. The more one processes into the contemporary AI environment, the more one experiences a growing unmanageability of overlapping arenas of conflicting or mixed interests, terminologies, and institutions: AI is too large to contain within specific borders. The generalised need for hype, replacing hype made out of motivation, is reflected in the need to write "sexy" grant proposals and the constant need for an AI-related newspaper article.

AI researchers have either to play the game of overpromising or to ally with doomsayers. In the first case, although they know overpromising is not likely to meet its goals, they take risk of either facing the consequences or delivering outcomes impressive enough to erase the memory of an initial promise. (Such funding strategies are explored in detail in the following chapter.) In the second case, they would

ally with institutions supporting "ethical AI" as experts who admit the dangerous potential of the technology and will advise those in power who promote a deterministic view of inherently dangerous technologies. Nevertheless, both sides serve the same beneficiary: an image of the State which is at the same time pioneering in terms of supporting visionary, "breakthrough" technologies but also protects its users from their hazardous outcomes (which, in turn, might be used against other States in their militarised form).

So far, this answered bits of the question about the impact of large expectational environments upon AI communities. But what does the sociology of promise and expectations learn from the case of AI? Typically, AI is used in such fields to speak about the high cost of promise and the narratives of rise-and-fall (the "AI winter"). However, the detailed examination of promises showed the important role of terminological branding: the masking of certain computer-related technologies under different guises and their attached promises (from AI and expert systems to machine learning and neural networks). This supports the view that the "winter" might be seen as a convenient fable in service of official histories, serving rhetorical purposes as much as overpromising does. Incremental development of various technologies which, under different circumstances, can be labelled as "AI," "ICTs," "computers," "smart technologies" (and so on), continues based on the efforts of individual researchers whose skill in adapting their research to funding schemes, or funding schemes to their research.

In a sense, AI continues to be caught in the same two traps of overpromising *as* AI, and masking of AI because of too much overpromising. As Robert Merton put it, "[a] small truth has a way of inducing a large exaggeration" (Merton 1965: 160) and AI's metaphorical association with animal intelligence and science fiction (sometimes fuelled by AI researchers, as shown in the case of Minsky) has an ever recurring way in evoking what Hubert Dreyfus called the "first-step fallacy" – one easily extrapolates by climbing the tree very quickly, that the moon can be reached as well (Dreyfus 2012). Then, once a winter becomes recognised by those in power who suggest its existence within certain regions and institutions, AI shows its transformative power; not by transforming societies, but by transforming itself. AI winters is form of negative hype which obscures incremental innovation happening behind the perceived notion of disillusionment. What was considered to be "AI" during an earlier moment of hype is not; and what was considered not to be "AI" during an "AI winter," now is; failure cannot be perceived as such if the expectation is not set from the beginning in the form of a terminology. According to Maturana and Varela: "The success or failure of a behaviour is always defined by the expectations that the observer specifies" (Maturana and Varela 1992: 138). Thus, in the case of Alvey, or more generally, computer science, Web 2.0 technologies, and ICTs, one cannot deny with ease that practical applications are not successfully implemented; such applications could be considered to be AI's progeny, should the expectations be differently set, or perceived as "motivational statements" at earlier stages. Or, should discord among enactors not have caused promissory disruptions and large fragmentation between branches of AI. Or, that contemporary fragmentation across industrial, military, or academic AI does not influence AI governance. Thus, given the strong impact of the discourse environment on AI communities and the rebranding of AI to meet societal expectations about it, it is shown that expertise and expectations exist in a mutual shaping fashion in the case of AI. AI's terminological flexibility allows an unmanageable amount of actors to build expectations and by doing so, construct their expertise. Expertise, showcased in the form of rhetorical certainty, is then able to set up agendas that are to be followed by the research communities. This role of certainty will be particularly examined in chapter 6. Below, the third iteration of the conceptualisation-expectation-motivation model, this time with weighting on promising.



Nonetheless, the expectation and promissory games in AI do not end with a historical examination and some experiential data from AI communities on the role of promises. With current debates in AI, this extends further to the inner strategies developed by practitioners in order to secure funds, examined in chapter 6 and the outer impact of the AI policy environment which aims at regulating practitioner communities (which is to be examined in future work). One final thing to consider prior to moving to the next chapter, is the possibility of a new hype, replacing current AI enthusiasm in a way similar to AI rhetoric replacing the Internet or ICT imaginaries. As mentioned by some interviewees, quantum computing is a good candidate and indeed, specialists from within the field have already caused alarm about the high expectations attached to the growth of quantum computers. For example, quantum computing specialist Scott Aaronson blogged the following passage in 2018, connecting AI hype cycles to the possibility of a quantum winter: "In any case, the history of AI reminds us that a crash could easily follow the current boom-time, if the results of quantum computing research don't live up to people's expectations" (Aaronson 2018). This renders the findings of the present chapter useful to researchers who will be interested in the hype management of quantum computing, should that grow into a separate entity if AI falls into another "winter," that is, become rebranded.

CHAPTER 6: EXPLORATION AND EXPLOITATION PRACTICES: FROM INTELLIGENCE UNDERSTANDING TO STRATEGIES OF FUNDING

6.1 Introduction, Theoretical Preliminaries, and Interview Approaches

The present chapter will suggest and explore a practice-based typology of AI expertise, by applying the SoE and introducing the relevance of the exploration/exploitation framework in assessing motivation for research based on expectations. Occasioned by the previous chapters' findings that contemporary AI researchers appear to be less visible in official AI shaping, or, if they are, they do represent ideological, corporate/industrial, or military interests, it is pertinent to look closer that what constitutes an AI research culture of the everyday. Particularly, how such researchers' expertise is interwoven with an expectational environment that is co-shaped by a variety of other types of expertise? To repeat some key points about SoE and introduce the exploration/exploitation framework:

> (1) Expectations and promissory behaviour are performative, that is, they have an impact on everyday practices, investments, and policies (Van Lente 2012; Pollock and Williams 2010:), not only in terms of attracting funds, but also causing their stagnation through the unrealisability of promises.
>
> (2) The concept of exploration/exploitation, borrowed from biological, cognitive, and organisational sciences, speaks of and investigates the trade-off between the seek for alternative choices and scaling new grounds (exploration), as opposed to seeking reward through existing choices: "Exploitation activates regions associated with reward seeking, which track and evaluate the value of current choices, while exploration relies on regions associated with attentional control, tracking the value of alternative choices" Laureiro-Martínez et al 2015: 319).

AI, looked at from the wider discourse perspective – without knowledge of the historical outlines discussed above –, AI appears to be an area of exploration; how does the human mind work? Can it be replicated by a machine? Can machines be intelligent in ways different that humans/other animals? For the AI specialist, however, it seems that AI becomes an area of exploitation. AI specialists are aware of the hurdles revolving around precise mappings and replications of the mind as conceived in terms of brain physiology with centres responsible for reason, intention and consciousness (let us call such endeavours "high epistemology AI") and become exploiters of available knowledge and funding opportunities to develop specific utilities employing AI techniques to develop applications-oriented products (let us call those "low epistemology AI")[96]. Low epistemology AI gains in terms of hype by the visionary connotations of high epistemology. Small applications are more easily funded if presented as parts of a larger scientific field called AI, than, say, recommendation software. As it was shown in chapter 2, AI's grand vision and narrative (a "cool term") is used to sustain approaches which have diverted much from this initial vision.

[96] For this tentative distinction, I am indebted to Prof Robin Williams who coined it during a supervision meeting.

The present chapter takes a closer look at my participants' declared motivations for research in the field of AI, or their respective specialisations that may broadly be presented as closely related to AI. Inspired by Van Lente's analysis of a bidirectional relationship of impact between agendas and expectations, I have been asking my interviewees what led them to and why (or whether) they prefer the institution they have been working while I interviewed them. I should remind the reader of the relatively untypical structure the examined institution has: a robotics centre being a collaboration between two prestigious universities of the same city with different prominent research department contributing to the centre as a whole. That means, every individual I asked was a member of a certain subject area/School department, being part of a certain School of a certain University, with each of the departments contributing to the centre for the purposes of events within the University or open to the public, publications, collaborative grant proposals, and so on. Moreover, some of the interviews examined come from respondents semi-external to the sample, in that they have been associated with the specific University in the past and their decision to leave (partly affected by funding rounds between 1990-2005 or other life choices) made them interesting candidates for comparison of results. With minor differences in terms of role and labour division, I can claim that despite the heterogeneous background of every participant, there is general agreement in their belonging to a joint effort, across the findings presented.

When it comes to expectations, then, I was interested to learn about their relationship to what I perceive as different levels of their academic identity. I asked the interviewees if they have their own personal agendas, if there is an institutional (interdepartmental, University-wide, or robotics centre-wide) agenda there are expected to follow, or whether these different levels of academic identity blend. Corollary to this was whether the good reputation of the universities they belonged to played a role in their decision to work there. I also asked, depending on their responses, whether such individual or institutional agendas, are built based on long- or short-term planning. This question was shaped again by Van Lente's work which makes a distinction between formal, strategic, usually beneficial, and generally short-term foresight of possible futures and informal, imaginary, usually distorting, and generally long-term expectations. I cannot claim that within that centre, or among the external interviewees I have spoken to, there is any sort of shared agenda; hence, although the centre's existence appears as a collective effort, its participants act in an individualist manner.

The vast majority of participants are referring to funding bodies' initials or funding schemes' names that did (or did not) support their applications. These funding bodies are presented on the list of abbreviations after the Preface. To pave the way for what is about to follow, it should be noted that while conducting fieldwork an emerging body of literature (e.g. Sutton 2020 on the journal *Radiography*), investigates the weaknesses of such schemes, such as the failure to measure impact that often goes beyond existing criteria. It is important to mention then, that the context under which I have collected the data presented below, was the one of an impending Brexit, the separation of UK from the EU, hence, when asked about funding schemes, opportunities, and challenges in research, several respondents referred to the still open challenges of a "chaotic no-deal Brexit" summarised in an article on *Nature* by Gibney (2019) which would determine the future of countless international research collaborations, as AI and robotics depends heavily on various layers of international knowledge and hardware exchange. In a sense then, UK AI scientists who would count as "explorers" would aim more towards European funding, whereas scientists who would count more as "exploiters" would aim more towards UKRI funding. Scientists who would aim at a balance between exploration and exploitation would ideally attract equal amounts of funds

from both sides. But this is only a hypothesis, and it will make much more sense to follow evidence-based arguments as expressed by the interviewed researchers.

6.2 Relevance – Chapter Research Questions

The novelty of the present chapter is its attempt at shedding light in the relationship between hype and everyday research practices and culture. In this way, it can be read as separate from the rest of the thesis. Within the scope of the thesis, however, it can be viewed as the final interview-based empirical chapter that takes a closer look at AI researchers' motivations to do research, with their research culture being a product of AI's history of expectations and expertise. Have AI scientists in the era of AI hype turned on their brain regions associated with reward-seeking means of survival? Or are they the romantic curiosity-driven explorers of alternative ways to describe and replicate intelligence? Can there be a balance between the two options? The following list of research questions specific to the chapter led the presentation of data and derived from iterative process between the more general questions asked of participants, further sharpened during analysis:

- How do scientists think about the institutions and the research culture they belong to?
- What is the relationship between institutional ecology and individual opportunism?
- Do scientists care more about advancing their field (explore), using their field for personal non-scientific benefits (exploit), or is there room for in-betweenness; and does this tension affect technological development?
- Does this tension tell us anything about hype? Does hype generate the tension? Does the tension sustain the hype?
- Does the specialisation of researcher impose a difference in perception of the tension?
- Is this tension specific to AI or do the present findings reflect a reality in broader academic and technological research circles?

In terms of contributions to theoretical STS, this chapter highlights how the SoE and exploration/exploitation frameworks are in interplay. This is an important lesson for STS scholarship, since, while existing literature focused on the way visions are employed in anticipatory strategies (for the better or for the worse), this study shows how visions are employed as funding attractors, revealing divisions and tensions between curiosity-driven and grant-driven research (for the better or for the worse).

As said, one of the central goals of SoE is to distinguish between the practice of formal foresight and the circulation of informal expectations. Van Lente employs the metaphor of "sea" of expectations, an inescapable necessity out which exercises of formal foresight emerge, as a strategic decision-making in order to "benefit from explorations of the future" (Van Lente 2012: 770). The objectives of systematic foresight include priority-setting for the identification of shared agendas, the building of networks in order to "reinforce the connectivity of the innovation system," and harmonisation of strategies of different stakeholder through the construction of a consensual vision (Van Lente 2012:.770). Van Lente's framework is mostly directed to policymakers who practice such future-oriented games and whose repertoire of future visions stems from committees who inform them, their personal research, and broader societal visions. When Van Lente speaks of exploration, he mostly refers to policy exploration of

technoscientific futures. What happens, however, when expectations are interplaying with scientific exploration for the advancement of technology? In other parts of Van Lente's work (e.g. Bakker et al 2011), the distinction between enactors and selectors (already encountered in chapter 3) is useful here. AI specialists in the present chapter are examined as enactors, i.e. promoters of certain technological options, candidates for selection by funders (and policymakers, although this will be the content of the next chapter) who are most sensitive to potential risk. Therefore, AI specialists as enactors would seek to convince funding bodies that their technological options are the less risky, in order to be selected. Ideally, this exploitation of the funding landscape will involve explorative innovation in the field. To refer again to the original exporation/exploitation framework as described in cognitive science: "The trade-off between the need to obtain new knowledge and the need to use that knowledge to improve performance is one of the most basic trade-offs in nature, and optimal performance usually requires some balance between exploratory and exploitative behaviors" (Berger-Tal et al 2014: n.p.). Is this true of AI specialists?

6.3 Findings from Interviews

This section is divided into two main parts. I will first offer an illustrative example, inspired by one particular quote that deserved investigation of background. When asked about the reason of current AI and robotics funding hype, a highly esteemed researcher, explicitly asking for confidence, shared with me an anecdote, revealing what appears to be a chanceful start of a round of hype, imbued with elements of politics (since the story involves a minister), history (since the story comes from 2013 and shaped all following decisions), and SoE (since a particular informal promissory game of convincing took place). However, the specific example is not enough. It is only one to generate extractable meaning, and one may argue that such rounds of funding would happen anyway, one way or another. Hence, the second part reflects on the everyday underpinnings of research culture. The main argument is: the anecdote referring to interactions of decision-makers that are high in their political and research positions analysed in 4.3.1 would have no significance if it were not for the underlying research culture that sustains such possibilities and is, in turn, affected by them[97]. To paraphrase William Shakespeare, "[t]here are more things in [empirical interviews], Horatio, than are dreamt of in your philosophy [, media, and science fiction]."

6.3.1 How the Seven Great Technologies Became Eight During a Train Ride, or, How Random Meetings Shape Scientific Funding

"The limitations accepted by the Royal Society drew increasing criticism from as early as the 1890s. Some critics feared the Grant would herald the end of the independent scholar." (MacLeod 1971: 357)

"What made them leave home and set sail upon dangerous seas […] was not […] the Heavenly Event by itself, but rather that unshining Assembly of Human Needs […] including certainly the Royal Society's need for ~~the~~

[97] Most of this chapter's empirical components and arguments are currently in press in a forthcoming book chapter (Galanos 2022), although employing a different theoretical framework to analyse the observed behaviour of AI specialists as "nomad science" being a reaction to "State science" (as in similar remarks made about early cyberneticians; Pickering 2009).

~~Solar Parallax~~ [winning the AI Race],- but what of the ~~Astronomers'~~ [AI Researchers'] own Desires, which may have been less than philosophical?" (Pynchon 1997: 102; my paraphrasing[98])

David Willetts, then a Member of Parliament, Minister for Universities and Science in the UK, published in 2013 a policy report entitled *Eight Great Technologies* through the UK's largest think tank, Policy Exchange, in which he recommended that the "[g]overnment should be promoting with further capital investment and technology support" (Willetts 2013: 9) with the aim to "make Britain the best place in the world to do science" (Willetts 2013: 7). The use of such wording ("best place in the world") to promote a country's technological development appears to be a rather political statement, masquerading as healthy international competition based on technological advancement. With emphasis on historicity (see the above epigram), it is important to notice that this "report was published alongside the Chancellor's important speech to the Royal Society" (Willetts 2013: 9). The Royal Society of England is in the UK context informally associated with allocation of governmental funds for research, as an authority over matters of science, and was eventually convinced to support Willetts' eight great technologies (among which were robotics and autonomous systems) by a "£1.5bn of extra science capital investment" (ibid) as part of the Tory government's attempt to programmatically invest in scientific and technological impact. Willetts asserts that the technologies were selected according to three main criteria: their importance in terms of scientific advance, Britain's distinction in each respective area, and the identifiability of commercial advantages for each of these areas' technological output (Willetts 2013: 7).

After comparing robotics[99] advancements in Japan and Germany, Willetts set a clear national agenda of competition between the UK and the US: "In the US as well as Government [sic] setting a regulatory environment DARPA has been promoting these technologies through sponsoring grand [grant?] challenges and funding them[100]. We are not leaving the development of these technologies to others." (Willetts 2013: 25-26). In the same style of national strategy rhetoric, the technological concepts of "algorithms," "autonomous systems," "data flows," or "clinical medicine" are used as assets of UK's "comparative advantage" to other countries; and to those "world-class" assets, the national superiority of the British Humanities are added: "Through the strength of the Humanities in our universities we also have a strong position in the ethical issues that arise – programming a scavenging robot and defining how it acts and in what circumstances should not be done in an ethical vacuum" (Willetts 2013: 27). A final point needs to be made about Willetts' 2013 report through the following quote, to situate the entire text more clearly within the question of exploration/exploitation framework of scientific advancement:

> "Research Councils provide grants for future research and training. Some of this funding pursues specific goals, much of it is driven entirely by the curiosity of researchers. Quite rightly, this adds up to

[98] Thomas Pynchon, in his novel *Mason & Dixon*, touches upon the relation between the Royal Society's funding power in the 18th century, and the inability of astronomers Mason and Dixon to secure fellowships with the Society (and thus funds), because of their will to satisfy their personal scientific curiosities but also bodily desires.

[99] From the phrasing used in Willetts' quotes, it becomes adequately clear that his conception of "robot" is one that encompasses the typical exo-scientific contemporary distinctions between robotics and AI. Hence, it is safe to say that his report, although not referring explicitly to "artificial intelligence," refers to this field as related to algorithm science and robotic autonomy. With effect: a series of AI funding initiatives were sparked after that report.

[100] The document's text has an abundance of misspellings and typos; possibly an indication of the rush under which it was produced and that it has not been thoroughly proofread or reviewed.

> a substantial level of freedom for academics to pursue curiosity driven research, while also making space for the pursuit of specific challenges." (Willetts 2013: 8).

To what extent this statement about exploratory, curiosity-driven research had actual application is to be discussed in the following section. For now, it is important to unveil the significance of this report in the development of AI research in the UK no matter if grant-oriented/exploitative or not. My interest in Willetts' report was sparked by one particular interview statement. In every interview I conducted, I included a question, with few changes depending on conversation context, across the lines of: "It seems that in the last years there's a resurgence in the field of robotics and AI with regard to research grants as well. Do you think that something has changed these years in your experience? Why is that?" Most respondents referred to technological advancements, such as computational speed and power and data availability allowing the wide-scale application of (deep) machine learning algorithms. It is interesting that one particular respondent who, specialises in microelectronics (presumably the "oil" fuelling the Moore's law engine), did not justify the hype based on technological achievements at all. His exact response to my question was:

> "So. A few years ago. So this is, you know, [pointing at the recorder] for the anonymising part on here right? A few years ago, one of my colleagues who is a very senior colleague in the field was on a train to London and he sat next to who was then the Science Minister guy called David Willetts, right? So, a man affectionately known as 'two brains' because he's a clever chap and he was at the time working on a document called the *Seven Great Technologies* which was about the sort of underpinning technologies in the UK, which sort of underpinned the UK's position in the world. And that, it was a white paper to go to Government to advise them to release money down through the research councils to fund investment into these technologies. So, it sort of became a key strategic driver. But between leaving Edinburgh and arriving in London my colleague most persuaded to turn into *Eight Great Technologies* where he included robotics and artificial intelligence on those *Seven Great Technologies*. And that sort of marked a key turning point in the Research Council delivering money for robotics and AI, right? Because from those, that sort of really, really, really high level strategic priorities document came out things like you know the Industrial Strategy Challenge Fund, but before they freed up a whole bunch of other moneys in robotics and AI and led to lots of these investments. In other places, you know, in other sectors, in manufacturing and this sort of stuff then it led to those similar sort of investments from the Research Councils and directly from government as well. But for robotics that was a huge place where it came in. So, I guess then the real question is, why did, you know, my colleague feel the need to inform Dave Willetts that robotics was such, it was such a big deal. And I guess that's to do with, you know, as you have here, you know, no it's not on here, you had it sort of as imagined possibilities or something..." (Lloyd Fluidic).

The interviewee, whose further specialisation I will only reveal below as part of less sensitive opinion-sharing, referred to my email invitation which included my initial proposal for a comparison between actual and imagined capabilities of AI. It is interesting that he tries to explain his colleague's persuasive attempt at convincing the Science Minister based on overpromising and imaginaries. However, I want to stay at this "why" question that he poses and explain it in a more mundane way, through the empirical interaction with AI practitioners. My sentiment is that the highly esteemed person who convinced Willetts to add a number in his report on technologies was mainly well aware of his AI/robotics "pack's" research culture which would benefit from the development of such schemes. By convincing the Science Minister about the high epistemology expectations, which led Willetts to speak about exploratory, high epistemology, curiosity-driven research, the anonymous prestigious researcher protected the space of his

low epistemology, grant-dependent, exploitative research being pursued by his colleagues. As presented by my interviewee, this appeared to be a meeting purely based on serendipity, which turned out to trigger a massive investment round. To what extent this was indeed a pure chance meeting, remains unknown.

To support my proposal, I will refer to AI researchers' everyday practices, to map the broader landscape of which Willetts and the influencing researcher happened to be visible mountain peaks. This will allow me to extract further themes in the study of everyday, low epistemology, research that unmask the messiness behind the high epistemology of mainstream perceptions of AI. To paraphrase a well-known aphorism on laws, often misattributed to Otto von Bismarck, "[research cultures] are like sausages; you have much more respect for them if you haven't actually seen how they're made."

6.3.2 Cutting the Edges of "Cutting Edge" Research: Interview Investigations of Underpinning Scientific Cynicism

As canned food is to the domesticated cat, infinite annual subscriptions are to the cinemagoer, and coffee is to the postgraduate student, research grants are to the AI researcher. And it seems that the good ol' days of the high epistemology AI enthusiast/worker who aimed at replicating intelligence and create general-purpose robots (good ol' fashioned AI) have given way to the low epistemology AI exploiters, in different degrees of funding priority hierarchy. This was already shown from historical views on AI, from the Lighthill report and the Mansfield amendment throughout the 1980s military funding of AI which paved both directly and indirectly an exploitative path for AI research. I want to begin this section by looking at two interview quotes of special cases of different types of researchers who, although having the same University as a starting point, their research trajectory and specialisation are quite different. They are both quite senior, with experience in previous rounds of AI hype, both of them exceptionally keen in reflecting on the history and social dimensions of AI, sometimes in written word to, escaping their technical specialisations. I use them as starting points, as they were among the first interviews I conducted for the project and therefore shaped significantly the research path followed later[101], providing me with sensitising concepts, leading to the emergent themes I will describe.

The first, computer vision for autonomous vehicles specialist, recently retired, reflects on the early days of researching autonomous vehicles and the tensions between (1) receiving grants that serve political purposes through military funding bodies, (2) the ways in which such grants could be used to conduct basic research which is the source of inspiration for the researcher, and (3) the rather practical-oriented apprehension of technology by the industry.

[101] It should be noted that these findings belong in the so-called "surprising" results. Nowhere in earlier descriptions of my research did I express an interest in investigating funding schemes in particular. But given that this area was something that my interviewees spoke about with passion (or critically) based on my generic questions about expectations and working routines, it is of crucial importance to report on those and problematise questions having to do with the future of curiosity-driven research. While my initial hypothesis revolved around the question of abstract expectations and visions about AI-as-redeemer or AI-as-condemner impacting research, distorting from actual needs for research, the question of expectations takes a rather complex shape: how are non-specialist visions and expectations employed by researchers to generate further expectations, in line however, governmental expectations, occasionally shaped by individuals and not the scientific community as a hypothetical whole.

> "And then I got money from the Defence Technology Centre and that was where we did the autonomous vehicle work. Then again, that was quite open ended; I am not a pacifist, I'm not a militarist; I don't like weapons. I think there's no point in having armed forces that are not the best in the world instead of having second best armed forces in the world, but I despise and loathe the arms trade and the selling of arms to third parties but it gave, doing autonomous vehicles research, gave us space to do the research we wanted to do. So we had, yeah, 4-5 years funding from there but most entirely it's been government funding with a little bit of industrial support, but largely the industrial support has been peripheral because they've been interested in exploiting the possible outcomes rather than actually being engaged with research right from the start." (Alfred Visionarius).

All these are reflections based on research culture contexts between 1970 and 1980. These were the times when lessons of overpromising have been learned, and the industry, back then largely separated from universities in lieu of initiatives such as the ISCF, was interested in applicable output. The second respondent, specialist in robot processing languages, design, intelligent knowledge-based systems, tries to tie early researchers' (with whom he has been moderately affiliated to) overpromising, which he justifies as "motivational statements" serving both explorative and exploitative purposes, with contemporary similar overpromising by deep learning developers; it is particularly interesting that he disambiguates between scientists who give open-ended motivational statements and politicians who perceive such statements as predictions expecting fulfilment:

> "[W]hat I realise is that a lot of what people say people like [highly esteemed roboticist, interviewee's close acquaintance], or Newell and Simon, and McCarthy in the original founders of AI and various others, you know, during the 80s and 90s, when they say, they make statements about what's going to happen, but these aren't predictions and it's unfair to take them as predictions. They might sound like predictions but what they are is, I think, much more motivational statements for the speakers themselves and others to hear and say, 'oh wow, if that's going to happen I want to be part of it.' And if this, I like this because it's quite hard to make effective motivational statements in a way it doesn't appear to be predictions. If you kind of soften it or qualify it, it just doesn't have the impact, and therefore it doesn't serve the purpose. […] It took a while to get there but, so, those predictions have come off but again the original statements I think are more motivational than prediction statements. And I think this is going on now with a lot of this deep learning stuff. And I think it could be that some of these deep learning guys are going to get caught out again because the politicians mistaking motivational statements for predictions will start to worry, well actually you haven't done all, we're getting into trouble with these things and you didn't tell us about this, and, you know, biases and stuff like this. I think the history of AI is that it's been slightly miscast in some ways. Now, of course it's not unique, maybe it's attracted more of this because of artificial intelligence and the somewhat emotional connotations that inevitably go along with anything like that as opposed to, say, cognitive science which has tended to keep his head down and maybe make better progress." (Aaron Auticous).

