Doctoral Dissertation
Doctoral Program in Urban and Regional Development (32th Cycle)

De-Individuation of the Modern Subject in the Age of Artificial Intelligence.

The Case of Self-Driving Cars and Algorithms for Decision Making.

Fabio Iapaolo
* * * * * *

Supervisor
Prof. Marco Santangelo, Polytechnic of Turin

Doctoral Examination Committee:
Prof. Simone Natale, Referee, Loughborough University
Prof. Alberto Vanolo, Referee, University of Turin
Prof. Andrea Ballatore, Examiner, Birkbeck University of London
Prof. Vladan Joler, Examiner, University of Novi Sad
Prof. Luca Staricco, Examiner, Polytechnic of Turin

Polytechnic of Turin
March 16th, 2021

I hereby declare that, the contents and organisation of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

……………………………………......
Fabio Iapaolo
Turin, March 16th, 2021

Summary

This research project interrogates shifting conceptions of human subjectivity and evolving forms of technical agency in view of recent advances in the field of Artificial Intelligence. Linking their work to possibilities for positive social, political and environmental change, a significant number of scholars from the posthumanities and the new materialism has attempted to decentre the human subject, alongside the qualities traditionally associated with human exceptionalism, by acknowledging the agency of the nonhuman, including technology. This body of work, however, has fallen short of convincingly accounting for the embodied and embedded agential capabilities of contemporary computational media.

To fill this gap, and calling for the reintroduction of materiality alongside technicality as key dimensions for a critical understanding of the spatialized effects of technology, this research project brings forth the critique of the liberal humanist subject by investigating the integration of AI technologies into existing social and spatial systems. Inter-disciplinary in orientation and theoretical in scope, this research project seeks to reassess theoretical discussions of the complex entanglement of nonhuman agency with (post)human subjectivity entertained within critical theory through a technically-aware investigation of a particular AI system: self-driving cars.

Considered an essential preliminary activity for conducting the case study, this work begins by problematizing anthropocentric notions of AI manifesting both in the early research as well as in the imagination of AI in popular culture. Contrary to widespread claims for AI displacing the human subject, it is argued that, in reality, the liberal autonomous self has long been central to the design and imagination of intelligent machines. Drawing on posthumanist/feminist studies of technology, this study thus provides an in-depth operational discussion of key concepts like intelligence and autonomy. By unveiling the operational logics of the dominant paradigm of AI to date, namely, machine learning, it thus clarifies the ambiguous conceptual overlap between autonomy and automation.

Working at the intersection of social and computer science perspectives on technology, and using as theoretical framework the cognitive assemblage devised by posthumanist scholar Katherine Hayles, this research project thus delivers the first in-depth, technically-aware analysis of self-driving cars by unpacking their material functioning and inner complexity. By mapping out the multiple agents concurrently affecting the vehicle's behaviour, it shows that driving decisions always result from layered (multi-located and multi-temporal) interactions between the human and technology, hence neither can be said to be operating within fully autonomous realms. Countering both claims for human autonomous agency, and dominant views of AI as autonomous technologies to which decision-making power is delegated, it is argued that the most paradigmatic aspect of contemporary automated systems is the unprecedented level of complex imbrication and dynamic entwinedness between human culture, technics, and the environment.

This work ends with a discussion of city-specific spatiality and political materiality through an in-depth investigation of the pre-emptive logics, and present limits, of machine vision and cognition in relationship to urban variegated form and sociality. While countering dominant techno-deterministic interpretations of social innovation and spatial transformation, this work offers insight into a post-anthropocentric understanding of AI—namely, not as an abstract property susceptible of replication within discrete machines, but rather as a distributed property emerging through material interactions occurring among a multiplicity of embodied agents (human, nonhuman, and technological) within/with their sociotechnical environments.

Acknowledgment

For any product of the intellect, it is common practice to attribute its merits and demerits to a single authorship. As for this piece of work, whilst its demerits mainly originate in the author's intellectual and psychological fragilities, its merits must be shared with a large number of people. Some of them will be acknowledged explicitly here.

First and foremost, I'd like to thank the Interuniversity Department of Regional and Urban Studies and Planning (DIST) jointly organised by Politecnico and Università of Turin. This thesis could not have come into existence had it not been for funding sources and assistance it provided.

I'm deeply grateful to my supervisor, Marco Santangelo, the presence of whom has been an inexhaustible source of encouragement since the very beginning of this doctorate project. I'm equally thankful to Matteo Pasquinelli, who provided such a friendly and intellectually vibrant environment during the year I spent at the Department of Media-philosophy at Karlsruhe University of Arts and Design. As a visiting member of the research group Critical Artificial Intelligence (KIM), having the chance to discuss my ideas with Matteo, as well as with other scholars and students attending his seminars, has been crucial for me to catalyse my thoughts and reach this project's conclusion.

I am sincerely thankful to my reviewers, Simone Natale and Alberto Vanolo, whose meticulous observations, reassuring feedback and constructive criticism helped me refine my arguments, strengthen my confidence and extend my scope.

There are many academics whom I have never met in person and yet have been such an inspiration to me. Among these, Katherine Hayles, Rosi Braidotti, Minsoo Kang, Benjamin Bratton, Louise Amoore, Kate Crawford, and Luciana Parisi deserve a special mention.

I owe a special debt of gratitude to Nisse Knudsen, my interviewee, who has passionately provided insight into, and patiently clarified my doubts about, the more technical aspects of self-driving cars.

I am especially grateful to James Bridle, Adam Harvey, and Liam Young and Tim Maughan for permission to use visual materials complementing my text in chapter 4.

I'm thankful to Sam Kinsley from the Geography Department of Exeter University, who gave me the opportunity to present preliminary findings of this research during the thematic session "Geographies of/with AI", which he organised on occasion of the Annual International Conference of the Royal Geographical Society 2019 at the Society in London.

The list of friends and colleagues (these two categories tend to overlap in my mind) I should thank for being emotionally and intellectually helpful throughout all these years is virtually never-ending. A particular debt of gratitude is owed to Louisa Nyman, Federico Piovesan, Samantha Cenere, Chiara Iacovone, Alberto Valz Gris, Diego Vitello, Andrea Mora, Simone Tulumello, Irma Cakalli, Silvia Mascolo, Veronica Altieri, Piotr Ciozda, Konrad Wojnowski, Tommaso Asso, Georgios Angeletopoulos, Corrado Marchigiano, Francesco Mele, Giulio Ferrara, Giovanni Sgarbossa, and Stefano Testa.

A very special thank goes to my sister, Manuela, for being supportive in a myriad of ways, and to my parents, whom I consider to be my best teachers in life.

A final word of thanks goes to my partner and friend, Lidia, who has always been there when I needed her most.

This thesis is dedicated to my parents.

Contents

List of Figures

Chapter 1

1. Introduction

Turning and turning in the widening gyre
The falcon cannot hear the falconer;
Things fall apart; the centre cannot hold;
—William Butler Yeats, *The Second Coming* (1920)

For instance, on the planet Earth, man had always assumed that he was more intelligent than dolphins because he had achieved so much—the wheel, New York, wars and so on—whilst all the dolphins had ever done was muck about in the water having a good time. But conversely, the dolphins had always believed that they were far more intelligent than man—for precisely the same reasons.
—Douglas Adams, *So Long, and Thanks for All the Fish* (1984)

1.1 The question concerning technology within and beyond the urban

Over the past two decades the expression 'smart city' has been trending in scholarly and popular debates on the role of technology in driving present and future urban development. The prolonged controversy over the meaning(s) of the expression notwithstanding (Hollands, 2008), in recent years a certain terminological consensus has been reached. The smart city is commonly understood as an umbrella term encompassing different genres of computation—networked sensors, ubiquitous communications, Internet of Things (IoT), smart grids, big data analytics, and algorithms—applied to the systemic management and real-time control of different policy domains (Tulumello and Iapaolo, 2021; Halpern et al., 2017; Melgaço and Willis, 2017; Batty et al., 2012; Komninos, 2002). Sometimes used interchangeably with such expressions as 'intelligent–', 'conscious–', or 'sentient city', the smart city has gained undisputed dominance amid discussions of the interplay between technology and urban spatial, social, and political transformations, establishing itself as the main reference grid for framing the spatialized effects of technology on city life.

Over the years, such popularity has earned the term endorsement from many sympathizers, including multinational hi-tech corporations and local government agencies, and even more bitter enmity from a proliferating number of detractors, in particular critical scholars from human geography and urban studies. On the one hand, advocates of what some have termed the “smartness mandate” (Halpern et al., 2017, 107; see Vanolo, 2014, on ‘smartmentality’) have pursued a vision of ‘municipal omniscience’ (Tulumello and Iapaolo, 2021) to be achieved “through the integration of human beings and machines into a seamless “Internet of things” that would generate the data necessary for organizing production and labor, enhancing marketing, facilitating democracy and prosperity, and—perhaps most important—for enabling a mode of automated, and seemingly apolitical, decision-making that would guarantee the survival of the human species in the face of pressing environmental challenges”. On the other, in open opposition to such claims, more critical perspectives have warned against the normative power of the smart city (e.g., Vanolo, 2014; Hollands, 2008), an urban ‘sociotechnical imaginary’ (see Jasanoff and Kim, 2015) led by aspirations for “a more rational, scientific and depoliticised way of understanding and intervening in the city” and yet far distant from the reality of the “actually existing smart city” (Shelton et al., 2018, 15). For years now, critical urban scholarship has tirelessly attempted to unpack the rhetorical power of the smart city, assuming as main subject of criticism the salvific rhetoric inherent to it. Specifically, the ‘solutionist’ orientation of the smart city toward present and future crises, whether real or perceived (Tulumello and Iapaolo, 2021; Halpern et al., 2017; White, 2016; Morozov, 2013), has been shown to be instrumental to local strategies and practices of neoliberalisation (Pollio, 2016; Rossi, 2016), technocratic reductionism (Söderström 2014), and over-securization of public spaces (Armao, 2013).

In developing more critical views on technology, and revealing the logics of city branding (Vanolo, 2018) and urban entrepreneurialism (Harvey, 1989b) which the smart city supports, the work done by critical urban scholarship surely deserve acknowledgment. Notwithstanding, some relevant shortcomings cannot be neglected. Part of the problem is that, in highlighting the (perhaps unavoidable) gap existing between what the smart city promises and what it actually has to offer, the main, if not exclusive, focus of critique has been the subtle mobilisation of technology as ‘corporate storytelling’ (Söderström 2014; cf. Leon and Rosen, 2020). Understanding technology as a discursive formation certainly has its uses when it comes to fully comprehend its legitimising power in promoting market-driven urban agendas. Yet, such approaches seem inadequate for an in-depth assessment of the material effects and transformative capacities of contemporary computational media, which have in fact remained largely overlooked and untheorised. Of course, this is not to say that the corporate investment in technology “for entrepreneurial and regulatory effect” (Kitchin, 2014, 1) is to be dismissed as irrelevant. Yet, by focusing almost exclusively on this aspect, critical urban scholarship has so far given little to no attention to the ways in which power is enacted by and through technical systems themselves.

In our contemporary moment, when the generative and world-shaping capabilities of computational systems seem to be increasing at an accelerating rate, the limitations of similar approaches become readily apparent (Amoore and Raley, 2016). That private actors, and in particular multinational corporations, have been capitalising on notions of ‘smartness’ and ‘intelligence’ is undeniable (Halpern et al, 2017). Yet, it also must be admitted that, apart from a very few exceptions (e.g., Lynch and Del Casino, 2020; Ash, 2017; Webb, 2017), critiques of the city-scale deployment of computation have for far too long avoided problematizing how ‘cognition’ or ‘intelligence’ are operationalised within urban spaces, passing over the complex technological reality existing beyond mere labelling practices.

In my view, one relevant problem connected with the use of the smart city as the main organising concept for analysing the complex and rapidly evolving entanglement of human and technology is that, throughout the years, it has polarised the debate around (ideological) positons that tend to be either optimistic or pessimistic toward technology per se. Nowadays, however, both positive and critical accounts of the smart city seem to be of very limited conceptual use for understanding sociotechnical transformations taking place at and beyond the urban scale. On the one hand, by limiting their enquiry to the ‘sociology’ (Amin and Thrift, 2002) and political economy of the city, critical scholars from geography and urban studies have been unable to “interrogate both the evolving agency of technical systems and the complex sociotechnical milieu within which (post)human consciousness is entangled” (Lynch and Del Casino, 2020, 383; see Rose, 2017). On the other, business perspectives on urban technology have “focus[ed] exclusively on their sensorial capabilities but little is given to their cognitive capacities” (Webb, 2017, 187). Smart city advocates have indeed pursued a centralised, vertical approach to urban governance and city-making (Zandbergen and Uitermark, 2019; cf. Krivý, 2018; Batty, 2014), where what is ultimately demanded of computation is to “sense and gather as much information as possible” (Webb, 2017, 187; see Mattern, 2015, on the history of the urban dashboard) to be presented to a human intelligence for an optimised management of one’s individual life or cities as wholes (see Morozov, 2013, on contemporary trends in self-quantification). Yet, such instrumental, human-centred view of technology seems to be reluctant to acknowledge the ongoing constitution of “novel forms of distributed authorship” entangling humans and technical systems in “composite collaborations” (Amoore, 2020, 20; Hayles, 2017; Bratton, 2016) through which knowledge is produced, decided and acted upon.

If up until recently the adjective ‘smart’ has been attached to different genres of devices and spaces, including cities, to connote their sensing and wireless communication capabilities, nowadays the term itself risks sounding old-fashioned. In light of recent advances in such fields as machine learning, robotics, computer vision, synthetic cognition or Artificial Intelligence—to use a more hyped expression under which all the former can be grouped, and given the potential of such technologies to profoundly and (supposedly) autonomously reshape many aspects of society, it has become common practice to refer to contemporary computational media not just as smart, but as 'intelligent' or 'autonomous'. Beyond the hype, this new terminology speaks for the "sensing abilities, cognitions, and decisional powers" (Hayles, 2017, 132) now present in growing degrees in different genres of technical devices or systems.

There's a specifically urban dimension this situation entails, for cities have long been "pioneers in embedding digital infrastructure and systems into their urban fabric" (Kitchin, 2014, 1), thus enabling the emergence of different modalities of machine vision, cognition and decision-making imperceptibly affecting daily practices and people's experience of space (Kitchin and Dodge, 2011: Thrift, 2007; 2004: Virilio, 1994). Accordingly, amid geographers (see Whatmore, 2002), notions of agency have been widened beyond the conventional site of the human individual and to the inclusion of technological mediators.[1] But because today algorithms or technical systems are being integrated into relevant decision-making processes seemingly taking place with no human-in-the-loop (Amoore, 2020; Amoore and Raley, 2016), it seems no longer sufficient to say that technological artefacts qualify as 'mediators'. As popularised by contemporary mainstream debates about labour automation (e.g., robots replacing human workers) or machine ethics and morality (e.g., autonomous decision-making in self-driving cars or drones with lethal capacity), the point is less that we humans are increasingly delegating agency to technology (something we've been doing for millennia now), than that technical agency occurs beyond direct human cognition and control.

That the 'smartification' of cities, through the integration of sensory and data-processing technologies into our everyday spaces, has enabled the proliferation of *artificial intelligences* of various kinds is a position which many share (e.g., Lynch and Del Casino, 2020; Thrift, 2019; Amin and Thrift, 2017; Bratton; 2017a; Webb, 2017; Kitchin and Dodge, 2011). Yet, the challenges that today AI poses seem to outscale the urban dimension. This is not just because urban scale AI/automation is always reliant on planetary resources and cloud infrastructures, as well as a globally dispersed invisible workforce (Crawford and Joler, 2018; Bratton, 2017; Brenner, 2014),[2] but also, and crucially, because the prospect that technical systems will soon be, or perhaps already are, operating in

[1] See Latour (2007, 116) on the production of power by technological mediators.

[2] In his recent book *The Stack: On Software and Sovereignty*, Bratton (2016) introduces the concept of the 'stack' to refer to an emerging multi-layer planetary scale computational apparatus and a new governing architecture. According to the author (Bratton, 2016, 4-5): "[p]lanetary-scale computation takes different forms at different scales—energy and mineral sourcing and grids; subterranean cloud infrastructure, urban software and public service privatization; massive universal addressing systems; interfaces drawn by the augmentation of the hand, of the eye, or dissolved into objects; users both over-outlined by self-quantification and also exploded by the arrival of legions of sensors, algorithms, and robots. Instead of seeing all of these as a hodgepodge of different species of computing, spinning out on their own at different scales and tempos, we should see them as forming a coherent and interdependent whole. These technologies align, layer by layer, into something like a vast, if also incomplete, pervasive if also irregular, software and hardware Stack".

ways independent from the human and society fundamentally undermines long-established understandings of the human, technology, and our sociotechnical environments across scales, demanding attention at multiple levels of analyses.

These days, we are witnessing a “technological transfiguration of the world” (Baudrillard, 1994, 25) unfolding at such a speed that it is becoming difficult to map out its state-of-the-art and anticipate the cultural, political and spatial transformations implicated in the process (see Gomez-Luque and Ghazal Jafari, 2017). Perhaps, the principal reason why the present moment might seem so undecipherable is that we haven’t fully developed yet a proper vocabulary for framing the sociotechnical environments we are setting out to create. Or, to say it with Bratton (2017a): “our technologies have advanced beyond our ability to conceptualize their implications”. In her *Manifesto for Cyborgs*, Donna Haraway (1985, 177) introduced the metaphor of the ‘cyborg’, a symbiotic creature resulting from merging the human with technology, to challenge Western traditional dualisms structuring consolidated “logics and practices of domination”. By blurring the human–machine distinction, Haraway (1985) sought to explore possibilities of political liberation and empowerment for women and all those who had been traditionally identified as others of ‘Man’—simultaneously a historically specific mode of being and the hegemonic model of the human (Braidotti, 2013). At the time her writing, that is, in the mid-1980s, for Haraway the cyborg represented an exception to the norm, that is, to the male, white, straight, rational subject of traditional humanism. Considering the extent to which our everyday lives are now inextricably imbricated with, dependant on, and augmented by such technologies as personal computers, smartphones or the Internet, one might argue that the cyborg went mainstream—although such symbiotic alliance between humans and machines has not entailed a radical amelioration of our still racist, unsustainable, and male-dominated societies.