As analysis of interviews took place simultaneously with data collection, such statements from early interviews (third and seventh) shaped what I was looking for during the following ones. The generic questions expressed in 4.2 were the outcomes of investigating the expectations debate with researchers. What emerged, however, were three more specific themes of existing tensions in AI/robotics research culture, which constitute the following analysis: (a) the impact of grant application schemes culture with its many unwritten laws about "who writes the best proposals" and "who knows who" on scientific development, (b) revisiting the role of active generation of expectations and overpromising in order to

achieve such successful applications, and (c) how this grant-oriented culture appears to differ across nations (e.g. in EU or the US) or across sectors (academia as opposed to industry).

a. Problematising the Restrictive Impact of "Impact" in Curiosity-Driven, Exploratory Research

For the present section I will draw material from 7 interviews that I find indicative of what was in one way or another mostly confirmed in the rest of the sample. I chose this particular set of 7 out of 25 interviews to demonstrate the subtle differentiation between area of research and connections to the theme. The first quote examined comes by Otto Sensious, a senior specialist in computer vision for autonomous vehicles. He was the first among my interviewees who made a specific case about the limitations posed by governmental funding bodies, sensitising me to look deeper into the problem. His seniority allowed him to draw comparisons between "good ol' days" and the hurdle of today:

> "[G]overnment organisations always have to be doing something different to justify their existence, I guess. So, you'll find you have a special programme, and let's say electric cars, and then they'll say 'right, that's been done, let's move on to the next one.' And the problem may not always be solved. So you are limited by what the Research Council is funding. Allied to that, [is] the balance between Research Council or indeed industry-driven research and the abstract maths I was talking about. You know, what would you call curiosity-driven research, nobody's interested [in] and has definitely changed; a lot. So it's harder now to have PhD students that are completely free in their topic." (Otto Sensious)."

I found it interesting that Otto extended his concern about funding schemes' limitations to selection of topics by PhD students, showing that, in a sense, new generations of AI/robotics scientists are entrenched from the beginning: they cannot progress into doctoral candidates if they do not comply with the governmental fashion-based funds; something which is likely to influence their career beyond their doctoral training. Douglas Medicliff, with background in biologically inspired computation, that is AI approaches which are motivated by looking at biological systems, is lucky enough to receive funding for the medical applications of his work, but admits, with healthy doses of humour, that much of his scientific curiosity is satisfied by unpaid work he conducts on his own and with his PhD students' assistance.

> "[M]y work is mostly being funded by EPSRC[102], some of it by [UK-based medical company], maybe other small things. Yeah, typically I got an idea. I write a grant, get some collaborators, and target the funding, you get rejected, but sometimes you get some money. [laughs] A lot of research I do is not funded too, so in my spare time I like to just try things out. And I have PhD students who work on ideas, some good, some not so good. [laughs] […] The things I'm really interested in are harder to get funded. So that tends to be stuff that's unfunded. I just kind of play around with, or get PhD students to work on. […] Well, funding determines what people can resource in terms of doing research. So, of course, you can do some research by yourself, but if you want to do something with lots of impact then you need more people to work on it and therefore you need funding to pay their salaries. And that tends to be determined by the priorities of the funding councils which are influenced by all sorts of things: by the media, by the government, by society, by academics with particular interests who are driving the process. [laughs] […] there is some drive to overemphasise what you think the outcomes will be;

[102] Refer to Appendix 1c for definitions and descriptions of the various funding bodies and schemes featured in this chapter. Earlier drafts of this dedicated a section to explaining those, however, this was found to be at the cost of the chapter's flow.

> because grant proposals will get assessed and you have to kind of fire the imagination of the people reading it. [laughs]" (Douglas Medicliff).

His assertions are of particular importance as, in a sense, probably without knowing, he partly alludes to a rather romantic view of the "underground" scientist, who takes advantage of governmental grants to access equipment, satisfies governmental or business needs to sufficient extent, but preserves the enthusiasm of insatiable scientific curiosity conducting possible groundbreaking, yet unpaid, work in after-office hours. In other words, while exploiting certain circumstances, he can still exercise some autonomy within the applications-driven funding landscape, part of his engagement with a multi-level game involving personal academic salary together with external sponsorships. Such a model was described by Pickering (2009, 2010) in his description of early cybernetics as nomad science. Another interviewee, Lloyd Fludic, self-identified as a crossdisciplinary researcher, offers a slightly similar yet more extreme view. While Medicliff tries to find a balance between exploiting the grants by serving their purpose and at the same time explore new areas, Fludic's experience refers to the inexistence of checkpoints once grants have been allocated, allowing the laboratories to explore science as they wish:

> "So the way that is at the moment is it corrals people for everyone to work on the same thing at the same time and penalises people who want to work on strange things that aren't in line with the pack, you know; but as we know from science all of the things which are the most interesting are the things which are disruptive which come from leftfield. So effectively what we have to do to make that work within my research group is we have to take the money, we have to say we're gonna do something and then when we've got the money do something else, right? And lots of the people who are actually pushing forward a good scientific agenda are working in that mode. They take the money to do something and then they actually use the money to do something else. […] So forcing collaboration, which is what a lot of the Research Councils trying do by nudging people towards interdisciplinary working is something which people do because it allows them to get the money but typically they take the money and they do what they were going to do anyway. […] But in fact what happens is the money comes in and because the way the funding works in the UK there's no checkpoints after that, the money is given as a block grant, it's not sort of doled out on a three monthly basis or on a six monthly basis; so there's no incentive really for people to collaborate. What there really is an incentive for people to say that they'll collaborate." (Lloyd Fluidic)."

A case can be made in that their specialisation makes them differ in their approach. Douglas Medicliff or Otto Sensious (both expressed the lack of funding towards abstract mathematics research) are so specific in their research that it seems impossible for their funded projects to pass unnoticed if they do not comply with their promises. Fluidic, on the other hand, using crossdisciplinarity as his starting point, earns the advantage of camouflage. Funders will be impressed by the language he is employing and the combination of highly scientific jargon, that by the time the research grant period is over, no form of specific evaluation will be able to assess the exactitude of match between promise and result. A similar case of crossdisciplinary researcher is Shengin Eering, cyber-physical systems specialist, who joined academia after a long career in the industry. His assertion acknowledges REF as an evaluation scheme, however, he deconstructs it by reducing it to a matter of personal connections and to a deeply entrenched problematic culture of corrupted academic (human) nature, nevertheless, within the context of contemporary academic environments being imbued with a plethora of funding opportunities of varying scale, allowing researchers to enact their niche strategies within them:

> "[…] we've got REF and all that, and you put names, and you get names to independently call up, you know, so we can call them up and ask 'Is this true or not, did you actually do this?' Well, if I'm my good buddy I'll say 'Yeah we did it.' Who's to say that's wrong? Who's to know whether that's right or wrong? OK? So it's human nature. And I doubt there is any, anyone within research that would, you know, sort of like, try to shoot somebody down unless they know that's either my, I really don't like that person or what, but it's personal. You see? So it's a very difficult culture to fix. […] But, you know, at the end of the day no one's a winner. No one's a winner with these sorts of hype curves. Even though if somebody's got funding because of it. There's no winning, in the sense, yeah, because nobody gained from it. Who gains? Who gains? That's the question. Who gains?" (Shengin Eering).

This last question of benefit ("who gains") is an interesting one. Indeed, who gains? I will return to this question regarding winners and losers of promissory work in this chapter's concluding remarks, but for now, let us continue the journey across tensions between government, industry and academia; and they weight which is placed on either exploration or exploitation. Ravi Autonomaskar, very senior in position and with an academic lineage leading back to AI's early pioneers, and robot learning specialist interested in AI/machine learning applications in assistive autonomous robots with several industrial partnerships in the construction of autonomous vehicles, confirms the above, but poses it in the form of a paradox: governments do not want to appear as directly associated with private interests and approaches to applied technology, however, for reasons of national strategy support, they will sponsor local industries. This allows certain freedoms in the researcher's life in the university, but the overall sentiment appears like a hide-and-seek mode of admitting who is less money-oriented and more curiosity-driven. Autonomaskar's remark can be further situated within the post-Lighthill/post-Mansfield, applications-oriented view of AI:

> "[…] the majority of my funding comes from fairly well understood kind of governmental and quasi-governmental sources. […] So, in the case of [collaborative industrial project], for instance, the objective is we want to build the technologies to have a self-driving car, we're going to partner with different people to make use of it, or maybe deploy a service, and then the money that comes out of this feeds to people, and it's very well understood. The point of academic research is slightly different. In principle, it's totally open-ended. In practice, it is not at all like that. In practice, what actually happens is that you have to convince these agencies that what you're doing is interesting. And my observation, perhaps of a personal observation and others might have a different view, is that the governments have increasingly gone away from open-ended ideas to very, very practical, you know, 'how is this going to make a difference' kind of ideas. But in a very kind of haphazard way. So that governments can never explicitly say that they're doing the same thing as companies; but actually they wish that they had products that are as well refined as that, and so we end up with projects that are strangely applied. But there is somewhat some element of freedom. But, if I speak from another perspective, as a researcher, the advantage of being in a university research department is that there is no authority that dictates what I do other than myself and the constraints placed by this generically 'how are you getting funding,' whereas in a company there's much more clear tone about 'why you're doing what you're doing'." (Ravi Autonomaskar).

Wolfgang Swarmroboter has thought deeply on the questions of expectations after reading my invitation letter, and while discussing on the role of institutional/governmental expectations, improvised, expressing both his distaste as to the way funding schemes work, serving political purposes (who do not take the reality of scientific advance into account). He further comes up with an alternative scheme of a system that rewards good research once it is produced, instead of offering grants to good promise-makers and proposal writers. His occasional uncertainty in certain points (I have preserved all pauses in the transcript) is

indicative of the chicken-egg type of question such policies entail. You cannot have good research without existing funds, but funds should go to good research. Hence, promises appear to be inevitable:

> "And it's not a good perspective to work under this kind of pressure, because it doesn't help and also, I mean, many people agree that this kind of expectations which are put out by politicians or other interest groups in the form of funding programmes, this is not necessarily helping advancing science, it's more the ability to recognise an opportunity, to find some kind of interesting relationship to be in a position to understand certain mechanisms or to produce, to start some mechanisms or to produce a certain type of dynamics. […] So, for example, computational neuroscience was heavily funded both in US and is still funded or was funded in Europe and the outcomes are very modest. Which is not so much that researchers are unsuccessful in a certain sense. It's just that these kind of hopes that somebody who doesn't understand the subject matter and has very simplistic ideas how is this should work, so if you tell the scientists 'go understand the brain, build an intelligent robot; I give you as much money as you like,' it's a bit, a bit weird, yeah? [laughs] It has nothing to do with how things are actually happening, how we approach an understanding and also... […] But as an institution, if you did something, then you should get funding, because then probably at least as an institution you will be able to do more, better research in the future as well. But most of the project-based funding is mainly based on promises and therefore it's not necessarily related to the outcome and just produces better promises. And I know a lot of people who have lots of very highly paid projects and they are not the most successful people, but they just know how to write proposals in order to set up a network and to produce some kind of activity to increase visibility and so on. But it's not necessarily directed towards progress in any sense." (Wolfgang Swarmroboter).

To what extent Swarmroboter's suggestion is shaped by his knowledge of punish/reward-based reinforcement learning is an unknown. However, this suggestion brings us back to the basic question of evaluating what kind of research should actually be funded. An alternative perspective is the one offered by Aslan Robotoglu, roboticist, specialist in human-robot interaction, with practical experience in many-legged robots and medical applications of robotics in areas such as laparoscopy. More of an altruist as a character, often expressing environmental and pacifist concerns during our interview, he is posing a distinction between what is funded and the mundane dimensions of possible useful applications that are invisible to funders and researchers. Such groups care increasingly less for daily life and human and interactions, so, we can speak of a process of alienation between mundane needs and funded research:

> "A robot that can talk, that can communicate with humans, that can interact, such kind of robots are useful in social life. They can help a lot or people in their homes, in a museum, in social environments, in a hospital and such. So, the thing is, [short pause] we are not very much aware or we don't have the channels to get into that kind of, or to be aware of that kind of needs. So, researchers are in their universities, in their labs and they try to get funding, and they respond to the calls about funding. OK? So these calls, or the funders do not necessarily always fund projects that are very, that will be very useful or that are addressing daily needs of people. OK? So, for example, industrial funding is an important thing. So, when you look at this funding, the major problem is what is needed in the factory floor, more automation and that sort of things. Or health care funding, when you look at this, you will see a lot of things towards the needs in the operation room. These are important things, of course, very good things, to develop medical robots and such or medical systems, and these are nice, but if you go to the street, or if you go to a school, or if you go to a library, if you go to a hospital, as a researcher, a robotics researcher, if you get into the procedures there, process there, one can perhaps observe or identify a lot of situations where some robotic systems could help people, could be very useful. So, I

> think robotics researchers do not have the ways or channels to perform this kind of observation, or there is not, let me put this in another way, there is not a sufficient amount of channels that will bring these actual daily needs of humans for robotics to our agenda or to our perspective, to our labs. So, that would be a nice thing, if we as researchers or, I don't know, as a society, we can develop such channels which will bring daily needs and researchers, robotics researchers together. And that would help introduce robots and robotic solutions to practical daily life problems." (Aslan Robotoglu).

It is interesting to note that Robotoglu referred to this lack of communication channels twice during the interview. Once, when I asked him about funding routines and challenges, and second time, during the last part of the interviewing process, in which I was asking interviewees to set their imagination free and speak about potential breakthroughs if no restriction whatsoever existed. While most interviewees spoke of specific technical advances so far unsolvable, Robotoglu referred again to the creation of such channels; that is, he saw within the context of funding the disparity among AI researchers, correlated, possibly to their lack of consensus about AI's conceptualisation.

Some carry-home messages from this section before we proceed to the second part of this chapter's empirical analysis: There is a generalised sentiment that explorative, curiosity-driven research remains unfunded and researchers struggle to exploit existing schemes (a) to survive, and (b) to conduct some of the explorative research if they have the time and volunteering PhD students. A hide-and-seek game is played between industry and government with academics in the middle, with occasional hide-outs called "curiosity-driven" research being the mask of applicable technologies and "applicable technologies" being the hideout of conducting research that the lab finds useful. Promises generate further loops of promise; not even promise-requirement cycles. They are rewarded more than outcomes and this creates a culture of writing the best promise-rich proposal. Researchers, with their focus placed on writing grant proposals (as well as governments and industries) are alienated from the real-world that can benefit from their findings. How much AI and robotics is used in developing countries? This relates to the open-ended question about who actually gains out of this messy network of funding. One hypothesis might be that colossal companies (e.g. Google, Amazon) benefit by these overall processes' by-products; open-access articles and tools that act as building blocks in advancing their systems; an institutional-scale application of what Zuboff terms "behavioural surplus" in the way such companies benefit from their users/customers everyday interaction with their product beyond the intended purpose – but this is a question to be addressed in future work. Although I briefly analysed the promissory value above, interviewees have provided me with more specific input that deserves a separate section of scrutiny. This is what follows.

b. Problematising (or Praising) Overpromising for Successful Grant Applications

The mantra of SoE as outlined above is that the more abstract the expectations and predictions, the worse the results for applicable anticipatory strategies. I will begin this section by examining a statement by Paolo Oceanio-Marinetti, whose background is in oceanography, and has only recently become specialist in data-driven machine learning applications in marine soft-robotics and fluid dynamics (e.g. predicting the movement of waves to sustain underwater vehicles). He is one of the interviewees who insisted many times that he is not an AI specialist per se, despite the fact that AI and ML figure on his institutional

webpage[103]. Interestingly, he admits that the strength of his institution in a certain field allows basic research and pinpoints to the advantages of hype in terms of gaining trust towards untried applications:

> "[…] to be honest, at the moment we are in a good spot because this thing, that machine learning is showing these almost superhero power in solving problems, people are putting the money into it. […] I have the feeling that many of those who put the money into machine learning think that they can basically give you the solution to a problem; which is not really the case. In my case, I'm basically travelling this wave of positiveness and of credibility in machine learning, and because I'm part of the data-driven innovation fund, it's easier for me to get some money out of these. So, even if this is a project that probably wouldn't be funded by an industry because it's too far-fetched, but it's still very hard to convince a company that you can do this stuff. […] now the [data-driven project] initiative is going strong here in [university name], so it's probably easier to get research at a more basic level, so doing basic research like this, because we don't know whether this is actually going to work; we're just starting now." (Paolo Oceanio-Marinetti).

I am not sure whether the oceanographer Oceanio-Marinetti made an intended pun by speaking about a *wave* of positiveness, but this poses yet another interesting riddle in the context of some available computational power and abundance of data: does overpromising and "far-fetched" research happen because of an existing hype wave or is the hype wave sustained by overpromising? Although Oceanio-Marinetti specifically distinguishes between industry-funded and government-funded projects, Marta Objectividez, a very successful computer vision researcher with specialty in object recognition who was recently employed at the University after a long career at a colossal company known for its AI applications for online services, expresses this question, explaining it as a possible combination between a market effect (a race to catch up) paired to technological progress:

> "I don't know if that's the issue or if it's just more of like a market effect. What's clear is that the area is growing a lot and really fast. Right? That's very clear. Whether the force underneath is overpromising or the force underneath is just this, like, general hype, of like "oh my gosh, those guys are hiring a ton of people, we should also hire a lot of people have an AI lab, you know, just in case," I don't know if, like, this is the effect of it [or if it] is the other one. But yeah, I do really think that it's growing really fast and I don't know if the technology is growing as fast. […] maybe the combination of the two." (Marta Objectividez).

Returning to some of the interviewees encountered before, Ravi Autonomaskar, fully aware of the complications following the Lighthill report, belongs to the AI scientists who learned to abstain from grandiose promises, however, his experience has shown that the iterative process of producing *something*, (e.g. advance some tools) can make up for the unrealised promises advancement in tools "makes up" for large promises; however, one should proceed with care:

> "I think they're being too arrogant and there are many, many groups that are arrogant like this and they're surely not solvent because they're not asking many important parts of that question. And so the question that I don't know the answer to is what will be the consequence of it; maybe nothing. I mean, maybe the technical tools and engineering that's produced will be enough for people to ignore the scientific promises but I don't know. I mean, I, as a personality, as a scientist, I would prefer it if we didn't make wild promises when not necessary." (Ravi Autonomaskar).

[103] More on such acts of humbleness and denial of certainty/expertise as an actual sign of expertise, in the following chapter.

It should be noted that Autonomaskar, aside of being a senior expert in AI applications in robotics, is also responsible for the sampled robotics centre public engagement projects. His views are imbued with the sense of responsibility of public speech that has to build confidence in the public mind but after careful consideration of previous lessons about distorting overpromising. Sensious reflects from a slightly different standpoint. Place emphasis on the rather natural way he shows that it is not only researchers making promises, but also research councils (who depend on what promises researchers give, bidding on science vote as currency of intended investment return), the inadequate evaluation criteria, and the income based on "being bullish":

> "I think research councils are under pressure, as I say, to get results. And perhaps they make promises that, you know, 'if we put an extra hundred thousand into electrical vehicles we will reduce pollution levels by...' and the question is, are they quoting academic studies? I guess the problem is the verifiability of those academic studies. So, [pause] when you're writing research proposals you have to write an impact study […] So the question is, when I'm writing a research proposal, I guess I have to be kind of bullish about what doing this research may achieve. […] The Research Council's expecting me to produce results. […] So if you can say this has generated a spinoff company, employed forty people and that's enabled us to go to Mars, that's good, you know, it's got to generate income, it's gotta generate income, and pay our salaries, it's gotta generate results to justify the government expenditure." (Otto Sensious).

In the following quote, I bring in together different parts of Shengin Eering's interview so that it makes more sense as he was referring to earlier statements. During the first part of the interview, that had to do with defining terminologies, he reduces all of his work (cybernetics, AI, virtual reality, robot sensing) to algorithms, as in his view, algorithm science is the major underlying science, and all the rest is what STS would call socially constructed terminologies satisfying the needs of relevant social groups; thus verifying the historical lesson that societal expectations shape the perceived fixity of technoscientific terminologies. These "semantics," as Eering called them, are the variable, but hype around different semantics is the constant; according to Eering, hype always exists, and its content changes just in order to fulfil the need for hype. Proposal writing follows the tropes of the hype and instructions on how to sell the exact same product with fashionable terminologies dominate the field. It's a game of salespersons stretching between the poles of honest sales and effective sales:

> "Terms have been conjoined, terms have been coined to make research more exciting. Ok? There's really no difference in my view in terms of whether it's machine learning or networks or whatever, genetic algorithms, whatever. They're algorithms at the end of the day. The complexity of the algorithm makes a difference. And of course the precision of the outcomes of the algorithm makes a difference. But other than that they're just algorithms. So, terminology is what humans would put onto something so that we can be unique. There's no difference from an ice cream and a frozen tub of cream. […] So the hype is always there. […] It's driven by human's curiosity […] You know, it's driven by our way of interpreting the world and putting our own stamp on it. Yeah? Similar to terminologies. It's how we view it. Someone views it in this way, "don't write it in that way." So it's about being a good salesman [sic] compared to being an honest salesman. And there are no honest salesmen around. There are very good salesmen around." (Shengin Eering).

How far can the salesperson's approach go? How much academic cynicism and tribalism is accepted in order to convince governments and funding bodies that one's research is worth the investment? The following long quote by Lloyd Fluidic, whom we encountered previously on the admittance of lack of

checkpoints once grants are allocated, offers what he calls "a very scathing view" of research realism. Because of tribal attributes humans express, the AI game is dominated by a relatively closed loop of successful researchers who are selected to evaluate research proposals, setting the trends of what is considered to be hyped terminology according to Eering, and increase the possibilities that the laboratory in which the research council advisor is working in will be benefited by the funds offered by the council[104]:

> "I mean, it's tribalism, it really is the answer. Academia is a sort of conglomerate of tribes who are all sort of existing in the same physical space and all have sort of chieftains which sort of are in charge of that one tribe and in order to succeed in academia, then you have to become the sort of tribal leader. So there's this sort of psychology which is about, you know, the way that people work on problems based upon not working the most interesting problems there are but working on the problems which will serve them the best benefit, you know, there's no such thing as altruism. Nobody's working on these projects because they think it's a good thing to do for society. They're working on it because they want to become a professor at a top university - is the honest, honest answer. And when people are working in that sort of mode and a sort of tribal mode they're not really open to collaboration because that weakens their position in their own particular tribe. […] So at the moment an example, a top, you know, example of this is in the Industrial Strategy Challenge Fund, right? So the Industrial Strategy Challenge Fund is not something which scientists are doing because they want to advance the economy in the UK. It's something which comes from a very very, very top level strategic position from the government, comes down through the base to RCUK, RCUK consults with members of esteem in the community, those members of esteem say these are the sorts of things which I or we collectively as a sort of council think we should be doing. They put out the call, the call comes down and then people, you know, effectively tend against that call, they put in their bid against that call. The problem with the system is that obviously the people are best placed to take the money are the people that advised the research councils in the first place. So if I advise the Research Council that we should definitely be working on soft robotics, for example, because it will help to underpin that UK strategic priorities in industrial innovation, they put out a call for robotics, for soft robotics in particular using wording that I developed in my scientific advisory panel, I then apply against it and I get the money, you know. The world is not a meritocracy." (Lloyd Fluidic).

If we assume that the Lloyd Fluidic's confidential confession is honest indeed, expressing the reality of funding criteria, at least in the UK, this still does not leave us with a binary ethical choice of good and evil. It could be the case that even if the system appears as corrupt (without "meritocracy"), it is so because the ones with better research output have also found means to adapt and flourish within hierarchical systems after all, ensuring their survivability. There is an inherent unknowability, however, about whether what is funded is so because of merit or not, since we will never have the opportunity to check for each and every "what if" of other, unsuccessful applicants. The question here takes another turn, and in order to summarise this section's messages:

1. Funding schemes and evaluation criteria are very likely to be impacted by informal structures shaping advisory committees of research councils, paired to an increasing involvement of industry and further stakeholder groups in research funding committees.

[104] The quote's first two sentences have been visited earlier, as part of the discussion on definitions and the meaning of AI. They are now set in better context, as the interviewee took advantage of the opportunity and speak about the research culture implications of terminological vagueness.

2. This creates self-propelling closed systems of terminologies that become hyped from period to period; creating winners out of those who are able to convince research councils that their terminology is best.
3. Institutions (such as universities) with long tradition of successful applications strengthen that loop, minimising the possibility of minor research institutions to enter the dance of promissory work, unless they quickly adapt to existing hyped terminologies.
4. The underlying technologies behind hyped terminologies are essentially expressions of the same principles with minor alterations or optimisations, being rebranded for the purposes of each generation's winners.
5. Overpromising (be it motivational statement of irresponsible exaggeration) is a technique that might ensure funding, and if performed with confidence of existing background (knowledge, equipment, institutional strength) may at least result in specific products, although different from what was promised (as also shown in the cases of the "AI effect" and interplay between internet and AI technologies). This might entail the danger of overlooking what can be done with existing tools.
6. Overpromising, then, blends the binary notion of curiosity-driven and industry-driven research and innovation, or of exploration/exploitation, as often, what appears to be explorative, given the rush of remaining reputable, results in exploitative products with practical applications.

The quotes selected for this section did not reflect much on the difference(s) between national and international, or sectorial (university/industry) approaches to exploration and exploitation. The following final section, preceding the chapter's discussion, will focus on these dimensions.

c. Problematising (or Praising) UKRI in Comparison to International Schemes or Industry-Oriented Mindsets

So far, I have outlined the underlying informal structures and overpromising tactics giving shape to contemporary impact-oriented research funds. This section will place more weight on how researchers think when comparing their governmental funds from the UK to the ones coming from industry or Europe. One of the last people I interviewed, Connie Converse, Professor, with background in information science, AI, and computer science, specialising in conversational AI (applications similar to Siri, Cortana, or Alexa), described an interesting tripartite taxonomy. In her experience, ERC tends to fund very visionary, "blue sky" and curiosity-driven research; EPSRC tends to fund more realistic projects, yet allowing certain degrees of freedom; whereas it is implied that the industry cares mostly about practical applications and requires more connections in order to collaborate:

> "I think so far I haven't received that much funding from industry to be able to say something. I mean, there are these open calls, for example, from Google or Facebook where they propose like a challenge and then you can write about something. It's very difficult to get anything funded without knowing someone who will fight for your project internally. [short pause] So the, I think, the nice thing about the EPSRC is that it funds quite, wouldn't say blue sky, but less... research which is maybe less directly applicable. So, you know, it's more basic research into like developing different techniques or algorithms rather than 'OK, here's a specific use case'; they do want to see a use case obviously but it doesn't like have to be so much applied. And that's what I really like about their funding is that it's

> more, um, less applied. […] With the EPSRC you need to be quite realistic. You can't just oversell your things, which again I like about the EPSRC, because if you start overselling yourself reviewers will say 'hold on a minute that's a bit much,' you know. And I think. you know. you have to sort of start from a concrete problem and then give a methodology where you can show 'OK...' And they love seeing a formula saying, you know, 'this is whatever, that's the algorithm and this is how we will extend it,' so you need to make it very plausible of how you tackle the problem. And so, you can't lie basically; and saying, you know, you would solve something [confidential type of hard AI problem] would be lying. [laughs] But on the other hand, you know, then you have funders like the ERC, which is from the European Commission which funds very blue sky research, very much, you know, visionary. And again, you know, I find it personally very hard to be visionary but realistic at the same time, so I never managed to get one of these grants [laughs] because, you know, you can't just say 'oh I solved this problem and I know how to solve it.' That wouldn't be visionary enough. So, if you know how to write a blue sky research proposal, I'd love to know. [laughs]" (Connie Converse).

Converse appears to show a world quite different that the world of Fluidic or the world of Sensious and others. In a sentence, the tension can be described as such: visionaries lack realism (Converse), but realists lack vision (Sensious). Later in the conversation, Converse referred to examples of visionary European projects that failed to deliver their promised results. Placing Converse and Fluidic's statements about institutional connections next to each other, it seems that this lack of meritocracy is uniform both in the industry and research council funding schemes. Nonetheless, Converse admitted little knowledge about the world of industry. Oceanio-Marinetti, surfing the waves of positiveness, draws a thick line between academic curiosity and entrepreneurship, being very critical of the way businesses treat scientific development:

> "[…] talking with companies is, to me, is a pain, maybe because I am very much an academic. So whenever I have to deal with entrepreneurs or stuff... I am sorry that you are recording this but I get bored immediately, I get bored because they feel like they can do stuff immediately, they can sort everything up in a second, and when you bring them back to reality, it's such a disappointment. So I'm just now speaking with people to design a... I cannot even talk about this because it's classified. But anyway, design a specific vehicle, and it seems like, the problem is already solved. As if, you know, buying stuff from different companies and putting it together, [dusting hands] you're done. That's not how it works. I mean, it's really, really complicated. And when it comes to problems like these, surely there is a reason why it's not been done so far." (Paolo Oceanio-Marinetti).

In a follow-up question, I specifically asked him about his views on curiosity, which explained his sentiments about the industry as relating to reputational benefits:

> "So, this is an easy question [laughs] for once; I'm totally driven by curiosity and, like I said before, a little bit of egotism, of course, the dream of being, you know, at least known in my niche for putting out some good work. So no, I think, if I will fail in my career as an academic will be because I don't get along very well with their entrepreneurial kind of approach. I'm not good at it. I don't care, and when I don't care about something, I just don't do it. Which is something really bad, because good people do it. I mean, good people are the ones who are at the same time able to be passionate about their work, really curious about it and find the way to get it across to industry and get good funding, even if it's a very, let's say, in the air idea. At the moment I still haven't developed the skills of selling well the products of my research; so it's still very much curiosity-driven." (Paolo Oceanio-Marinetti).

To showcase the diversity of opinions about the industry-applicability/government-curiosity spectrum, even among very proximal researchers, let us examine a quote by Maurice Constructeur, whose office was located only few doors away from Oceanio-Marinetti's. Like Oceanio-Marinetti, Constructeur did not admit long-time expertise in AI or ML, being a construction engineer with specialty in computer vision applications for construction tools. He became increasingly acquainted with ML applications in construction engineering in the last 3 to 5 years because of a number of projects he was participating in or supervising involving ML and due to a surprising growth of ML-related articles submitted to a prestigious journal on automation in construction he is editor of. His approach to industry is a far milder one and, in a sense, we could attribute to him an archetype the honest salesperson, as defined earlier by Eering. Business can be good, if the promissory work takes place responsibly as the work progresses, with short rounds of investment, as opposed to block grants offered by governmental funding:

> "So, I mean, like everybody I try to attract EPSRC, UKRI funding or EU fundings or..., but most of my funding so far has been through actually other sources. So we had some big funding from the [public construction funding body] to develop technology force and training of trades in particular. So we've done virtual reality and things like this, and in that case I worked very closely with the funder because obviously these funders tend to be a little bit more, I think more like a client, a customer let's say of some research, although, I mean, I think it's good working with them, it's not recognised to be as prestigious but it can be good funding and if you build a good relationship with people in industry if they like what you do they tend to be pretty easy at coming back and extending and giving you more funding. So the effort to get the funding is generally not necessary that much upfront compared to UKRI and the likes. But I think it even goes down as things progress. If you do a good job. It's like everything. If you don't give a good product you lose a customer. But the, if you do give a good product and these customers they fund you, they are happy to continue fund you if they have the resources and training." (Maurice Constructeur).

I was particularly impressed by the following part of Eering's self-introduction. While most academics so far express a rather critical stance towards industry, Eering's approach is that knowing the fast-pace rhythms under which industry performs can benefit education. Hence, he knows that curiosity can flourish within academic circles, but academia lacks the speed of industry to actually produce the curiosity-driven innovation. Moreover, his industrial background allows him to be flexible in terms of opportunities and follow the money:

> "I came from industry and we see from industry there's a lot of push unlike education in the sense where push is normally very slow. Industry it's very fast pace. And I was hoping that my research could drive that pace even further so that's the whole reason behind it. […] And, yeah, so kinds of funding can be industry, can be European, I would say, more European than EPSRC. [when asked about Brexit's impact] I think if there's a threat, I'll, I'll just move. You know, I'll go where the money is. Yeah [laughs]" (Shengin Eering).

It is interesting to notice the very diverse views on ERC as well. While Arendt made a case for the dangerously visionary European approach, Eering, with his industrial background claims that ERC funds most of his projects. One of my first respondents, Dalia Virtualia, a highly positioned specialist in intelligent graphical characters, virtual agents, in the overlap between AI and virtual reality confirms such a view, claiming that no matter whether it is ERC or EPSRC, "funders follow fashion" – however, it is further interesting to compare her experience with that of Fluidic's as she adds to his argument about lack of checkpoints of realised promises by mentioning the unavailability of further resources for evaluation,

since all expenditure was based on promise (following Stoker's argument). Virtualia's suggestion when it comes to industry is that the field of technology transfer should be empowered, and bridges should be made between industrial application and academic research. Finally, she makes a case for involving end-users in the research, slightly echoing Robotoglu's earlier statement on communication channels. The fact that her response begins with reference to Brexit, is quite telling of the uncertainty felt by researchers, no matter the typology across the exploration/exploitation spectrum, since much of the AI/robotics development depends on simultaneous UK and EU collaborations:

> "Well, until recently the EU funded my work, thank you Brexit, though not just that; the EU swung into one of their risk-averse, lets-fund-industry modes as well which makes it hard to get money out of them for the kind of thing I'm interested in and they're not very into interactional systems. Funders follow fashion and that's not the fashion at the moment. So, in the EU there's a formal framework for accounting for what you're doing to your funders which consists of review meetings and things like this. They hire reviewers who come and assess what the project is doing. These projects have gone very well, I've ran quite a few of them, I think I've done a decent job of running the projects. And we produced interesting results both scientifically and, I might say from an application viewpoint in the sense of interesting systems, research systems. None of these made it through into products but it's kinda not surprising really, depending on researchers [to do technology transfer], this is really asking for trouble. There's no reason why researchers should be the best people to take research into actual products out in the field. Technology transfer is a whole different ball game. Anyway don't get me started on technology transfer. So with UK ones, the funders don't in fact these days get to ask for much feedback, they expect you to produce publications and so on, you will produce an impact plan which says how you hope to make impact with your research. All this is assessed when you're funded. There's very little follow up at the moment, they don't have the resources to do it[105]. At one point they used to ask people to write final reports. They went through a phase of assessing final reports but they didn't have the resources to do it. These are all things that we have to do for them - they don't have the staff so panels are unpaid efforts and assessment. So there just weren't enough people to go around to make that work basically. And then, in the longer term there are research assessment exercises which assess research in a much more global level. Other than that, stakeholders, you know, they vary. We've always tried to involve people in our applications when we're developing them. And I'm sure there's some benefit to our stakeholders in doing that, we can't give them a product in exchange, it's just not what we do. We try to make sure that we engage with our end-users when we're designing things. Because if your view of intelligence is that it's interactional then you have to interact with people to do it. [laughs]." (Dalia Virtualia)

Virtualia's multilevel, UK, European, and global experience, invites us to think in more international contexts. A good comparison was made by Gerald Compulagnus, one of my interviewees who was external to my sample, who used to work at one of the two universities I examined but left, with background in pure mathematics, and developer of a historically significant AI computer language, applying AI in finance. He admits that the UK universities do not have enough financial power to allow the strategic "escape" of practitioners who move to industry, and compares this with American contexts:

[105] To quote an anonymous peer with experience in such funding schemes who offered feedback on this passage: "With EPSRC funds you are required to draft an impact plan – but our evaluations of this work showed that recipients were allowed to use the belief of researchers in the utility of their work in lieu of actual evidence of utility! ESRC was much stricter about how it assessed impact…"

> "The University's role very often is to be bought out eventually and that makes it tricky and awkward for angel investors. In America certainly the richer universities are generally very happy to say, you know, 'bless you, go and do something and if you're successful, you'll remember us.' British universities don't have enough money to make that a successful strategy. So they tend to make life a bit more difficult for angel investors and angel investors get used to it but it's still a bit painful and it was particularly painful with [university in which researcher was previously employed]." (Gerald Compulangus).

This suggests, then, a tension between industry and academia in UK contexts, possibly reflecting the competition between the domain of commerce and the domain of education; each side aims at protecting forward-thinking individuals within their premises, to showcase their value in the eyes of the government, which, in turn, aims at incentivising collaboration between the two. This section offers probably the most conflicting messages so far when it comes to researchers' views on different sources of funding. Some carry-home messages before this chapter's discussion:

1. There is room to investigate the transformation of academic exploration *into* industrial exploitation through technology transfer; which poses another question of expertise: who is the technology transfer expert?
2. A horizon of UK-EU geopolitical uncertainty paired to the very diversified views of researchers as to the research freedoms enabled by any of the two parties. According to researchers, the hype appears to shape research priorities both in the UK and EU, hence UKRI and ERC are both impacted by the vagueness of terminological rebranding.
3. Industry can be perceived as a very different sector, depending on every researcher's background and tradition: from a machine that cares only about numbers to a good responsible collaborator, or even a source of inspiration for a faster and more effective academic culture.

In the following section, I will aim at offering a more generalisable message based on the above empirical observations.

6.4 Discussion

"As they sat together in the candlelight, these neurosurgeons, senior academics and stock-brokers displayed all the talents for intrigue and survival exercised by years of service in industry, commerce and university life. For all the formal vocabulary of agendas and minutes, proposed and seconded motions, the verbal paraphernalia bequeathed by a hundred committee meetings, there were in effect tribal conferences. Here they discussed […] their plans for alliance and betrayal." (J. G. Ballard, High-Rise*. Ballard 1975: 136)*

One of the first questions asked in the beginning of this chapter was the apparent chicken-egg question on whether hype is generated and sustained by the tension between opportunistic exploitation of existing resources and exploration of alternative new choices for research practice; or whether this tension is the product of the hype. My suggestion, based on the above is that there is a constant element of hype which mobilises promising and expectation-setting. While SoE, so far, suggests the mobilisation of current decisions based on future expectations, I add that a constant need for hype is associated with new technologies. This can work at different levels and scales; from the everyday funding strategies of an individual researcher, who abides by or generates hype, to the collective efforts in forming a field or a new

niche. An example of the latter scale can be drawn if we connect these findings to historical elements. During a 1973 BBC show in which four AI pioneers were invited to defend AI against Sir Lighthill's criticisms, the following dialogue occurred. Sir Lighthill's critique was interrupted by John McCarthy, inventor of the term AI, admitting with humorous cynicism why AI could have *any* been named otherwise if were not for certain practical disambiguation and academic competition. This further proves the diachronic nature of the use of terminologies by different research groups in order to attract funds, at least in the case of AI with its underlying imaginaries:

> "**James Lighthill:** Now, what are the arguments for not calling this computer science, as I did in my talk and in my report, and calling it artificial intelligence? It's because one wants to make some sort of analogy, one wants to bring in, what one can gain by a study of how brains of living creatures operate, this is the only possible reason for calling it artificial intelligence…
>
> **John McCarthy:** Excuse me, excuse me, I invented the term artificial intelligence, I, um [laughs by McCarthy, Lighthill, and the audience], I invented it because we had to do something when we were trying to get money for a summer study [much louder laughs] in 1956, and I had a previous bad experience, the previous bad experience concerned, occurred in 1952, when Claude Shannon and I decided to collect a bunch of studies which we hoped would contribute to launching this field, and Shannon thought that artificial intelligence was too flashy a term and might attract unfavourable notice and, so, we agreed to call it 'automata studies,' and I was terribly disappointed when the papers we received were about automata, and very few of them had anything to do with the goal that at least I was interested in, so, I decided not to fly any false flags anymore, but to say that this is a study aimed at the long-term goal of achieving human-level intelligence [camera turns to Lighthill who is grinning with irony]. Since that time, many people have quarrelled with the terms, but ended up using it, uhm, Newell and Simon, the group at Carnegie Melon University, tried to use 'complex information processing,' which is certainly a very neutral term, but the trouble was that it didn't identify their field because everyone would say 'well, my information is complex [mild laughs], I don't see what's special about you'." (BBC TV 1973: 00:44:37-00:46:45, original verbal emphasis)

This is in agreement with Konrad's (2006) notion of the dynamics between collective and individual expectations, although, I would prefer to place emphasis on the hype (more present-oriented) rather than the expectation, in such cases. Let us be reminded that official history (Olazaran 1996) of AI accepted the predominance of McCarthy/Minsky's flavour of symbolic AI at the time. The following example, showing how Hinton, in his migratory period from Edinburgh's AI to the US, found shelter at Carnegie Melon, under Alan Newell's hospice, following the underpaid, exploration path:

> "The following afternoon [1981], Newell offered him a job in the department, though Hinton stopped him before accepting.
>
> 'There is something you should know,' Hinton said.
>
> 'What's that?' Newell asked.
>
> 'I don't actually know any computer science.'
>
> 'That's okay. We have people here who do.'
>
> 'In that case, I accept.'
>
> 'What about salary?' Newell asked.
>
> 'Oh, no. I don't care about that,' Hinton said. 'I am not doing it for the money.'
>
> Later, Hinton discovered he was paid about a third less than his colleagues." (Metz 2021: 41)

Hinton's neural network research would develop only in Canada, after his disappointment with DARPA being the only source of AI funding, itself a result "of the grant money that Minsky pulled away from Rosenblatt and other connectionists" (Metz 2021: 44), hence, I suggest that expertise/expectations dynamics make inextricable the relation between conceptualisation, politics, and exploration/exploitation practices. While space does not allow for further explications of what can be termed "machine intelligence terminology wars," it is shown that at AI's grandiose explorative project ("the long-term goal of achieving human-level intelligence") is proven thus far to be a standard card to be played to exploit short-term goals of achieving minor optimisations and assist incremental innovation.

Returning to the exploration/exploitation model and trying to extract and visualise the observations of this empirical journey, I aimed looking at different respondents' specialisations and derive meaningful conclusions about their research cultures based on that. I was impressed finding the following distinction of "AI workers" by Joseph Weizenbaum, early AI pioneer, Professor of Computer Science who developed the famous ELIZA program, in his influential 1976 book *Computer Power and Human Reason*: "Workers in AI tend to think of themselves as working in one of two modes, often called *performance mode* and *simulation mode*." To explain this, Weizenbaum employs the metaphor (or previous example in technological innovation) of flight simulation. Early attempts at simulating flight (and intelligence) followed the example of existing flying creatures, such as birds (or intelligent beings such as humans). These were eventually replaced by further pioneers in flying who cared about flying as a practical performance, considering "that their task was to build flying [or intelligent] machines based on whatever principles they could discover" (Weizenbaum 1976: 164-165; relating to the old form/function pair). Such a dividing line cannot be absolute, as researchers write computer programs based on human tasks that they simulate or wish to assist/augment, and likewise, discovering new principles that might lead to intelligent machinery might reveal principles about human intelligence as well (including nonhuman intelligence in the cases of bio-inspired robotics). Does the simulation/performance distinction tell us anything about the high epistemological values of simulating the human brain or creating general purpose robots and the low epistemological values of making small-scale applications of AI systems? Only partly. However, there is certainly direct resonance between the simulation/performance modes and the exploration/exploitation model since the exploitation of whatever available resources to provide with performance (of at least short-term, low epistemological value) can be contrasted with the long-term exploration of uncharted areas that will allow exact intelligence simulation (high epistemological value).

Nonetheless, this distinction had to do mainly with AI per se, with little or no reference to embodiment. For several authors, intelligence cannot exist without embodiment, and hence AI's physical support, a computer, a smartphone, a robot, or an autonomous vehicle, is crucial in judging whether AI has achieved its goal as a scientific enterprise. Wolfgang Swarmroboter, the computational neuroscientist who applies his knowledge in swarm robotics, when asked how he would explain the difference between AI and robotics to a layperson, gave the following answer:

> So it's more like the difference between an [pause] a scientist and, I don't know, a teacher maybe; a researcher and a teacher. So teachers, they have all kinds of facts and they need to work with the facts and to make some kind of meaningful whole out of facts; in robotics, it's always about establishing facts. It's more the process how to get this information rather than how to use information. Obviously, artificial intelligence is very important in robotics so once the robot has acquired some kind of information from its sensors you want to do all kinds of reasoning or predictions or any kind of inferences from the data. So then you need artificial intelligence but that's not the point in robotics, in

> robotics the point is to make sure that whatever you believe is true or whatever the computer on the robot holds true is actually in sync with the environment, it is this kind of tension between an ideal world which you can simulate in the computer and the real world, that's the battleground of robotics. (Wolfgang Swarmroboter)

To simplify as much as possible:

AI: attempts at simulating or intelligence – leaning towards exploration

Robotics: attempts at performing intelligence by applying AI in physical space – leaning towards exploitation

It is no wonder then, that most of my interviewees from the examined prestigious robotics centre were mostly keen in showing how, despite their curiosity, their findings have practical applications in robotics or in other similar, quasi-physical performance outputs. It was true that people with seniority, not necessarily in academic position, but mostly in age and the generation they belonged to, did comparisons between the early visions of AI, visions that have now been replaced by the small building blocks of incremental innovation.

But what do such divisions tell us about the exploitation of funding schemes? In a sense, the entire discussion above on the distinction between simulation and performance can be viewed as *an aspect of broader AI exploration*, while the empirical discourse on funding schemes can be viewed as *an aspect of broader AI exploitation*. That means that exploration and exploitation work on two levels of division allowing various intersections.

The table below synopsises my argument visually. The columns A and B signify the difference between high and low epistemology AI: the vision of achieving AI either by simulation or performance as opposed to the practical means to conduct research on that field – therefore, the exploitation of sources (B) to achieve scientific exploration (A). The main contribution of the table is the further subdivision into types of sub-exploration and sub-exploitation within the columns. From the empirical journey in 4.3.2 we encountered interviewees who placed more emphasis in exploring possibilities of partnership as means to attract funds from various sources, as opposed to governmental sponsorship chasers who aim at sustaining their individual agendas (either as persons or laboratories). The table can be perceived as a spectrum from romanticism to cynicism with researchers finding themselves, at different times on different areas of it, although, at least during anonymised interviews, admitting their identity as more curiosity-driven or industry-driven. Below the table, I will offer more specific descriptions of how these ideal types framework is to be understood in a combined form.

A1-B1 (explorers-explorers): visionary researchers who will try and attract funds from any source possible. They mostly care about the advancement of science and they are not afraid of offering grandiose promises (motivational statements) to convince funders who support blue sky research. Their approach is rather long-term and their findings, usually differing from intended output, can be of use among exploiters.

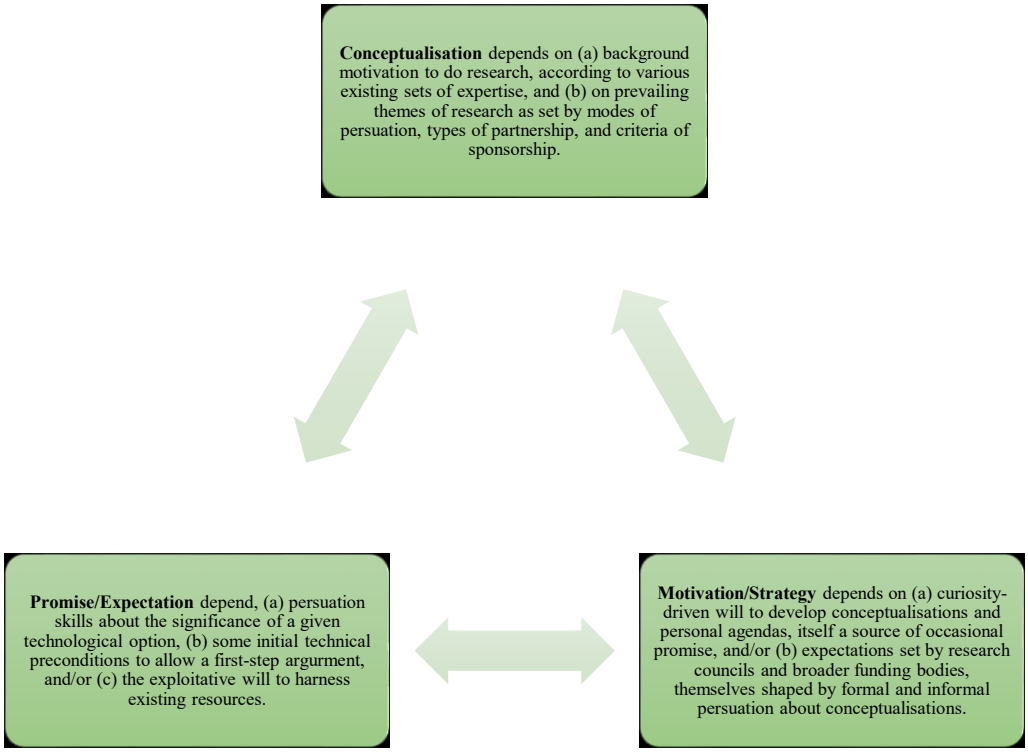
A1-B2 (explorers-exploiters): researchers who employ grandiose visions and rhetoric to attract funds, however, with limited possibilities of successful applications in the context of "impact," REF, and so on, which treats breakthrough promises with scepticism. They will eventually learn to adapt to other ideal types.

A2-B1 (exploiters-explorers): flexible researchers capable of applying their research across a range of fields, aiming at attracting various forms of funding; they are less interested in "breakthroughs" but are visionary in transferring knowledge from academic environments to industry and vice versa. Their promissory games are based on promoting the virtue of transferability.

A2-B2 (exploiters-exploiters): researchers with very specific contributions to the field, or even the generation of their own niche, aiming at influencing advisory committees about the importance of their research, setting funding trends that eventually benefit themselves as individuals or as laboratories.

AI practitioners			
A. High Epistemology Explorers (proposing new projects, interested in answering "big" questions of AI; visionaries)		**B. Low Epistemology Exploiters** (applying to existing calls, interested in AI-related applications)	
A1. Explorers (performance mode)	**A2. Exploiters** (simulation mode)	**B1. Explorers** (partnership mode)	**B2. Exploiters** (sponsorship mode)
They make use of computational models to contribute to the understanding of intelligence	They make use of existing models of mind and computation to create "intelligence"	They make use of funds for AI with an interest in contribution to the field as a whole	They make use of funds for AI to sustain individual agendas and creation of niches
E.g. whole brain simulation, computational neuroscientists, intelligence architecture	E.g. Artificial life, intelligent robotics, biorobotics, autonomous vehicles, computer vision	E.g. moderate proposal writers keen to industrial and international partnerships	E.g. good proposal writers or influencers keen to governmental sponsorships

These four "ideal types" are not clear cut – it is their interactions sustaining the taxonomy. Another simpler metaphor to capture this tension, revealed after personal communication with Robin Williams, is the division between "gentlemen" and "players" in cricket of the 19th century England: gentlemen who played for leisure (amateurs) and players who played for money (professionals), often so that the gentlemen had someone to compete with (Coleman 1973). This simple metaphor, the division between high epistemologists, but also researchers with fixed contracts and early career researchers or computer scientists who use the term AI as a label to maximise chances to be recruited, offers a solution to the question: are there explorers-explorers nowadays, and if yes, who are they and how are they funded? In my view, the older generation of AI pioneers whose grandiose visions have led to the AI winter turbulence but also in some existing applications of their work, is now replaced by the new wave of artificial intelligentsia – mainstream AI authors with sufficient experience in and undeniable contributions to the field, who have now, closer to their retirement, become public speakers of AI, promoting generic future visions. This allows further appreciation of the longitudinal continuity in AI research communities' terminological strategies: terms of AI-level explorative magnitude allow continue being exploited on the basis of high interpretative flexibility according to political and social contexts of given times. To close this chapter, I am revisiting the proposed concepts-promises-motives scheme, with added weight on motivations and how they influence, and are influenced by, conceptualisations and expectations.


Conceptualisation depends on (a) background motivation to do research, according to various existing sets of expertise, and (b) on prevailing themes of research as set by modes of persuation, types of partnership, and criteria of sponsorship.
Promise/Expectation depend, (a) persuation skills about the significance of a given technological option, (b) some initial technical preconditions to allow a first-step argurment, and/or (c) the exploitative will to harness existing resources.
Motivation/Strategy depends on (a) curiosity-driven will to develop conceptualisations and personal agendas, itself a source of occasional promise, and/or (b) expectations set by research councils and broader funding bodies, themselves shaped by formal and informal persuation about conceptualisations.

CHAPTER 7: CONCLUDING SUMMARY, DISCUSSION, AND FUTURE WORK

Figure 4 "Cours Informatique: Village Internet 2000." Picture taken by the author in Brussels, November 2021.

This thesis' introduction began with a picture I took in 2019 Athens, during my initial data analysis, depicting the leftover signage of a now closed small company selling computer products called 'Future." I took the above picture while taking a break from drafting the final version of this thesis, when I was invited to offer advice at a training course on 'AI in the Workplace' organised by the European Trade Union Institute[114]. It depicts a similar piece of signage remnant of a business that offered courses in informatics, attracting potential customers' interest to sharpen their skills for the future by its name, which translates as 'Internet Village 2000.' A possible allusion to Marshal McLuhan's 1964 popular theorisation of electronic and networked technologies resulting, deterministically, into a Global Village (McLuhan 1964)? An all-too familiar metaphor for those who grew up in and before the 1990s about the forthcoming millennial change? Be what it may, 22 years after 2000, the business's main specialty now (photocopying services for 20 eurocents) appears to be competitive, given the living costs of central Brussels. Would that

[114] https://www.etui.org/training/ai-workplace-data-and-algorithms

shop's situation be very different now if the label read 'AI courses: Smart Cities 3000'? Maybe. The purpose of including these two images is to show how the high-level (as in HLEGAI 2019) policy and technical debates about AI expectations are and have been embedded in everyday, mundane elements not only in contemporary newspapers, but as deep as in material infrastructures that now breed nostalgic visions of days of past imagined futures. I used these images of closed or repurposed shops as windows to look into and embark into a journey of mapping expectations and expertise in time and space. In the following sections, I will summarise the findings of this thesis, revisit the research questions posed in section 2.1, and based on these findings I will offer.