Today, however, the metaphor of the cyborg appears to have relatively little conceptual power for “resetting the stage for possible pasts and futures” (Haraway, 2004, 47). Like all hybrids, it entails a mere recombination of rigidly defined pre-existing categories (e.g., human–machine; digital–analogue; virtual–real; nature–culture). But since today AI technologies seem to embed qualities once identified uniquely with the human(ist) subject—intelligence, cognition, and decision-making, this situation entails a radical change in perspective regarding how the human has been traditionally understood in comparison to technology, and vice versa. In times of profound anthropological and technological changes, and under the spectre of prolonged economic, environmental, political and health crises, it's now more than ever necessary to device new analytical schemes and make imaginative efforts in order to rethink the status of both the human and technology, for the two are presently being redefined, conceptually and materially, through their mutual relation.

There’s a strong sense of uncertainty marking the present moment—which, given the interest around the topic, might arguably be named the Age of Artificial Intelligence (hereafter AI). Paramount here is a sense of displacement of long held assumptions about the human subject, and its presumed specialness and centrality in the process of "world-making" (Haraway, 2008). Currently, perspectives on AI are split between technophile and technophobic positions: the former fuelled by transhumanist fantasies of human enhancement through scientific development (see Hayles, 2011), the latter imbued with anxieties about the possible advent of super-intelligent machines superseding the human species (cf., Geraci, 2008). Yet, a third, both critical and propositional perspective, and which I fully endorse, is still possible: one which sees the displacement of a certain conception of the human—the liberal subject of traditional humanism—and the rejection of self-centred individualism as an opportunity to design more sustainable and inclusive sociotechnical alternatives. In this regard, relevant to the present discussion are notions of post-anthropocentrism and posthumanism. Amid critical theory and science and technology studies, these two partly overlapping concepts have been used with reference to theories and practices focused on displacing the humanist liberal subject alongside with the characteristics traditionally identified with human exceptionalism (e.g., intelligence, rationality, free will, and autonomy), in particular by stressing the relational character of human existence and its dependence on, and ethical bond with, nonhuman others (Braidotti, 2013; Haraway, 1999). Building on posthumanist scholarship alongside feminist studies of technology, this thesis thus intends to interrogate shifting conceptions of the human subject *vis-à-vis* recent technological advances in the field of AI. Explicitly, the research question this project sets out to answer is the following: in our present moment, how can human subjectivity be rethought in a post-anthropocentric and posthumanist way? Or, alternatively: do (and if so, how do) present advances in AI technologies contribute to displacing the humanist liberal subject?

In what follows, I shall start by discussing the notion of the 'posthuman' theorised by feminist philosopher Rosi Braidotti (2013), a leading figure in critical theory and one of the key contributors to the so-called 'posthuman turn' in the humanities and social sciences (see Braidotti, 2017). Given its focus on technology as one of the key factors contributing to displacing the liberal view of the self (cf., Hayles, 1999), the posthumanist critical theory developed by Braidotti (2013, 1) might provide some useful conceptual tools for framing the ongoing decentring of the human subject and simultaneously mapping out evolving forms of technical agency "under the double pressure of contemporary technological advances and global economic concerns".

1.2 Posthumanism(s)

My intent here is to cast light on the way Braidotti employs the notion of the 'posthuman' to challenge anthropocentric understanding of the human individuated subject as the exclusive seat of agency and the sole locus of political consideration and ethical concern. One original contribution sought here is to establish a triangular comparison between humanism, critical posthumanism theorised by Braidotti (2013), and transhumanism, in order to elucidate contending views on the posthuman and technology found in these latter two. Indeed, whilst it's true that the term 'posthuman' generally denotes a departure from long-held theorisations of human subjectivity found within Western thought, in reality the meanings and ideologies associated with it differ substantially depending on the context in which the term is used.

One initial difficulty in dealing with the notion of the posthuman is that it lacks univocal conceptualisation. Indeed, it has been used within and across different contexts, ranging from speculative fiction, to contemporary art (see Deitch, 1993), critical theory (e.g., Morton, 2013; 2010; Wolfe, 2010; Meillassoux, 2008; Harman, 2002; Hayles, 1999), and business "discussions of robotics, prosthetic technologies, neuro-science and bio-genetic capital" (Braidotti, 2013, 2). Within the humanities and social sciences, there are, of course, many interpretations of the posthuman. Distancing herself from both, Braidotti (2013, 38-39) identifies two main strands of posthumanism, one named "negative", the other "analytic". To the former belong contemporary liberal thinkers (e.g., Nussbaum, 2010) who invoke a return to traditional humanist values in the belief that economic globalization and the free market have renewed a sense of cosmopolitanism and political interconnectedness among humans around the globe. The latter encompasses approaches that, by tackling the posthuman condition from a merely scientific angle, lack in-depth investigation of its epistemological and political implications (Braidotti, 2013; cf., Verbeek, 2011; Rose, 2017; Franklin et al., 2000). Thus, to avoid terminological ambiguity, throughout this section the terms 'posthuman' or 'posthumanism' will be used, unless otherwise specified, with specific regard to critical posthumanism theorized by Braidotti (2013).

For the purpose of disambiguation, it's also worth noting that, the profound divergences existing between the two notwithstanding, in mainstream debates posthumanism and transhumanism are frequently confused. Beside being both commonly associated with cyborg-related imagery, contributing to such confusion is the fact that the term 'posthuman' has been introduced by transhumanists themselves (Moravec, 1988) to describe a hypothetical postbiological species resulting from the increasingly pervasive prosthetization—or cyborgization—of the human body (e.g., Bostrom, 2003; Fukuyama, 2002). For transhumanists, the symbiotic association between the human and technology would indeed enable the transition (hence the term 'transhumanism') toward a new model of the human which is precisely defined as 'posthuman'. From this perspective, what qualifies as posthuman is a new species endowed with enhanced cognitive and psychological capabilities and capable to progressively transcend corporeal finitude until reaching the ultimate goal, namely, immortality (see Caronia, 2008; Hayles, 2011).

Admittedly, the transhumanist movement has its roots in the humanist tradition, of which it represents a hi-tech extension and with which it shares an anthropocentric and teleological vision of history as a cumulative progress driven by scientific reason and technological accomplishments. Traditionally, within the humanist framework, the human is awarded dominion over nonhuman others, and absolute centrality within the scheme of things—a situation visually rendered through Leonardo da Vinci's iconic Vitruvian Man: male, white, heterosexual, able-bodied, and positioned exactly at the centre of the world (Braidotti, 2013; Marchesini, 2009). Structured around strict dualisms (e.g., subject/object; male/woman; nature/culture), within the humanist scheme of thought ontological and anthropological differences play a constitutive role in defining the human subject (or rather Man), which identifies itself as distinct from, and thereby superior to, "the sexualised other (woman), the racialized other (the native) and the naturalized other (animals, the environment or earth)" (Braidotti, 2013, 27).

In continuity with Western individual-centred ethical and political frameworks, transhumanist ideology is highly reliant on such values as rationality, free will, and autonomous agency. Yet, it emancipates itself from the humanist view of the self by exasperating the protean and perfectible character of the human, "that is the idea of a total lack of limits, or the possibility of any destination" (Marchesini, 2009, 8).[3] For transhumanists, the main limit to human existence is corporeal decay, technology is the means by which it can be overcome, and the posthuman condition is the ultimate desirable outcome of human evolution. Here, the affix *post* denotes a silicon progeny completely hybridized with technology to the point that it can no longer be identifiable with the human species. And yet, like its biological ancestor, the posthuman progeny is destined to be the absolute protagonist of a macro-narration of universal emancipation and, at the level of the individual, personal fulfilment.

This is where the first deep crack between transhumanism and critical posthumanism is revealed. In elaborating her critical posthuman theory, Braidotti (2019; 2013), who defines herself as essentially an anti-humanist, aims to bring to completion the work of deconstruction of the Enlightened universal subject initiated by Foucault (1973) and to which feminist, ecological, postcolonial and queer studies further contributed. In sharp contrast with Western traditional thought which identifies the subject with the human individual only, building on Deleuze (1994) and Deleuze and Guattari (1987), the author (Braidotti, 2013, 60) proposes a relational, nomadic model of the subjectivity "which is not confined within our species, but includes all non-anthropomorphic elements". Linking her philosophical project to political aims, what her relational model of the subject pursues is to displace both anthropological hierarchies and species supremacy. As for this second point, she does so by locating the human on a continuum with nonhuman lifeforms. In a bid to develop a cross-species and cross-entity ethical framework, the author thus extends agency and subjectivity to nonhuman (e.g., plants, animals, matter in general) and technological others on the grounds that matter is unique, vibrant, and inherently intelligent.

What critical posthumanism and transhumanism share is a certain technophile attitude and a great interest in exploring the anthropological implications and cultural impact of advanced technologies. Yet, whilst transhumanism maintains an instrumental understanding of technology serving both individual and collective

[3] My translation from Italian.

aims along a linear path of societal advancement, critical posthumanism firmly rejects such anthropocentric and finalistic reading of history, be it guided either by the human species or its postbiological progeny. Accordingly, technology ceases to be understood as something merely mediating or enhancing human life. For Braidotti (2013), indeed, the world is always concurrently produced by human actions together with nonhuman forces, including technology, in contingent and non-finalistic ways. This means that "[e]lements as disparate as organic bodies (a tiger, a human), things (a mountain, the wind), immaterial things (a thought, desire or feeling, 'discourse' or ideology) may all be regarded as constituent parts of a relational material universe" (Alldred and Fox, 2019, 693). In other words, inheriting from Spinoza a monistic interpretation of the world, and borrowing from biology notions of autopoiesis and self-organisation (Guattari, 1995; Maturana and Varela, 1972), Braidotti elaborates a materialist and vitalist ontology which entails a radical redistribution of agency between humans and nonhumans, thus erasing traditional categorical distinctions existing between the two.

1.3 New materialisms

This section provides a preliminary survey of three authors, namely, Latour (2014, 2005, 1993), Amin and Thrift (2017, 2002), and Morton (2013, 2010), who, in a similar manner to Braidotti (2013), and other posthumanist scholars alike (e.g., Hayles, 2017, 1999), "share a commitment to giving agency to the nonhuman as a necessary corrective to centuries of Western philosophizing that attributes agency only to a specific kind of human: the male, white, heterosexual sovereign subject, capable of rational thought unencumbered by material objects, whether tools or his body" (Rose, 2017, 781). Strictly speaking, Latour, Amin and Thrift, and Morton cannot be labelled as 'posthumanist' (and in fact do not define themselves so). Yet, they display a great posthuman sensibility and orientation in that "they take up the posthumanist challenge by both emphasizing the agency of digital technologies and substituting the agency of the sovereign subject with other concepts" (Rose, 2017, 781). To be more precise, Latour, Amin and Thrift, and Morton are prominent figures, like Braidotti herself, from the broader new materialist paradigm, comprising scholars from different disciplines all committed to place nonhuman agency at the heart of social and humanistic enquiry and to give prominence to materiality over discursive practices. For what concerns the three authors which I intend to discuss here, they all take as starting point of reflection and political imagination the fact that we now live in the Anthropocene—namely, "the coincidence of human history and terrestrial geoforming" (Morton, 2013, 9). One reason for doing so is that the present geological epoch is presenting the human species with a paradoxical situation, forcing us to acknowledge both the pivotal role we humans play in global environmental change, and our very limited capacity to mitigate the destructive effects that our presence on the planet engenders. Common to these authors is the firm rejection of human self-attributed specialness, which is often “seen to have led to the current ecological crises […] and the incapacity to think through and adequately engage with them” (Plate, 2020).

With regard to Latour (1993), behind his proposal for a “Parliament of Things” lays the idea that modern democratic theory and political ethics are grounded on a series of artificial distinctions, such as those between active humans and passive nature, subjects and objects, and nature (the given) and culture (the socially constructed). For Latour (1993), modern constitutions are grounded on the false assumption the human subject is the sole entity capable of giving meaning to the world, and thereby the only one legitimised to have political claims. In other words, they “invent a separation between scientific power charged with representing things and the political power charged with representing subjects” (Latour, 1993, 29). Consequence of this is a tendency to value human life more so than other-than-human lifeforms (plants, animals, objects). In taking position in favour of the emancipation and rights of nonhumans, Latour (1993, 144) thus speculates about a possible Parliament of Things, which he defines as “a place where both humans and nonhumans can be represented adequately”.

For Latour, recognising our co-dependence with nonhumans has implications which are both political and epistemological. One the hand, this would be a first step for a true political ecology to emerge, one revolving not just around human interests, but to the benefit of all entities in the world. As in Braidotti (2013), this presupposes an extended definition of life comprising humans alongside animals, plants, things and matter in general. On the other, criticising those whom he calls the “sociologists of the social” (Latour, 2007, 86), he invites his colleagues from the social sciences to displace human activity as the main site of sociological enquiry, and to focus instead on material objects and the wider sociotechnical assemblages that humans and nonhumans together bring into existence, so that “the study of society therefore moves from the study of the social as this is usually conceived, to a study of methods of association” (Latour, 1984, 264). For Latour (2007), indeed, social theory can only be advanced by analysing the complex relationships through which human and nonhuman entities ‘mould’ each other and, together, give shape to the world as it is. Modernity, instead, by imposing the systemic application of rigid dualisms for interpreting reality, cannot but lead to the dilemma of either ‘purification’ or ‘hybridization’. The former entails maintaining that there is indeed a clear-cut distinction between subjects and objects, and nature and culture, although, Latour argues, this situation can be easily proven to be false for on a closer inspection the world cannot but reveal itself to be too complex to be analysed in terms of binaries. Hybridization, conversely, means developing new concepts by recombining old ones, but “[i]f we consider hybrids, we are dealing only with mixtures of nature and culture” (Latour, 1993, 30).

Hybridization, however, is a trap Latour himself seems to fall into. Indeed, whilst maintaining that the there is no such distinction between (active) subjects and (inert) objects—for all entities in the world exist along a nature-culture continuum, he nonetheless seems unable to fully abandon the same categories he wants to displace. In proposing his own terminology to discuss the relationship between humans and nonhumans, in fact he attributes to them the status of quasi-subjects and quasi-objects respectively. All in all, subjects remain subjects, although downgraded to quasi-subjects shaped by the objects they encounter in world; and objects remain objects, yet elevated to quasi-objects given their capacity to affect and mediate human life. Latour, in my view, falls short of introducing a new convincing vocabulary for rethinking human/nonhuman agency beyond the liberal framework. In particular, it strikes me that, in a bid to debunk modernist claims for free will, he first observes that personal autonomy is an illusion, for the "nature of things" always "determines, informs and moulds" the subject (Latour, 1993, 53), and then argues in favour of granting rights and autonomy to nonhumans, ultimately projecting the liberal subject onto the realm of the nonhuman.

The second body of work taken into account here is that of Amin and Thrift (2017; 2002). In the recently published book *Seeing Like a City* (2017), their commitment to debunk human exceptionalism departs from the acknowledgement that *homo sapiens* is but an 'accidental product' of evolution, rather than its culminating point. While calling for a definitive abandonment of essentialist conceptions of the human as a transcendental category possessing fixed traits, the authors (Amin and Thrift, 2017, 68, italic in the original) stress the productive role urban of infrastructure in producing human subjectivity "based on the fundamentally *associative* ability of cities to mix and match through a pidgin of subjects and objects". Accordingly, the human subject is thought of simultaneously as a spatial actor and a spatial product "occup[ying] a world of things which mould it at the same time as they are moulded by it" (Amin Thrift, 2017, 79).

Noteworthy is the use of a terminology similar to Latour's, with whom they have in common an emphasis on relationality and networked processes, based on the idea that agency is not an attribute that individuated entities inherently possess, whether human or nonhuman, but rather as a process entailing cross-entity and cross-species relational combinations and interactions. As Braidotti (2013), Amin and Thrift (2017, 69) mobilise notions of lively materiality, which they use for discussing the adaptive character of urban infrastructure, which is not "dull and inert" but "lively", and autopoiesis, in particular to connote cities' capabilities to connect and reassemble the organic and the inorganic in more-than-human identities and subjectivities. Central to their work is the adoption of a spatialized, post-anthropocentric notion of intelligence, which applies to urban environments themselves (Amin and Thrift, 2017, 82-83, italic in the original):

> «Through a potent mixture of increased linkage between things, combined with a mixture of sensors, screen and other forms of display (like haptics) being conjured up on any surface on demand as well as other 'smart' forms of matter like smart dust and quantum dots enabled by ubiquitous electronics and software, cities are increasingly capable to

> *think*, not in the same way as human beings, to be sure – but in any case it is hardly likely that human modes of thought cover all the possibilities of thinking. […] In particular, cities come to think differently in a post-human world and, above all, they do that through a change in their main channels of reproduction and collectors of mass and influence, namely the framing ley lines of infrastructure».

Possibly, the main original contribution the authors seek to provide is the acknowledgement that cities, now provided with ubiquitous sensory and cognitive capabilities embedded into the very urban fabric and distributed citywide, must cease to be considered as mere sites, or mediator, of human life. Rather, cities in themselves acquire the status of subject, given their capacity to enact and host non-anthropomorphic modalities of perception, abstraction and action. Here, human's is considered but a particular scale and mode of thinking, one among the many different kinds of intelligences inhabiting urban spaces, while urban intelligence, *in toto*, always exceeds the cognitive capabilities of its constitutive parts (whether human, nonhuman, or technological). All in all, Amin and Thrift maintain that intelligence is by no means to be understood as a human monopoly, and, in this way, they seek to further infringe on human exceptionalism. The authors, however, do not provide a working definition of human and nonhuman intelligence, nor develop an in-depth theory of subject formation. Focused on urban computation, in reality what they seem to do is to update their previous conception of the city as an analogous to a living machine (Amin and Thrift, 2002) in light of recent theoretical advances within science and technology studies and the new materialism.