7.1 Summary and Research Questions Revisited

This work began with the presentation of current AI environments as they are expressed within a range of competing expectations to the non-specialist. I highlighted the theological/mythological undertones informing everyday expectations and presented instances of how these shape what is considered to be practical AI today and what kind of speculations accompany it. In reviewing STS (and social science, more broadly) research on AI, a gap was identified in that contemporary studies have not benefitted from the insights that could be derived from AI's long history and practical knowledge. Firstly, there is lack of empirical research with AI specialist views, and even if there is, such research is cut from the historical origins of AI's social, economic, and political shaping. Secondly, current AI debates, are largely shaped by sensationalist media, further influenced by age-old narratives of fascination and fear about the implications of imitating human intelligence. Thirdly, early research showed that AI's history can offer lessons about similar interplays of hype and practice, yet, with alternative arrangements, based on different socio-political contexts. However, there is something to be learned by viewing AI in its historical emergence and by investigating views from those who advance it. This led to the following research questions which I hereby repeat verbatim as outlined in section 2.1, offering short responses based on concluding remarks discussed in the chapters above, and further discussed in the remainder of this final chapter:

- What is the relationship between AI development and the oscillating promissory environment, including periods of hype and disillusionment (AI winters), in terms of AI conceptualisation, funding, from practitioners' perspective?

My main finding in addressing the question is that the very idea of a hype cycle is as reductionist as that of a linear technological trajectory. In part confirming the notion of the "AI effect," and favouring it against the notion of the "AI winter," I have enriched its sociological significance by showing the complex interplays between the needs for new terminologies during periods of loss of trust, industrial and military pressures to meet institutional goals, and the simultaneous will to mark AI's territory with potential intent to highlight and fix its limitations. Promises about, conceptualisations of, and motivations for doing AI are deeply linked and nearly impossible to disentangle.

- To what extent does informal and non-developer assessment of expectations influence formal articulations of AI communities' practices and strategic foresight? In other words, do non-specialist commentaries impact what scientists are expected to do?

Following from the above, certain influential early and recent AI community members' intent to conduct curiosity-driven research has left them vulnerable to non-technical assessment and criticism. Early

researchers focused more on their promises in order to establish their field, but stayed shortly on defining it. It was easy for critics, then, to take advantage of the unrealised promises and define AI heteronomously. Numerous examples in the thesis demonstrated the role of serendipity and informal communications in decision-making: from the misunderstanding of Minsky and Papert's criticism of the perceptron model and Feigenbaum's visit to Japan and its consecutive international AI race domino effect, to Li's role in developing ImageNet and standing between the academy-military-industrial complex and the train incident with David Willetts. This acts as an additional call to pay closer attention to informal structures behind the shaping of formal articulations of science, technology, and their associated forecasting.

- What can historical examinations of AI's conceptual and promissory settings tell about the current rebranding of AI and how can historical assessment help in better understanding of expectations as an evolving collective process of sociotechnical shaping?

An important finding was that when a hyped object as wide as AI becomes less hyped in the technological domain, it will migrate to less technical ones which will preserve the hype during a period of apparent dormancy, that is, in effect, a period of continued technological development under different names. It did not matter much whether big data generation, collection, and use for pattern recognition was performed under the scope of a Web 2.0 imaginary of perfectly annotated data or an AI revolution. As much as Google Photos performed badly as an "online service," so also did the AI technologies of the Gender Shades experiment.

The combination of two existing STS frameworks, the study of expectations and the study expertise and experience proved to be a useful tool to provide with effective equipment for studying AI's broad environment through a thorough historical overview of AI concepts and interplaying promises and interviews with AI specialists. It was impossible to cover the entirety of AI history given my focus on the English-speaking world, and it was impossible to represent all possible views of present researchers' attitudes towards AI, as I have examined but a few academic researchers out of an exponentially growing and expanding AI community which, in itself is not unified. Different subfields' members (such as object or emotion detection and recognition, intelligent robotics, or generative AI) could tell entirely different stories about their historical beginnings and their motivations for doing research. But it is my hope that I have brought a convincing argument about the persistence in time of tension between different arenas that continue shaping, in their different reincarnations, what often appears to be a neutral tool.

Historical analysis of AI enactors and selectors and the various oscillatory formations and mixings of practitioner and non-practitioner arenas showed how longitudinal analysis of expectations and promises is useful for a better understanding of the circumstances which bring large scale technoscientific projects into being. The expertise-expectations framework offers sufficient classification to see who appears as a technical expert or not at a formal level, yet allows mapping of informal elements such as off-the-record communications, personal beliefs, ideologies, or vested interests at individual or collective levels, given that such data can become available to the socio-historical investigator. Existing scholarship on longitudinal analysis of expectations and expertise/experience studies is benefitted from AI's historical assessment of promises and arena negotiations. Such a historical sociology of expectations and expertise allows the researcher to think of expectations and expertise not only as instances of the present (susceptible only to performativity, fulfilment, or credentials scrutiny), but also as products of long historical moulding, and a multiplicity of complex interactions between forgetfulness, hype migration, alliances, and discord. The legacy of such events is contained within the contemporary AI communities, either of the high or the

low epistemological registers, through a series of research and promising norms that need to be observed in order to maintain relevance within the AI community.

Historical examination enabled the better understanding of specialists' consensus in that there is no consensus about what constitutes AI: on the one hand, for multiple reasons, AI cannot be defined because it is too broad a field; on the other, for some researchers AI is reduced to a certain combination of component technologies – this can be viewed as a direct consequence of the different waves of transformation of AI from a field of science studying the artificial imitation of thought into a practical application of algorithmic predictive analytics. The vagueness of AI as a hyped concept allows its high interpretative flexibility, and, as per a recent optimistic proposal, it might offer opportunity for creativity and appreciation of technoscience as a fluid entity:

> "Perhaps concepts can be useful precisely in being ambiguous. Their flexibility allows operational definitions to be constructed pragmatically and to evolve with new knowledge. In many areas of science, only concepts that are effectively dead have fixed and immutable definitions. Ambiguous concepts can be useful in communication because they tap into a body of implicit knowledge in the receiver's mind, a body of knowledge that cannot be fixed but is in continual flux, and this fuzzy communication is a source of creativity in science." (Leng and Leng 2020: 75).

Although Leng and Leng refer to the usefulness of vagueness as a tool to allow scientific progress, based on my interviewees' suggestions, a more critical view sees it as a method to take advantage of the existing hype. Based on the previous arguments, hype may mobilise expectations, and related rebrandings, about technoscientific developments and related outcomes. Outcomes may not immediately materialise, and evidence of this in the short-term will allow hype to attach itself to another technology, possibly of very similar content. With AI, this happened several times, the same technical field changing names, being hyped and disillusioned, with the latest instantiation of hype being attached to one of the very first technical options of AI. Perceptron enactors from the 1950s-1960s had to wait until the 2010s to receive the blessings of official history (and, alas, policy's regulatory) selectors. Conceptual brandings are variables in the constancy of hype which, in turn, is a mechanism for competing promises towards allocation of funds. To quote an interviewee, when I asked him whether we live in a moment of hype now: "We always live in the world of hype" (Shengin Eering).

AI promising is deeply associated with AI's convenient convenient terminological flexibility, which can be used as a buzzword by experts within multiple areas of computing in order to precisely sustain their expertise and generate more promises, in a feedback loop. History and interviews showed that, in addition, AI researchers oscillate between enthusiastic "motivational statements" and will to exaggerate results in order to establish a field or secure funds in an exploitative fashion. This often happens in response to public and political discourse which shapes the demands and expectations, and might even result in such claims' assessment, leading to "AI winters" of funding cuts. Would that be an unjust assessment, given the continuity of research within rebranded versions of AI (presented as novel niches within computer science) and the occasional successes after many years in hibernation? Perhaps success is a virtue coalescent with patience. Hinton and LeCun's resurfacing in the 2010s after the 1980s was an extremely revealing case for this, and while the success of deep convolutional networks might have revived the theological faith in AI, this might be hindered due to relatively unforeseen problems relating to data protection or context processing, as flagged by some practitioners.

Moving to empirical insights on funding strategies and AI governance, I come to suggest that the current AI paradigm, a strong ML-oriented, statistical reasoning approach to extracting patterns has its roots in two significant points in time – the Mansfield's Amendment (1969) and the Lighthill report (1973) which demanded an applications-orientation of science, that is, science in the service of practical technology. This is equally reflected in the ways scientists are applying for funding but also in the way governments seek to regulate research. Contemporary researchers generally tend to accept this practical orientation as it allows them to enter and play in the funding game, commissioned by governments, industries, and the military. This acceptance, turning researchers more into "exploiters" of existing resources, rather than romantic "explorers" of scientific frontiers shapes their promising strategies which, by default then, mutually shapes AI's conceptualisation back, as a practical application. This has become now crystallised in official, policy-level, descriptions of AI that define it as per its applications. Most researchers have become less active in shaping AI's concept, in contrast to the AI communities from three decades ago. Instead, they are eager to comply and shape AI policies revolving around responsible pursuit of their field, however, as the field is determined by the policy understanding. This, I would argue has very little relation, for example, to the Computer Professionals for Social Responsibility initiative from the 1980s (Suchman 1984) where AI specialists actively voiced their genuine community concerns against undesired uses of their research funded by governments and other parties. Funding had always been a core incentive and mechanism to pursue AI research. However, the increasing demand for practical applications paired to additional funding opportunities for researchers who collaborate with industry creates a distinction between (a) a majority of low epistemology, downstream, research communities, usually with little or no awareness of AI's long trajectory, and (b) a high epistemology minority of either romantic researchers with narrow funding opportunities or prestigious researchers whose double position as both specialists and public influencers allows them to make more theoretical statements about AI's nature, scope, and future, however, often in tentative alliances with big players from private industries.

Throughout its history, negotiations of its capabilities and purpose, led to the frequent rebranding of various AI approaches or branches as new fields dissociated with AI (the "AI effect"). The "AI winter," then, is an opportunity for the STS scholar to look at the ways in which fluidity of path investment operates. Researchers within the domains currently recognised as AI are eager to appear proactive in defending their field and involve themselves in discussions about social implications of AI[115]. Overpromising seems like a necessary evil within this environment, even for those who are not in favour of such a strategy. Promising can take place within research grant applications which might often promise more than what can be achieved, however, within the confinements of practical applications. Contemporary lack of ad hoc assessments is paired to serendipitous acts of convincing (as in the *Eight Great Technologies* illustrative example) and a general defence of opportunism as an approach to doing science. Generalised alienation between research communities, policymakers, end-users (the public), and other actors results into the preservation of a very vague perception of AI which, in turn, allows various

[115] However, one should not be surprised if the recent ensemble of virtual reality investment and promotion through Facebook and Microsoft's "metaverse" strategies paired to the popular uprising of the novel Matrix sequel in the post-COVID-19 era of habitualisation of the virtual (Park and Kim 2022; Taiwo 2022) results in a new round of hype. A thesis similar to the present one could be imagined, where the early fascination with digital worlds and virtual reality are being assessed as troughs of disillusionment (from the 1990s fascination to the Google Glass market failure; Kudina and Verbeek 2019) as part of another waiting game (Budde and Bakker 2012) for the right sociotechnical conditions to emerge and enable a new VR hype.

dynamics of individual vested interests to manoeuvre within collective expectations, national strategies, and public imaginaries.

7.3 Limitations Revisited and Future Work

I have mentioned in section 2.3.1 a number of limitations concerning the present thesis. Briefly, these include regional and sectoral concerns. While I had excellent local access to the researcher community I studied, this is always a partial sample of the broader practitioner communities of non-Anglophone and non-academic backgrounds, although the vast majority have worked with industrial and international partnerships. The broader arena analysis was conducted by inferring their significance based on my historical and media assessment of published works, as well as based on views that practitioners shared. It went beyond the viable scoping of the present thesis to engage empirically with wider expectational dynamics, although I admit that this could add further nuances to the understanding of expectation-expertise structures. This work hopes to be part of a broader network of emerging critical, historical and empirical social science investigations of AI. As shown in the introductory chapter, STS and other neighbouring branches of social science, have already begun looking at such sociological mechanisms of AI, however, this subfield is in its infancy. The present PhD is part of the first generation of empirical doctoral research during this new round of AI hype, opening the path to investigations of more specific sociotechnical configurations, some of which have been covered in bodies of STS literature cited above, or are currently in the making[118]. Moreover, it is my hope that the expertise-expectations arena approach will be useful to research, as much as the tripartite analysis of conceptualisations-expectations-motivations. As mentioned in few instances throughout this thesis, an initial version of the text referred to a quadripartite analysis that also included policy/regulation as part of the multifactor shaping process. While space limitations did not allow for the incorporation of this dimension at length here, I contemplate near-future expansion on this. Motivational variations such as exploration and exploitation are not only inspired or constrained by funding schemes but also dependent on what is encouraged, allowed, or prohibited by regulatory bodies – a topic which becomes increasingly significant in the field of AI, and the AI researchers' experience as dialectic outcome of their expectations and expertise should be examined in this light. More questions follow this: if vested interests guide AI research communities, should expertise strategies be reconsidered in the cases of scientific advisory boards? And if such scathing images of informal influence and relatively random allocation of grants are real, then what should be the responsible anticipatory strategies when it comes to AI technologies? At a more basic level: do AI researchers, swamped in their funding competitions, care about AI governance?

Lastly, this thesis examined AI as a broad sociotechnical construct with its vagueness and breadth being constitutive of its complex trajectory. An additional future direction of potential applicability of the historical sociology of expectations and expertise approach and the conceptualisation-expectation-motivation(-regulation) framework, is the examination of more specific applications of AI. For example, two potential areas are medical uses of machine learning AI or the use of generative AI techniques for art and entertainment purposes. Both of these domains carry long histories of interaction with computer-based

[118] Through networking on various occasions, I have met scholars whose doctoral dissertations, covering such topics, are expected to be submitted roughly at the same period as mine. Notably, the works of SJ Bennett and Chris Baird from the University of Edinburgh ought to shed light on corners I have not touched presently.

techniques and contemporary studies of them would benefit from historical lessons from previous rounds of hype, stories of waiting games, and current reporting on motivations for enacting upon them.

BIBLIOGRAPHY

Aaronson, S. (2018). Quantum computing for policymakers and philosopher-novelists. [blog post] *Shetl-optimized*. 06 June 2018. Retrieved 18-07-2018 from: https://www.scottaaronson.com/blog/?p=3848

Abrishami, P., Boer, A., & Horstman, K. (2014). Understanding the adoption dynamics of medical innovations: affordances of the da Vinci robot in the Netherlands. *Social science & medicine*, *117*, 125-133.

Andreski, S. (1972). *Social Sciences as Sorcery*. London: Andre Deutcsh.

Adams, D. (1979). *The Hitchhiker's Guide to the Galaxy*. London: Pan Books.

Adorno, T. (1951[2020]). *Minima Moralia*. Trans. E.F.N. Jephcott. London, New York: Verso.

Agar, J. (2020). What is science for? The Lighthill report on artificial intelligence reinterpreted. I(3): 289-310.

Artificial Intelligence Channel, The (2017). Artificial Intelligence Debate - Yann LeCun vs. Gary Marcus - Does AI Need More Innate Machinery?. October 5, 2017. Online video, posted October 20, 2017. Retrieved 15-11-2022 from: https://www.youtube.com/watch?v=aCCotxqxFsk

AI Community (2022). Creation of a Taxonomy for the European AI Ecosystem: A report of the Cross-KIC Activity 'Innovation Impact Artificial Intelligence'. European Institute of Innovation and Technology. Report. Retrieved 13-11-2022 from: https://eit.europa.eu/sites/default/files/creation_of_a_taxonomy_for_the_european_ai_ecosystem_final.pdf

AI Effect. (2019). In: *Wikipedia*. Accessed 17-12-2019 from: https://en.wikipedia.org/wiki/AI_effect

AI Next Campaign (2018). [Online page] Defense Advanced Research Projects Agency (DARPA). Accessed 21-09-2019 from https://www.darpa.mil/work-with-us/ai-next-campaign

Al-Sibai, N. (2022). MIT Researcher: Don't Ignore the Possibility That AI Is Becoming Conscious. *Futurism*. 16 February 2022. Retrieved 17-02-2022 from https://wordpress.futurism.com/mit-researcher-conscious-ai

Alciné, J. (2015). Google Photos, y'all fucked up. My friend's not a gorilla. *Twitter Thread*, June, 28. https://twitter.com/jackyalcine/status/615329515909156865/

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives, 31*(2), 211-36.

Algorithm Watch (2021). The AI Ethics Guidelines Global Inventory. https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/.

Allen, A. (2019). *Imagining intelligent artefacts: Myths and a digital sublime regarding artificial intelligence in Swedish newspaper Svenska Dagbladet*. MA Thesis. Department of Media Studies. Stockholms Universitet. Accessed 01-05-2020 from https://su.diva-portal.org/smash/get/diva2:1349751/FULLTEXT01.pdf

Allen, G. C. (2019). *Understanding China's AI Strategy: Clues to Chinese Strategic Thinking on Artificial Intelligence and National Security*. Washington, DC: Center for a New American Security. Accessed 18-03-2020 from https://nsiteam.com/social/wp-content/uploads/2019/05/CNAS-Understanding-Chinas-AI-Strategy-Gregory-C.-Allen-FINAL-2.15.19.pdf

Allyn, B. (2021, October 5). Here are 4 key points from the Facebook whistleblower's testimony on Capitol Hill. NPR. https://www.npr.org/2021/10/05/1043377310/facebook-whistleblower-franceshaugen-congress

Alvesalo-Kuusi, A., & Whyte, D. (2018). Researching the Powerful: A Call for the Reconstruction of Research Ethics. *Sociological Research Online, 23*(1), 136-152.

Åm, H. (2019). Ethics as ritual: Smoothing over moments of dislocation in biomedicine. *Sociology of Health & Illness, 41*(3), 455-469.

Anderson, J. (2018). *The Future of the World: Futurology, Futurists, and the Struggle for the Post–Cold War Imagination*. New York: Oxford University Press.

Annoni, A., Cesar, R.M. Anzai, Y., Hall, W., Hardman, L., Van Harmelen, F., Heintz, F., Motta, E., De Heaver, M., Ten Holter, C., Keene, P., Ott, I., Perrault, R., De Prato, G., Sun, Z., Tan, T., Tang, C., Zhao, Z. (2008). *ArtificiaI Intelligence: How Knowledge Is Created, Transferred, and Used*. Executive Summary. Elsevier AI Resource Center. Accessed 17-01-2019 from https://www.elsevier.com/?a=827872

Araujo, L., Mason, K., & Spring, M. (2014, September). Expectations in networks: market shaping devices of the driverless car. In *30th IMP Conference, Kedge Business School*. Bordeaux, France. 4th-6th September 2004.

Arras, K. & Cerqui, D. (2005). Do we want to share our lives and bodies with robots? A 2000-people survey. Tech Rep 0605–001, *Autonomous Systems Lab (ASL)*. Swiss Federal Institute of Technology Lausanne (EPFL).

Ashby, W. R. (1954). *Design for a Brain*. New York: John Wiley and Sons Inc.

Aylett, R. & Vargas, P. (2021). *Living with Robots: What Every Anxious Human Needs to Know*. Cambridge, Massachusetts, London, England: The MIT Press.

Bakker, S. & Budde, B. (2012). Technological hype and disappointment: lessons from the hydrogen and fuel cell case. *Technology Analysis & Strategic Management 24*(6): 549-563.

Bakker, S., van Lente, H., & Meeus, M. T. H. (2011). Arenas of expectations for hydrogen technologies. *Technological Forecasting and Social Change, 78*(1), 152-162.
Ballard, J. G. (1975 [2006]). *High-Rise*. London, New York, Toronto and Sydney: Harper Perennial.
Barbrook, R. & Cameron, A. (1996) The California ideology. *Science as Culture. 6*((26), 44–72.
Bartneck, C. (2004, April). From Fiction to Science–A cultural reflection of social robots. In *Proceedings of the CHI2004 Workshop on Shaping Human-Robot Interaction* (pp. 1-4).
Bareis, J., & Katzenbach, C. (2021). Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics. *Science, Technology, & Human Values*. [Online ahead of print].
Bartneck, C., & Forlizzi, J. (2004, September). A design-centred framework for social human-robot interaction. In *Robot and Human Interactive Communication, 2004. ROMAN 2004. 13th IEEE International Workshop on* (pp. 591-594). IEEE.
BBC News. (2002, December). Fire Destroy's Librarian's Work. *BBC News*. 09 December 2002. Retrieved 16-07-2020 from: http://news.bbc.co.uk/1/hi/scotland/2558655.stm
BBC TV. (1973, June). The General Purpose Robot is a Mirage.(with Professor Sir James Lighthill, Professor Donald Michie, Professor Richard Gregory and Professor John McCarthy. *Controversy*. Royal Institution. Accessed 06-11-2019 from: http://www.aiai.ed.ac.uk/events/lighthill1973/
Barnes, B., Bloor, D., Henry, J. (1996). *Scientific Knowledge: A Sociological Analysis*. Chicago: The University of Chicago Press.
Beckett, S. (1955, 1956, 1958). *Three Novels: Molloy, Malone Dies, The Unnamable*. New York: Grove Press.
Bedard, S., Tack, D., Pageau, G., Ricard, B., & Rittenhouse, M. (2011). *Initial Evaluation of the Dermoskeleton Concept: Application of Biomechatronics and Artificial Intelligence to Address the Soldiers Overload Challenge*. Defence Research and Development Canada Valcartier (QUEBEC).
Bengio, Y. (1993). A connectionist approach to speech recognition. In *Advances in Pattern Recognition Systems Using Neural Network Technologies* (pp. 3-23).
Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Cambridge: Polity.
Berghahn, K. L. & Grimm, R. eds. (1990). *Utopian Vision, Technological Innovation, and Poetic Imagination*. Heidelberg: Winter.
Bell, A. J. (1999). Levels and loops: the future of artificial intelligence and neuroscience. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 354*(1392), 2013-2020.
Bellet, C. (2019). The Future of Animal Health: How Digital Technologies Reconfigure Animal Healthcare in Farming. *Discover Society 7*.
Berger, J., Sorensern, A. T., & Rasmussen, S. J. (2010). Positive effects of negative publicity: When negative reviews increase sales. *Marketing Science 29*(5), 815-827.
Berger-Tal, O., Nathan, J., Meron, E., & Saltz, D. (2014). The Exploration-Exploitation Dilemma: A Multidisciplinary Framework. *PLoS ONE, 9*(4), e95693.
Bernal, J. D. (1929 [2017]). *The World, the Flesh and the Devil: An Enquiry into the Future of the Three Enemies of the Rational Soul*. London and Brooklyn: Verso Books.
Bernal, J. D. (1954 [1971]). *Science in History*. Vol 1. Cambridge and Massachusetts: MIT Press.
Bietti, E. (2020). From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 210-219
Bijker, W. E., Hughes, T. P., & Pinch, T. (1987). *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. Cambridge, Massachusetts, London, England: The MIT Press.
Bird, J. (1991). Britain picks wrong way to beat the Japanese: an analysis of Britain's Alvey program shows that support for precompetitive research does not equal economic success. *Science, 252*(5010), 1248-1249.
Bischoff, R., Guhl, T., Wendel, A., Khatami, F., Bruyninckx, H., Siciliano, B., ... & Ibarbia, J. A. (2010, June). euRobotics-Shaping the future of European robotics. In *Robotics (ISR), 2010 41st International Symposium on and 2010 6th German Conference on Robotics (ROBOTIK)* (pp. 1-8). VDE.
Boden, M. (1987). *Artificial Intelligence and Natural Man*. (2nd edition, expanded). London: The MIT Press.
Boden, M. (1989). *Artificial Intelligence in Psychology: Interdisciplinary Essays*. Cambridge, Massachusetts, and London: The MIT Press.
Boden, M. A. (2016). *Artificial Intelligence: Its Nature and Future*. Oxford: Oxford University press.
Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., ... & Sorrell, T. (2017). Principles of robotics: regulating robots in the real world. *Connection Science*, *29*(2), 124-129.
Bogost, I. (2015). The Cathedral of Computation. *The Atlantic*, January 15, 2015. https://www.theatlantic.com/technology/archive/2015/01/the-cathedral-of-computation/384300/
Booch, G. (2015). I, for One, Welcome Our New Computer Overlords. *IEEE Software*, *32*(6), 8-10.
Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.
Bowen, G. A. (2009). Document Analysis as a Qualitative Research Method. *Qualitative Research Journal*, *9*(2), 27-40.

Brady, M., Gerhardt, L., & Davidson, H. F. (Eds.). (2012). *Robotics and Artificial Intelligence* (Vol. 11). Heidelberg: Springer Science & Business Media.
Brennen JS, Howard, PN, and Nielsen, RK (2020) What to expect when you're expecting robots: Futures, expectations, and pseudo-artificial general intelligence in UK news. *Journalism*. Epub ahead of print, 05 August 2020. DOI: 10.1177/1464884920947535.
Brewster, T. (2021). Project Maven: Startups Backed By Google, Peter Thiel, Eric Schmidt And James Murdoch Are Building AI And Facial Recognition Surveillance Tools For The Pentagon. *Forbes*. September, 8. Retreved 17-11-2022 from: https://www.forbes.com/sites/thomasbrewster/2021/09/08/project-maven-startups-backed-by-google-peter-thiel-eric-schmidt-and-james-murdoch-build-ai-and-facial-recognition-surveillance-for-the-defense-department/
Brinkmann, S. & Kvale, S. (2015). *InterViews: Learning the Craft of Qualitative Research Interviewing*. Los Angeles, London, New Delhi, Singapore, Washington DC: Sage.
Brooker, P., Dutton, W., & Mair, M. (2019). The new ghosts in the machine: 'Pragmatist' AI and the conceptual perils of anthropomorphic description. *Ethnographic Studies, 16*, 272-298.
Brooks, R. (2002). *Robot: The Future of Flesh and Machines*. London: Penguin.
Brown, N., & Michael, M. (2003). A Sociology of Expectations: Retrospecting Prospects and Prospecting Retrospects. *Technology Analysis & Strategic Management*, *15*(1), 3-18.
Brown, N., & Michael, M. (2004). Risky creatures: institutional species boundary change in biotechnology regulation. *Health, Risk & Society, 6*(3), 207-222.
Bruckenberger, U., Weiss, A., Mirnig, N., Strasser, E., Stadler, S., & Tscheligi, M. (2013, October). The good, the bad, the weird: Audience evaluation of a "Real" robot in relation to science fiction and mass media. In *International Conference on Social Robotics* (pp. 301-310). Cham, Heidelberg, New York, Dordrecht, London: Springer.
Bryman, A. (2008). *Social Research Methods*. Oxford: Oxford University Press.
Bryson, J. J. (2010). Robots should be slaves. *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues, 8*, 63-74.
Bryson, J., & Winfield, A. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. *Computer, 50*(5), 116-119.
Bryson, J. (2018). The Moral, Legal, and Economic Hazard of Anthropomorphizing Robots and AI. In: Coeckelbergh, M., Loh, J., & Funk, M. (Eds.). (2018). *Envisioning Robots in Society–Power, Politics, and Public Space: Proceedings of Robophilosophy 2018/TRANSOR 2018 (Vol. 311)*. Amsterdam, Berlin, Washington: IOS Press.
Budde, B., & Konrad, K. (2019). Tentative governing of fuel cell innovation in a dynamic network of expectations. *Research Policy, 48*(5), 1098-1112.
Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlstrom, P., Henke, N., Trench, M. (2017 June). *Artificial Intelligence: The Next Digital Frontier?* McKinsey Global Institute. [Online]. Retrieved 05-11-2017 from https://www.mckinsey.com/~/media/mckinsey/industries/advanced%20electronics/our%20insights/how%20artificial%20intelligence%20can%20deliver%20real%20value%20to%20companies/mgi-artificial-intelligence-discussion-paper.ashx
Bultitude, K., Grant, L., Burnet, F. and Johnson, B. (2007) Robot Thought: final evaluation report. Project Report. University of the West of England.
Buolamwini, J. A. (2017). *Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers* (Master of Arts dissertation, Massachusetts Institute of Technology).
Callon, M. (2007). What Does it Mean to Say that Economics is Performative? In: MacKenzie, D. Muniesa, F. & Siu, L. (eds). *On the Performativity of Economics: Do Economists Make Markets*. Princeton: Princeton University Press, 311-57.
Calvert, J. (2006). What's special about basic research?. *Science, Technology, & Human Values, 31*(2), 199-220.
Campbell, M., Hoane, A. J., & Hsu, F. H., (2002). Deep Blue. *Artificial Intelligence*, *134*(1), pp. 57-83.
Campion, A., Gasco-Hernandez, M., Jankin Mikhaylov, S., and Esteve, M. (2020). Overcoming the Challenges of Collaboratively Adopting Artificial Intelligence in the Public Sector. *Social Science Computer Review*. Epub ahead of print, 20 December 2020. DOI: 10.1177/0894439320979953.
Campolo, A. and Crawford, K. (2020). Enchanted determinism: Power without responsibility in artificial intelligence. *Engaging Science, Technology, and Society, 6*, 1-19.
Campolo, A., Sanfilippo, M. R., Whittaker, M., & Crawford, K. (2017). AI Now 2017 report. *The AI Now Institute*. Retrieved 15-11-2022 from: https://ainowinstitute.org/AI_Now_2017_Report.pdf
Charles, J. (1995). US Commerce Department assesses the US artificial intelligence sector. *IEEE Expert 10*(1), 70-72.

Chollet, F. (2019). On the measure of intelligence. arXiv preprint arXiv:1911.01547. Epub ahead of print, 5 November 2019. Retrieved 30-03-2021 from: https://arxiv.org/pdf/1911.01547.pdf
Calvert, J. (2010). Synthetic biology: constructing nature?. *The sociological review, 58*(s1), 95-112.
Cantwell Smith, B. (2019). *The Promise of Artificial Intelligence: Reckoning and Judgement*. Cambridge, Massachusetts, London: The MIT Press.
Capurro, R., Hausmanninger, T., Weber, K., Weil, F., Cerqui, D., Weber, J., Apel, M. (2006). Ethics in Robotics. *International Review of Information Ethics, 6*(12/2006). Retrieved 05-02-2015 from http://www.i-r-i-e.net/inhalt/006/006_full.pdf
Capurro, R., Nagenborg, M. (2009). Ethics and Robotics. Amsterdam: IOS Press.
Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2017). Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. *Science and Engineering Ethics*, 1-24.
Cellan-Jones, R. (2014). Stephen Hawking Warns Artificial Intelligence Could End Mankind. *BBC News* [online journal]. Retrieved 12-05-2015 from http://www.bbc.com/news/technology-30290540
Chalmers, D. (2010). The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies*, *17*(9-1), 7-65.
Cheon, E., & Su, N. M. (2017, February). Configuring the User: "Robots have Needs Too". In: *CSCW '17*, February 25-March 01, 2917, Portland, Oregon, USA, 191-206.
Churchill, W. (1943). House of Commons Rebuilding. *Commons Sitting HC Deb, Hansard 1803–2005*, *393*, 403-73.
Clark, J. (2016). Artificial intelligence has a 'sea of dudes' problem. *Bloomberg Terminal*. June 16, 2016. Retrieved 14/11/2022 from: https://www.bloomberg.com/professional/blog/artificial-intelligence-sea-dudes-problem/
Coleman, D. C. (1973). Gentlemen and Players. *The Economic History Review, 26*(1), 92-98.
Collingridge, D. (1980). *The Social Control of Technology*. London: Frances Pinter (Publishers) Limited.
Collingridge, D. & Reeve, C. (1986). *Science Speaks to Power: The Role of Experts in Policy Making*. New York: St. Martin's Press.
Collins, H. M. (1987). Expert Systems and the Science of Knowledge. In: Bijker, W. E., Hughes, T. P., & Pinch, T. (1987). *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. Cambridge, Massachusetts, London, England: The MIT Press.
Collins, H. M. (1990). Artificial Experts: Social Knowledge and Intelligent Machines. Massachusetts: MIT Press.
Collins, H. (2018). *Artifictional Intelligence: Against Humanity's Surrender to Computers*. Medford: Polity.
Collins, H. M., and Evans, R. (2002). The Third Wave of Science Studies: Studies of Expertise and Experience. *Social Studies of Science, 32*(2), 235-296.
Collins, H., & Evans, R. (2008). Rethinking expertise. Chicago: University of Chicago Press.
Collins, H., Evans, R., Durant, D., & Weinel, M. (2020). *Experts and the Will of the People: Society, Populism, and Science*. Cham: Palgrave Macmillan.
Collins, H. M. & Pinch. T. (1998). *The Golem at Large: What You Should Know about Technology*. Cambridge: Cambridge University Press.
Conger, K. & Metz, C. (2018). Tech Workers Now Want to Know: What Are We Building This For? *New York Times (Online)*, New York: New York Times Company. October 7. Retrieved 17-11-2022 from: https://www.nytimes.com/2018/10/07/technology/tech-workers-ask-censorship-surveillance.html
Contrada, N. (1995). Golem and Robot: A Search for Connections. *Journal of the Fantastic in the Arts*, *7*(2/3 (26/27), 244-254.
Corke, P. I. (2010). A Chat with Matt Mason [Turning Point]. *IEEE Robotics & Automation Magazine*, *17*(1), 136-132.
Costandi, M. (2015). Fragment of Rat Brain Simulated in Supercomputer: Blue Brain Project Announces Results of a Decade's Work. *Nature: International Weekly Journal of Science,* (October 08, 2015). Retrieved 25-10-2015 from http://www.nature.com/news/fragment-of-rat-brain-simulated-in-supercomputer-1.18536
Crawford, K. (2021). *The Atlas of AI*. New Haven and London: Yale University Press.
Crawford, K., Whittaker, M., Elish, M. C., Barocas, S., Plasek, A., & Ferryman, K. (2016). The AI now report. The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term. *The AI Now Institute*. Retrieved 15-11-2022 from: https://artificialintelligencenow.com/media/documents/AINowSummaryReport_3_RpmwKHu.pdf
Creighton, J. (2018, February 14). The "Father of Artificial Intelligence" Says Singularity Is 30 Years Away: All evidence points to the fact that the singularity is coming (regardless of which futurist you believe). *Futurism* [online magazine]. Accessed 22-02-2019 from: https://futurism.com/father-artificial-intelligence-singularity-decades-away

Crevier, D. (1993) *AI: The Tumultuous History of the Search for Artificial Intelligence*. New York: Basic Books.
Dandurand, G., Claveau, F., Dubé, J.F., and Millerand, F. (2020). Social Dynamics of Expectations and Expertise: AI in Digital Humanitarian Innovation. *Engaging Science, Technology, and Society, 6*, 591-614.
DARPA (2021). AI next campaign. Defense Advanced Research Projects Agency. Retrieved 17-05-2020. Available at: https://www.darpa.mil/work-with-us/ai-next-campaign (accessed 12-01-2021).
Debord, G. (1988[1998]). *Comments on the Society of the Spectacle*. Trans. Malcolm Imrie. London, New York: Verso.
Deleuze, G. (1986[2013]). *Foucault*. London, New Delhi, New York, Sydney: Bloomsbury Academic.
Deleuze, G. & Guattari, F. (1988[2012]. *A Thousand Plateaus: Capitalism and Schizophrenia*. London, New Delhi, New York, Sydney: Bloomsbury Academic.
De Landa, M. (1991). *War in the Age of Intelligent Machines*. Cambridge, Massachusetts, and London, England: The MIT Press.
de Luca, A. & Beltran, G. (2019). The Decade Tech Lost its Way. *The New York Times*. August 15, 2019. Retrieved 15-11-2022 from: https://www.nytimes.com/interactive/2019/12/15/technology/decade-in-tech.html
de Sola Pool, I., & Abelson, R. (1961). The Simulmatics Project. *Public Opinion Quarterly, 25*(2), 167-183.
De Waard, M., Inja, M., & Visser, A. (2013, April). Analysis of Flat Terrain for the Atlas Robot. In: *3rd Joint Conference of AI Robotics and 5th RoboCup Iran Open International Symposium (RIOS)* (pp. 1-6).
Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Li, F. (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255).
Dexe, J. and Franke, U. (2020). Nordic lights? National AI policies for doing well by doing good. *Journal of Cyber Policy*, Epub ahead of print, 09 December 2020. DOI: 10 10.1080/23738871.2020.1856160.
Dias, W.P.S. (2002). Reflective practice, artificial intelligence, and engineering design: Common trends and interrelationships. Artificial Intelligence for Engineering Design, Analysis and Manufacturing. *AI EDAM, 16*(4): 261-271.
Dignum, V. (2019) *Responsible artificial intelligence. How to develop and use AI in a responsible way*. Cham: Springer
Directorate General XIII. Telecommunications, Information Industries and Innovation Commission of the European Communities. (1989). *ESPRIT. European Strategic Programme for Research and Development in Information Technology. The project synopses. Index of projects and programme overview*. Volume 1 of a series of 8. September. EU Commission - Working Document.
Doctorow, C. (2001). *Metacrap: Putting the torch to seven straw-men of the meta-utopia*. Personal blog. Retrieved 12-12-2015 from: https://chnm.gmu.edu/digitalhistory/links/pdf/preserving/8_17.pdf
Dreyfus, H. L. (1965). *Alchemy and Artificial Intelligence*. Papers Series. P-3244. Santa Monica: RAND Corporation.
Dreyfus, H. L. (2006). Overcoming the Myth of the Mental. *Topoi 25*. 43-49.
Dreyfus, H. L. (2012). A History of First Step Fallacies. *Minds and Machines, 22*(2), 87-99.
Dreyfus, H. L., & Dreyfus, E. S. (1995). Making a mind vs. modeling the brain: AI back to a branchpoint. *Informatica, 19*(4), 425-441.
Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... & Galanos, V. (2019). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management 57*, 101994, 1-47.
Eden, A. H., Steinhart, E., Pearce, D., and Moor, J. H.. (2012). Singularity Hypotheses: An Overview. In: Eden, A. H., Steinhart, E., Pearce, D., and Moor, J. H.. (eds.). *Singularity Hypotheses* (pp. 1-12). Springer Berlin Heidelberg, 2012.
Edwards, P. N. (1996). *The Closed World: Computers and the Politics of Discourse in Cold War America*. Cambridge, Massachusetts, London: MIT Press.
Edwards, P. N. (2019). Infrastructuration: On habits, norms and routines as elements of infrastructure. In: Kornberger, M. et al (eds.). *Thinking Infrastructures*. Emerald Publishing Limited.
Edwards, L., Schäfer, B., & Harbinja, E. (Eds.). (2020). *Future Law: Emerging Technology, Regulation and Ethics*. Edinburgh: Edinburgh University Press.
Eisenhower, D. (1961). President Dwight Eisenhower Farewell Address. *C-Span*. January 17, 1961. Retrieved 02-12-2021 from https://www.c-span.org/video/?15026-1/president-dwight-eisenhower-farewell-address
Eliot, T.S. (1936). *Murder in the Cathedral*. New York: Harourt, Brace & Company.
Else, H. (2018). AI conference widely known as 'NIPS' changes its controversial acronym. Nature News. 19 November 2018. Retrieved 13-09-2021 from https://www.nature.com/articles/d41586-018-07476-w
Emmott, S. J. (1995, ed). *Information Superhighways: Multimedia Users and Futures*. London: Academic Press.

Engelberger, J. F. (2012). *Robotics in Practice: Management and Applications of Industrial Robots*. Springer Science & Business Media.
Epstein, S. (1995). The construction of lay expertise: AIDS activism and the forging of credibility in the reform of clinical trials. *Science, Technology, & Human Values, 20*(4), 408-437.
European Commission (2021, April 21). *Proposal for a 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*. SEC(2021) 167 final, SWD(2021) 84 final, SWD(2021) 85 final.
European Parliament (2017, February 16). Robots and Artificial Intelligence: MEPs call for EU-wide Liability Rules. [Plenary session. Press release]. *European Parliament News*. Retrieved 17-02-2017 from http://www.europarl.europa.eu/news/en/news-room/20170210IPR61808/robots-and-artificial-intelligence-meps-call-for-eu-wide-liability-rules
European Parliament. Committee on Legal Affairs (2016, May 5) Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)). Retrieved 12-06-2017 from http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML%2BCOMPARL%2BPE-582.443%2B01%2BDOC%2BPDF%2BV0//EN
European Parliament. Directorate-General for Internal Policies. Policy Department C. Citizens' Rights and Constitutional Affairs (2016, October). European Civil Law Rules for Robotics: Study for the Juri Committee. Retrieved 10-12-2016 from http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf
European Parliament (2017, February 16). Robots and Artificial Intelligence: MEPs call for EU-wide Liability Rules. [Plenary session. Press release]. *European Parliament News*. Retrieved 17-02-2017 from http://www.europarl.europa.eu/news/en/news-room/20170210IPR61808/robots-and-artificial-intelligence-meps-call-for-eu-wide-liability-rules
Fahlman, S. (1981). Computing facilities for AI: a survey of present and near-future options. *AI Magazine, 2*(1), 16-23.
Fang, L. (2018). Leaked Emails Show Google Expected Lucrative Military Drone AI Work To Grow Exponentially: Google reportedly played down the size of the contract in discussions with uneasy employees, but leaked emails show the company expected revenue to grow rapidly. *The Intercept*. June, 1. Retrieved 17-11-2022 from: https://theintercept.com/2018/05/31/google-leaked-emails-drone-ai-pentagon-lucrative/
Feigenbaum, E. A. (1992). A Personal View of Expert Systems: Looking Back and Looking Ahead. *Expert Systems with Applications, 5*(3-4), 193-20.
Feigenbaum, E. A. and McCorduck, P. (1983). *The Fifth Generation: Artificial Intelligence and Japan's Computer Challenge to the World*. London and Sydney: Pan Books.
Feynman, R. P. (1974). Cargo Cult Science. *Engineering and Science*, *37*(7), 10-13.
Fanon, F. (1952[1986]). *Black Skin, White Masks*. London: Pluto.
Fijalkow, J. (2013). Neoliberal and neoconservative literacy education policies in contemporary France. In: Goodman, K. S., Calfee, R. C., & Goodman, Y. M. (Eds.). *Whose Knowledge Counts in Government Literacy Policies?: Why Expertise Matters*. New York: Routledge, pp. 47-66.
Fleming, N. (2009). Interview: Robotic futures. *New Scientist*, *203*(2723), 28-29.
Fleck, J. (1979, October). Artificial Intelligence: A Case Study in Scientific Development. *AISB Quarterly, (Newsletter of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour), 35*, 3-6.
Fleck, J (1982). Development and Establishment in Artificial Intelligence. In Elias, N., Martins, H., & Whitley, R. (eds) *Scientific Establishments and Hierarchies*. Dordrecht: D. Reidel, 169–217.
Fleck, J. (1984). Artificial Intelligence and Industrial Robots: An Automatic End for Utopian Thought? In Mendelsohn, E., and Nowotny, H., (eds) *Science Between Utopia and Dystopia: Sociology of the Sciences*, Vol VIII, 1984, 189-231.
Fleck, J. (1988). Innofusion or Diffusation? The Nature of Technological Development in Robotics. Edinburgh PICT Working Paper No. 4.
Fleck, J. (1992). The Effective Management of Available Expertise in Artificial Intelligence. Final Report to the ESRC. Feb. 1992. Available as Edinburgh University Department of Business Studies Working Paper Series No. 92/6.
Fleck, J. (1993). The generation of Knowledge in Artificial Intelligence. In: *Proceedings of the European Conference on Computer Science, Communications and Society: A Technical and Cultural Challenge*. Neuchatel, Switzerland, 241-250.
Fleck, J. (1994). Knowing Engineers: Response to Diane Forsythe's "Engineering Knowledge: the construction of knowledge in Artificial Intelligence", *Social Studies of Science, 24*(1), 105-113.
Flichy, P. (2007). *The Internet* Imaginaire. Cambridge: The MIT Press.
Floridi, L. (2015). Singularitarians, AItheists, and Why the Problem with Artificial Intelligence is H.A.L.

(Humanity At Large), not HAL. In: Sullins, J. (ed.). *Philosophy and Computers, 14(2)*, 8-11.
Floridi, L. (2020). AI and its New Winter: From Myths to Realities. *Philosophy & Technology, 33*(1), 1-3.
Floridi, L. (2021). The European Legislation on AI: a Brief Analysis of its Philosophical Approach. *Philosophy & Technology, 34*: 215-222.
Flusser, V. (1985 [2011]). *Into the Universe of Technical Images*. Trans. Nancy Ann Roth. Minneapolis, London: University of Minnesota Press.
Flusser, V. (1991 [2014]). *Gestures*. Trans. Nancy Ann Roth. Minneapolis, London: University of Minnesota Press.
Forester, T. (1985). *The Information Technology Revolution* (ed.). Oxford: Basil Blackwell.
Fouse, S., Cross, S., & Lapin, Z. J. (2020). DARPA's Impact on Artificial Intelligence. *AI Magazine, 41*(2), 3-8.
Freund, E. (1982). Fast Nonlinear Control with Arbitrary Pole-Placement for Industrial Robots and Manipulators. *The International Journal of Robotics Research*, *1*(1), 65-78.
Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation?. *Technological Forecasting and Social Change, 114*, 254-280.
Frude, N. (1983). *The Intimate Machine: Close Encounters with the New Computers*. London: Century Publishing.
Galanos, V. (2014). Beyond Information Revolution: Postlude to a Past Future. Master Thesis. Retrieved 03-09-2021 from https://www.academia.edu/12120426/Beyond_Information_Revolution_Postlude_to_a_Past_Future_Vassilis_Galanos
Galanos, V. (2017). Singularitarianism and Schizophrenia. *AI & Society 32*, 573–590.
Galanos, V. (2018). Artificial Intelligence Does Not Exist: Lessons from Shared Cognition and the Opposition to the Nature/Nurture Divide. Kreps et al. (eds) *13th IFIP TC 9 International Conference on Human Choice and Computers. HCC13 2018. Held at the 24th IFIP World Computer Congress. WCC 2018. Poznan. Poland. September 19–21. 2018. Proceedings*. Switzerland: Springer Nature. 359-373.
Galanos, V. (2019a). Exploring expanding expertise: artificial intelligence as an existential threat and the role of prestigious commentators, 2014–2018. *Technology Analysis & Strategic Management 31*(4), 421-432.
Galanos, V. (2019b). Teratological Aspects in Artificial Intelligence and Robotics: From Monstrous Threats to Rorschach Opportunities" In Diego Compagna and Stefanie Steinhart (eds.) *Monsters, Monstrosities, and the Monstrous in Culture and Society*. Delaware and Malaga: Vernon Press. pp. 103-129.
Galanos, V. (2020). Tekken's Mokujin and the Disjunctive Synthesis of Gender Performativity. *Press Start 6*(1), n.p.
Galanos, V. (2022a). Nomadic artificial Intelligence and Royal Research Councils: Curiosity-Driven Research Against Imperatives Implying Imperialism. In: Tinnirello, M. (ed.). *The Global Politics of Artificial Intelligence*. Taylor & Francis – CRC Press.
Galanos, V. (2022b). Longitudinal Hype: Terminologies Fade, Promises Stay – An Essay Review on The Robots Are Among Us (1955) and 2062: The World that AI Made (2018). *Interfaces: Essays and Reviews on Computing and Culture 3*, Charles Babbage Institute, University of Minnesota, 73-87.
Galanos, V. (2022c). Why so Few AI Practitioners in Ai Policy? Specialist Views on Questions of Control, Regulation, and an Updated Paradox of Participation. Specialist Views on Questions of Control, Regulation, and an Updated Paradox of Participation. *SSRN preprint*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4213120
Galison, P. (1997). *Image & logic: A material culture of microphysics*. Chicago: The University of Chicago Press.
Garcia, M. (2016). Racist in the Machine. *World Policy Journal, 33*(4), 111-117.
Garud, R., Schildt, H. A., & Lant, T. K. (2014). Entrepreneurial storytelling, future expectations, and the paradox of legitimacy. *Organization Science, 25*(5), 1479-1492.
Garvey, C. (2019). Artificial intelligence and Japan's fifth generation: The information society, neoliberalism, and alternative modernities. *Pacific Historical Review, 88*(4), 619-658.
Gebru, T., Hoffman, J., & Fei-Fei, L. (2017). Fine-grained recognition in the wild: A multi-task domain adaptation approach. In *Proceedings of the IEEE international conference on computer vision* (pp. 1349-1358).
Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86-92.
Geraci, R. M. (2007). Robots and the sacred in science and science fiction: theological implications of artificial intelligence. *Zygon®, 42*(4), 961-980.
Gerghiou, L. (2018). To Every Thing There is a Season – lessons from the Alvey Programme for Creating an Innovation Ecosystem for Artificial Intelligence. *Manchester Policy Blogs*. 15 May 2018 [Blog post] Retrieved via WayBackMachine 25-09-2021 from https://blog.policy.manchester.ac.uk/posts/2018/05/to-every-thing-there-is-a-season-lessons-from-the-alvey-programme-for-creating-an-innovation-ecosystem-

for-artificial-intelligence/
Ghosh, S. (2017, January 5). A Supercomputer Just Made the World's First AI-Created Film Trailer – Here's How Well It Did. In: *IFLSCIENCE*. Retrieved 06-01-2017 from http://www.iflscience.com/technology/a-supercomputer-just-made-the-worlds-first-aicreated-film-trailer-heres-how-well-it-did/all/
Gibbs, S. (2014, October 27). Elon Musk: Artificial Intelligence is Our Biggest Existential Threat. *The Guardian*. Retrieved 25-11-2014 from https://www.theguardian.com/technology/2014/oct/27/elon-musk-artificial-intelligence-ai-biggest-existential-threat
Gibney, E. (2019). How UK Scientists are preparing for a Chaotic No-Deal Brexit. *Nature 565*(7740), 408-410.
Gieryn, T. F. (1983). Boundary-work and the demarcation of science from non-science: Strains and interests in professional ideologies of scientists. *American Sociological Review, 48*(6), 781-795.
Goertzel, B., De Garis, H., Pennachin, C., Geisweiller, N., Araujo, S., Pitt, J., ... & Huang, D. (2010). OpenCogBot: Achieving Generally Intelligent Virtual Agent Control and Humanoid Robotics via Cognitive Synergy. In *Proceedings of ICAI* (Vol. 10).
Gomes, L. (2015). Facebook AI Director Yann LeCun on His Quest to Unleash Deep Learning and Make Machines Smarter. *IEEE Spectrum*. 18 February 2015. Retrieved 16-07-2020 from: https://spectrum.ieee.org/facebook-ai-director-yann-lecun-on-deep-learning
Gong, J. (2018). Examining Henry Kissinger's Uninformed Comments on AI: Another public figure with no expertise on AI issues sweeping, unfounded statements about it threatening humanity. *Skynet Today*. 06 September 2018. Retrieved 18-07-2020 from: https://www.skynettoday.com/briefs/kissinger-ai
Good, I. J. (1965). Speculations Concerning the First Ultraintelligent Machine. *Advances in Computers*, *6*, 31-88.
Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*. Retrieved 14-11-2022 from: https://arxiv.org/pdf/1412.6572.pdf
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139-144.
Gorey, C. (2020). 'People of colour aren't empowered to make changes they're brought in to make' *SiliconPublic*. September 2, 2020. Retrieved 14-11-2022 from:
https://www.siliconrepublic.com/machines/deborah-raji-ai-racial-bias
Gray, M. L., & Suri, S. (2017, January 09). The Humans Working Behind the AI Curtain. *Harvard Business Review*. Retrieved 23-11-2020 from https://hbr.org/2017/01/the-humans-working-behind-the-ai-curtain
Gregory, R. (1974). *Concepts and Mechanisms of Perception*. London: Duckworth.
Grand View Research (2017, July). Artificial Intelligence Market Analysis By Solution (Hardware, Software, Services), By Technology (Deep Learning, Machine Learning, Natural Language Processing, Machine Vision), By End-use, By Region, and Segment Forecasts, 2014 – 2025. [Online]. Retrieved 05-11-2017 from http://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market
Graubard, S. (1988, Ed.) *The Artificial Intelligence Debate: False Starts, Real Foundations*. Cambridge: MIT Press.
Grossman, L. (2011). 2045: The Year Man Becomes Immortal. *Time Magazine*, *177*(7), 42-49.
Grudin, J. (2009). AI and HCI: Two fields divided by a common focus. *AI Magazine, 30*(4), 48-57.
Haddow, G., Bruce, A., Calvert, J., Harmon, S. H., & Marsden, W. (2010). Not "human" enough to be human but not "animal" enough to be animal–the case of the HFEA, cybrids and xenotransplantation in the UK. *New Genetics and Society, 29*(1), 3-17.
Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines 39*, 99-120.
Hagendorff, T., & Wezel, K. (2019). 15 challenges for AI: or what AI (currently) can't do. AI & Society 35, 355–365.
Hajič, J., & Hajičová, E. (2007). Some of Our Best Friends Are Statisticians. In: Matoušek, V. & Mautner, P. (eds.) *International Conference on Text, Speech and Dialogue* (pp. 2-10). Springer, Berlin, Heidelberg.
Han, S, Kelly, E, Nikou, S, and Svee, E. (2020). Reflections on Artificial Intelligence Alignment with Human Values: A Phenomenological Perspective. *ECIS 2020 Research Papers*. 92. Accessed 09-06-2020 from https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1091&context=ecis2020_rp
Hanson, D., Olney, A., Prilliman, S., Mathews, E., Zielke, M., Hammons, D., ... & Stephanou, H. (2005, July). Upending the uncanny valley. In *Proceedings of the national conference on artificial intelligence* (Vol. 20, No. 4, p. 1728). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.
Hanson, D., Bar-Cohen, Y., & Marom, A. (2009). *The Coming Robot Revolution Expectations and Fears About Emerging Intelligent, Humanlike Machines*. New York: Springer Science and Business Media, LLC.
Haraway, D. (2016). *Staying with the Trouble: Making Kin in the Chthulucene*. Durham and London: Duke University Press.
Harington, J., Sir. (1618). *The most elegant and witty epigrams of Sir Iohn Harrington, Knight digested into foure bookes: three vvhereof neuer before published*. London: Printed by G[eorge] P[urslowe] for Iohn