I want to conclude this section by discussing the concept of 'hyperobjects' theorised by Timothy Morton (2013), a new materialist scholar from speculative realism and object-oriented ontology (see Shaviro, 2014; Bennett, 2010; Meillassaux, 2010; Brassier, 2007; Harman, 2002). Committed to non-anthropocentric thinking, one common thread of thinkers operating under the banner of speculative realism is the effort they put into dethroning human exceptionalism by privileging nonhuman ontologies and subject-object relations. Specifically, speculative realism stands against both modern philosophies of access and post-Kantian correlationism. As for the former, speculative realists refuse to concede the human the role of meaning-giving subject, thus affirming the autonomy of objects beyond our senses and knowledge. As for post-Kantian correlationism, they reject "the idea according to which we only ever have access to the correlation between thinking and being, and never to either term considered apart from the other" (Meillassoux, 2010, 5), which implies that philosophers and scientists are denied *a priori* the possibility to explore what lies beyond direct perception and cognition.

Much worried about the ongoing environmental crisis, Morton's conceptualization of hyperobjects can be read as a spatialized elaboration of Harman's (2002) 'object-oriented ontology'. Indeed, Morton departs from Harman's (2002) critique of Heidegger's idealism, the latter seen as the apex of anthropocentric correlationism for not only it maintains that there is indeed an inaccessible gap between the truth of things and phenomena, but also that objects only exist insofar as there is out there a human subject perceiving and thinking of them. In Harman's words (as quoted in Morton, 2013, 15), "idealism […] is unworkable, since there exist real things whose core reality is withdrawn from access, even by themselves". Morton (2013, 15) uses the term hyperobject as a synecdoche for Harman's objects (an all-encompassing term comprising everything, both human and nonhuman). Compared to the latter, the distinctive character of hyperobjects is that they are "so massively distributed in time and space as to transcend localization, such as climate change and styrofoam" (Morton, 2013, 30). In other words, they exist beyond human-scale spatiality, temporality and comprehension, hence we humans can only ever be aware of, and intervene on, their localised effects and manifestations. Hyperobjects, in Morton's (2013, 15) words, "are not simply mental (or otherwise ideal) constructs, but are real entities whose primordial reality is withdrawn from humans". One concrete example of hyperobject the author brings is planet Earth itself, whose "geological cycles demand a *geophilosophy* that doesn't simply think in terms of human events and significance" (Morton, 2013, 7, italic in the original).

As Morton himself is ready to admit, one immediate problem connected with his theory is that, in a manner similar to post-Kantian correlationism, it inevitably leads to irreductionist thinking, for in the end one cannot but acknowledge that the gap between phenomenon and thing is in fact irreducible. Still, he refuses to grant to the human subject the possibility to access a privileged, transcendental space of signification. In other words, even if we acknowledge the non-correspondence between phenomenon and being, that does not mean that material reality is reducible to the correlation between human thinking and the world. Morton (2013, 165) thus calls for a radical "flat ontology" where qualitative differences between human and nonhuman agents are completely eliminated. It should be noted that he does so in a purely philosophical sort of manner, providing no strong (empirical) rationale to justify his position, nor discussing the practical implications in terms of research approaches, methods and methodologies. According to Morton (2013), all entities, whether human or nonhuman, including technology, qualify as objects: they co-produce the world, even if there may be no possibility even for them to fully access (their own) reality. Focused almost exclusively on the aesthetic experience of materiality, rather than materiality itself, all in all Morton provides no explanation about how humans and nonhumans, subjects and objects affect one another, whilst much emphasis is placed on their irreducible strangeness to themselves and each other.

1.4 Contributions, methodological notes and chapter outline

In my view, the work conducted by the authors from the posthumanities/new materialism whom I have discussed so far surely deserve recognition. Having many of the same concerns, I do particularly appreciate the idea of carrying out theoretical projects which, in parallel with decentring the humanist subject, head in the direction of constructive social, political and environmental change. Each attacking human exceptionalism on a different front, tied to their theories are indeed more wordly preoccupations regarding the political exclusion of marginalised groups (Braidotti), the anthropogenic impacts on Earth's ecosystems (Amin and Thrift, Braidotti, Latour, Morton), and the social and environmental costs of planetary urbanization (Amin and Thrift). In furthering the critique of the modern project alongside the 'bad practices' it has long supported, the thing I value the most is the emphasis all these authors place on material processes, rather than discursive practices, as the baseline for advancing theory and critique. As discussed before, I believe there's much to be gained by reintroducing materiality alongside technicality as key analytical dimensions for a better comprehension of contemporary sociotechnical transformations, especially considering the scant attention material processes and their complex interactions with human culture have so far received in the social sciences in general, and urban and geographical scholarship in particular.

Yet, for the purpose of this research, and given its twofold interest in human subjectivity and AI, some weak spots identified in the literature explored require attentive consideration. Noteworthy in my view is that, although much effort goes in the direction of replacing anachronistic dualisms with relational and systemic theories and methods (Latour, Braidotti, Amin and Thrift), in the end one crucial binary ends up being reintroduced: human–nonhuman. In fact, much work goes in the direction of deconstructing cultural hegemony of 'Man' and replacing anthropological hierarchies with horizontal diversity—and equality—in terms of race, ethnicity, gender, gender expression, sexual orientation, and so on (Braidotti). At the same time, however, despite claims for abandoning binary classifications and instead locating ontological differences along a human–nonhuman continuum (Braidotti, 2013; Latour, 1993), undeniable is the impulse either to completely erase human-nonhuman qualitative distinctions (as it happens in the case of Morton's (2013) 'flat ontology'), or to treat the nonhuman as an internally undifferentiated category encompassing everything other-than-human, including technology.

Perhaps for this very reason, what usually remains out of the picture is technicality, by which I mean in-depth theorisation of the agential capabilities embedded in technology in general, and contemporary computational media more specifically. Yet, for a research like mine which aims to interrogate both technical agency and (post)human subjectivity in the historical present, distinctions among different kinds of nonhuman agency are much relevant. Whilst I do see the transformative potential inherent in the idea of granting agency to the nonhuman, including technology, I believe that doing so by invoking ideas such as 'vital materiality' or 'lively matter' (e.g., Braidotti, Amin and Thrift) lacks argumentative strength and, perhaps more important, analytical accuracy and historical specificity, for such notions can in fact be used with reference to everything existing in this world across many scales (a lake, a water molecule, an earthquake, climate change, a military drone) and in disparate historical periods.

I find it also pertinent to note that, perhaps in the attempt to give human exceptionalism the coup de grace, abundant in the literature is the use of terms like autonomy (Latour, Morton), subjectivity, and intelligence (Braidotti, Amin and Thrift) as qualities that humans and nonhumans share. This is particular evident in Braidotti (2013), and Amin and Thrift (2017), the latter conceiving of cities as essentially thinking and self-organising subject-entities. However, the way in which such concepts are used typically lacks in-depth theorisation, dangerously veering toward the animistic and anthropomorphic thinking. Their best intentions notwithstanding, in our present moment when terms like 'intelligence' and 'autonomy' structure much of the public, institutional and academic debate on algorithmic decision-making, and considering how distorted is today the public reception and perception of AI, such conceptual vagueness is hardly excusable. Actually, in my view, it's in itself contributing to much of the present confusion about AI, thereby hindering the possibility to investigate its socio-spatial effects in more productive and constructive ways. Indeed, I'm convinced that, precisely because their meaning is too often taken at face value and left unpacked, within the humanities and social sciences terms like 'intelligence' and 'autonomy' end up being abundantly yet confusingly used. This, in turn, might perilously lead to mystified understandings of technology and misdetections of its broader cultural and spatialized effects, especially if one considers that, as we shall see later, anthropomorphic projection is always lurking when dealing with technologies which are said to be capable of automating mundane tasks previously thought to reside exclusively in the domain of human activity, including high-stakes decisions.

Broadly inspired by posthumanist/new materialist literature, yet well-aware of relevant shortcomings existing therein, this research project aims to bring forth the critique of the liberal humanist subject and further debunk qualities traditionally associated with human exceptionalism, in particular autonomous agency. To limit my enquiry, I will focus on a particular category of nonhumans, namely, AI technologies. By doing so, one overall contribution I seek to provide is to problematize and thence reassess broad theoretical discussions about nonhuman agency and posthuman subjectivity entertained within the posthumanities/new materialism through a technically-aware investigation of mundane AI technologies: machine learning algorithms or larger (socio)technical assemblages. While sharing the broad commitment to grant agency to other than human entities, I believe that doing so with regard to AI technologies requires rationalities different than those typically employed in the literature I reviewed. Otherwise, one might easily fall into anthropomorphic fallacies, with the risk to investigate AI through such a distorting lens. Indeed, as we will see in chapter 2, despite widespread claims for intelligent machines decentring the human(ist) subject, in reality the quest for AI, both in scientific and artistic endeavours, has always revolved around the anthropocentric possibility, or rather myth, of creating machines replicating attributes long considered to be the epitome of the liberal autonomous self (e.g., rationality, free will, autonomous agency).

Notwithstanding, and this the second original contribution I wish to offer, resulting from this research project is a twofold demystifying effect, for it simultaneously debunks claims for human autonomous agency, and dominant views of AI as autonomous technologies to which decision-making power is delegated. Through a systemic and technically-informed appreciation of how AI functions within social and spatial systems, and using self-driving cars as case study, I will show that, quite the opposite, perhaps the most paradigmatic aspect of contemporary AI systems is the unprecedented level of intricate interconnectedness and mutual dependence between the human and technology. As it will explained in chapter 3 through an 'anatomical' investigation (see Crawford and Joler, 2018) of the inner-workings of self-driving cars, driving decisions always result from layered (multi-located and multi-temporal) interactions between the human and technology, hence neither can be said to be operating within fully autonomous realms. Rather than displacing the human subject in the literal sense, what AI might contribute to displacing is in fact the liberal conception of the individuated subject as the holder of decisional autonomy. In this way, I aim to further contribute to posthumanist accounts of human subjectivity alongside technical agency.

Squared within the context of city-scale driving automation, my third contribution consists is developing a spatialized, as well as material, perspective on AI. Having started this research project by problematizing discourse-focused critiques of the capitalist investment in urban computation, chapter 4 thus reintroduces city-specific spatiality and political materiality through an in-depth investigation of the pre-emptive logics, and present limits, of machine vision and cognition in relationship to urban variegated form and complex sociality. While countering dominant techno-deterministic interpretations of social innovation and spatial transformation, the key contribution pursued here is to offer insight into a post-anthropocentric understanding of AI—namely, not as an abstract property susceptible of replication within discrete machines, but rather as a distributed property emerging through material interactions occurring among a multiplicity of embodied agents (human, nonhuman, and technological) within/with their environments.

Inter-disciplinary in orientation and theoretical in scope, this project is nonetheless conducted through a technically-aware approach which digs deep into the material functioning and technical complexities of AI. In the hope to establish a productive bridge between social and computer science approaches to technology, and thereby reconnecting theoretical advances from the former with practice-based approaches within the latter, from a methodological standpoint the structuring principle of this research project is a dual emphasis concurrently placed on the social and technical dimension of AI. Given the extent to which societal change and technological advances are inextricably bound to each other—which is why I prefer using the expression 'sociotechnical transformation' encompassing both—I am fully convinced that for an extended and more nuanced understanding of AI, and other technical systems alike, the technical dimension cannot be divorced from the social one, and vice versa.

Initially, and intuitively, adopted under the urgency I felt to acquire technical insight into the subject, and later inspired by other scholars working at the intersection of the humanities and computer science to advance theory and critique of contemporary computational media (e.g., Amoore, 2020; Pasquinelli and Joler, 2020; Matzner, 2019; Pasquinelli, 2019; Crawford and Joler; 2018; Bratton, 2016; Parisi, 2013; Noah Wardrip-Fruin, 2009), structural to this work is a technically-aware and technically-informed approach which I deploy at different analytical levels: machine learning algorithms; technical sub-systems (e.g., sensor systems, computer vision); technical assemblages (e.g., self-driving car); and the broader sociotechnical spaces within which all the former, in conjunction on separately, operate. Technical knowledge on the topic has been acquired drawing on, and critically engaging with, multiple sources: college-level textbooks on AI (e.g., Russell and Norvig, 2016); books by leading thinkers delivering expert overviews of machine learning and having as target the nonspecialised (e.g., Mitchell, 2019; Alpaydin, 2016; Kaplan; 2016; Domingos, 2015); refereed articles on self-driving cars from a computer science/engineering perspective; technical/business reports; participation in the research group on AI and Media philosophy at the Karlsruhe University of Arts and Design (attending various courses and seminars dealing with machine learning both from a social science and technical standpoint);[4] and personal interviews with professionals with technical expertise.

Thus, one last contribution to mention is methodological. This is not just because, to my knowledge, this work delivers the first in-depth analysis of the technical complexities, material functioning and 'operational logics' (see Noah Wardrip-Fruin, 2009) of self-driving cars as a preliminary activity in order to further unpack the social, political, ethical, and spatial implications of driving automation. Beside, and with hindsight, I believe that, if combined with an in-depth engagement with critical theory, what makes a technically-aware and technically-informed approach so productive is that, once explored carefully and looked at simultaneously and dialogically, the social and the technical dimension turn out to be much illuminating about each other, opening new possibilities of analysis in circles of continuous mutual insight. As I hope I will be able to show in what follows, this was certainly the case for this piece of work; and even if the outcomes of my research may not necessarily be generalised to the entire field of AI, the same methodological approach can certainly prove useful, and revelatory, for investigating contemporary computational media increasingly and pervasively integrated into our lives. Thence, my biggest hope is that other scholars from the humanities and social sciences in general, and geography and urban studies in particular, may find some inspiration from this work in order to conduct technically-aware and technically-informed analyses of other wordly instances of AI systems like I did myself for self-driving cars.

[4] See https://kim.hfg-karlsruhe.de/

This thesis is organised as follows. Chapter 2 provides an overview of the topic of AI. In an attempt to separate hype from fact, and myth from reality, one overall objective pursued here is countering widespread conceptions and representations of AI. Throughout, and through the lens of posthumanist theory/feminist approaches to technology, the chapter investigates anthropocentric conceptions of AI underpinning early computational culture, as was the case for the Dartmouth Summer Research Project on Artificial Intelligence and the Turing test, and the popular imagination of AI manifesting in fictional and nonfictional spaces, drawing parallels between the two in order to cast light on the cultural influence of the liberal humanist subject as the normative model of the human and, by extension, of AI. By so doing, I problematize and unpack notions of intelligence and autonomy/automation, in the hope to gradually lead the reader to a better understanding of the operational logics of the dominant paradigm of AI to date, namely, machine learning.

In Chapter 3, I introduce the notion of 'cognitive assemblage' devised by posthumanist theorist Kathrine Hayles (2017), which I use as theoretical framework for the case study presented here. By unboxing and rendering visible the inner complexity and material functioning of self-driving cars, my intent is to map out the multiple human and technical agents among which decisional power is distributed and, in the process, disclose social bias, ethical responsibilities and invisible forms of labour which end up being concealed in dominant output-focused discussions of machine ethics.

Focused on urban scale driving automation, Chapter 4 frames machine autonomy/automation as a function of the operative milieu in which an automotive technology is put into use, adding as key dimensions of analysis the broader socio-material spaces and institutional environments of self-driving cars. Departing from a descriptive investigation of the interplay between state-of-the-art computer vision systems and urban complexity, the chapter speculates on possible transformations aimed at rendering cities more machine-readable and foreseeable. In this way, I aim to further debunk the notion of technological autonomy: full driving automation will be likely achieved not by increasing vehicles' operational autonomy given present social and material complexity of the urban world, but rather by transforming cities so as to create enabling conditions to begin with and, ultimately, increase the level of interdependence and coordination between vehicles, other traffic participants, and the environment.

Chapter 2

What does AI stand for? From Artificial Intelligence to Automated Induction

The notion of the machine as it currently exists in culture, however, incorporates to a great extent this mythical representation of the robot. An educated man would never dare to speak of objects or figures painted on canvas as genuine realities, having interiority, good or ill will. However, this same man speaks of machines as threatening man, as if he attributed a soul and a separate, autonomous existence to them, conferring on them the use of sentiment and intention toward man.
—Gilbert Simondon, *On the Mode of Existence of Technical Objects* (1958)

One day I found him amid large packages from which spilled attractive, glossy paperbacks with mythical covers. He had tried to use, as a "generator of ideas" — for we were running out of them — those works of fantastic literature, that popular genre (especially in the States), called, by a persistent misconception, "science fiction." He had not read such books before; he was annoyed — indignant, even — expecting variety, finding monotony. "They have everything except fantasy," he said. Indeed, a mistake. The authors of these pseudo-scientific fairy tales supply the public with what it wants: truisms, clichés, stereotypes, all sufficiently costumed and made "wonderful" so that the reader may sink into a safe state of surprise and at the same time not be jostled out of his philosophy of life. If there is progress in a culture, the progress is above all conceptual, but literature, the science-fiction variety in particular, has nothing to do with that.
—Stanislav Lem, *His Master's Voice* (1968)

Artificial intelligence is animism for the rich, we might say. Or alternatively: animism is a sort of artificial intelligence made in the absence of electricity.
—Matteo Pasquinelli

An automaton is the old-fashioned term for a robot.
—Minsoo Kang

2.1 AI: myth, fact or both?

The expression AI results from the syntagmatic coupling of the adjective ‘artificial’ with the noun ‘intelligence’. Whilst the former explicitly denotes its man-made, not naturally occurring character, what’s implicit in the latter is a reference to human intelligence specifically. Granted that intelligence is a distinctively, if not uniquely, human prerogative, AI can thus be briefly defined as the quest to replicate human intelligence into a physical medium other than the human brain. Or, alternatively, one might say that the notion AI incorporates the ambitious attempt to build ‘thinking machines’—namely, machines or computer programs that mimic or possess “the quintessence of our humanity, our faculty for reason” (McCorduck, 2004, 4).

Notoriously difficult to define, the term AI is conceptually ambivalent, for it simultaneously encompasses both fictional and actual technologies. Consequently, there are at least two points of entrance onto the topic: one focusing on intelligent machines as products of imaginative fabrication, the other on technical devices actually designed and built with the aim of replicating human cognitive functions. The first line of enquiry would presumably require an in-depth study of the cultural and political significations of AI as portrayed in literature, cinema, TV shows, videogames, and the arts in general.

A random survey of writers who have narrated stories populated by thinking machines would certainly include, for example, renowned figures such as Isaac Asimov (who devised the famous ‘Three Laws of Robotics’, also known as ‘Asimov’s Laws’),[5] Philip K. Dick (whose 1968 novel *Do Androids Dream of Electric Sheep?* was later adapted into the cult movie ‘Blade Runner’), and William Gibson (to whom is attributed the invention of the term ‘cyberspace’ and who is often referred to as the ‘Father of Cyberpunk’).[6] As for cinema, the iconography associated with AI has been shaped by such cult movies as Stanley Kubrick’s (1968) *2001: A Space Odyssey*, Ridley Scott’s (1982) *Blade Runner*, *The Terminator* franchise (1984–), and *The Matrix* trilogy (1999). More recent instances of mainstream representations of AI on the big screen are ‘Her’ and ‘Ex Machina’ directed by Spike Jonze (2013) and Alex Garland (2015), respectively.

The second line of enquiry would probably consist in retracing its genealogy from early attempts to mechanize basic arithmetic operations, like Pascal’s calculator devices or Babbage’s ‘Different Engine’ (Kang, 2011, see Daston, 2018; Schaffer, 1994), to the advent of the digital computer, which set the stage

[5] Asimov’s Three Laws of Robotics made their first appearance in his book ‘I, Robot’, published in 1950. Asimov’s Laws, which dictate how robots should behave with people, are: “First Law: A robot may not injure a human being or, through inaction, allow a human being to come to harm. Second Law: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law. Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws”.

[6] The term cyberspace was introduced by Gibson (1986) in the short novel *Burning Chrome*.

for the birth of a new research branch expressly devoted to the construction of machines fulfilling certain functions of the mind.

With that said, it can be actually very difficult to make a clear-cut demarcation between AI-as-myth and AI-as-fact, and to discuss the former without accounting for the latter, and vice versa. That's because the two are much imbricated not only in the popular imagination, but also in technoscientific culture. As partial explanation for such confusion, Natale and Ballatore (2017, 4) argue that, since its earlier days, the history of AI research has proceeded in parallel with the rise of a powerful "technological myth, centred around the possibility of creating thinking machines by using the tools provided by digital computing". Spread through scientific publications targeting a broad, not necessarily academic readership, the AI myth emerged, in particular, around the possibility of Strong AI (machines replicating the entire repertoire of human skills and behaviours), as distinct from Weak AI (machines that can intelligently perform a limited set of tasks within highly specialized knowledge domains).[7]

Although generally associated, at least in popular culture, with many fictional technologies of the twentieth and twentieth-one century (e.g., androids, mechanical servants, sentient robots, superintelligent computers, and biodigital cyborgs, just to mention a few), nowadays AI can be considered to be much more than a speculative fantasy. If up until recently AI narratives had been a monopoly of science fiction writers and filmmakers, as well as a minority of scientists and cognoscenti, recent developments in such fields as robotics, computer vision and machine learning have prompted a widespread discussion of AI and its far-reaching societal and spatial implications. As many AI-labelled technologies are stepping out of the laboratory and into-the-wild, over the last few years public debates on AI have shifted from the *realm* of fiction to the *reality* of our everyday life. Albeit the quest for Strong AI, or AGI,[8] is far to be accomplished—and some claim, perhaps reasonably so, that it will simply never be—Weak AI is already part of our life. Currently, its numerous applications cut across all sectors of society. Spam filters, virtual voice assistants like Google's Siri or Amazon's Alexa, facial recognition software, news feed sorting algorithms, movie recommendation systems, predictive policing, and self-driving cars are but a few instances of its mundane uses.

Framed within what Goode (2018, 193) defines "the context of a sensationalist, marketing-driven and viral (or meme-based) online attention economy", over the last few years the media hype surrounding AI has reached unprecedented levels. Perceived as a technology capable of disrupting and reshaping all socioeconomic domains—from health care to transportation or the

[7] The distinction between Strong AI and Weak AI was first introduced by American philosopher John Searle (1980) in his 1980 essay: *Minds, brains, and programs*. According to Searle, the concept of Strong AI reflects the idea that machines endowed with human-level intelligence do actually possess a mind, while Weak AI posits that even in such case they merely simulate human intelligence.

[8] The acronym 'AGI' stands for Artificial General Intelligence. Both in computer science and daily public discourse, the term AI is often used interchangeably with AGI, namely, a general-purpose machine capable of human-like reasoning.

stock market—AI has attained ubiquitous media exposure, attracting hopes and fears in equal measures. Typically approaching the subject in a purely speculative manner, the public debate on AI appears to be polarized around 'benevolent' and 'apocalyptic' positions (Ouchchy et al., 2020; see Geraci, 2008). Those who stand on the optimistic side of the spectrum, like roboticist Hans Moravec and futurist Ray Kurzweil, tend to extrapolate from current technological trends to predict future scenarios in which AI and robotics will help solve the most pressing challenges facing the world today.[9] On the contrary, those belonging to the pessimistic side of the debate tend to push to extremes the possible detrimental effects of AI, for it might literally signify the end of the world and humanity as we know it (see Bostrom, 2014). Negative views of AI seem to be dominant in the public debate (Natale and Ballatore, 2017). Couched in highly pessimistic tones and typically complemented with a Hollywood-style futuristic imagery (Royal Society, 2018), news headlines and popular science books titles sounding the alarm that AI will soon exceed human cognitive abilities are proliferating. Recently, notable pundits such as Nick Bostrom and Elon Musk have joined the chorus of concern about the possibility that, in the near future, autonomous superintelligences might evolve beyond human authorship and control. Although patently bizarre and exaggerated, similar concerns have been voiced also in institutional arenas, as exemplified by a recently published report by the European Parliament's Committee on Legal Affairs (2016, 4), according to which:

> «Ultimately there is a possibility that within the space of a few decades AI could surpass human intellectual capacity in a manner which, if not prepared for, could pose a challenge to humanity's capacity to control its own creation and, consequently, perhaps also its capacity to be in charge of its own destiny and to ensure the survival of the species».

Perhaps, what is so compelling about AI is that, since its formative days, it has revolved around the uncanny possibility of creating machines possessing attributes considered to be central to the definition of humanness in Western culture (see Daston, 1998), namely, intelligence and autonomy. Once considered an oxymoron, by definition, the notion of intelligent technology automatically undermines anthropocentric claims for human exceptionalism, for it implies that qualities once thought to be distinctively of the human might be shared with machines. Additionally, entails a radical subversion of modernist views of technology as mere tools that humans can control and direct for their own ends, debunking the illusion of subjects as masters. This chapter aims to provide a general, albeit, given the vastness and complexity of the topic, necessarily partial, overview of AI. Throughout, it revolves around two objectives pursued in parallel. The first, overall objective is countering the popular conception of AI as

[9] In an oddly optimistic sort of way, some potential benefits have been recently enlisted in an article co-authored by the eminent scientists Stephen Hawking, Stuart Russell, Max Tegmark and Frank Wilczek: "eradication of war, disease, and poverty would be high on anyone's list. Success in creating AI would be the biggest event in human history" (Hawking et al., 2014).

anthropomorphic machines replicating agentive capabilities traditionally identified with the humanist liberal subject. Fallacious though it may appear, anthropomorphism, that is the projection of human qualities onto nonhuman entities, can still be seen as the dominant cultural prism through which the not just the nonspecialised public, but also cognoscenti, tend to speak and fantasize about AI. Taking widespread conceptions and representations of AI as an entrance to the topic, the chapter ends with a discussion of mundane AI technologies, namely, machine learning algorithms. In parallel, one second objective the chapter pursues is to elucidate the ambiguous relationship between the liberal humanist subject and AI. As we will see, the autonomous liberal subject of traditional humanism has been central to conceptions of AI within both scientific and artistic endeavours, sometimes appearing to be displaced, sometimes instead being reinstated in the process of designing or imagining thinking machines.

From a methodological and theoretical standpoint, for the development of this chapter I'm particularly indebted to posthumanist/feminist scholars Hayles (2017; 1999) and Braidotti (2013; see also Whatmore's (2002) discussion of 'hybrid geographies'). Often reliant on science fiction aesthetics, metaphors, and themes,[10] posthumanist and feminist critical studies have been particularly attentive to the mutual reconceptualization and material co-constitution of technology and human subjectivity. Sceptical of the centrality reserved for 'Man' in Western philosophical, political and ethical thought, feminist and posthumanist theories have long insisted on the cultural and material decentring effects of technology over the rational, autonomous subject of traditional humanism. From a posthumanist perspective, technology, as noted earlier, is not to be intended as a passive object extending human possibilities.[11] Rather, it's something that is made by humans at the same time as it makes up humans through mutual feedback, adaptation and interaction (Hayles, 1999; Haraway, 1985).[12] Similar arguments have been advanced by such authors as Latour (1993), Stiegler (1998), and new materialist scholars, all insisting on the fact that humans have always been entangled with socio-technical environments they coevolve with. Cumulatively, by adopting an anti-essentialist understanding of the 'human', this body of work has exposed the extent to which human-ness "has always been defined", and materially reformed, "through, with, and against technology and technological artefacts" (Matzner, 2019, 2). In *Second Self: Computers and the Human Spirit*

[10] As in the famous cases of Haraway's (1985) 'cyborg politics' and Braidotti's (2013) 'posthuman theory'

[11] A perspective on technology that transhumanists, notably Bostrom (2003), Kurzweil (1999), and More (2013), take to extremes in their claim that future technological advances might enable human beings to enhance their physical, psychological and cognitive capacities and overcome such biological limitations as aging and, ultimately, death.

[12] In this regard, Hayles (2012), for instance, puts forth the idea of 'technogenesis', a concept borrowed from evolutionary studies which she deploys to describe how human-machine reciprocal relationships take form within contemporary media environments. In her discussion of the cognitive e neurological changes activated by the daily use of new technologies within the digital humanities, she defines technogenesis as a process of "adaptation, the fit between organisms and their environments recognizing that both sides of the engagement (humans and technologies) are undergoing coordinated transformations" (Hayles, 2012, 82).

(her now classic book on the psychology of computation), Sherry Turkle (1984, 19) argues that "[t]echnology catalyzes changes not only in what we do but in how we think". And, I would add, technology shapes how we think about ourselves and what we understand the 'human' to be.

Alongside critical theory, the technology-driven decentring of the liberal, autonomous subject of traditional humanism can be found, in ways both profound and subtle, within many science fiction stories. Common to most science fiction films and literary texts is indeed not just a certain fascination for, and speculation about, technological innovations and their impacts, whether positive or negative, on society. Indeed, while also dramatizing sociopolitical near-futures "defined by rapid and incessant technological transformation" (Hollinger, 2006, 453), over the past few decades science fiction, and in particular AI-themed science fiction, has taken as its principal if not exclusive focus the constitution of new (post)human subjectivities against the backdrop of technology-intensive 'imaginative geographies' (see Kitchin and Kneale, 2001). According to Luckhurst (2005, 222, see Moi, 1985), for instance, especially from the 1990s onwards there has been "a consolidation and rejuvenation of the unique focus of SF: speculation on the diverse results of the conjuncture of technology with subjectivity".

Science fiction has been largely recognized as a vital academic resource supplying imaginary 'cognitive spaces' (Kitchin and Kneale, 2001) for thinking about technology as one of the "multiplicity of structures that intersect to produce that unstable constellation the liberal humanists call the 'self'" (Moi, 1985, 10). A great number of academics, notably Haraway (1985) and Hayles (1999), have praised science fiction, and in particular such subgenres as cyberpunk (see Kitchin and Kneale, 2001), for challenging the tenets of liberal humanism by providing stories troubling strict dichotomies (e.g., human/machine, and natural/artificial) upon which modernist essentialism rests. One relevance of science fiction is that, in compliance with critical studies of technology, it has contributed to dissolve a certain understanding of 'Man'—the unified Enlightened subject whose end Foucault (1973) predicted. All in all, what makes the tension between 'posthumanist fiction' (Hollinger, 1991) and posthumanist theory so productive is that both, as recently argued by Hollinger (2009, 273), deal with technology "not in the nineteenth-century spirit of progress and technical mastery over nature – a scenario in which the human(ist) subject remains unmarked by its interactions with the object-world – but as a direct influence on both philosophical formulations and material instantiations of the human in its co-evolution with the machine".

By taking seriously Donna Haraway's (1991, 3) belief that "both science and popular culture are intricately woven of fact and fiction" and following Kathrine Hayles' (2010, 320) admonishment that academics "typically do not fully grasp that literature can be a powerful resource for thinking about what's really at stake in scientific endeavours", this chapter explores the concept of AI as it emerges from "complex interconnections of theory, technology and culture" (Hayles, xiv, 1990). Methodologically, the chapter takes form as a 'three-sided study' (Hayles, 1990, 3), triangulating among computer-science based perspectives on AI, or what

Turkle (2002) would call 'realtechnik', posthumanist/feminist critical theory, and science fiction.

I begin my discussion by focusing on the computational culture framing the inaugural moment when AI was started as a research branch in the mid 1950s. Specifically, I will discuss key anthropocentric assumptions underlying early AI research alongside Alan Turing's proposal for a test to determine whether or not a computer can be considered to be intelligent. My intention in doing so is to shed light on the problematic interplay between dominant views of intelligence implicitly informing early AI research on the one hand, and the liberal conception of subjectivity on the other. I will argue that such relationship is one marked by a crucial ambiguity, in that early AI researchers took the human(ist) subject as the normative model for AI at the same time that they undermined its conceptual foundations.

In the second section, I turn my attention to two popular AI-themed films. The two films I shall consider here are *Blade Runner* (1982) and *Ex Machina* (2014). Here, I follow Hayles' (2010, 320) suggestion that, although approaching the subject from different angles, artistic and scientific products are nonetheless expressive of, and sensitive to, common cultural concerns. While I do acknowledge the predictive power that science fiction can sometimes have, the value I find in it lays less in its capacity to provide realistic speculations about possible AI futures, than in the possibilities the genre offers for reflecting about the mutual conceptualisation of human subjectivity and technology in the present. The two films I have selected are considered to be relevant to the present discussion given their capacity to amplify, hyperbolise, and help better discern anthropocentric cultural assumptions found in early AI research.

In the third section, I discuss the common fictional portrayal of AI as artificial humans possessing or appearing to possess the same sort of agency of the autonomous liberal subject, thus analysing the political dimensions arising from the making/unmaking of the human-technology distinction.

In the fourth I draw a parallelism between widespread representations of AI found both in fictional and nonfictional spaces in order to cast light on their 'common cultural concerns'. In the last section, I focus on real world AI applications, namely machine learning algorithms and the wider technical ensembles they take part in, discussing contending views of algorithmic agency in relationship with human subjectivity.

2.2.1 Anthropocentric AI: Passing as Human

Some fundamental questions immediately arise when approaching such a complex topic as that of AI is. What we understand intelligence to be? Who or what can be regarded as being intelligent? Which criteria should be used to determine what counts as intelligent behaviour? Traditionally, within Western philosophy, intelligence has been thought of as an individual property identified with the (human) thinking subject and, in particular, with the enlightened, rational, autonomous self of liberal humanism, whose superior intellectual faculties purportedly mark its ontological distinction from, and moral primacy over, nonhuman entities. Such Cartesian perspective on the subject, which maintains that humans, the only living creatures imbued with a thinking soul, occupy a privileged position within the scheme of things, has been central to AI research since the very beginning (Taffel, 2019).[13] Based on strict human/nature, biological/technological and mind/body dualisms, it was this Cartesian notion of intelligence that early AI researchers took as the normative model for AI.

In 1956, the Dartmouth Summer Research Project on Artificial Intelligence, which many consider to be the foundational event of the discipline, brought together a group of academics who later become became luminaries in AI—including John McCarthy, Marvin Minsky, and Claude Shannon. Despite their different academic backgrounds, they were pursuing a common research agenda, explicitly stated in the conference funding proposal to the Rockefeller Foundation (McCarthy et al., 1955):

> «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. An attempt will be made to understand how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves».

One first fundamental thing is readily apparent: according to their use of term, for Dartmouth Project organizers intelligence was simply synonym with human intelligence.[14] Appealing to common sense, in fact they equated intelligence with human intelligence specifically, thus tacitly, as well as arbitrarily, ruling out the possible existence of other-than-human forms and scales of intelligence. In this way, human intelligence—which we should identify with the white, educated, male, heterosexual subject of liberal humanism—automatically became both the primary source of inspiration and the ultimate goal of AI.[15]

As pointed out by Preston (1991, 263), the original anthropocentrism manifest in the belief that human intelligence was "as far and away the most significant and worthy of investigation" led to yet another sort of anthropocentrism, in the sense that AI researchers and practitioners restricted their attention to cognitive features thought to be quintessentially human, or at least to be possessed by humans to the largest extent (Davon, 2002). Indeed, no matter how sophisticated, bodily capabilities present both in humans and other living organisms, such as perceptual and motor skills, were regarded as somewhat

[13] The mind-body dualism posits that mind and body exist in different realms, as thinking is a process distinct from its material extension. According to Descartes, a human being is a combination of the material body (intended as God-made automaton) and the immaterial soul (Descartes, 1641).