Budge. Retrieved 27-07-2021 from: https://quod.lib.umich.edu/cgi/t/text/text-idx?c=eebo;idno=A02647.0001.001
Harmon, S. H., Laurie, G., & Haddow, G. (2013). Governing risk, engaging publics and engendering trust: New horizons for law and social science?. *Science and Public Policy, 40*(1), 25-33.
Harris, M. (2017a, September 27). God is a Bot, and Anthony Levandowski is His Messenger. In *Wired* [Online]. Retrieved 28-09-2017 from https://www.wired.com/story/god-is-a-bot-and-anthony-levandowski-is-his-messenger/
Harris, M. (2017b, November 15). Inside the First Church of Artificial Intelligence. In *Wired* [Online]. Retrieved 17-11-2017 from https://www.wired.com/story/anthony-levandowski-artificial-intelligence-religion/
Hassabis, D. (2017). Artificial Intelligence: Chess match of the century. *Nature*, *544*(7651), 413-414.
Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. Cambridge, Massachusetts: MIT Press.
Hawking, S., Russell, S., Tegmark, M., & Wilczek, F. (2014). Stephen Hawking: 'Transcendence Looks at the Implications of Artificial Intelligence-But Are We Taking AI Seriously Enough?'. *The Independent*, *2014*(05-01), 9313474. Retrieved 12-05-2015 from http://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html
Hayes-Roth, F., & Jacobstein, N. (1994). The state of knowledge-based systems. *Communications of the ACM, 37*(3), 26-39.
Hayes-Roth, F., Davidson, J. E., Erman, L. D., & Lark, J. S. (1991). Frameworks for developing intelligent systems: The ABE systems engineering environment. *IEEE Expert, 6*(3), 30-40.
Heidegger M (1977) The question concerning technology and other essays. New York: Harper and Row.
Hendler, J. (1994). Beyond the fifth generation: Parallel AI research in Japan. *IEEE Expert, 9*(1): 2-7.
Hendler, J. (2008). Avoiding Another AI Winter. *IEEE Intelligent Systems, 23*(2), 2-4.
Hern, A. (2014, June 18). Elon Musk Says He Invested in DeepMind over 'Terminator' Fears. *The Guardian*. Retrieved 30-07-2014 from https://www.theguardian.com/technology/2014/jun/18/elon-musk-deepmind-ai-tesla-motors
Hern, A. (2017, May 24). China Censored Google's AlphaGo Match Against World's Best Go Player. *The Guardian*. Retrieved 25-05-2017 from https://www.theguardian.com/technology/2017/may/24/china-censored-googles-alphago-match-against-worlds-best-go-player
Hielscher, S. & Kivimaa, P. (2019). Governance Through Expectations: Examining the Long-Term Policy Relevance of Smart Meters in the United Kingdom. *Futures 109*, 101-107.
High-Level Expert Group on Artificial Intelligence (HLEGAI) (2019, April 8). A definition of AI: Main capabilities and scientific disciplines. Published 8 April 2019. European Commission: B-1049 Brussels. Accessed 18-07-2019 from https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines
Highnam, P. (2020). The Defense Advanced Research Projects Agency's Artificial Intelligence Vision. *AI Magazine, 41*(2), 83-85.
History of artificial intelligence (2021). Wikipedia. [Online page]. Accessed 01-06-2020 from: https://en.wikipedia.org/wiki/History_of_artificial_intelligence
Heinz Nixdorf MuseumsForum (2017). *Die Roboter Sind Unter Ins*. Blog post. November 7, 2017. Retrieved 18-06-2021 from: https://blog.hnf.de/die-roboter-sind-unter-uns/
Hoffmann, U. (1992). Aviation, androids, and artificial intelligence: the intricate paths of literary imagination and technological development. WZB Discussion Paper. FS2/92-109. Berlin: Social Science Research Center, Berlin, Wissenschaftszentrum Berlin für Sozialforschung gGmbH (WZB).
Hofstader, D.R. (1979 [1989]). *Gödel, Escher, Bach: An Eternal Golden Braid*. New York: Vintage Books.
Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: University of Michigan Press.
Holmqvist, K., Holsanova, J., Barthelson, M., & Lundqvist, D. (2003). Reading or Scanning? A Study of Newspaper and Net Paper Reading. In: Radach, R, Hyona, J., Deubel, H. (eds). *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, 657-670.
Holsanova, J. (2014). Reception of Multimodality: Applying Eye Tracking Methodology in Multimodal Research. *Routledge Handbook of Multimodal Analysis*, 285-296.
Hooton, C. (2015, July 20). A robot has passed a self-awareness test. *The Independent*. Retrieved 25 July 2015 from http://www.independent.co.uk/life-style/gadgets-and-tech/news/a-robot-has-passed-the-self-awareness-test-10395895.html
House of Commons. Science and Technology Committee (2016, October). Robotics and Artificial Intelligence: Fifth Report of Session 2016-17. Report Together with Formal Minutes Relating to the Report. Retrieved 20-10-2016 from https://www.publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf

House of Lords. Select Committee on Artificial Intelligence. (2018, April 16). *AI in the UK: Ready, Willing, and Able? Report of Session 2017-19*. Ordered to be printed 13 March 2018 and published 16 April 2018. The Authority of the House of Lords. Accessed 16-04-2018 from: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

Howe, J. (2007). Artificial Intelligence at Edinburgh University: A Perspective. *The University of Edinburgh, School of Informatics*. [University website]. Retrieved 22-07-2019 from: http://www.inf.ed.ac.uk/about/AIhistory.html

Hudson, A. D., Finn, E., & Wylie, R. (2021, e-published, ahead of print). What can science fiction tell us about the future of artificial intelligence policy?. AI & Society, 1-15.

Huggler, J. (2015, July 2). Robot Kills Man at Volkswagen Plant in Germany. *The Telegraph.* Retrieved 3-07-2015 from http://www.telegraph.co.uk/news/worldnews/europe/germany/11712513/Robot-kills-man-at-Volkswagen-plant-in-Germany.html

Huyssen, A. (1986). *After the Great Divide: Modernism, Mass Culture, Postmodernism*. Bloomington and Indianapolis: Indiana University Press.

Hyysalo, S. (2006). Representations of Use and Practice-Bound Imaginaries in Automating the Safety of the Elderly. *Social Studies of Science, 36*(4): 599–626.

IBM Research Editorial Staff (2011). Dave Ferrucci at Computer History Museum: How it all began and what's next. Blog post. *IBM Research Blog*. December 01, 2011. Retrieved 11-09-2021 from: https://www.ibm.com/blogs/research/2011/12/dave-ferrucci-at-computer-history-museum-how-it-all-began-and-whats-next/

Imbrie, A., Dunham, J, Gelles, R., and Aitken, C. (2020). *Mainframes: A Provisional Analysis of Rhetorical Frames in AI*. CSET Issue Brief. Center for Security and Emerging Technology. Accessed 16-08-2020 from: https://cset.georgetown.edu/wp-content/uploads/CSET- Mainframes-A-Provisional-Analysis-of-Rhetorical-Frames-in-AI.pdf

International Data Corporation (2017 April). Worldwide Spending on Cognitive and Artificial Intelligence Systems Forecast to Reach $12.5 Billion This Year, According to New IDC Spending Guide. [Online]. Retrieved 05-11-2017 from https://www.idc.com/getdoc.jsp?containerId=prUS42439617

Jackson, M. (2002). Biotechnology and the Critique of Globalisation. *Ethnos*, *67*(2), 141-154.

Jasanoff, S. & Kim, S. H. (2009). Containing the Atom: Sociotechnical Imaginaries and Nuclear Power in the United States and South Korea. *Minerva 47*: 119–146.

Jensen, C. B. (2014). Continuous variations: The conceptual and the empirical in STS. *Science, Technology, & Human Values*, *39*(2), 192-213.

Jordan, B., Wasson, C., & Roth-Lobo, H. S. (2015). Ethnographic Study Lifts the Hood on what REALLY Goes On inside that Car. *EPIC: Advancing the Value of Ethnography in Industry*. Retrieved 11-11-2017 from https://www.epicpeople.org/ethnographic-study-lifts-the-hood/

Kanakia, H., Shenoy, G., & Shah, J. (2019). Cambridge Analytica–A Case Study. *Indian Journal of Science and Technology, 12*(29), 1-5.

Katz, Y. (2017). Manufacturing an artificial intelligence revolution. *Social Science Research Network*. Accessed 01-02-2020 from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3078224

Keller, I., & Lohan, K. S. (2016, August). Analysis of Illumination Robustness in Long-Term Object Learning. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on* (pp. 240-245). IEEE.

Kelly, K. (2014). The Three Breakthroughs that have Finally Unleashed AI on the World. *Wired*. 27 October 2014. Retrieved 04-09-2020 from: https://www.wired.com/2014/10/future-of-artificial-intelligence

Kelly, K. (2017). *The Inevitable: Understanding the 12 Technological Forces that will Shape Our Future*. London: Viking.

Kerr, A., Barry, M, & Kelleher, J. (2020). Expectations of AI and the Performativity of Ethics: Implications for Communication Governance. *Big Data & Society* (in press). https://doi.org/10.1177/2053951720915939

Kerssens, N. (2019). De-Agentializing Data Practices: The Shifting Power of Metaphor in 1990s Discourses on Data Mining. *Journal of Cultural Analytics*, 16 May. 1-26.

Keyes, O. (2018). The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction, 2(CSCW)*, 1-22.

Kim, E., Paul, R., Shic, F., & Scassellati, B. (2012). Bridging the research gap: Making HRI useful to individuals with autism. *Speech-Language Pathology Faculty Publications*. Paper 38.

Kirkels, A. (2016). Biomass boom or bubble? A longitudinal study on expectation dynamics. *Technological Forecasting and Social Change 103*, 83-96.

Kissinger, H. (2018). How the Enlightenment Ends: Philosophically, intellectually – in every way – human society is unprepared for the rise of artificial intelligence. June 2018. *The Altantic*. Retrieved 08-06-2019 from: https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/

Kissinger, H., Schmidt, E., & Huttenlocher, D. (2019). The Metamorphosis: AI will bring many wonders. It may also destabilize everything from nuclear détente to human friendships. We need to think much harder about how to adapt. August 2019. *The Altantic*. Retrieved 02-09-2019 from: https://www.theatlantic.com/magazine/archive/2019/08/henry-kissinger-the-metamorphosis-ai/592771/

Kissinger, H., Schmidt, E., & Huttenlocher, D. (2021, forthcoming). *The Age of A.I. and our Human Future*. Little, Brown & Company.

Kling, R., & Scacchi, W. (1982). The web of computing: Computer technology as social organization. In Yovits, M. (Ed.) Advances in Computers, Vol. 21, 1-90.

Kok, J. N., Boers, E. J., Kosters, W. A., Van der Putten, P., & Poel, M. (2009). Artificial intelligence: definition, trends, techniques, and cases. In: Kok. J. N. (ed). *Artificial Intelligence, vol. 1,* Paris: United Nations Educational, Scientific, and Cultural Organization/Encyclopedia of Life Support Systems, 270-299.

Konrad, K. (2006). The social dynamics of expectations: the interaction of collective and actor-specific expectations on electronic commerce and interactive television. *Technology Analysis & Strategic Management, 18*(3-4), 429-444.

Konrad, K. & Böhle (2019). Socio-technical futures and the governance of innovation processes – An introduction to the special issue. *Futures 109*, 101-107.

Konrad, K., Truffer, B., & Voß, J. P. (2008). Multi-regime dynamics in the analysis of sectoral transformation potentials: evidence from German utility sectors. *Journal of Cleaner Production, 16*(11), 1190-1202.

Konrad, K., Van Lente, H., Groves, C., & Selin, C. (2017). Performing and Governing the Future in Science and Technology. In: Miller, C. A., Felt, U., Fouché, R. & Smith-Doerr, L. (eds). *The Handbook of Science and Technology Studies* (4th edition) Cambridge: MIT Press. 465-493.

Korosec, K. & Harris, M. (2020). Anthony Levandowski sentenced to 18 months in prison as new $4B lawsuit against Uber is filed. *TechCrunch*. 5 August 2020. Retrieved 18-09-2020 from: https://techcrunch.com/2020/08/04/anthony-levandowski-sentenced-to-18-months-in-prison-as-new-4b-lawsuit-against-uber-is-filed/

Kotliar, D. M. (2020). The return of the social: Algorithmic identity in an age of symbolic demise. *New Media & Society, 22*(7), 1152-1167.

Kriz, S., Ferro, T.D., Damera, P., Porter, J.R. (2010). Fictional robots as a data source in HRI research: Exploring the link between science fiction and interactional expectations. In *2010 IEEE RO-MAN*, pp. 458–463

Krizhevsky A, Sutskever I, and Hinton GE (2012) ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*. 1097-1105.

Kubicek, H. & Dutton, W. H. (1997). The Social Shaping of Information Superhighways: An Introduction. In: Kubicek, H., Dutton, W. H. & Williams, R. (eds). *The Social Shaping of Information Superhighways*. Frankfurt and New York: Campus Verlag and St. Martin's Press, pp. 9-44.

Kudina, O. & Verbeek, P. P. (2019). Ethics from within: Google Glass, the Collingridge dilemma, and the mediated value of privacy. *Science, Technology, & Human Values, 44*(2), 291-314.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Kuipers, B., McCarthy, J., & Weizenbaum, J. (1976). Computer power and human reason. *ACM SIGART Bulletin,* (58), 4-13.

Kurzweil, R. (2005). *The Singularity is Near: When Humans Transcend Biology*. London: Viking Books.

Kurzweil, R., & Kapor, M. (2009 [2001]). A Wager on the Turing Test. In *Parsing the Turing Test* (pp. 463-477). Springer Netherlands.

Kvale, S. (2006). Dominance through interviews and dialogues. *Qualitative inquiry*, *12*(3), 480-500.

Lachney, M. & Foster, E.K. (2020). Historicizing making and doing: Seymour Papert, Sherry Turkle, and epistemological foundations of the maker movement. *History and Technology, 36*(1): 54-82.

Lancaster, K. (2017). Confidentiality, anonymity and power relations in elite interviewing: conducting qualitative policy research in a politicised domain. *International Journal of Social Research Methodology, 20*(1), 93-103.

Laplace, P.S. (1814[1901]). *A Philosophical Essay on Probabilities*. Trans. Frederick Wilson Truscott and Frederick Lincoln Emory. London: John Wiley & Sons.

Laterza, V. (2021). Could Cambridge Analytica Have Delivered Donald Trump's 2016 Presidential Victory? An Anthropologist's Look at Big Data and Political Campaigning. *Public Anthropologist, 3*(1), 119-147.

Laureiro-Martínez, D., Brusoni, S., Canessa, N., & Zollo, M. (2015). Understanding the exploration–exploitation dilemma: An fMRI study of attention control and decision-making performance. *Strategic Management Journal, 36*(3), 319-338.

Lawler, D., Shute, J., Sanchez, R, Alexander, H. (2016, July 9). Barack Obama Condemns 'Deranged' Gunman, Says US Won't Return to Sixties-Era Race Riots. *The Telegraph*. Retrieved 08-05-2017 from http://www.telegraph.co.uk/news/2016/07/09/dallas-police-shooting-us-investigators-probe-links-with-black-p/

Le Doeuff, M. (1989). *The Philosophical Imaginary*. London: The Athlone Press.
LeCun, Y., Boser, B, Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation, 1*(4): 541–551.
LeCun, Y., & Bengio, Y. (1994, October). Word-level training of a handwritten word recognizer based on convolutional neural networks. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, vol. 3-Conference C: Signal Processing* (Cat. No. 94CH3440-5) (Vol. 2, pp. 88-92). IEEE.
Lefebvre, H. (1987). The Everyday and Everydayness. *Yale French Studies 73*, 7-11.
Legard, R., Keegan, J., & Ward, K. (2003). In-Depth Interviews. In: Ritchie, J. & Lewis, J. (Eds.). *Qualitative Research Practice: A Guide for Social Science Students and Researchers*. London, Thousand Oaks, New Delhi: Sage Publications.
Leifer, M.S. & Pusey, M.F. (2017). Is a Time Symmetric Interpretation of Quantum Theory Possible Without Retrocausality? *Proceedings of the Royal Society A 473*(20160607): 1-25.
Lem, S. (1975). *The Cyberiad: Fables for the Cybernetic Age*. (Trans. Michael Kandel). London: Futura Publications Limited.
Leng, G. & Leng, I. R. (2020). *The Matter of Facts: Skepticism, Persuasion, and Evidence in Science*. Cambridge, Massachusetts, London, England: The MIT Press.
Lepore, J. (2020). *If Then: How One Data Company Invented the Future*. London: John Murray.
Lessig, L. (1998). Open code and open societies: Values of internet governance. *Chi.-Kent L. Rev., 74*, 1405.
Levin, S. (2016, September 08). A Beauty Contest was Judged by AI and the Robots didn't like Dark Skin. *The Guardian*. Retrieved 15-09-2016 from https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people
Lewis, D. (2016, March 18). An AI-Written Novella Almost Won a Literary Prize. *Smithsonian.com. Smart News*. Retrieved 28-04-2016 from http://www.smithsonianmag.com/smart-news/ai-written-novella-almost-won-literary-prize-180958577/
Li, F., Deng, J., & Li, K. (2009). ImageNet: Constructing a large-scale image database. Journal of vision, 9(8), 1037-1037.
Li, R. Y. M. (2017). *An Economic Analysis on Automated Construction Safety: Internet of Things, Artificial Intelligence and 3D Printing*. Netherlands: Springer.
Liang, Y., & Lee, S. A. (2017). Fear of Autonomous Robots and Artificial Intelligence: Evidence from National Representative Data with Probability Sampling. *International Journal of Social Robotics*, 1-6.
Lighthill, J., Sir. (1973). *Artificial Intelligence: A General Survey.* In: *Artificial Intelligence: A Paper Symposium*. London: Science Research Council.
Lin, P., Abney, K., and Bekey, G. (eds) (2012). *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, Massachusetts, London: MIT Press.
Lin, C., Chang, P., & Luh, J. (1983). Formulation and Optimization of Cubic Polynomial Joint Trajectories for Industrial Robots. *IEEE Transactions on automatic control*, *28*(12), 1066-1074.
Lincoln, Y. S., & Guba, E. G. (1986). But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. *New Directions for Program Evaluation, 1986*(30), 73-84.
Linden, A., & Fenn, J. (2003). Understanding Gartner's hype cycles. *Strategic Analysis Report Nº R-20-1971. Gartner, Inc*.
Liu, Z. (2021). Sociological perspectives on artificial intelligence: A typological reading. *Sociology Compass*, *15*(3): 1-13.
Lloyd, L., Kechagias, D., & Skiena, S. (2005). Lydia: A System for Large-Scale News Analysis. In: *International Symposium on String Processing and Information Retrieval* (pp. 161-166). Berlin, Heidelberg: Springer.
Loizos, C. (2017, July 19). This Famous Roboticist Doesn't Think Elon Musk Understands AI. *TechCrunch*. Retrieved 20-07-2017 from https://techcrunch.com/2017/07/19/this-famous-roboticist-doesnt-think-elon-musk-understands-ai/
Lorenčík, D., Tarhaničová, M., & Sinčák, P. (2013, January). Influence of Sci-Fi films on artificial intelligence and vice-versa. In *Applied Machine Intelligence and Informatics (SAMI), 2013 IEEE 11th International Symposium on* (pp. 27-31). IEEE.
Lorge Parnas, D. (1988). Why engineers should not use artificial intelligence. *INFOR: Information Systems and Operational Research, 26*(4): 234-246.
Lösch, A. (2008). Anticipating the futures of nanotechnology: Visionary images as means of communication. In *Presenting Futures* (pp. 123-142). Springer Netherlands.
Lyall, C., & Tait, J. (2019). Beyond the limits to governance: new rules of engagement for the tentative governance of the life sciences. *Research Policy, 48*(5), 1128-1137.
Lyon, D. (1988). *The Information Society: Issues and Illusions*. Cambridge: Polity Press & Basil Blackwell.
MacDorman, K. F., Vasudevan, S. K., & Ho, C. C. (2009). Does Japan really have robot mania? Comparing attitudes by implicit and explicit measures. *AI & society*, *23*(4), 485-510.

Macnaghten, P. (2010). Researching technoscientific concerns in the making: narrative structures, public responses, and emerging nanotechnologies. *Environment and Planning A*, *42*(1), 23-37.
MacKenzie, D. (1990). *Inventing Accuracy: A Historical Sociology of Nuclear Missile Guidance*. Cambridge, Massachusetts: The MIT Press.
MacKenzie, D. (1998). The Certainty Trough. In: Williams, R., Faulkner, W., and Fleck, J. (eds). *Exploring Expertise: Issues and Perspectives*. London: MacMillan Press.
MacKenzie, D. (2006). Is Economics Performative? Option Theory and the Construction of Derivatives Markets. In: MacKenzie, D. Muniesa, F. & Siu, L. (eds). *On the Performativity of Economics: Do Economists Make Markets*. Princeton: Princeton University Press, 54-86.
MacKenzie, D. and Wajcman, J. (1999). Introduction. In: MacKenzie, D. and Wajcman, J. (eds). *The Social Shaping of Technology*. Buckingham: Open University Press.
March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization science, 2*(1), 71-87.
Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*. Retrieved 15-08-2022 from: https://arxiv.org/ftp/arxiv/papers/1801/1801.00631.pdf
Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the society for research in child development, 57*(4, Serial No. 228).
Markoff, J. (2015, May 25). Relax, the Terminator is Far Away. *The New York Times*. Retrieved 11-11-2017 from https://www.nytimes.com/2015/05/26/science/darpa-robotics-challenge-terminator.html
Markram, H. (2006). The Blue Brain Project. *Nature Reviews Neuroscience*,*7*(2), 153-160.
Markram, H., Muller, E., Ramaswamy, S., Reimann, M. W., Abdellah, M., Sanchez, C. A., ... & Muñoz-Céspedes, A. (2015). Reconstruction and Simulation of Neocortical Microcircuitry. *Cell*, *163*(2), 456-492.
Marris, C. (2015). The construction of imaginaries of the public as a threat to synthetic biology. *Science as Culture, 24*(1), 83-98.
Mason, J. (2002). *Qualitative Researching*. London: Sage.
Matheny, M., S. Thadaney Israni, M. Ahmed, and D. Whicher, (eds) (2019). Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. NAM Special Publication. Washington, DC: National Academy of Medicine.
Matney, L (2018, December 05). Where Facebook AI research moves next: The company's chief AI scientist reflects and predicts. *TechCrunch*. Retrieved 22-10-2022 from: https://techcrunch.com/2018/12/05/where-facebook-ai-research-moves-next/
Maturana, H. R. & Varela, F. J. (1992). *The Tree of Knowledge: The Biological Roots of Human Understanding*. Revised Edition. Boston, Massachusetts: Shambhala Publications.
Mayer-Schönberger, V. (2011). *Delete: The virtue of forgetting in the digital age*. Princeton University Press.
Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.
McCarthy, J., Minsky M., Shannon, C.M., *et al*. (1955 [2006]). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955, In: *AI Magazine (27)*4. [online] Retrieved 02-02-2016 from: http://www.aaai.org/ojs/index.php/aimagazine/article/view/1904/1802
McCauley, L. (2007). AI armageddon and the three laws of robotics. *Ethics and Information Technology*, *9*(2), 153-164.
McCorduck, P. (1979). *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. New York: W.H. Freeman.
McCorduck, P. (2004). *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. Natick, Massachusetts: A. K. Peters, Ltd./CRC Press.
McCorduck, P. (2019): *This Could Be Important: My Life and Times with the Artificial Intelligentsia*. Pittsburgh, PA: Carnegie Mellon University, ETC Press, Signature.
McDermott, D., Waldrop, M. M., Chandrasekaran, B., McDermott, J., & Schank, R. (1985). The dark ages of AI: a panel discussion at AAAI-84. *AI Magazine, 6*(3), 122-122.
McFarland, M. (2016, July 11). Robot's Role in Killing Dallas Shooter is a First. *CNN Tech*. Retrieved 29-04-2017 from http://money.cnn.com/2016/07/08/technology/dallas-robot-death/index.html
McFarland, D. & Bösser, T. (1993). *Intelligent Behavior in Animals and Robots*. Cambridge, MIT Press.
McGoogan, C. (2016, July 13). Robot Security Guard Knocks Over Toddler at Shopping Centre. *The Telegraph*. Retrieved 20-09-2017 from http://www.telegraph.co.uk/technology/2016/07/13/robot-security-guard-knocks-over-toddler-at-shopping-centre/
McLuhan, M. (1964). *Understanding media: the extensions of man*. New York: Signet.
McNamara, P. (2007). Google Earth and 'collateral damage'. *NetworkWorld*. January, 19. Retrieved 17-11-2022 from: https://www.networkworld.com/article/2303152/google-earth-and--collateral-damage-.html
McRobbie, L. R. (2017, August 1). Should We Stop Keeping Pets? Why More and More Ethicists Say Yes. *The Guardian*. Retrieved 03-08-2017 from https://www.theguardian.com/lifeandstyle/2017/aug/01/should-we-

stop-keeping-pets-why-more-and-more-ethicists-say-yes
McTaggart, J. M. E. (1964). *A Commentary on Hegel's Logic*. New York: Russell & Russell.
Melton, N., Axsen, J. & Sperling, D. (2016). Moving beyond alternative fuel hype to decarbonize transportation. *Nature Energy 1*: 16013.
Merleau-Ponty, M. (1945[2014]). *Phenomenology of Perception*. Trans. Donald. A. Landes. London and New York: Routledge.
Merleau-Ponty, M. (1964). *Signs*. Trans. Richard McCleary. Evanston III.: Northwestern University Press.
Merton, R. K. (1948). The Self-Fulfilling Prophecy. *The Antioch Review. 8*(2): 193–210.
Merton, R. K. (1965[1993]). *On the Shoulders of Giants: A Shandean Postscript – The Post-Italianate Edition*. Chicago and London: The University of Chicago Press.
Metta, G., Sandini, G., Vernon, D., Natale, L., & Nori, F. (2008, August). The iCub Humanoid Robot: An Open Platform for Research in Embodied Cognition. In: *Proceedings of the 8th workshop on performance metrics for intelligent systems* (pp. 50-56). ACM.
Metz, C. (2021). *Genius Makers: The Mavericks Who Brought AI to Google, Facebook and the World*. Dublin: Penguin.
Mialet, H. (2012). *Hawking Incorporated: Stephen Hawking and the Anthropology of the Knowing Subject*. Chicago: University of Chicago Press.
Michie, D. (ed.) (1979). *Expert Systems in the Micro-electronic Age*. Edinburgh: Edinburgh University Press.
Michie, D. (1982). *Machine Intelligence and Related Topics: An Information Scientist's Weekend Book*. New York: Gordon and Breach Science Publishers.
Miller, S. J. (2009, June 04). Evolutionary Robotics and Battlestar Galactica: an Interview with Hod Lipson. [Web Blog Post]. Retrieved 11-11-2017 from http://galacticasitrep.blogspot.co.uk/2009/06/evolutionary-robotics-and-battlestar.html
Millward, D. (2015, July 18). Robot Passes Self-Awareness Test: A Simple Experiment has Shown that Robots have Greater Self Awareness and Deductive Powers than Previously Thought. *The Telegraph*. Retrieved 25-07-2015 from http://www.telegraph.co.uk/news/worldnews/northamerica/usa/11748084/Robot-passes-self-awareness-test.html
Milne, G. (2020). *Smoke & Mirrors: How Hype Obscures the Future and How to See Past It.* Great Britain: Robinson.
Minsky, M. (1960). Steps toward artificial intelligence. *Proceedings of the IRE, 49*(1), 8-30.
Minsky, M. (1968). *Semantic Information Processing*. Cambridge, Massachusetts, and London: The MIT Press.
Minsky, M. (1975). A framework for representing knowledge. In: Winston, P. (Ed.) *The Psychology of Computer Vision*. New York: McGraw Hill, 211-217.
Minsky, M. (2007). *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. New York, London, Toronto, Sydney: Simon & Schuster.
Minsky, M.L. & Papert, S.A. (1969[1988]). *Perceptrons An Introduction to Computational Geometry* (Expanded Edition). Cambridge, Massachusetts, and London: The MIT Press.
Mishra, A. (2021). AI as a Hype Tool: To look beyond the hype, we must understand its genesis and propagation. 24 August 2021. *Skynet Today*. Retrieved 01-09-2021 from: https://www.skynettoday.com/editorials/ai-hype
Misselhorn, C. (2009). Empathy with inanimate objects and the uncanny valley. *Minds and Machines*, *19*(3), 345-359.
Mitchell, M. (2019). *Artificial Intelligence: A Guide for Thinking Humans*.UK, USA, Canada, Ireland, Australia, India, New Zealand, South Africa: Pelican Books.
Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, D. & Gebru, T. (2019, January). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).
Montaña, R. C. (2017). *Portable moving images: a media history of storage formats*. Berlin and Boston: Walter de Gruyter GmbH & Co KG.
Moore, G. E. (1965). Cramming more components onto integrated circuits. In *Electronics 32*(8).
Morse, J. (1991). Subjects, respondents, informants and participants? *Qualitative Health Research, 1*(4), 403-406.
Morton, T. (2016). Hyperobjects. *CSPA Quarterly, (15)*, 7-9.
Moshkina, L. V., & Arkin, R. C. (2008, June). Lethality and autonomous systems: The roboticist demographic. In *Technology and Society, 2008. ISTAS 2008. IEEE International Symposium* (pp. 1-9). IEEE.
Müller, V. C. (2020) Ethics of Artificial Intelligence and Robotics. In Zalta, E. N. (ed.). The Stanford Encyclopedia of Philosophy (Summer 2021 Edition). Retrieved 18-08-2021 from: https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/
Müller, V. C. (2021). Is it time for robot rights? Moral status in artificial entities. *Ethics and Information Technology*, (E-publication ahead of print): 1-9.