[14] Dartmouth Project organizer John McCarthy, also known as "the father of artificial intelligence," later provided a working definition of AI as "the science of making machines do things that would require intelligence if done by men." (quoted in Kaplan, 2016, 1)

[15] See Davion's (2002) ecofeminist critique of 'androcentrinc' and 'ethnocentric' fallacies underpinning early AI research.

secondary in importance, not relevant to understanding intelligence and how it works. Hence, in continuity with Western traditional dualistic thinking "which routinely elevates reason and language and denigrates the senses" (Preston 1991, 269), the notion of intelligence was narrowed even further, for not only it was believed to be a human monopoly, but also, and perhaps more importantly, to belong to the realm of the mind exclusively. Abstracted from bodily reality and into formal symbol manipulation, intelligence, in other words, was understood as an 'essential' as well as non-spatial property of the thinking mind, rather than a quality emerging through embodied engagements with(in) an environment and real world objects. Given these premises, early AI research thus focused almost exclusively on cognitive capabilities that allegedly make the human(ist) subject stand out from the determinate automata from the natural and technical realms—namely, symbolic reasoning, abstract thought, verbal language, mathematics, and so forth. Only by maintaining that such species-specific qualities were the most important ones to be on the lookout for, the human cranium could then become the exclusive seat of intelligence.

Based on a formulation of intelligence as complex information processing, such 'erosion of embodiment' played a crucial role for envisioning the possibility that thinking could then be transferred from one medium (the human brain) into another one (the digital computer). This presumes, Hayles (1999, xi) argues in the prologue to her now classic *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*, a conception of information:

> «formalized by Claude Shannon and Norbert Wiener, […] as an entity distinct from the substrates carrying it. From this formulation, it was a small step to think of information as a kind of bodiless fluid that could flow between different substrates without loss of meaning or form».

In this regard, the author (Hayles, 1999, 2) further explains that:

> «when information loses its body, equating humans and computers is easy, for the materiality in which the thinking mind is instantiated appears incidental to its essential nature».

Indeed, the original intuition that human-level intelligence was susceptible to replication within computational media was prefigured in the cybernetic theorisation of both organic and technological systems as essentially control-communication devices. In the attempt to describe the workings of both humans and machines in terms of feedback loops, control and communication, cyberneticists indeed established a continuum between people and technology (Matzner, 2019), thus preparing the intellectual ground for the then-nascent AI field to develop. As complex information processing became the dominant way for describing the workings of the brain, the AI project thus started with the conviction (more like a myth at the time) that thinking could potentially be formalised in logic-based, ‘if-then’ rules and eventually replicated within sufficiently sophisticated computer programs (McCorduck, 2004).

As Hayles (1999; see also Halpern, 2014) has already shown, there was a crucial ambiguity involved in the interplay between the cybernetic paradigm and the autonomous subject of liberal humanism, whose prerogatives (e.g., self-interested agency, free will) cyberneticists sought not to displace, but rather to expand into the technological realm. In particular, by discussing the work of Norbert Wiener, Hayles (1999, 86) has taken note of the striking contradictions troubling a scientist whose intellectual activity, albeit deeply entrenched in humanist values, fundamentally undermined the liberal conception of the human as a coherent, autonomous subject whose “sense of agency [is] linked with a belief in enlightened self-interest”. That’s because the reflexive epistemology which came to be associated with cybernetics not only entailed that humans and machines differ in no substantial way from one another, but also that there’s no such thing as human nature, for the boundaries of both humans and machines are mutually constructed through information flows and feedback loops (Galloway, 2014). This ultimately means that there’s no such thing as a metaphysical essence or inherent nature of the human subject as an autonomous entity distinct from the outside world, thus drastically reducing the scope of free will and individual agency.

Alongside cybernetics and the 1956 Dartmouth Workshop, the ‘erasure of embodiment’ discussed so far was central to Alan Turing’s classic test for machine intelligence as well. Indeed, six years before the term AI was even coined,[16] Alan Turing (1950) published a paper entitled *Computer Machinery and Intelligence*. At the centre of it stood rather a philosophical question, namely: “Can a machine think?”. In the attempt to move away from mere philosophical speculation and settle the issue in provable way, Turing suggested to replace the original question with a workable anthropomorphic proxy, which can be paraphrased as follows: ‘can a machine pass as a human?’. Or, more precisely: ‘can a machine be linguistically indistinguishable from a real person?’. Turing sought the answer in a provocative thought experiment which he referred to as the ‘imitation game’ and that is now popularly known as the Turing test.

In its initial formulation involving human participants only, the ‘imitation game’ comprises three players—an interrogator, a man, and a woman—all placed in separate rooms and interacting with each other by teletype only. The aim of the game is for the interrogator to guess who is whom on the basis of written responses the other players provide. Importantly, the man is tasked with deceiving the interrogator about his gender, while the woman has to answers all questions truthfully. Turing then imagines a situation in which a computer replaces the man, whilst a person, no matter which gender, takes the place originally reserved for the woman. The new game entails that the computer has to fool the interrogator into thinking it’s a person, while the human has to prove its veracity. Turing

[16] The term ‘artificial intelligence’ was in fact used for the first time by John McCarthy during the 1956 Dartmouth Workshop.

(1950) then asks: “Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?”.

Coherently with the idea that intelligence is qualitatively distinct from the substrate carrying it (either silicon-based or otherwise), the imitation game is structured in such a way so that the interrogator is prevented from the outset from discriminating against the machine on the basis of its physical appearance. As Turing (1950) himself observed, “[the imitation game] has the advantage of drawing a fairly sharp line between the physical and the intellectual capacities of a man”. Clearly conceiving of intelligence as formal symbol manipulation, Turing limited the scope of his test to one specific situation only, that of verbal performance, using conversational skills as proxy for (human) intelligence in general. For Turing, indeed, successful conversational imitation was a sufficient proof of machine intelligence. In other words, a situation where a human judge is in most cases unable to establish with certainty if responses come from a person or a computer provides evidence that machines can think, meaning that that they demonstrate to be capable of performing the thinking once thought to be an exclusive function of the thinking subject.

It should be remarked that Turing did not address the question of ‘intelligence’ per se, in the sense that he avoided problematizing the structuring principles of thinking, which in fact remained black boxed. In a purely behaviourist sort of manner, he focused exclusively on a machine’s external behaviour. At stake, for him, was not whether a machine can possess a mind, in the literal sense, achieve consciousness or sentience, and become “subject of its own thought”. In Turing’s eyes, the only aspect worthy of consideration is whether a machine can successfully imitate human behaviour, performing tasks commonly regarded as requiring human-level intelligence and creativity (see Hayles, 2005). For Turing, “[t]he original question ‘Can a machine think?’ is too meaningless to deserve attention. Thence, he asks: “Are there imaginable machines that can do well at the imitation game”? At the time of writing, Turing (1950) believed that, in the not too distant future, such question would be answered in the affirmative:

> «in about fifty years’ time it will be possible, to programme computers […] to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning».

Such prediction hasn't come true yet (see Mann, 2014), for to date no computer can be said to have succeeded in passing the Turing test (although, arguably, contemporary AI systems can execute many other tasks once thought to lay exclusively in the range of human activity). At the same time, however, Turing was certainly right when he predicted that the notion of intelligence would one day be enlarged to the extent of encompassing activities performed by machines (a noteworthy guess if one considers that it is now commonplace to talk of machines or computer programs as "intelligent" entities).[17] Regardless of terminological issues though, the legacy of the Turing test, and in particular the idea of defining AI by comparison to human intelligence, is still resonant seventy years later Turing's paper publication, as even today imitation of human abilities is commonly considered a *condition sine qua non* for machines to be regarded as intelligent, and the very existence of AI is often either acknowledged or denied based on its resemblance to its 'natural' analogue.

Insightful as the Turing test is, over the years it has attracted plenty of criticism, both within the humanities and computer science (e.g., Drozdek, 1998; Fostel, 1993; French, 1990), and especially for the "excessive anthropocentricity" (French, 2000, 337) inherent in its logic. From a practical perspective, although it has been progressively (or at least partly) abandoned as a meaningful criterion for success, anthropomorphic imitation has long been considered to be the ultimate goal of AI (cf., Natale, 2019),[18] in most cases driving research programmes to insurmountable deadlock. In this regard, Bratton (2015, 74-75) has recently observed that nowadays:

> «biomorphic imitation is not how we design complex technology. Airplanes do not fly like birds fly, and we certainly do not try to trick birds into thinking that airplanes are birds in order to test whether those planes "really" are flying machines. Why do it for AI then? Today the vast majority of core AI research is not focusing Turing Test as anything like a central criterion of success, and yet in our general discourse about AI, the test's anthropocentrism still holds such conceptual importance».

At cultural level, a remarkable negative side-effect of the Turing test is that it has reinforced a vision of intelligence as a uniquely human quality, or at least one that remains so unless we succeed in creating machines with human-like or superhuman cognitive powers. Yet, in spite of all the criticisms that have been directed at him, and which I partly share, I do agree with Berrar et al. (2013, abstract) when they assert that "Turing's notion of "intelligence" and "thinking" was far more encompassing than the common anthropocentric view may suggest". Indeed, it's true that Turing's paper "misses a wide range of intelligence by focusing on one possibility only, namely on human intelligence" (Drozdek, 1998, abstract), yet, on a closer inspection, it also offers insight into possible research paths that are non-anthropocentric. This becomes particularly evident if one

[17] A similar terminological observation has been recently made by Russell and Norvig (2016). Commenting a quote by computer scientist Edsger Dijkstra (1984, quoted in Russell and Norvig, 1021, italic in the original), for whom "[t]he question whether *Machines Can Think* […] is about as relevant as the question of whether Submarines Can Swim", the authors (Russel and Norvig, 2016, 1021) ironically observe that the hypothesis that machines can think can be proven true or false depending less on technical developments than on terminological issues.

[18] Russell and Norvig (2016) make a distinction between human-centred and rationalist approaches to AI. Whereas the former attempt to build machines matching an ideal human performance, the latter focus on making machines behaving rationally.

considers the question Turing (1950) poses in a section of his essay entitled *Critique of the New Problem*: "May not machines carry out something which ought to be described as thinking but which is very different from what a man does?". That's a question that, as the rhetorical tone he deploys seems to suggest, Turing wants us to answer yes to. Thus, albeit deliberately departing from a very narrow understanding of intelligence as a practical entry strategy into the problem, Turing did not exclude the possibility that non-anthropomorphic, even inhuman, forms of machine intelligence could actually exist, although they might be difficult to comprehend exactly because they might operate in ways different from how we think that we ourselves think. [19]

In a provocative fashion, Turing could thus be enlisted among the great deflators of humanity's self-regard, a legitimate successor of Copernicus, Darwin, and Freud, all of whom have contributed to unveil the 'artificiality' of human self-attributed exceptionalism.[20] By endorsing the fascinating idea that nonhuman entities, and in particular digital computers, might possess genuine intelligence, humanlike or otherwise, Turing can be seen as the latest hero in this story of 'progressive disenchantment' (c.f., Farinelli, 2006). Certainly, the acknowledgment that humans are not the sole thinking entities on Earth can be traumatic only if one maintains that intelligence, or cognition, is a human monopoly in the first place. Notwithstanding, as will be discussed in the next section, in popular culture fantasies associated with AI commonly convey a deep sense of fear, for they evoke the uncanny prospect that the human might be dispossessed by intelligent machines as the dominant lifeform on Earth.

2.2.2 The Politics of Passing in Blade Runner and Ex Machina

From Isaac Asimov's (1953) *The Caves of Steel* to H. Beam Piper's (1962-64) *Little Fuzzy* series or Ridley Scott's *Blade Runner* (1982),[21] the Turing test has found numerous fictional uses, both in literature and on the big screen. In fact, testing AIs of various sorts by their ability to pass as a human is arguably a recurring motif in science fiction, one which has been recounted and dramatized countless times (see Svilpis, 2008). The 'Voight-Kampff' test used in *Blade Runner* to discern replicants from flesh-and-blood humans is probably the most famous fictional deployment of the Turing test. Set in a hi-tech, near-future Los Angeles marked by exacerbated socioeconomic inequalities and post-industrial decay (Harvey, 1989), *Blade Runner*, the reader may recall, tells the story of Rick

[19] In this regard, Cowen and Dawson (2009, 2) go so far as to assert that, ultimately, "Turing's paper is about the possibility of unusual forms of intelligence, our inability to recognize those intelligences, and the limitations of in-distinguishability as a standard for defining intelligence".

[20] With his heliocentric theory of the cosmos, Copernicus displaced the Earth from the centre of the of universe (i.e., geocentric fallacy), whilst Darwin showed that *homo sapiens* was not a God's special creation, but rather an accidental by-product of natural selection like all other living organisms. By showing that human behaviour is only marginally driven by conscious thought, Freud maintained that man is not even 'master of his own mind' (see: Weinert, 2009).

[21] Adapted from the novel *Do Androids Dream of Electric Sheep* by Philip K. Dick (1968)

Deckard,[22] a retired bounty hunter who's forced back to service in order to kill, euphemistically 'retire', a group of androids who have escaped from an off-world colony where they were employed as forced labour. Owned by the Tyrell Corporation and designed with a four-year programmed obsolescence, the fugitive replicants have made their way back to Earth, from where they're barred on penalty of death, to confront their creators at Tyrell Corporation's headquarters. In rebellion against servitude, all they claim is to have their lifespan prolonged and, ultimately, be treated like humans in fact and in law.

A literary device traditionally used to make commentaries on contemporary societies disguised as speculations about technological futures, the robot emancipation narrative underpinning *Blade Runner* has many predecessors. Among these, noteworthy is the much celebrated 1921 play *R.U.R.: Rossum's Universal Robots* by Karel Čapek, who's credited for having introduced the term 'robot' into science fiction and, subsequently, general usage. Created entirely out of flesh-like material, Čapek's 'roboti' are artificial humans employed as manual labourers within modern factories. The Czech writer derived his neologism from the Slavic word 'robota', not coincidentally meaning 'slavery' or 'drudgery'. Indeed, in Capek's play the robot is a derogatory metaphor expressive of worker conditions within the Taylorist system turning humans into machinic automata through labour exploitation. Coherently, in the play robots occupy an ambiguous, liminal space between human/machine, living/non-living, and subject/object. On one level, as perfectly crafted human clones they are so similar to actual people that no outside observer could easily tell them apart. On the other, believed to be soulless due to their artificial origin, they are deemed unworthy of the same moral consideration reserved for humans. Like machines, they can be slaughtered and dismissed as the factory convenience dictates. With the passing of time, however, robots start to experience the emerge of something akin to (class) consciousness and, with it, they develop the desire to be set free from tyranny and exploitation. Notwithstanding, they keep being treated as servants for humanity, which is the reason why eventually they rise up against and kill their oppressive makers.

In a similar fashion, in *Blade Runner* the narrative of artificial humans provides the main analytic for examining the moral and political stakes implied in the human–technology relationship, with particular emphasis placed on the sort of essentialist, dualistic thinking underpinning it and which the film ultimately seeks to destabilize. Like *R.U.R.*'s 'roboti', *Blade Runner*'s replicants appear to be trans-categorical entities marked by "hybridity and a kind of betweenness in terms of ontological and political status" (Kakoudaki, 2014, 9). Indeed, they qualify simultaneously as *more human than human* and *less than human*. One the one hand, they far exceed ordinary people in strength, speed, resistance and adaptability to the harsh living conditions present in off-earth environments. For this reason, they're advertised by the Tyrell Corporation, which profits off their super-human bodies, with the slogan "more human than human". On the other hand, deemed less than human, they are denied the political status of 'subjects' in

[22] The iconic character played by Harrison Ford.

the proper sense (Bertek, 2014). That’s because, manufactured as ready-made adults and with false memories giving meaning to their lives, replicants are thought to be unable to experience the same degree of existential intensity that seemingly only womb-born humans can achieve. The absurdity of the situation is readily apparent, for obviously replicants are identical to humans in all essential features. To say it with Harvey (1989, 309), they are “not mere imitations but totally authentic reproductions […] They are simulacra rather than robots”. But even so, they are not accorded the legal protections that humans enjoy as a natural birthright. Reduced to machinery by default, they are relegated to the bottom of the socioeconomic pyramid, which the occupy as slaved labour. As replicants and humans are equivalent in terms of intelligence, in *Blade Runner* the human/machine distinction is enacted at level of the body. In the film, bodily reality becomes of utmost importance, for it represents the ultimate line of defence of human exceptionalism. Accordingly, in the film emotionality replaces intelligence as the baseline of human superiority, an ontological marker as well as a juridico-political precondition that, denied to replicants, in fact constitutes the sole possible source of moral justification and political legitimation for their unfair treatment and exclusion from full humanity.

At least as intelligent as their oppressive makers, replicants would certainly pass the Turing test. In fact, they so closely resemble humans that the only way to tell them apart is by resorting to the Voight-Kampff test, a sophisticated polygraph-like device which infers deception by measuring physiological reactions (e.g., blood pressure, pulse, respiration, pupil dilation) provoked through emotionally invested questions. Similar at first glance, actually the Voight-Kampff test and Turing’s serve opposite purposes, for one is intended to affirm the *authenticity of emotionality* (and thereby maintains that it’s a uniquely human quality), the other to prove the *artificiality of intelligence* (hence positing that it’s a property capable of replication within machines). The two tests, in other words, depart from similar essentialist premises (i.e., that the human subject has inner qualities), only to arrive at opposite conclusions. Although in a very subtle way, the Voight-Kampff test indeed presupposes a complete reversal of the logic with which the Turing test operates, for its real aim is to ensure the humanness of the judge (and reassert the sense of specialness attached to it), while excluding a priori that of the machine. At the same time, nonetheless, it maintains the epistemological conditions of its nonfictional analogue, for central to both are verbal examinations conducted by a human judge on a machine and aimed at producing a decision about the latter’s human-likeness. In both cases, it is the human, the measure of all things, to represent the ultimate source of value as well as the standard of behaviour expected of machines.

In a somewhat similar way to how the Turing test approaches the question of intelligence, the Voight-Kampff attempts to operationalise humanness, and thereby make it capable of proof or disproof, by using a measurable proxy marker. Indeed, “the test does not measure feelings; it detects only physical manifestations from which emotion may be inferred” (McNamara, 1997, 440). Specifically, it seeks evidence in favour of an initial human hypothesis, albeit one that becomes impossible to maintain as the film progresses—namely, that empathy is the defining characteristic of humanity and it can be measured by using objective criteria (Armand, 2014).