Müller, V. C., & Bostrom, N. (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In: *Fundamental Issues of Artificial Intelligence* (pp. 553-570). Springer International Publishing.
Multi-Annual Roadmap (MAR) for Horizon 2020 (2016). SPARC Robotics, euRobotics AISBL, Brussels, Belgium, 2017.
Murgia, M. (2016, May 14). Humans Versus Robots: How a Google Computer Beat a World Champion at this Board Game – and what it Means for the Future. *The Telegraph*. Retrieved 20-09-2016 from http://s.telegraph.co.uk/graphics/projects/go-google-computer-game/index.html
Murphy, R., & Woods, D. D. (2009). Beyond Asimov: the three laws of responsible robotics. *IEEE Intelligent Systems*, *24*(4).
Murphy, T. (1985). Artificial intelligence topics at IBM. *Simulation, 44*(1): 33-37.
Musil, R. (1978[1995]). *The Man Without Qualities. Volume II*.Trans. Burton Pike. New York: Vintage International.
Natale, S., & Ballatore, A. (2017, forthcoming). Imagining the thinking machine: technological myths and the rise of Artificial Intelligence. *Convergence: The International Journal of Research into New Media Technologies*.
Naudé, W. (2021). Artificial intelligence: neither Utopian nor apocalyptic impacts soon. *Economics of Innovation and New Technology, 30*(1): 1-23.
Neisser, U. (1963). The imitation of man by machine. *Science, 139*(3551), 193-197.
Nelkin, D. & Lindee, S.M. (1995). *The DNA Mystique: The Gene as a Cultural Icon*. New York. W.H. Freeman and Company.
Newman, N., Fletcher, R., Levy, D., Nielsen, R. K. (2016). *Reuters Institute Digital News Report 2016*. Oxford: University of Oxford. Reuters Institute for the Study of Journalism. Available online at: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Digital-News-Report-2016.pdf
Nickinson, P. (2016, November 4). Google Home review: Taking back the living room. *AndroidCentral*. Retrieved 05-01-2017 from http://www.androidcentral.com/google-home
Nilsson, N.J. (2010). *The Quest for Artificial Intelligence: A History of Ideas and Achievements.* Cambridge: Cambridge University Press. [online version] Accessed 18-11-2018 from: https://ai.stanford.edu/~nilsson/QAI/qai.pdf
Nordmann, A. (2007). If and then: a critique of speculative nanoethics. *Nanoethics*, *1*(1), 31-46.
Nordmann, A., & Rip, A. (2009). Mind the Gap Revisited. *Nature Nanotechnology*, *4*(5), 273-274.
Northrop Grumman Corporation (2017). *Northrop Gramman Remotec – Robotic Platforms and Sub-Systems*. Retrieved 01-06-2017 from http://www.northropgrumman.com/Capabilities/Remotec/Pages/default.aspx
Novet, J. (2015, December 10). Facebook Open Sources its Artificial Intelligence Server. *Venture Beat*. Retrieved 13-12-2015 from http://venturebeat.com/2015/12/10/facebook-open-sources-its-artificial-intelligence-server/
Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic Analysis: Striving to Meet the Trustworthiness Criteria. *International Journal of Qualitative Methods 16*(1), n.p.
O'Connell, M. (2017). To be a Machine: Adventures Among Cyborgs, Utopians, Hackers, and the Futurists Solving the Modest Problem of Death. London: Granta.
O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown.
O'Reilly, T. (2005). What is Web 2.0. *Communications & Strategies 65*(1), 17-37.
Oakley, B.W. (1983). Great Britain. *IEEE Spectrum, 20*(11): 69-71.
Oakley, B. and Owen, K. (1989). *Alvey: Britain's Strategic Computing Initiative*. Cambridge, Massachusetts and London: The MIT Press.
Oakley, B. W. (1990). Intelligent Knowledge-Based Systems – AI in the UK. In: Kurweil, R. (ed.) *The Age of Intelligent Machines*. Cambridge: MIT Press.
Obozintsev, L. (2018). *From Skynet to Siri: An exploration of the nature and effects of media coverage of artificial intelligence*. Doctoral Thesis. University of Delaware: ProQuest.
Ochigane, R. (2019). The invention of 'Ethical AI': How Big Tech manipulates Academia to avoid regulation. *The Intercept*. December 20. Retrieved 04-01-2020 from: https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/
OECD (2019). *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449. [Online document] Accessed 01-10-2019 from: https://www.oecd.org/going-digital/ai/principles/
Oh, C., Lee, T., Kim, Y., Park, S., & Suh, B. (2017, May). Us vs. Them: Understanding Artificial Intelligence Technophobia over the Google DeepMind Challenge Match. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 2523-2534). ACM.
Olazaran, M. (1996). A sociological study of the official history of the perceptrons controversy. *Social Studies of Science, 26*(3): 611-659.

Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity*. London, New Delhi, New York, Sydney: Bloomsbury Publishing.
Overbury, R. E. (1969). Technological forecasting a criticism of the Delphi technique. *Long range planning*, *1*(4), 76-77.
Owen, R., Mcnaghten, P. & Stilgoe, J. (2012). Responsible research and innovation: From science in society to science for society, with society. *Science and Public Policy 39*, 751-760.
Papert, S. (1988). One AI or Many? In Graubard, S. (Ed.) *The Artificial Intelligence Debate: False Starts, Real Foundations*. Cambridge: MIT Press. 241–267.
Park, S. M., & Kim, Y. G. (2022). A Metaverse: taxonomy, components, applications, and open challenges. *IEEE Access*. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9667507
Partridge, D. & Wilks, Y. (1990). The Foundations of Artificial Intelligence: A Sourcebook. New York: Cambridge University Press.
Parviainen, J., & Coeckelbergh, M. (2020). The political choreography of the Sophia robot: beyond robot rights and citizenship to political performances for the social robotics market. AI & Society. https://doi.org/10.1007/s00146-020-01104-w
Pasquale, F. (2020). *New Laws of Robotics: Defending Human Expertise in the Age of AI*. Cambridge, Massachusetts, and London, England: The Belknap Press of Harvard University Press.
Pavlik, J. V. (2023). Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education. *Journalism & Mass Communication Educator*, 10776958221149577.
Pedersen, D. B., & Hendricks, V. F. (2014). Science bubbles. *Philosophy & technology*, *27*(4), 503-518.
Penn, J. (2021). *Inventing intelligence: on the history of complex information processing and artificial intelligence in the United States in the mid-twentieth century*. PhD Thesis, University of Cambridge, UK.
Phillips, A. (1996). *Monogamy*. London: Faber and Faber.
Pias, C. (2016, ed.). *Cybernetics: The Macy Conferences, 1946-1953 – Transactions*. Zurich and Berlin: Diaphanes.
Pickering, A. (2009). Cybernetics as Nomad Science. *Deleuzian Intersections in Science, Technology and Anthropology*, 155-162.
Pickering, A. (2010). *The Cybernetic Brain: Sketches of Another Future*. Chicago: University of Chicago Press.
Pielke, R. A. Jr. (2007). *The Honest Broker: Making Sense of Science in Policy and Politics*. Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, Sao Paolo: Cambridge University Press.
Pierce, J. R., Carroll, John B., Hamp, E. P. Hays, David G., Hockett, C. F., Oettinger, A. G. and Perlis, A. (1966). *Languages and machines: Computers in translation and linguistics*. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council.
Pinch, T. J., & Bijker, W. E. (1984). The Social Construction of Facts and Artefacts, or, How the Sociology of Science and the Sociology of Technology Might Benefit Each Other. *Social Studies of Science, 14*(3), 399-441.
Pixel (1987 [in Greek]). Oi Athlothetouses Etaireies. *Pixel 34*, 82-82. Retrieved 05-06-2020 from: https://issuu.com/dimitrispanokostas/docs/pixel_issue_034_ocr
Planck, M. K. (1950). *Scientific Autobiography and Other Papers*. New York: Philosophical library.
Plasek, A. (2016). On the Cruelty of Really Writing a History of Machine Learning. *IEEE Annals of the History of Computing 38*(4), 6-8.
Pollock, N. and Williams, R. (2011). Who decides the shape of product markets? The knowledge institutions that name and categorise new technologies. *Information and Organization, 21*(4): 194-217.
Poole, B. (2018). A nice review of the challenges in deep learning… Tweet, January 4, 2018. Retrieved 15-11-2022 from: https://twitter.com/ylecun/status/949032334011092994
Principia Cybernetica Web (2020). *Combinatorial Explosion*. [Online Encyclopedia]. Accessed 20-05-2020 from http://pespmc1.vub.ac.be/ASC/COMBIN_EXPLO.html
Pynchon, T. (1973). *Gravity's Rainbow*. London: Picador.
Pynchon, T. (1997). *Mason & Dixon*. New York: Henry & Holt Company.
Rabinow, J. (1962). Developments in Character Recognition Machines at Rabinow Engineering Company. In: Fischer jr, G. L. *et al* (eds.). *Optical Character Recognition*. Washington, DC: Spartan Books.
Raffaelli, T. (1994) The early philosophical writings of Alfred Marshall. Part II: Marshall's papers. *Research in the History of Economic Thought and Methodology, Archival Supplement 4*: 95-159.
Raji, I. D., & Buolamwini, J. (2019, January). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 429-435).
Raji, I. D., Kumar, I. E., Horowitz, A., & Selbst, A. (2022, June). The fallacy of ai functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 959-972).

Rayner, S. (2004). The novelty trap: why does institutional learning about new technologies seem so difficult? *Industry and Higher Education, 18*(6), 349-355.
Reed, M. S., Graves, A., Dandy, N., Posthumus, H., Hubacek, K., Morris, J., ... & Stringer, L. C. (2009). Who's in and why? A typology of stakeholder analysis methods for natural resource management. *Journal of Environmental Management, 90*(5), 1933-1949.
Reichardt, J. (1978). *Robots: Fact, Fiction, Prediction*. London: Thames and Hudson, Ltd.
Resnick, M. , Port, O. and Hall, A. (1982) `Artificial intelligence: the second computer age begins'. *Business Week* (March 8): 66-75
Revell, T. (2017). Google DeepMind's NHS data deal 'failed to comply'with law. New Scientist. July 3, 2017. Retrieved 13-11-2021 from: https://institutions.newscientist.com/article/2139395-google-deepminds-nhs-data-deal-failed-to-comply-with-law/
Ribes, D. (2019). STS, Meet Data Science, Once Again. *Science, Technology, & Human Values, 44*(3), 514-539.
Rip, A. (1986). Legitimations of science in a changing world. In: Bungarten, T. (ed.) *Wissenschaftssprache und Gesellschaft: Aspekte der wissenschaftlichen Kommunikation und des Wissenstransfers in der heutigen Zeit*. Hamburg: Edition Akademion.
Rip, A. (2006). Folk theories of nanotechnologists. *Science as culture*, *15*(4), 349-365.
Rip, A. & Kemp, R. (1998). Technological change. In: Rayner, S. Malone, E.L. (Eds.). *Human Choice and Climate Change*, Volume 2, Columbus, OH: Battelle Press, pp. 327-399.
Rip, A. & Robinson, D. K. R. (2014). Constructive Technology Assessment and the Methodology of Insertion. In: Doorn, N., van de Poel, I., Schuurbiers, D., & Gorman, M. E. (eds.). *Early Engagement and New Technologies: Opening Up the Laboratory*. Dodrecht: Springer Science + Business Media, pp. 37-53. Quoting the reprint in Rip, A. (2020). *Nanotechnology and its Governance*. London and New York: Routledge, pp. 128-144.
Roberts, L. (1988). Expanding AI Research. In: Bartee, T. C. (ed.). *Expert Systems and Artificial Intelligence: Applications and Management*. Indianapolis: Howard W. Sams & Co.
Robertson, J. Robo Sapiens Japanicus: Robots, Gender, Family, and the Japanese Nation. Oakland: University of California Press, 2018.
Robson, C. (2002). *Real World Research*. Malden, Oxford, Melbourne, Berlin: Blackwell Publishing.
Rodik, P. & Primorac, J. (2015). To use or not to use: Computer-assisted qualitative data analysis software usage among early-career sociologists in Croatia. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research 16*(1), Art. 12.
Roland, A. & Shiman, P. (2002). Strategic computing: DARPA and the quest for machine intelligence, 1983-1993. Cambridge, Massachusetts, London, England: MIT Press.
Rose, F. (1985). *Into the Heart of the Mind: An American Quest for Artificial Intelligence*. New York: Vintage Books.
Rosen, P. (1993). The Social Construction of Mountain Bikes: Technology and Postmodernity in the Cycle Industry. *Social Studies of Science 23*(3), 479-513.
Rosenblatt, F. (1957). *The Perceptron–A Perceiving and Recognizing Automaton*. Cornell Aeronautical Laboratory. Report 85-460-1. Accessed 16-12-2019 from: https://blogs.umass.edu/brain-wars/files/2016/03/rosenblatt-1957.pdf
Rosenthal, S., Biswas, J., & Veloso, M. (2010, May). An Effective Personal Mobile Robot Agent Through Symbiotic Human-Robot Interaction. In: *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1 (pp. 915-922).* International Foundation for Autonomous Agents and Multiagent Systems.
Rotman, D. (2013). How Technology is Destroying Jobs. *Technology Review*, *16*(4), 28-35.
Rowen, N. (1992). The Making of Frankenstein's Monster: Post-Golem, Pre-Robot. *State of the Fantastic: Studies in the Theory and Practice of Fantastic Literature and Film*, 169-177.
Royal Society (2018a). *AI Narratives and Why They Matter*. Executive report. Accessed 12-12-2018 from: https://royalsociety.org/topics-policy/projects/ai-narratives/
Royal Society (2018b). Portrayals and Perceptions of AI and Why They Matter. Accessed 02-12-2021 from: https://www.repository.cam.ac.uk/bitstream/handle/1810/287193/EMBARGO%20-%20web%20version.pdf?sequence=1
Rubin, H. J., & Rubin, I. S. (2005). *Qualitative interviewing: The art of hearing data*. (2nd ed.) Sage.
Ruef, A., & Markard, J. (2010). What happens after a hype? How changing expectations affected innovation activities in the case of stationary fuel cells. *Technology Analysis & Strategic Management, 22*(3), 317-338.
Rumelhart, D. E., Hinton, G. E., Williams, R. J. (1986). Learning Representations by Back-Propagation Errors. *Nature 323*(6088), 533-536.
Rusconi, E., & Mitchener-Nissen, T. (2014). The Role of Expectations, Hype and Ethics in Neuroimaging and Neuromodulation Futures. *Frontiers in systems neuroscience, (8)*214.

Russell, N. M. (2011). *Black students and mathematics achievement: A mixed-method analysis of in-school and out-of-school factors shaping student success*. PhD thesis. University of Washington.
Russell, S. (2019). *Human Compatible: AI and the Problem of Control*. UK: Penguin.
Russell, S. & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach* (1st edition). New Jersey: Prentice-Hall.
Russell, S. & Norvig, P. (2016). *Artificial Intelligence: A Modern Approach* (3rd edition). Essex: Pearson Education Limited.
Ryder, C. (2017). New Artificial Intelligence Research Institute Launches: First of Its Kind Dedicated to the Study of Social Implications of AI. *NYU Tandon School of Engineering News*. Retrieved 15-11-2022 from: https://engineering.nyu.edu/news/new-artificial-intelligence-research-institute-launches
Ryfle, S. (1998). *Japan's Favorite Mon-Star: The Unauthorized Biography of the Big G*. Toronto: ECW Press.
Salvini, P., Laschi, C., & Dario, P. (2007). Roboethics in biorobotics: discussion of case studies. In *International Conference on Robotics and Automation (ICRA 2007) Workshop on Robo-Ethics, Rome, Italy*.
Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-15).
Schaller, R. R. (1997). Moore's Law: Past, Present and Future. *IEEE spectrum*, *34*(6), 52-59.
School of Informatics (2002). "The New Town Fire of 7th December 2002…" *Bulletin* 09 December 2002. Retrieved 15-07-2020 from: https://www.inf.ed.ac.uk/emergency/bulletins/bulletin-09122002.html
Schot, J.W. (1992). Constructive Technology Assessment and Technology Dynamics: The Case of Clean Technologies. *Science, Technology & Human Values 17*, 36-56.
Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417-424.
Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society, 4*(2): 1-17.
Selbst, A.D., boyd, d., Friedler, S.A., Venkatasubramanian, S., Vertesi, J., (2018) Fairness and abstraction in sociotechnical systems. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ACM Press, New York, NY,. 59-68.
Selin, C. (2008). The sociology of the future: tracing stories of technology and time. *Sociology Compass*, *2*(6), 1878-1895.
Selwyn, N., & Gallo Cordoba, B. (2021, e-published ahead of print). Australian public understandings of artificial intelligence. *AI & Society*, 1-18.
Shapin, S. (1988). Following scientists around. *Social Studies of Science, 18*(3), 533-550.
Shapiro, E. and Warren, D.H. (1993). The 5th Generation Project: personal perspectives. *Communications of the ACM, 36*(3): 47-49.
Shapiro, C. (2000).Artificial Intelligence. In: Ralston, A., Reilly, E. D., and Hemmendinger, D. (Eds) *Encyclopedia of Computer Science*. Fourth Edition. New York: Van Nostrand Reinhold.
Shead, S. (2020) Researchers: Are we on the cusp of an 'AI winter'? BBC. 12 January 2020. https://www.bbc.co.uk/news/technology-51064369 (accessed 12 January 2021).
Shneiderman, B. (2022a) *Human-Centered AI*. Oxford: Oxford University Press.
Shneiderman, B. (2022b). 79th note on Human-Centered AI. Online mailing list message. 3 November 2022. Retrieved 13-11-2022 from: https://groups.google.com/g/human-centered-ai/c/Q7H8i2a_S1M/m/ixDN2nhuBAAJ?utm_medium=email&utm_source=footer
Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... & Dieleman, S. (2016). Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature*, *529*(7587), 484-489.
Simon, H. A., & Newell, A. (1958). Heuristic problem solving: The next advance in operations research. *Operations research*, *6*(1), 1-10.
Simonite, T. (2021). What really happened when Google ousted Timnit Gebru: She was a star engineer who warned that messy AI can spread racism. Google brought her in. Then it forced her out. Can Big Tech take criticism from within?. That's a problem for all of us. WIRED magazine, June 8, 2021. Retrieved 15-11-2022 from: https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/
Simpkins, A. M. & Simpkins, C. A. (2011). Zen Meditation in Psychotherapy: Techniques for Clinical Practice. Hoboken: John Wiley & Sons, Inc.
Singer, P. W. (2009). *Wired for War: The Robotics Revolution and Conflict in the 21st Century*. London: Penguin.
Sismondo, S. (2011). *An Introduction to Science and Technology Studies*. West Sussex: John Wiley & Sons.
Sloman, A. (1978). *The Computer Revolution in Philosophy: Philosophy, Science and Models of Mind*. Brighton: Harvester Press.
Sloman, A. (2003). *What is Artificial Intelligence?*. University of Birmingham School of Computer Science. 29 April, 2003. [from personal homepage] http://www.cs.bham.ac.uk/~axs/misc/aiforschools.html
Sloman, A. (2006). Why Asimov's three laws of robotics are unethical. [Personal Webpage, online]. Retrieved

19-11-2017 from http://www.cs.bham.ac.uk/research/projects/cogaff/misc/asimov-three-laws.html
Sloman, A. (2011). John McCarthy – Some Reminiscences. Retrieved 12-01-2021. Available at: https://www.cs.bham.ac.uk/research/projects/cogaff/sloman-jmc-aisb.pdf.
Sluckin, W. (1960). *Minds and Machines*. Middlesex: Penguin Books.
Smith, K. E. (2006). Problematising power relations in 'elite' interviews. *Geoforum*, *37*(4), 643-653.
Smith, M. L. (2006). Overcoming theory-practice inconsistencies: Critical realism and information systems research. *Information and Organization, 16*(3), 191-211.
Smith, B. C. (2019) *The Promise of Artificial Intelligence: Reckoning and Judgement*. Cambridge, Massachusetts, London: The MIT Press.
Smith, R.D.J., Scott, D., Kamwendo, Z.T., Calvert, J. (2019) An Agenda for Responsible Research and Innovation in ERA CoBioTech. Swindon, UK: Biotechnology and Biological Sciences Research Council and ERA CoFund on Biotechnology.
Snow, Charles Percy (1959 [2001]). *The Two Cultures*. London: Cambridge University Press.
Snow, J. (2018). "We're in a diversity crisis": cofounder of Black in AI on what's poisoning algorithms in our lives. *MIT Technology Review*, February 14, 2018. Retrieved 13-11-2022 from: https://www.technologyreview.com/2018/02/14/145462/were-in-a-diversity-crisis-black-in-ais-founder-on-whats-poisoning-the-algorithms-in-our/
Solon, O. (2017, January 30). Oh the Humanity! Poker Computer Trounces Humans in Big Step for AI. *The Guardian*. Retrieved 01-02-2017 from https://www.theguardian.com/technology/2017/jan/30/libratus-poker-artificial-intelligence-professional-human-players-competition
Sparrow, R., & Sparrow, L. (2006). In the hands of machines? The future of aged care. *Minds and Machines*, *16*(2), 141-161.
Spinardi, G. & Williams, R. (2005). The governance challenge of breakthrough science and technology. In: Lyall, C. & Tait, J. (eds). *New modes of governance: developing an integrated policy approach to science, technology, risk and the environment*. Aldershot, Ashgate, pp.45-66.
Steels, L. and Brooks, R. (2018, Eds). *The artificial life route to artificial intelligence: Building embodied, situated agents*. Hillsdale, New Jersey, Hove, UK: Lawrence Erlbaum Associates, Publishers.
Stilgoe, J. (2018). Machine learning, social learning and the governance of self-driving cars. *Social Studies of Science, 48*(1), 25-56.
Stilgoe, J., Owen, R and Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy 42*(9): 1568–1580.
Stix, C. and Maas, M.M. (2021). Bridging the gap: the case for an 'Incompletely Theorized Agreement' on AI policy. *AI and Ethics*. Epub ahead of print, 15 January 2021. DOI: 10.1007/s43681-020-00037-w.
Stonehouse (2019). *Stonehouse's Poems for Zen Monks*. Translated by Red Pine. Anacortes, Washington: Empty Bowl.
Strehl, R. (1952 [1955]). *The Robots are Among Us*. London and New York: Arco Publishers.
Strubell, E., Ganesh, A. & McCallum, A. (2020). Energy and Policy Considerations for Modern Depp Learning Research. In *Proceedings of the AAAI Conference on Artificial Intelligence 34*(09), 13693-13696.
Suchman, L. (1984, Spring). DARPA Strategic Computing Initiative: A progress report on the CPSR response. *The CPSR Newsletter 2*(2). Available online: http://cpsr.org/prevsite/publications/newsletters/old/1980s/Spring1984.txt/
Suchman, L. (2007). *Human-machine reconfigurations: Plans and situated actions*. Cambridge: Cambridge University Press.
Suchman, L. and Trigg, R. (1993). Artificial intelligence as craftwork. In: Chaiklin, S. & Lave, J. (Eds.), *Understanding Practice: Perspectives on Activity and Context* (Learning in Doing: Social, Cognitive and Computational Perspectives). Cambridge: Cambridge University Press, pp. 144-178.
Summer, J. (2014). Defiance to compliance: Visions of the computer in postwar Britain. *History and Technology 30*(4), 309-333.
Sutton, E. (2020). The Increasing Significance of Impact within the Research Excellence Framework (REF). *Radiography*. [In press, corrected proof]. https://doi.org/10.1016/j.radi.2020.02.004
Sutton, R. and Barto, A. (2015, in progress). Reinforcement Learning: An Introduction. Cambridge, Massachusetts, London, England: The MIT Press.
Szollosy, M. (2016). Freud, Frankenstein and our Fear of Robots: Projection in Our Cultural Perception of Technology. *AI & Society*, *32*. 433-439.
Taiwo, O. (2022). Embodying Ubuntu, through the Physical Journal as an antidote to the effects of the Anthropocene. *Body, Space & Technology, 21*(1).
Takayama, L., Ju, W., & Nass, C. (2008, March). Beyond dirty, dangerous and dull: what everyday people think robots should do. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction* (pp. 25-32). ACM.