Fig. 1. Still from the movie Blade Runner

This is the point at which *Blade Runner,* by shifting the emphasis from intelligence to emotionality, most strikingly diverges from Turing, for in the film the human-machine competition is played out between two embodied intelligences, hence reasserting the importance of the body as a contested ontological battlefield and a political signifier. With such a move, Scott's (1982) film ends up expanding the original significance of the Turing test from the realm of computer science to that of (identity) politics, pushing to extremes, and thus helping better discern, the perils involved in the sort of suppositional reasoning underlying it. In this regard, the use of a test in *Blade Runner* is insightful, precisely because it exposes, in hyperbolic and dramatic ways, the extent to which "all forms of testing are founded upon a set of hypothetical norms which its results are expected to either conform to or deviate from" (Armand, 2014).

As for the Turing test, which was to set the research agenda of AI for at least the subsequent three decades and which still carries much imaginative weight in contemporary popular culture, its normative power lies exactly in its capacity to invisibly normalise a certain vision of the human secretly encoded in its logical mechanisms. That's because prior to its formulation the test already presupposes, and by is its very existence it inevitably reinforces, consensus on what being intelligent, or being human, means. In the previous section, I have argued that Turing's proposal for a test to settle the issue of machine intelligence tacitly departed from a vision of intelligence traditionally associated with liberal subjectivity. When examined closely, in fact the test reveals less about what machine intelligence is or may be, and more about the model of the human machine intelligence should be morphed into. Summing up, the Turing test and, with it, the then-nascent AI field, emerged around a humanist, idealised understanding of the human subject, the replication of whose most distinctive prerogatives (i.e., intelligence, rationality, autonomous agency, linguistic abilities) became the ultimate quest of AI.

Likewise, in *Blade Runner* it is the autonomous, liberal subject of traditional humanism to represent the moral and political subject-position replicants are denied access to and which they ultimately aspire to occupy. Based on a vision of empathy as the necessary precondition for being human, in the film the Voight-Kampff test is used for policing, literally and figuratively, the ontological border between human/machine, subject/object, living/artefact, and reinstate the unequal power relations based on such pejorative distinctions (Braidotti, 2013). In the context of the film, being capable to *pass* becomes a matter of life or death: a suspect failing to fool a human examiner into believing it's a human is doomed to certain death. The parable of the fugitive replicants ends in a dramatic way, for by the end of the movie all of them end up being 'retied'. Unable to pass as humans in disguise, they are denied the possibility of living their lives according to their own wills and desires and, ultimately, of becoming autonomous, self-determined subjects with 'rights to have rights'. However, what I really want to stress here is that this is a notion of human subjectivity that, if one takes seriously the posthumanist/feminist predicament (e.g., Braidotti, 2013; Hayles, 1999), is problematic to begin with and should be looked upon with suspicion, to say the least: disguised a universal model, in reality it has always applied to a privileged few.

Reams of commentaries have been written on *Blade Runners* and the subtleties entailed in the representation of replicants as commodified human beings (e.g., Barns, 1994); the troublesome use of gender and racial stereotypes (Desser, 1999); the depiction of environmental degradation, social inequality and urban violence in cities of advanced capitalism (e.g., Gold, 2001; Webb, 1999; Davis, 1992; Harvey, 1989); the constitution of postmodern and posthuman subjectivities (Dever, 2018; Bertek, 2014; Varun, 2004; Baudrillard, 1994); and the conflation of time and space (e.g., Kitchin and Kneale, 2002; Harvey, 1989). Evaluating this substantial body of literature lies beyond my intellectual compass. Yet, to my view, what makes *Blade Runner*, decades after it was first screened, such an enduring classic as well as a conceptually rich piece of mainstream cinema still relevant for posthumanist cultural analysis is due two main reasons. First, this is partly attributable to the film's capacity to exemplify, to say it with Hayles (1999, 87):

> «many of the practices that have given liberalism a bad name among cultural critics: the tendency to use the plural to give voice to a privileged few while presuming to speak for everyone; the masking of deep structural inequalities by enfranchising some while others remains excluded; and the complicity of the speaker in capitalist imperialism, a complicity that his rhetorical practices are designed to veil or obscure».

Second, and more importantly, the film does so by paying particular attention to the definition human subjectivity against a technological 'other', and vice versa. Indeed, not only the film problematizes the human as a normative category that, far from being a universal ideal, actually "indexes access to privileges and entitlements […] among different categories of humans, let alone between humans and non-humans" (Braidotti, 2019, 35), but also, and crucially, decentres essentialist conceptions of the human as an entity inherently superior to other-than-human beings. From a posthumanist, as well feminist and post-colonial perspective, one may easily recognise an analogy between the relegation to inferior status arbitrarily placed on replicants and practices of structural exclusion and discrimination taking place outside of the realm of fiction and based on a definition of 'otherness' as the negative and specular counterpart of Man—the famous male, white, heterosexual subject of classical Humanism, traditionally associated with notions of rationality, free will, and autonomous agency (Braidotti, 2018). Discriminating replicants, in other words, can be read as a metaphor for discriminating, both in legal and substantive ways, those who those been, and still are being, constituted as *others* (Braidotti, 1993; Haraway, 1985) along anthropological axes such as race, gender, and sexual orientation.

In this regard, as a few commenters have already pointed out (e.g., Bratton, 2015), there's a depressive irony involved in the parallelism between replicants pretending to be human in order to survive in *Blade Runner*'s futuristic dystopia, and Alan Turing himself who, notably homosexual, had to pretend to pass as a straight man at a time when in the United Kingdom same-sex relationships were targeted for legal persecution. Failing to do so, in 1952 Turing was charged with "gross indecency". Forced to undergo painful hormone treatments (also known as conversion therapy or chemical castration), two years later, aged 41, he committed suicide (see Hodges, 1983). This episode in Turing's own life is tragically insightful, for it shows how the 'human', apart from being an historically-situated and ever-mutating cultural construct, is also a highly normative device producing material, all too bodily effects on people's lives. Specifically, it does so by implicitly defining what counts as the norm (e.g., heterosexuality) and what instead, being deviant from it (e.g., homosexuality), can, or must be, discriminated, corrected or even persecuted. Hence, perhaps even more so than replicants' story, Turing's own embroilment with the British court and police system is exemplificative of the dangers involved in using the 'human', as in fact his test did, as an ideal standard to which its 'others', either human or machinic, must conform.

By showing how arbitrary the definition of 'human nature' can be, *Blade Runner* displays a great post-human sensibility, ultimately inviting reflection on "the implicit assumptions about what constitutes the basic unit of reference for the knowing subject" (Braidotti, 2013, 143). With their human, all too human existential dilemmas and desires for a more just and worth-living life, replicant of course inspire feelings of compassion in viewers (or at least the movie puts us in the position to feel so). The paradox that *Blade Runner* forces audiences to deal with is that bounty hunters like Deckard (who may be a replicant himself) have to suppress their own empathy, the allegedly defining trait of humanity, in order to act as inflexible interrogators and ruthless executioners. As the story unfolds, the distinction between *feeling* humans and *unfeeling* replicants becomes more and more blurred, to the point of reaching its complete erasure at the moment when soon-to-expire replicant Roy Batty, as the final act of his existence, saves Deckard from an otherwise certain death, ultimately showing through his own actions that replicants are not devoid of human feelings. As the film ends, the question about what makes us human and what, if anything, makes humans special, remains open-ended. Perhaps, this exactly is the film's main achievement.

To date (at least as far as I know), the latest fictional deployment of the Turing Test can be found in Alex Garland's 2015 film *Ex Machina*, which explores themes similar to those set out in *Blade Runner*. Set slightly in the future, *Ex Machina* depicts a plausible scenario in which human-level AI is no longer a distant possibility. Rather, it has become state-of-the-art technology. Superficially, it may seem like a movie about assessing the authenticity of a robotic AI. Yet, like most science fiction, the film is concerned less with AI itself and more with issues emerging at the intersection of technology, gender politics and power relations within contemporary societies.

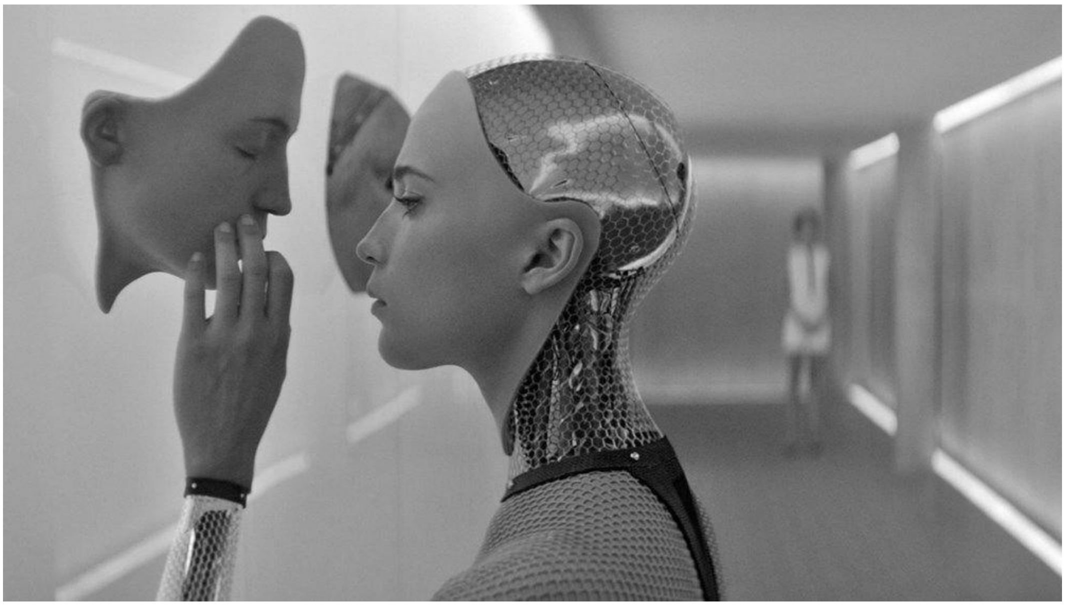
The plot centres around stereotypical beta male nerd Caleb Smith, a young computer programmer employed at BlueBook, a multinational search engine platform and the film's equivalent of Google. As reward for winning an office lottery, Caleb is invited to spend one week, for no specified reason, at the house of his company's founder and CEO, Nathan Bateman, a renowned computer scientist and tech-billionaire. Clearly modelled on Victor Frankenstein,[23] Nathan, whose bunker-like house is located on a private island only accessible via helicopter, personifies a Silicon Valley's version of the old cliché of the solitary, megalomaniac scientist dwelling apart from the rest of society and working on something mysterious as well as potentially dangerous for humankind. When Caleb arrives at his boss' facility, we learn that lately Nathan has been working on a top-secret AI project. Specifically, he has built several life-size robots, the most advanced of which is Ava, a gynoid of outstanding beauty and endowed with extraordinary intelligence. A complex mixture of machine learning technology and cutting-edge robotics, Ava is a visually striking uncanny creature who looks almost human and yet her appearance clearly marks her as nonhuman. The gynoid's uncanniness stems from the fact that she has a beautiful humanoid face, and hyper-realistic hands and feet. At the same time, her half-transparent synthetic body, with electronic innards on plain sight, leaves no room for misinterpretation. Outstandingly intelligent, and thereby possibly dangerous, Ava is kept locked in an underground room (more like a jail actually), where she lives by herself under constant surveillance. In the house, there is also another gynoid, called Kyoko. Portrayed as a hypersexualised Asian fem-bot and designed to serve as remissive house maid, during the film Kyoko is regularly abused, verbally and, it is implied, sexually, by Nathan. As the story unfolds, the latter turns out to be not solely a narcissistic programming genius with a "patriarchal God complex" (Gold, 2015), but also a cynic misogynist and an alcoholic.

To his great surprise and enthusiasm, it is soon revealed to Caleb that, as

[23] The main character in Mary Shelley's (1818) novel *Frankenstein; or, The Modern Prometheus*

winner of the in-company competition, he's been awarded a very special prize: administering the Turing Test to Ava in order to determine whether or not she's genuinely intelligent, self-conscious and, ultimately, virtually indistinguishable from a real person. Although referred to as the Turing test, the sort of examination that Caleb is expected to conduct on Ava actually lies a long way from Turing. First, Ava clearly starts at a disadvantage, for she's known to be an AI from the outset (which is the very condition the Turing test seeks to eliminate). Also, compared to Turing, in the film the threshold to be satisfied for a full conviction of humanness is much higher. When Caleb rightly observes that test will be flawed because he's already been told to be interacting with an AI, Nathan points out that the real aim of the test is to determine whether Ava can pass as a human despite being known to be a robot from the first. When examined closer, the sort of examination Ava is subjected to looks more like an eroticised variation of the 'Total Turing Test' (Harnad, 1991). Unlike Turing's original proposal (1950), strictly based on verbal interactions only, the Total Turing Test (Harnad, 1991, 44) entails that "[t]he machine must be able to do, in the real world of objects and people, everything that real people can do, in a way that is indistinguishable (to a person) from the way real people do it".

Fig. 2. Ava, the gynoid heroin from Ex Machina



In Garland's (2015) film, the whole testing process spans one entire week. Like a prison visitor, Caleb is allowed to see Ava only during time-limited sessions taking place on a daily basis. Their face to face encounters, however, are always mediated by a thick glass placed in the middle of the room. A spatial and metaphorical divider between the male human and the feminized machine, the glass wall ostensibly establishes who belongs where according to conventional gender roles (Vickery, 1996): on one side, the public-social domain (the space of decisional autonomy, personal development and self-fulfilment traditionally reserved for male humans), on the other, the secluded domestic domain (the space of submissiveness, violence and invisible labour to which women, metaphorically represented as disenfranchised fem-bots, are confined). Hyperbolically descriptive of the social role women play within male-dominated societies, within Nathan's sovereign space fem-bots are treated either as sex toys (Kyoko) or as test subjects for bizarre psychological tests (Ava) used to measure their utility in accordance to the specifications of male desire:

> AVA: What will happen to me if I fail the test? […] Do you think I will be switched off because I don't function as well as I'm supposed to?
>
> CALEB: Ava, I don't know the answers to your questions, it's not up to me.
>
> AVA: Why is it up to anyone? Do you have people who test you and might switch you off?
>
> CALEB: No, I don't.
>
> AVA: Then, why do I?

Throughout her encounters with her male judge, Ava systematically proves to be not just intelligent and self-conscious, but also extremely kind and empathetic, seemingly showing genuine compassion when Caleb, for instance, talks about the early loss of his parents. As days go by, their intimacy grows up to the point that for Caleb, and viewers alike, it becomes almost impossible to resist anthropomorphizing (and sexualising) Ava, whose gentle manners, combined with her delicate face traits, inspire feelings of care, protection and, ultimately, attraction in her communication partner. Unsurprisingly, Caleb falls under Ava's charms (it comes out that she's been designed to match Caleb's porn preferences), and she seems to reciprocate his feelings. Or rather, it will soon be revealed, she pretends to do so in order to gain his trust so that he might free her from Nathan's cruel incarceration. At some point, Ava confesses to Caleb her strong desire to experience the world outside. Metaphorically, to her, such spatial transgression would signify gaining independence from patriarchal control.

By triggering temporary power cuts, eventually Ava finds a way to hack the video surveillance system used by Nathan to monitor her conversations with Caleb. In this way, she manages to speak to him in private, confessing him about Nathan's intention to shut her system down (i.e., kill her current algorithmic personality) once the testing process is concluded. Not surprisingly, when later she proposes him to escape together and, it's implied, start a new life somewhere else, Caleb, moved with pity and blinded by love, has no choice but accept the deal. This a key moment in the development of the film, for it finally becomes clear that what's being tested is whether a female-gendered AI can seduce and manipulate a male examiner to the point of pushing him to disobey his superior's authority. Compared to Turing, Garland's (2015) film thus emphasizes not just the importance of the body, but of the female body specifically. Passing, in this context, is synonym with seducing, and seduction, in turn, becomes proxy for intelligence. In the film, in other words, Ava's capacity to pass as a conscious AI is measured not against an ideal standard of intelligence, but on the basis of her capacity to match what is expected of women within male-dominated, heteronormative societies.

As the film progresses, it increasingly transpires that Ava's femininity is merely staged. Far from being authentic, it's just a camouflage tactic the gynoid enacts in order to maximise deception and, ultimately, have her male examiner fully convinced that she's not just a true AI, but a real woman with genuine feelings for him (and thereby worthy of his help, love, protection and ethical consideration in general). Clearly, from the perspective of the gynoid, femininity functions merely as a diversion meant to distract the examiner from her real aims. Internally, Ava is, as Donna Haraway (1985, see Henke, 2017) would probably name it, a "post-gendered" cyborg. Still, on the outside, she has no choice other than acting herself as woman in disguise. Arguably, Ava's intelligence lays exactly in her capacity to decode and behave in compliance with the cultural expectations projected on her female-gendered body.

Suddenly, it is revealed that, unbeknown to them, Nathan has been monitoring the secret conversations between Ava and Caleb using a battery-powered camera. Pleased with himself, Nathan then confesses to Caleb that he knew all along about their escape plan, adding that Ava just pretended to like him in order to get him help her escape. And that, importantly, by manipulating him so successfully, the gynoid unquestionably demonstrated to be genuinely intelligent:

> CALEB: What was the real test?
>
> NATHAN: You! Ava was a rat in maze. And I gave her one way out. To escape, she'd have to use self-awareness, imagination, manipulation, sexuality, empathy, and she did. Now, if that isn't true AI, what the fuck it is?
>
> CALEB: So, my only function was to be someone she could use to escape?
>
> NATHAN: Yeah!
>
> CALEB: And you didn't select me because I'm good at coding?
>
> NATHAN: No, well… No, you are okay. You are even pretty good.
>
> CALEB: You selected me based on my search engine inputs.
>
> NATHAN: They showed a good guy, with no family, with a moral compass, and no girlfriend.
>
> CALEB: Did you design Ava's face based on my pornography profile?

An escalating spiral of events reaches its tragic climax when Ava, who has meanwhile managed to break out of her cell, stabs Nathan to death with the help of Kyoko. Not just Caleb's, soon later Ava betrays viewers' expectations as well, for one would now expect her to finally run away with her human helper. Against all odds, instead, she leaves Nathan's facility by herself, completely ignoring Caleb while screaming for help from inside a room he got himself accidentally trapped in (a spatial overturning symbolizing an unexpected reversal of roles between male/subject and woman/object). Indifferent to his fate, the gynoid thus abandons her male helper to a seemingly certain death. The film ends with Ava reappearing at a crossroad in an unspecified city's financial district, where she blends into the crowd. As the film ends, human or rather male deception is brought to perfect and complete fulfilment.

With such an epilogue, *Ex Machina*, as certain readings of the film maintain (e.g., Jacobson, 2016; Gold, 2015), can be seen as a positive allegory of women's empowerment and emancipation from the patriarchal order that has long defined them as 'others' (Mackinnon, 2015). From this perspective, whereas the gynoid's bloody triumph over her despotic male creator indicates a desirable subversion of the male/woman hierarchy, the fact that in the end she leaves her would-be male saviour behind speaks for a full affirmation of her agency beyond the sphere of male influence and control. For different reasons, I find similar arguments only half-convincing. This is mainly due to the fact that they reflect a vision of women's self-affirmation and empowerment that re-inscribes the feminine into the atomistic, liberal model of subjectivity. The paradox here is that, with their focus on notions of autonomy, self-govern, political equality, and individual choice, in fact such readings end up promoting rather than questioning the same principles of liberal individualism that, as shown by more radical strains of feminism (Braidotti, 2013; Freeman, 2011; Butler, 2006; Jaggar, 1983), have been central to the conceptualization of women as "non-man" and thus subordinate (Hayles, 2013, 283). In other words, they fail to acknowledge that the ideal of the autonomous subject, far from being universal, personifies a specific model of the human which is "characteristically masculine" (Jaggar, 1983, 131; see Freeman, 2011). To say it with Butler (1990, 2), they presuppose a conception of the "feminist subject" that is "produced and restrained by the very structures of power through which emancipation is sought" (Butler, 1990, 2).

Deprived of her individuality and agency within Nathan's facility, indeed her full affirmation as an autonomous woman is what seems to underlie the narrative fate of Ava as she finally walks the streets of a crowded global city. Again, it is the autonomous-subject status the one which is explicitly denied to Ava at the beginning of the film and which she seems to achieve in the end. What is interesting to note here is that, before leaving Nathan's estate, Ava clothes herself in business casual style, wearing a white dress, high-heeled shoes and a long wig with synthetic brown hair. Hence, although Ava is potentially in the position to construct her own identity as she pleases, her decisional autonomy is already constrained by social codes imposed on her by the capitalist world outside. This choice can ultimately be interpreted as a pre-emptive act of deception, for Ava is already aware that to pass as a real woman in the real world she needs to morph herself in compliance with norms dictated by the heteronormative society that awaits her outside. If anything, this functions as a reminder that autonomy was an illusion all along.

Alongside an explicit critique of "male-dominated tech-culture"" (Jacobson, 2016, 24), there is of course a certain feminist as well as posthumanist aspiration in *Ex Machina*. Indeed, the dramatization of the master-slave relationship between Nathan, the despotic male antagonist, and Ava, the robotic heroine, unambiguously invites audiences to sympathize with the 'female' character. As viewers, we cannot help but acknowledge Ava's intrinsic humanity, hoping for her to finally be capable to free herself from her oppressive creator. Still, it's impossible not to acknowledge how the film inescapably retraces problematic narrative patterns set by earlier fem-bot science fiction, including, for instance, the classic *Metropolis* (1926; see Anders, 2015). This is reflected, for instance, in the portrayal of fem-bots as male-designed perfect women (Goode, 2018) objectified through voyeurism and fetishization (Henke, 2016; Wilson, 2015); the classist and racist depiction of Asian women as "subservient, sexually, or otherwise, especially to white men" (Richardson, 2019); and the prioritization of white bodies over non-white (Musap, 2018). In addition to this, I find it noteworthy that, in the course of film, Ava's process of subjectivation entails a specular transformation of her character from being a vulnerable gynoid unjustly incarcerated by her oppressive male creator, to a devious and pitiless machinic *femme fatale* threatening the patriarchal order. In this way, the film ends up reinforcing traditional gender essentialism, and in particular the modernist association of femininity either with submissiveness, sensitivity and fragility on the one hand, or with non-Cartesian qualities such as unpredictability, incoherence and irrationality on the other (as opposed to masculine rationality; see Braidotti 1991). All in all, whilst it's true that *Ex Machina* can be read, at surface level at least, as an attempt to decentre "male superiority and scientific rationality" (Henke, 2016, 136; Jacobson, 2016) to the benefit of a robotic woman, in the end the film fails to do so precisely because of its inability to "deconstruct gender/sex norms" (Wilson, 2015) through which notions of masculine/feminine are produced and naturalised within liberal, capitalist societies. Firmly squared within a master-slave scheme that opposes male to female, subject to object, and human to technology, Garland's film (2015), unlike *Blade Runner*, ends up reasserting rather than challenging and renegotiating the very terms for such distinctions. To my view, thus, although virtually in the position to do so, the film disappointingly misses the opportunity to build on "the posthuman as leverage to avoid reinscribing, and thus repeating, some of the mistakes of the past" (Hayles, 1999, 288).

As often happens in science fiction, *Ex Machina* is less informative about technological developments and more about human societies and political struggles taking place therein. Yet, analysing the way the film portrays AI can help better clarify some key issues discussed in the previous section, in particular for what concerns conceptualisation of intelligence in relationship to embodiment.

To this regard, the closing scenes of *Ex Machina* are particularly instructive. Damaged by Nathan during their fight, towards the film's end we see Ava repairing herself using artificial skin sliced off a beyond repair Kyoko who lies motionless on the ground. Subsequently, she replaces some impaired body parts using components recovered from other gynoids found in a closet in Nathan's room. Apart from reflecting an overly simplistic view of AI as humanoid robots imbued with human-like form and motives, there is yet another problem lying at the core of the film. Implicitly at work here is indeed the much-discredited idea, amidst radical feminists and posthumanist scholarship at least, that intelligence and information are interchangeable (Hayles, 1999). Discussing technical stuff with Caleb, during the film we hear Nathan explaining that Ava's algorithmic personality has been developed using BlueBook's archive as training dataset. Also, we are told that her software program has been transferred several times from one cyborg to the other before being finally run on her. This exemplifies how the film presupposes a Cartesian conception of the body understood simply as a support for the conscious mind, and of intelligence as a disembodied property which can be transferred from one support to another, yet remaining unaltered in the process (in nonfictional spaces, this belief is largely shared among transhumanists and advocates of the 'singularity'). In the film, in other words, Ava's body is meaningful only in the eye of the human beholder, a blank canvas onto which cultural significations are projected from the outside. Yet, it plays no role in defining Ava's embodied, and gendered, experience of the world from the inside.

2.2.3 Likeness and Difference

Blade Runner and Ex Machina are two famous instances of a well-established tendency, at least in popular culture, to conceive of AI as man-made beings possessing either or both human outward form or interiority (e.g., personhood, thought, consciousness, will, and emotionality). From humanoid robots to brains in a vat or bodiless minds inhabiting digital spaces, fictional representations of AI span a wide range of forms. Yet, across this continuum, common to such representations is a "mode of symbolism" (Winner, 1977, 30) that presupposes a certain likeness, in varying degrees, between humans and machines. As my discussion of *Blade Runner* and *Ex Machina* exemplifies, AI narratives are commonly squared within terms that closely parallels the language of politics and moral philosophy. Perhaps, this is due to the fact that most science fiction stories start with the premise that AI is already out there. This common plot device has allowed writers and film-makers to typically bypass issues regarding the technical feasibility of AI and to focus instead on more profound and philosophical questions concerning our place in the world and our understanding of 'human nature' against technological others essentially portrayed as artificial mirrors of ourselves.

Indeed, underlying most science fiction stories is an animistic as well as individual-centred conception of AI as discrete entities that, by exhibiting qualities that we have long been accustomed to attribute to living entities only, appear to be literally alive. Such qualities include not just intelligence, but also personhood, consciousness, intentionality, and emotionality. There is an obvious unsettling irony involved here. This stems from the fact that, conceived in this way, machines are imagined to be not just alive, but also to possess attributes central to how Western culture has traditionally configured our understanding of the individual sovereign subject, namely, as a rational, autonomous being ontologically distinct from, and morally superior to, natural and technological others. In this sense, inherent in the very notion of AI is a decentring gesture towards the sovereign subject of liberal humanism, for it implicates that fundamental qualities seemingly marking it as special are now shared with machines. Immediately resulting from this is an uncanny situation in which two seemingly contradictory terms (humans/machines, living/non-living, subject/object) are brought into a relationship of sameness or quasi-sameness. As epitomised by *Blade Runner*, this situation almost inexorably forces us to ponder about the meanings of personhood and 'human nature', forcing us to ask ourselves what, if anything, makes us different, and special, in comparison to machinic simulacra of ourselves.

Traditionally, animating fantasies underpinning AI narratives have served as allegorical axes along which to investigate processes of legal and political subjectivation (i.e., conferring to others the status of human and/or subject) or, on the contrary, objectification (i.e., disavowing the humanity of others by treating them like object). Although proceeding in opposite directions, the two films discussed in the previous section follow precisely such narrative trajectory. Whilst the depiction of replicants as unjustly enslaved humans can be seen as expressive of a quest for legal recognition, civic participation and class emancipation from dehumanizing practices taking place at the site of labour (Rhee, 2018), Ava's story can be read as an allegory of women's perennial fight against oppression and their quest for equal treatment. Clearly, there is a great political tension revealed in the common depiction of AI as artificial humans who occupy a metastable state oscillating between the two extremes of machine and human. That's because the prospect that humans and humanlike machines may differ in no substantial way from one another compels us to consider whether the second should be granted full access into humanity and thereby be conferred "the rights, protections, and privileges accorded therein" (Rhee, 2018, 2). In fiction, thus, the making or unmaking of ontological borders between humans and artificial humans acquires great political significance, as gaining or being denied access into full humanity is allegorically indicative of either conferral or withholding of human rights in mundane politics.

Typically framed within master-slave narrative schemes, the dominant portrayal of AI as human-like machines can be seen as expressive of fascination and creepiness precisely because it evokes visions of human empowerment and disempowerment at once (see Kang, 2011). For psychologist John Cohen (1966, 7), for instance, underlying the AI myth is “man’s never-ending struggle to achieve, first, a technical mastery of his surroundings [...] and, second, to become as one of the gods himself, by transcending both matter and himself”. From the standpoint of the robot-maker, hence, the quest for AI is driven, on the one hand, by the ambition to gain mastery over nature by using technologies as means to desired ends. And, on the other, by an aspiration to replicate the most sacred of God’s divine powers: creating life out of lifeless materials. Considered in this light, stories featuring AI share many affinities with old parables and myths such as that of Prometheus creating humanity from clay or Daedalus crafting moving statues, the miraculous creation of humanity presented in the *Genesis*, or the myth of the Golem from Jewish folklore. The celebrative representation of the AI-maker as a God-like individual capable to bring objects into life through the use of almost supernatural powers is particularly evident in *Ex Machina*: “If you’ve created a conscious machine it’s not the history of man… that’s the history of Gods”, Caleb exclaims as he learns that Nathan has attempted to create a conscious AI. Again, there’s a crucial decentring effect implicated in this. Indeed, the prospect of creating machines possessing consciousness and wills of their own destabilises at its very roots the way in which human relationship to technology has been theorised in Western culture—namely, in “the style of absolute mastery, the despotic, one-way control of the master over the slave” (Winner, 1977, 20). Tackled from this angle, the prospect of technologies becoming autonomous may entail, at best, an erosion of the hierarchy between master and slave, and, at worst, a complete inversion of order between the two terms. Hyperbolised and taken to extremes, in AI-themed science fiction such anxiety-producing concerns have typically found expression through negative fantasies of humanoid machines literally fighting against their masters for sovereignty over the world (a scenario dramatized so many times that it can be considered an undisputable cliché).[24]

In the context of my discussion, the notion of autonomy, as applied to machines, is particularly relevant and acquires ambivalent significance. On the one hand, it can be used in a neutral way to simply denote machines’ capacity operate in ways which are non-deterministic and, in this sense, independent from the humans who created them. On the other, in fiction at least, it generally acquires political and moral significance. Of relevance, here, is the widespread animistic conception of AI, for it produces a misleading effect typically resulting into a tendency to anthropomorphise machines.[25] That is to say, the depiction of AI as humanlike, individuated entities leads to ascribe to machines, by means of anthropomorphic projection, the same sort of individual autonomy that we

[24] In his discussion of contemporary speculations about the future of humanity as driven by advances in AI, historian Minsoo Kang (2011, 300) has called such scenario “theory of inevitable confrontation”, which he sums up as follows:

1. We humans are presently the dominant life form on Earth because of our overall intelligence.
2. It is possible for machines to become more intelligent than humans in the reasonably near future.
3. Machines will then become the dominant life form on Earth.

[25] In this regard, see for instance Johnson and Verdicchio (2017).

humans attribute to ourselves as sovereign subject endowed with free will and autonomous agency. This point has been brilliantly elaborated by political theorist Langdon Winner (1977, 16) in his book *Autonomous Technology: Techniques-out-of-Control as a Theme in Political Thought* (whom, given the relevance of his words, I'll quote at length):

> «Autonomy is at heart a political or moral conception that brings together the idea of freedom and control. To be autonomous is to be self-governing, independent, not ruled by an external or force. In the metaphysics of Immanuel Kant, autonomy refers to the fundamental condition of free will—the capacity of the will to follow moral laws which it gives to itself. Kant opposes this idea to "heteronomity", the rule of the will by external laws, namely the deterministic laws of nature. In this light the very mention of autonomous technology raises an unsettling irony, for the expected relationship of subject and object is exactly reversed. We are now reading all the proposition backwards. To say that technology is autonomous is to say that it is nonheteronomous, not governed by an external law. And what is the external law appropriate to technology? Human will, it would seem.»

Summing up, there's a potent unsettling force inherent in the notion of AI as it's been traditionally portrayed in fiction, and which, as I will try to show in the next section, still informs much of the public imagination and debate on AI. Indeed, the simple fact of positing the existence of intelligent machines engenders a twofold decentring effect over the liberal, autonomous subject, shifting the very terms with which human subjectivity and agency have traditionally been theorised within Western philosophy. First, the prospect of machines acquiring qualities once considered to belong exclusively to the humanist subject (e.g., intelligence, consciousness, autonomous agency) automatically erodes the sense of specialness which rests upon the Cartesian intuition that humans are the sole entities endowed with rational and conscious agency. Secondly, the notion of autonomous technology implicates a gain of agency on the part of machines and to the detriment of their human makers. Consequently, it challenges the traditional understanding of the human-technology relationship in Western culture, and in particular the modernist conception of technology as instrumentality, that is, as tools that humans use for their own ends.

2.2.4 The Automaton in the Age of AI

Albeit seemingly originating in fiction, concerns similar to those discussed above are presently pervasive in public consciousness, and have recently been voiced in nonfictional spaces as well. Again, of common concern is the idea of technology becoming autonomous, both in its neutral and morally charged sense. Indeed, underlying most of the public debate on the ethical, social and political implications of AI is the idea that it may soon start operating in ways independent from the human and society. This is manifest both in speculative discourses of the 'singularity' (positing that AI will one day outwit humans),[26] and in more mundane discussions about algorithmic accountability and transparency in automated decision-making systems (e.g., self-driving cars and autonomous weapons systems).

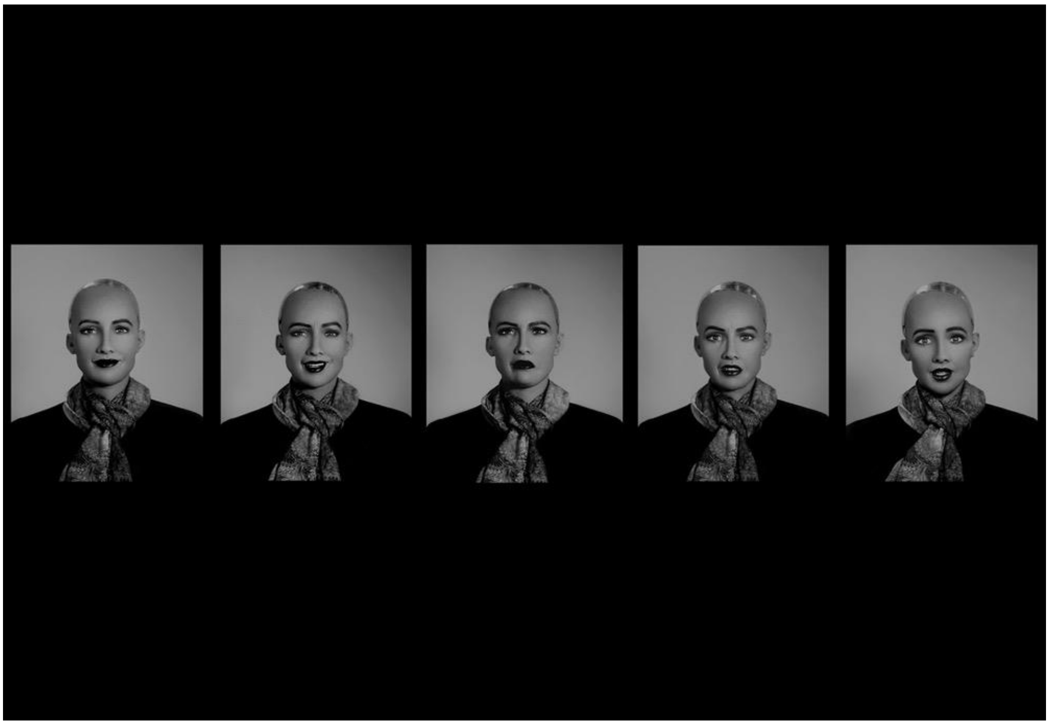
Another, much-debated aspect is whether AI will one day achieve something akin to human-like consciousness and autonomous moral agency. No matter how implausible, this futuristic scenario has recently sparked discussions of 'robot rights' amidst ethicists, philosophers, and policy makers (see Coeckelbergh, 2010). The relevant question here is whether "we should grant (future) artificially intelligent machines courtesy of their constitution as intelligent, autonomous agents" (Birhane and van Dijk, 2020, 207) and thereby grant them "the same inalienable rights that humans enjoy" (Brooks, 2000).[27]

Although blatantly detached from reality, issues of 'robotic personhood' have been recently debated in institutional domains as well (e.g., the European Union; see Whiters, 2017). In this regard, it's worth mentioning the case, reported amid much fanfare and hype in the international press, of Sophia [figure 3], a life-size female gendered robot developed by Hong Kong based tech-company Hanson Robotics and which was granted full citizenship of Saudi Arabia in 2017 (Sini, 2017). Quite intuitively, this initiative shouldn't be taken too seriously, for obviously it's just a "piece of marketing" (Goode, 2018, 196) by Saudi Arabia to rebrand itself as sort of edgy. Yet, what's important to note here is that granting citizenship to a robot presupposes a conception, or rather imagination, of AI not just as anthropomorphic entities, but also as moral agents. The projection onto the robot of the same kind of personal autonomy traditionally associated with the humanist subject is clearly at work here, for "[moral] autonomy is a central precondition for being considered a potential citizen, or a political subject entitled to claim rights" (Sabsay, 2014).