Tarkkala, H., Helén, I., & Snell, K. (2019). From health to wealth: The future of personalized medicine in the making. *Futures, 109*, 142-152.
Taylor, T. & Dorin, A. (2020). Rise of the Self-Replicators: Early Visions of Machines, AI and Robots that Can Reproduce and Evolve. Switzerland: Springer Nature.
Tedre, M. (2014). *The Science of Computing: Shaping a Discipline*. Boca Raton: Taylor and Francis / CRC Press.
Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. New York: Alfred. A. Knopf.
Tesler, L. (2019). *CV: Summary*. [Online document]. Accessed 17-12-2019 from http://www.nomodes.com/Larry_Tesler_Consulting/CV.html
Toffler, A. (1980). *The third wave*. New York: Bantam books.
Track, E., Forbes, N., & Strawn, G. (2017). The End of Moore's Law. *Computing in Science & Engineering*, *19*(2), 4-6.
Treblin, N. (2016, October 22). Robots Could be the Worst Thing Ever for Humanity, Warns Stephen Hawking. *RT*. Retrieved 01-11-2016 from https://www.rt.com/uk/363502-artificial-intelligence-stephen-hawking/
Truby, J., Brown, R. and Dahdal, A. (2020). Banking on AI: mandating a proactive approach to AI regulation in the financial sector. *Law and Financial Markets Review, 14*(2): 110-120.
Turek, F. (2011). Machine Vision Fundamentals, How to Make Robots See. *NASA Tech Briefs Magazine*, *35*(6), 60-62.
Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *59*(236), 433-460.
Turing, A. M. (1951). Intelligent Machinery: A Heretic Theory. *BBC, The '51 Society Radio Broadcast*. Accessed through the Turing Digital Archive: http://www.turingarchive.org/browse.php/b/4
Turkle, S. (1981). Computers as Rorschach: Subjectivity and Social Responsibility. Bo Sundin (ed.). *Is the Computer a Tool?* Stockholm. Almquist and Wiksell. 81–99.
Turkle, S. (1984). *The Second Self: Computers and the Human Spirit*. London, Toronto, Sydney, and New York: Granada.
Tzafestas, S.G. (2016). *Roboethics: A Navigating Overview*. Cham, Heidelberg, New York, Dordrecht, London: Springer.
Ulam, S. (1958). Tribute to John von Neumann. *Bulletin of the American Mathematical Society 64*(3): 1-49.
Ulnicane, I., Knight, W., Leach, T., et al. (2020). Framing governance for a contested emerging technology: insights from AI policy. *Policy and Society*, Epub ahead of print, 17 December 2020. DOI: 10.1080/14494035.2020.1855800.
UK AI Council (2021). *AI Roadmap*. 06 January 2021 [Report]. Retrieved 18-02-2021 from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/949539/AI_Council_AI_Roadmap.pdf
United Kingdom Parliament (1943). *House of Commons Rebuilding*. HC Deb 28 October 1943 vol 393 cc403-73. [Commons Sitting]. Retrieved 25-07-2017 from http://hansard.millbanksystems.com/commons/1943/oct/28/house-of-commons-rebuilding
United States. Congress. House. Committee on Science and Technology. Subcommittee on Investigations and Oversight (1983). *Japanese technological advances and possible United States responses using research joint ventures: hearings before the Subcommittee on Investigations and Oversight and the Subcommittee on Science, Research, and Technology of the Committee on Science and Technology*, U.S. House of Representatives, Ninety-eighth Congress, first session, June 29-30, 1983.
United States Department of Labor. Occupational Safety and Health Administration (2017). *Fatality and Catastrophe Investigation Summaries*. Retrieved 28-04-2017 from https://www.osha.gov/pls/imis/accidentsearch.html
Uribe, R., & Gunter, B. (2004). Research note: The Tabloidization of British Tabloids. *European Journal of Communication*, *19*(3), 387-402.
Vainio, A. (2013). Beyond research ethics: Anonymity as 'ontology', 'analysis' and 'independence'. *Qualitative Research, 13*(6), 685-698.
Vallès-Peris, N., & Domènech, M. (2020). Roboticists' imaginaries of robots for care: the radical imaginary as a tool for an ethical discussion. *Engineering Studies, 12*(3), 157-176.
Vallor, S. (2016). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford: Oxford University Press.
Van Est, R., Gerrutsen, J. & Kool, L. (2017). *Human Rights in the Robot Age: Challenges Arising from the Use of Robotics, Artificial Intelligence, and Virtual and Augmented Reality. – Expert report written for the Committee on Culture, Science, Education and Media of the Parliamentary Assembly of the Council of Europe (PACE)*. The Hague: Rathenau Instituut.
Van de Riet, R.P. (1993). An overview and appraisal of the fifth generation computer system project. *Future Generation Computer Systems, 9*(2): 83-103.

Van Deursen, A. J., & Helsper, E. J. (2018). Collateral benefits of Internet use: Explaining the diverse outcomes of engaging with the Internet. N*ew Media & Society, 20*(7), 2333-2351.
Van Hove, P. (1991) ESPRIT, the European strategic programme for research and development in information technology. In: *Proceedings of the workshop on Speech and Natural Language (HLT '91)*. Association for Computational Linguistics, USA, pp.34–48.
Van Lente, H. (2012). Navigating Foresight in a Sea of Expectations: Lessons from the Sociology of Expectations. *Technology Analysis & Strategic Management*, *24*(8), 769-782.
Van Lente, H., Spitters, C. & Peine, A. (2013). Comparing technological hype cycles: Towards a theory. *Technological Forecasting and Social Change 80*(8), 1615–1628.
Van Lieshout, R. (2016, June 19). Tech Revolution vs Human Evolution: A Call for Smarter Marketing. In *The Marketing Technologist*. [Online]. Retrieved 02-11-2017 from https://www.themarketingtechnologist.co/tech-revolution-vs-human-evolution/
Vardi, M. Y. (2013). The great robotics debate. *Communications of the ACM*, *56*(7), 5.
Vasterman, P. L. (2005). Media-hype: Self-reinforcing news waves, journalistic standards and the construction of social problems. *European Journal of Communication, 20*(4), 508-530.
Veale, M., & Borgesius, F. Z. (2021). Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International, 22*(4), 97-112.
Verran, H. (1998). Re-imagining land ownership in Australia. *Postcolonial Studies: Culture, Politics, Economy 1*(2): 237-254.
Veruggio, G., & Operto, F. (2006). Roboethics: a Bottom-Up Interdisciplinary Discourse in the Field of Applied Ethics in Robotics. *International Review of Information Ethics*, *6*(12), 2-8.
Veruggio, G., Operto, F. (2008). Roboethics: Social and Ethical Implications of Robotics. In: Siciliano, B., Khatib, O. (eds.). *Springer handbook of robotics*. Berlin: Springer, 1499–1524.
Veruggio, G., Operto, F., & Bekey, G. (2016). Roboethics: Social and Ethical Implications. In: *Springer handbook of robotics* (pp. 2135-2160). Springer International Publishing.
Vinge, V. (1993). The Coming Technological Singularity. *VISION-21 Symposium, NASA Lewis Research Center and Ohio Aerospace Institute*. March 30-31, 1993. Available at: https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022856.pdf
Vinge, V. (2008). Signs of the Singularity. *IEEE Spectrum*, *45*(6).
Waldrop, M.M. (1984). Artificial intelligence (I): into the world; AI has become a hot property in financial circles: but do the promises have anything to do with reality?. Science, 223: 802-806..
Wallach, H. (2020). *Keynote*. At panel Navigating the Broader Impacts of AI Research, NeurIPS 2020 Workshop, December 12, 2020. Retrieved 15-11-2022 from: https://ai-broader-impacts-workshop.github.io/#Recordings
Walsh, T. (2018). *2062: The World that AI Made*. Carlton: La Trobe University Press, Black Inc.
Wark, M. (2021). *Capital Is Dead: is this something worse?* London: Verso.
Warwick, K. (1998). *In the Mind of the Machine: The Breakthrough of Artificial Intelligence*. London: Arrow.
Warwick, K. (2000). *QI: The Quest for Intelligence*. London: Judy Piatkus.
Way of the Future (2017). *What is this all about?* [Website]. Retrieved 17-11-2017 from http://www.wayofthefuture.church/
Weaver, W. (1949). Translation. Memorandum. Reprinted. In: Locke, W.N. & Booth, A.D. (eds.). *Machine Translation of Languages: Fourteen Essays*, (pp. 15–23). Cambridge: MIT Press.
Weber, J. (2005). Helpless machines and true loving care givers: a feminist critique of recent trends in human-robot interaction. *Journal of Information, Communication and Ethics in Society*, *3*(4), 209-218.
Weber, J. (2011). Black-Boxing Organisms, Exploiting the Unpredictable: Control Paradigms in Human–Machine Translations. In *Science in the Context of Application* (pp. 409-429). Springer Netherlands.
Webster, G., Creemers, R., Triolo, P., Kania, E. (2017, August 1). Full Translation: China's 'New Generation Aritificial Intelligence Development Plan [2017]. *New America, DigiChina* [Blog Post]. Accessed 10-03-2020 from https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/ [Original can be found as "State Council Notice on the Issuance of the New Generation AI Development Plan" [国务院关于印发新一代人工智能发展规划的通知]. "MIIT's Notice Regarding the Release of the Three Year Action Plan to Promote the Development of New-Generation Artificial Intelligence Industry (2018-2020) [工业和信息化部关于印发《促进新一代人工智能产业发展三年行动计划（2018-2020年）》的通], December 14, 2017, http://www.miit.gov.cn/n1146295/n1652858/n1652930/n3757016/c5960820/content.html]

Weinberg, J. (June 21, 2019). Searle Found to Have Violated Sexual Harassment Policies. *Daily Nous*. June 21 2019. Retrieved 03-09-2021 from https://dailynous.com/2019/06/21/searle-found-violated-sexual-harassment-policies/
Weiss, A., Igelsböck, J., Wurhofer, D., & Tscheligi, M. (2011). Looking forward to a "robotic society"?. *International Journal of Social Robotics*, *3*(2), 111-123.
Weizenbaum, J. (1976). *Computer Power and Human Reason: From Judgement to Calculation*. New York and San Francisco: W.H. Freeman & Company.
Winfield, A. F., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133).
Whitby, B. (2003). *Artificial Intelligence: A Beginner's Guide*. Oxford: Oneworld.
Wiener, N. (1950 [1989]). *The Human Use of Human Beings: Cybernetics and Society*. London: Free Association Books.
Wikipedia (2021a) AI Effect. https://en.wikipedia.org/wiki/AI_effect (Accessed 30 January 2021).
Wikipedia (2021b) AI Winter. https://en.wikipedia.org/wiki/AI_winter (Accessed 30 January 2021).
Wiles, R., Charles, V., Crow, G., & Heath, S. (2006). *Researching researchers: lessons for research ethics*. Qualitative Research, 6(3), 283-299.
Wiles, R., Crow, G., Heath, S., & Charles, V. (2008). The management of confidentiality and anonymity in social research. *International Journal of Social Research Methodology, 11*(5), 417-428.
Willetts, D. (2013). *Eight Great Technologies*. London: Policy Exchange.
Williams, K. (1989). Researching the powerful: Problems and possibilities of social research. *Contemporary Crises, 13*(3), 253-274.
Williams, R. (1997). The Social Shaping of Information and Communication Technologies. In: Kubicek, H., Dutton, W. H. & Williams, R. (eds). *The Social Shaping of Information Superhighways*. Frankfurt and New York: Campus Verlag and St. Martin's Press, pp. 299-338.
Williams, R. (2006). Compressed Foresight and Narrative Bias: Pitfalls in Assessing High Technology Futures. *Science as Culture*, *15*(4), 327-348.
Williams. R. (2019). European Perspectives on the Anticipatory Governance of AI. In: Qian, S. (ed.) *AI Governance 2019: A Year in Review*. *Observations of 50 Global Experts*. Shanghai Institute for Science of Science. April 2020. [Online Report]. Accessed 15-05-2020 from: https://www.aigovernancereview.com/
Williams, R. (2019). Why science and innovation policy needs Science and Technology Studies?. In: Canzler, W., Kuhlmann, S., Simon, D. (eds.) *Handbook on Science and Public Policy*. Cheltenham, Northampton: Edward Elgar Publishing. (pp. 503-522)
Williams, R., & Edge, D. (1996). The Social Shaping of Technology. *Research Policy, 25*(6), 865-899.
Williams, R., Faulkner, W., and Fleck, J. (1998). Exploring Expertise: Issues and Perspectives. In: Williams, R., Faulkner, W., and Fleck, J. (eds). *Exploring Expertise: Issues and Perspectives*. London: MacMillan Press.
Williams, R., Stewart, J. & Slack, R. *Social Learning in Technological Innovation: Experimenting with Information and Communication Technologies*. Aldershot: Edward Elgar.
Winner, L. (1977). *Autonomous Technology: Technics-out-of-Control as a Theme in Political Thought*. Cambridge, Massachusetts, and London, England: The MIT Press.
Winner, L. (1980). Do Artifacts Have Politics? *Daedalus 109*(1), 121-136.
Winner, L. (1984). Mythinformation in the high-tech era. *Bulletin of Science, Technology & Society, 4*(6), 582-596.
Winograd, T. (1972). *Understanding Natural Language*, New York: Academic Press.
Winograd, T. & Flores, F. (1988). *Understanding Computers and Cognition: A New Foundation for Design*. New Jersey: Ablex.
Winston, P. H. & Holmes, D. (2018). The Genesis Enterprise: Taking Artificial Intelligence to another Level via a Computational Account of Human Story Understranding. Report. *MIT Libraries DSpace*. Retrieved 20-09-2020 from http://hdl.handle.net/1721.1/119651
Woolgar, S. (1985). Why not a sociology of machines? The case of sociology and artificial intelligence. *Sociology*, *19*(4), 557-572.
Woolgar, S. (1987). Reconstructing Man and Machine: A Note on Sociological Critiques of Cognitivism. In: Bijker, W. E., Hughes, T. P., & Pinch, T. (eds). *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. Cambridge, Massachusetts, London, England: The MIT Press.
Woolgar S (2014) Struggles with representation: Could it be otherwise? In Coopmans C, Vertesi J, Lynch M and Woolgar, S (Eds) Representation in scientific practice revisited. Cambridge: The MIT Press.
Wyatt, S. (2008). Technological Determinism is Dead; Long Live Technological Determinism. In: Hackett, E., Amsterdamska, O., Lynch, M. and Wajcman J. (eds.), *The Handbook of Science and Technology Studies*. Cambridge, MA: MIT Press, pp. 165-181.

Wynne, B. (2014). Further disorientation in the hall of mirrors. *Public Understanding of Science, 23*(1), 60-70.
Xiang, F. (2018). AI will spell the end of capitalism. 3 May 2018. *Washington Post.* Retrieved 08-09-2020 from: https://www.washingtonpost.com/news/theworldpost/wp/2018/05/03/end-of-capitalism/
Yampolskiy, R. V. (2015). *Artificial Superintelligence: a Futuristic Approach*. Boca Raton, London, New York: CRC Press.
Yu, K. (2008). Confidentiality revisited. *Journal of Academic Ethics, 6*(2), 161-172.
Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J. C., Sellitto, M., Shoham, Y., Clark, Y. & Perrault, R. (2021). *The AI Index 2021 Annual Report*. AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA.
Zhu, J. (2009). *Intentional Systems and the Artificial Intelligence (AI) Hermeneutic Network: Agency and Intentionality in Expressive Computational Systems*. [Doctoral dissertation] Georgia: Georgia Institute of Technology.
Zuboff, S. (1988). *In the Age of the Smart Machine: The Future of Work and Power*. Oxford: Heineman Professional Publishing.
Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. London: Profile books.
Zunt. D. (2002). *Who did Actually Invent the Word "Robot" and What Does it Mean?* [Online article]. Retrieved 10 May 2016 from http://capek.misto.cz/english/robot.html
Zwart, H. (2020). Coming to terms with technoscience: The Heideggerian way. *Human Studies 43*, 385–408.

APPENDICES:

Appendix 1: List of Abbreviations and Glossary

I summarise below brief explanations of (a) technical terms, (b) 1980s AI-related research initiatives, (c) funding bodies and schemes relevant to contemporary support of AI, and (d) theoretical terms, to offer a brief orientation tool only in assistance to the detailed explanations offered throughout the thesis' narrative.

a. Technical Terms

AI: Artificial intelligence. A scientific field which aims at the replication or augmentation of intelligent processes through computational means, either by deductive instructions to the computer or by algorithmic extraction of patterns based on the availability of digital data[119].

ML: Machine learning. An approach within AI which allows (mostly online) computer software systems to improve their operations; for example, the process of customization of book recommendations on a large online bookstore based on customer behaviour.

Digital data (simply: data): Stored digital bits associated with a user or a system's actions within software, expressed in alphanumerical values, ranging from clicks or time spent on a website to age, location, and other labels of an individual, a process's, or an event's representation on the computing device (computer, smartphone, or otherwise). Capturing and storing data enables ML applications.

[119] A thesis which critically explores AI practices might seem in need of a section of definitions. The more I delved into AI research, the more I understood, however, that AI's high degree of "interpretative flexibility" (Pinch and Bijker 1984) is a crucial agent in the field's consistent expectation dynamics. This led me to conduct an historical/archival review of the history of AI's definitions (section 1.3.2), in assistance to everyday examples of what is commonly understood by "AI" today (1.1). Therefore, the reader has to patiently wait until these sections to obtain a more honest understanding of the phrase.

Robotics: A field which shares certain initial aims with AI in mimicking intelligent behaviour, however, mostly focused on the physical support of a system (hardware). Like "AI," "robots" also suffer and benefit by deeply contextual semantics. Generally, depending on AI's terminological strictness of use, it is not necessary that an industrial robot crane employs AI techniques, or that an online recommendation algorithm is considered to be a "robot[120]."

b. 1980s AI-related Programmes

FGCS: Fifth Generation Computer Systems (1982-1993). With heavy financial support from the Japanese Ministry of International Trade and Industry (MITI), this project promised the development of high level AI systems which would develop contextual reasoning, sparking a series of response research programmes around the globe.

SCI: Strategic Computing Initiative (occasionally found as "SC," 1983-1993). Supported by the US's Defense Advanced Research Projects Agency (DARPA), SCI was the American response to Japan's FGCS, promising to create an large-scale AI system equivalent to the telephone.

Alvey Programme: The UK's response to the FGCS (1984-1990), a collaborative effort between government, academia, and industry to compete on the 1980s AI race.

ESPRIT: The European Strategic Program of Research in Information Technology (1983-1998), being the European Economic Community (predecessor to the EU), also a partial response to the FGCS, however, with less emphasis on "AI" per se.

c. Funding Bodies and Schemes Relevant to Contemporary AI Strategies

UKRI: UK Research and Innovation, the main body responsible for fostering partnerships between universities, businesses, research organisations, government, and charities. Its previous equivalent, that several researchers are used to refer to UKRI using the previous body's name, is RCUK (Research Councils UK).

EPSRC: The Engineering and Physical Sciences Research Council, covering fields such as structural engineering, chemistry, advanced materials, healthcare technologies, mathematics, and manufacturing.

ISCF: The Industrial Strategy Challenge Fund, commissioned by the UKRI, is part of UK's industrial strategy to raise "productivity and earning power in the UK" incentivises researchers to apply for research grants in line with the collaborative purposes of UKRI.

REF: The Research Excellence Framework, is a process of expert review undertaken by the UK's higher education funding bodies (hence, UKRI) to assess each research institution's output (such as publications, events, or exhibitions), impact beyond academia, and research environment.

ERC: The European Research Council complements funding activities around Europe, assisting existing national funding bodies and being the main component of the EU's Research Framework Programme. The main difference, according to the ERC website with other national bodies with more nation-specific

[120] Nevertheless, consider the "I am not a robot" captcha method used on several websites which tacitly considers online malicious software as "robots."

funding agendas is that ERC claims to be "investigator-driven" and "bottom-up," ensuring that "funds are channelled into new and promising areas of research with a greater degree of flexibility."

More information about these funding bodies can be found on their respective websites[121]. While a thorough examination of these schemes and bodies' history would exceed the scope of the thesis, it is worthwhile to mention that ERC's finding in 2007 was partly in continuation of the European approach to science and technology funding which dominated the spirit of ESPRIT. Similarly, the approach held by the UKRI for ISCF is very similar to the UK tradition (as expressed in the Alvey Programme) of at once separating industrial and academic interests, however, incentivising collaborations between them.

d. Theoretical Terms

SoE: Sociology of expectations. A field within science and technology studies (STS) which interrogates the impact of future-orientation in science and technology; how futures are negotiated, and how imagined futures shape contemporary decisions. Precursors, variations and adjacent terminologies are explored in section 2.1.

SEE: Studies of expertise and experience. Another subfield of STS that studies the credibility and social construction of expertise, shaping and shaped by experiential traits within a given field. This framework allows for critical investigations on who can or should be considered a credible and/or trustful spokesperson of a field, and how this credibility/trustworthiness is obtained or impacts the field.

[121] UKRI: https://ukri.org
EPSRC: https://epsrc.ukri.org/about/
ISCF: https://ukri.org/innovation/industrial-strategy-challenge-fund/
REF: https://ref.ac.uk
ERC: https://erc.europa.eu

Appendix 2: Interview Schedule

1. **Contextual information and background (Information about the respondent/institution)**

- Can you briefly explain your background and what brought you into the field? Can you give me an example of your recent projects?
- How did you arrive in this department? Why did you choose this department?
- Does your institute have a long-term view on its research objects?
- How do you work with stakeholders and funding bodies? Do you share the same goals?

2. **Ground mapping questions (Nature of AI, controversies, promises)**

- How do you define AI/Machine Learning/Robotics? What do such terms mean to you?
- What has changed in AI while you have been working in the area?
- In your view, does the recent development in AI change society, for example, humans' relationship with technology, between each other, or anything else?
- Are there any obstacles to AI innovation?

- Are you aware of what have been the promises of AI in the past decades?
- What is expected to be delivered by AI now? Who expects this? How likely are such expectations, in your view, to be met?
- Concerns have been expressed in the past about the unrealisability of promises in AI, such as the Lighthill report. Are you aware of those?

3. **Content mining (ethics/challenges, governance, public portrayals)**

- Does AI pose any foreseeable threat? How do you feel about Elon Musk or Stephen Hawking's publicly expressed concerns?
- In your view, is it important to include ethical discussions in AI?
- As you may be aware, the UK has produced with a number of reports concerned with the hopes and fears about AI, which, at the moment suggest that it would be inappropriate to introduce regulatory frameworks; something which differs from EU's stance suggesting a strong regulatory framework of legal liability of robots. Is this important/relevant or not in your view?
- How accurate is the media coverage of AI in relation to your work? What, if anything, is missing or should be added from such portrayals?
- So far there appears to have been limited public engagement with AI. What do you think about public engagement with such technologies? Is it important? Does your institution communicate your work to wider audiences?
- Do you read/watch science fiction? Did it ever inspire your work? Does it distort the understanding of your work, for example, with blockbuster films, etc?
- What do you think is generally missing from AI development in order to achieve a massive breakthrough?
- In the last minutes of our conversation, based on what we have already said, is there anything else you would like to add?

Appendix 3: Consent form

THE UNIVERSITY
of EDINBURGH

Vassilis Galanos
Institute for the Study of Science, Technology and Innovation.
The University of Edinburgh,
Old Surgeons' Hall
High School Yards
Edinburgh EH1 1LZ
Vassilis.Galanos@ed.ac.uk
+7506887026

Date: __/__/20__

Informed Consent for the Project *Expectations in Artificial Intelligence and Robotics: Specialists' Views*

Please tick the appropriate boxes	**Yes**	**No**
1. Taking part in the study		
I consent voluntarily to be a participant in this study.	☐	☐
I understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	☐	☐
I understand that taking part in the study involves audio recording of the interview, and that recordings will be transcribed by the researcher.	☐	☐
I understand I can request that I will be sent the audio and transcription files in order to review that for verification of accuracy.	☐	☐
2. Use of the information in the study		
I understand that information I provide will be anonymised and used for the completion of the researcher's doctoral thesis as well as scholarly publications, and similar research outputs.	☐	☐
I understand that personal information collected about me that can identify me, such as my name or where I live or work, will not be shared beyond the researcher's knowledge.	☐	☐
I agree that my information can be quoted in research outputs.	☐	☐
3. Future use and reuse of the information by others		

I give permission for the recordings and transcripts of the interview that I provide to be deposited in the researcher's personal archive within encrypted folders and in an anonymised manner so it can be used for future research and learning. Audio file names will have no identification signifiers and any potential identification traits will be removed from the transcripts, replaced by pseudonyms.	☐	☐

4. Signatures

I have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

Name of participant [IN CAPITALS] Signature Date_____

Name of researcher [IN CAPITALS] Signature Date_____

5. Study contact details for further information

If you have any further queries about the project, please contact the researcher at Vassilis.Galanos@ed.ac.uk or +7506887026. Alternatively, you can contact my PhD supervisor, Prof Robin Williams at +44 131 650 6387 or R.Williams@ed.ac.uk.

Appendix 4: Invitation letter

THE UNIVERSITY
of EDINBURGH

Vassilis Galanos
Institute for the Study of Science, Technology and Innovation.
The University of Edinburgh,
Old Surgeons' Hall
High School Yards
Edinburgh EH1 1LZ
Vassilis.Galanos@ed.ac.uk
+7506887026

Date: __/__/20__

Information about the project *Expectations in Artificial Intelligence and Robotics: Specialists' Views*

Thank you for considering participating in this study. My name is Vassilis Galanos and I am a PhD student based in the department of Science, Technology and Innovation Studies at the University of Edinburgh. My research focuses on expectations about artificial intelligence and robotics and how these expectations may influence technological developments.

Why is your participation important in this research?

I am keen to involve the views of practical specialists in the field. The study seeks to redress the apparent lack of social science research into the perceptions and practices of AI and robotics researchers. A stronger evidence base here may contribute to more effective public and policy debates around developments in this area which otherwise may be unduly influenced by science fiction or media accounts. There are no right or wrong answers; but your opinion as an expert is extremely valuable to this research and integral to the conclusions that can be drawn.

What is the length of this interview and what is involved?

Our conversation will last approximately an hour and will revolve around your own work, AI/robotics and its capabilities, and the relationship of AI research with governmental institutions, as well as ethical challenges of AI.

Are there any benefits or disadvantages for you?

Beyond an opportunity for AI and robotics communities to be voiced, there are no further direct benefits to you if you choose to take part. Moreover, all of the responses as well as your institute and/or laboratory will be fully anonymised so that what you say will serve solely educational purposes. Therefore, your data are fully protected. Your participation is fully voluntary and you have the right to withdraw at any time from the interview.

What about confidentiality?

You can be assured that confidentiality of your views is my main priority and hence, in the consent form provided, I am asking you to agree with the full anonymisation of the information you give to me. I will safely store data and recording in encrypted folders and recordings will be destroyed after the final submission of my thesis. I will send you a copy of the transcription as soon as I produce it and you can have the opportunity to check for accuracy. I can also send you at a later stage a summary of the findings.

How long will the data be kept for?

As said, all quotes used in my research and potential future reports and publications will be fully anonymised. After our interview, I will take the time to transcribe it and send you, if you want to, the transcript, in case you want to check for potential inaccuracies. In case you wish to add something important in your views that has been missing, feel free to let me know. If you so wish, I will be more than happy to send you further outcomes of the research such as digital copies of my thesis when published and/or other research outputs in which your interview has shaped my views on.

Any further questions or concerns?

If you have any additional questions, you can use the contact details presented on the top of this information sheet. Alternatively, you can contact my PhD supervisor, Prof Robin Williams at +44 131 650 6387 or R.Williams@ed.ac.uk.

Once again, I am grateful for your participation in this project.