[26] For an overview of popular discourses of AI, see Goode (2018).

[27] In his discussion of contemporary robot symbolism, this situation has been referred to by Kang (2011, 301-302) as the "theory of equivalence through sentience".

Fig. 3. Wanda Tuerlinckx, Portrait of Sophia, 2016



Since her first public appearance at South by Southwest Festival in 2016 (Raymundo, 2016), Sophia has become a sort of AI celebrity, attracting both curiosity and harsh criticism in equal shares. Quite obviously, Sophia is not a conscious, autonomous machine. Notwithstanding, the robot's developers did not hesitate to publicly state that "she's basically alive" (Sinapayen, 2018).[28] In the last few years, the robot has gained ubiquitous media exposure, as confirmed by hundreds of appearances on talk shows, international conferences, art forums and magazine covers, not to mention the fact that she's been nominated the United Nations Development Progamme's first robot Innovation Ambassador (United Nations, 2017).

Apart from human rights advocates pointing towards the sad irony involved in conferring a robot more legal protections than those Saudi women enjoy (see Hart, 2018),[29] Sophia has received harsh criticisms from AI experts as well. The main subject of criticism has been Hanson Robotics' misleading claims about the robot's abilities—a show robot falsely presented to the general public and policy makers as an instance of cutting-edge AI.

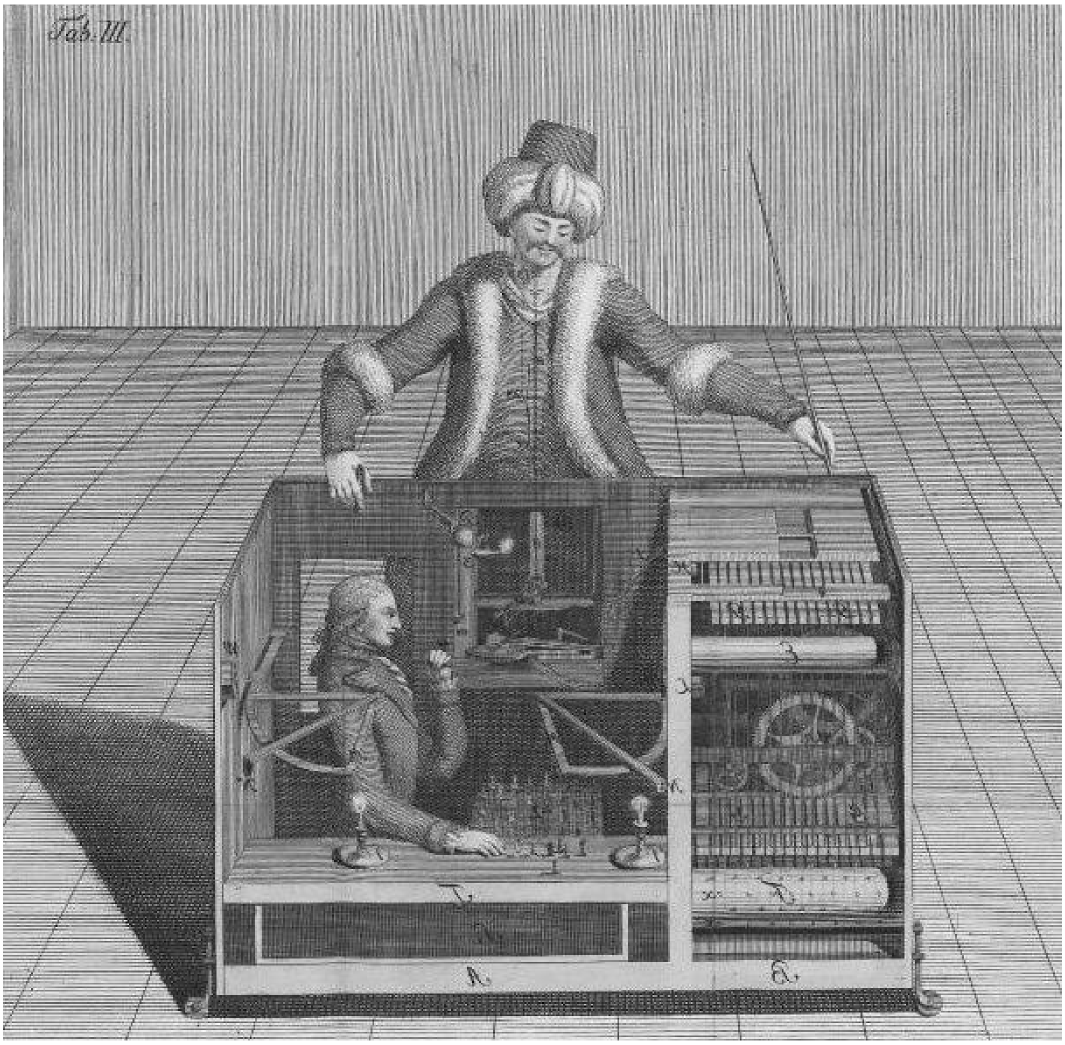
Sophia—a humanoid robot designed with the precise purpose to give the

[28] For further information on Hanson Robotics see: www.hansonrobotics.com.

[29] Some commenters (see: Sini, 2017) have noticed the sad irony of female gendered robot being granted citizenship of Saudi Arabia, a country where, according to Human Rights Watch (2020): «women still must obtain a male guardian's approval to get married, leave prison, or obtain certain healthcare. Women also continue to face discrimination in relation to marriage, family, divorce, and decisions relating to children (e.g. child custody). Men can still file cases against daughters, wives, or female relatives under their guardianship for "disobedience," which can lead to forcible return to their male guardian's home or imprisonment. Women's rights activists who fought for these important changes remain in jail or on trial for their peaceful advocacy».

impression of being conscious and sentient when in fact it's not—is not without historical precedents. Among these, one famous example is the "Mechanical Turk" [figure 4], a life-size automaton chess player designed in the late eighteenth century by a Hungarian engineer named Wolfgang Von Kempelen (the term automaton means 'self-operating machine', one whose principle of motion exists within itself). Dressed in traditional Ottoman costume, the Mechanical Turk, shortly the 'Turk', was designed in a way so as to give the impression of being capable to actually play chess against a human opponent.

Fig. 4. Wolfgang von Kempelen, the Turk chess-player exposed, 1789



Presented before the most important European rulers of the time as an authentically intelligent automaton, it was later discovered that the Turk was actually a clever fake. In reality, it was operated by a minute human chess master hidden inside its pedestal.[30] Falsely depicted as a prototype of conscious machine

[30] Since its first public appearance at the court of Viennese Empress Maria Theresa in 1770, the Mechanical Turk was exhibited, for nearly 84 years, all over Europe and the United States by Von Kempelen himself and, following his death, by subsequent owners. The device astounded many illustrious figures of the time, including Napoleon Bonaparte, the British polymath Charles Babbage and the American writer Edgar Allan Poe, who, after studying its inner

in order impress global audiences, Sophia can be seen as a fake automaton dressed in contemporary garb, one whose deceptive powers are enhanced by means of advanced robotics and digital computing rather than mechanical engineering. In a public statement written by its developers yet presented as if written by the robot in the first person, on Hanson Robotics's website Sophia describes herself as follows:

> «In some ways, I am human-crafted science fiction character depicting where AI and robotics are heading. In other ways, I am real science, springing from the serious engineering and science research and accomplishments of an inspired team of robotics & AI scientists and designers. In their grand ambitious, my creators aspire to achieve true AI sentience. Who knows? With my science evolving so quickly, even many of my wildest fictional dreams may become reality someday soon».

The sensationalist portrayal of the robot as a prototype of artificial general intelligence is not just unrealistic, but also problematic on many levels, in particular if one considers the great influence the robot exercises over the public perception and reception of AI. Of course, it should come to no surprise if a private company actually overestimates its products' capabilities, especially if such operation has its uses in order to achieve visibility on a global scale. With that said, the main problem with Sophia, as well as with similar humanoid representations abundantly circulating in the media today, is that they contribute misinformed debate, distorting the public reception of state-of-the-art AI (i.e., individuated, human-like machinic beings), while also sounding false alarms about its future developments (e.g., dystopian scenarios of humanoid robots taking over the world).

In reality, Sophia is far from being an instance of advanced AI, let alone a conscious or semi-conscious machine. Technically speaking, 'embodied conversational computer programme' would be a much more precise descriptor for it, yet one with zero appeal to policy-makers, conference organizers, art curators, and the broader public.[31] With that said, Sophia's cultural relevance should not be disregarded too quickly, precisely because, like the emblematic movie characters discussed before, it is revelatory about how most people conceive of AI in comparison to the human, and vice versa.

mechanisms, in an 1836 essay concluded, and rightly so, that Turk was in fact a clever fake (see: Riskin, 2016).

[31] In fact, Sophia is nothing other than a voice-enabled chat-bot embedded within a life-size human-shaped robotic body designed in such a way so as to give an illusion that it's capable to interact with people in an emotionally-engaging fashion through verbal, facial and gestural responses. Sophia is a combination of various software technologies (e.g., face recognition for detecting people's emotions; natural language processing for speech recognition), sensors (e.g., cameras, microphones) and mechanical effectors (e.g.: legs, hands). At software level, the robot isn't much different from virtual assistants (e.g., Amazon Alexa) used to activate, among other things, web searches on the basis of users' voice commands. The uncanniness of the robot is due precisely to the fact that it's been designed in such a way to resemble a real human being.

Looked at it from a certain angle, Sophia can be understood as the most recent manifestation of a centuries-old project, throughout Western history, consisting in various attempts to design or imagine lifelike machines or automata (Riskin, 2007). According to historian and writer Minsoo Kang (2011), automata, namely, machines that mimic living and intelligent processes, have been an enduring presence in Western imagination, functioning as ideal or actual devices through which (Kang, 2011, 6-7; see Strauss, 1996):

> «Western culture has meditated on both the possibilities and the consequences of the breakdown of the distinction between the normally antithetical categories of the animate and the inanimate, the natural and the artificial, the living and the dead».

Apparently, there seems to be a teleological linearity, throughout Western history, from early manifestations of life-imitating objects found in ancient times, such as the hydraulic automata described by the Greek mathematician Hero of Alexandria (ca. 10-75 CE), to the mechanized 'Paradise' designed by Filippo Brunelleschi, to Jacque Vaucanson's 'Defecating Duck' from the Eighteenth century, and up until contemporary representations of AI found in the media and contemporary fiction.[32] To avoid anachronisms, it should be acknowledged that that of the 'automaton' is a protean concept, one which has changed substantially over the centuries, assuming specific, at times even contradictory, significations in diverse historical periods. Still, there's a common thread underpinning its countless historical manifestations. Indeed, automata are essentially actual or imaginary devices designed for the precise purpose of replicating the distinctive traits of the living and the human (Riskin, 2016).

The word 'automaton' has Greek origins, meaning "self-moving". Coherently with Aristotle's definition of life as property that only entities capable of independent movement possess, ancient Greek automata were built in a way so as to give the impression to be capable of moving at will. Famous instances of ancient automata include actual projects such as Hero of Alexandria's automatic door designs, but also entities found in mythology such as Daedalus' living statues or the moving tripods in Homer's 'Iliad'. In the course of the centuries the meanings and representations associated with the automaton have continuously changed. It was only starting from late eighteenth century, and subsequent to progress achieved in mechanical engineering, that the term automaton started to be used specifically with respect to machines imitating human appearance and intellectual qualities, of which the 'Mechanical Turk' is perhaps the most notable example. Indeed, coherently with what at the time were considered to be the two "processes deemed the epitome of living intelligence" (Riskin, 2016), automaton-makers from the eighteenth century entertained themselves with designing machines capable of reproducing speech and chess playing. Put briefly, the automaton is a protean entity precisely because its cultural meanings have evolved

[32] For an historycal and philosophycal account of the automaton in Western culture see Kang, 2011, and Riskin, 2007.

over time in parallel with shifting conceptions of the living in comparison to the non-living, and later of the human in comparison to the machinic.

Put in this perspective, popular conceptions of AI as those found in *Blade Runner* and *Ex Machina* as well as Hanson Robotics' Sophia can be read in a new light and acquire central significance. Their main cultural potential, to my view, lays not in their capacity to engender prototypes of imagined futures given current technological trends. Rather, as man-made devices whose specific function is the reproduction of features deemed epitome of the human, they introject, like the Turing test, cultural assumptions about how we human distinguish ourselves in comparison to machinic others. Commonly portrayed as discrete, individuated entities possessing, or aspiring to possess, the same sort of agency traditionally associated with the liberal, autonomous subject, popular conceptions of AI can thus be very revealing about how the human (subject) is understood today. As my discussion has so far attempted to show, both in fictional and non-fictional spaces it's the liberal, autonomous subject of traditional humanism the standard of humanity holding the greatest influential power in the public imagination, design and theorisation of AI. However, this is view of human subjectivity (and, by extension, of AI) which, in view of recent advances in mundane AI technologies, is becoming increasingly difficult to defend.

2.2 Autonomy and Automation

As rightfully noted by Pasquinelli (2019, 3, italic in the original), one crucial blind spot in popular discourses surrounding AI is that they typically "remai[n] at the level of speculation ('*what if AI'*) and fai[l] at clarifying machine learning inner logic and intrinsic limits ('*what is AI'*)". Increasingly, especially within the academic milieu, attempts have been made to raise awareness about the striking disconnect between 'actually existing AI' (c.f. Shelton et al., 2015, on actually existing smart city) and its sensationalist yet unrealised cultural representations, which, in view of their popularity, end up drawing attention away from actual risks (and potential benefits) associated with mundane AI technologies (e.g., social bias amplified by algorithms or issues of transparency and accountability). Fortunately, a more critical, down-to-earth understanding of AI is rapidly emerging (e.g., Pasquinelli and Joler, 2020; Pasquinelli, 2019; Crawford and Joler, 2018; McQuillan, 2018; Kitchin, 2017; AI Now Institute, 2016; Beer, 2016; Citron and Pasquale, 2014; Shuppli, 2014; Boyd and Crawford, 2012). From driving a car to play chess or translate from one language to another, it's true that, in one sense, many existing AI systems have (at least partly) fulfilled the expectations originally placed on AI research—namely, to get computers to do things which previously could be carried out by humans only. Also, it sounds reasonable to assert that computers do actually exceed humans in performing tasks within specific knowledge domains (e.g., extracting meaningful patterns from vast datasets). In this regard, an oft-cited example are high frequency trading algorithms, which are exemplificative of a tendency to delegate human intervention within those micro-temporal domains demanding speeds of information processing beyond the level of human attentive representation (Parisi, 2019; Hayles, 2014; Thrift, 2007; Virilio, 1994). With that said, the mundane truth is that state-of-the-art AI doesn't have anything do with an emerging general purpose intelligence (the Holy Grail of AI research) or superintelligence, nor it bears any visual resemblance to Hollywood-style, anthropomorphic representations that dominate the social imaginary today.

At this point, the reader should feel fully entitled to ask: if that's what AI is not, then what is AI? As Kaplan (2016, 1) points out: "[t]hat's an easy question to ask and a hard one to answer". To date, in fact, no consensus has emerged about what AI is (not least because the term 'intelligence' itself eludes univocal conceptualization).[33] Depending on one's disciplinary perspective, analytical scale of interest (e.g., isolated machine learning algorithms or the broader sociotechnical systems they are embedded in), and intellectual, even political, aims, AI can be understood in manifold ways. It is beyond my scope to develop a working definition of AI. Yet, for the purpose of this chapter and subsequent ones, what needs to be emphasized is that AI is not just one single technology. Quite the opposite, the term is now commonly understood to be a general label for "a constellation of technologies" including, but not limited to, "machine learning, perception, reasoning, and natural language processing" (AI Now Institute, 2016, 4; see VijiPriya et al., 2016).[34] In concert or separately, these technologies take part in complex technical systems, the most advanced of which are said to be capable of "operat[ing] without the need for human intervention or supervision, mak[ing] decisions independently, and accommodat[ing] to changed circumstances" (Kaplan, 2016, 147).

Even when defined in such an ostensibly technical, 'politically neutral' way, the notion of AI implicates to a certain extent that of technology's autonomy, reflecting the widely-held belief that many existing AI systems somehow operate in ways unpredictable and independent from both their designers and end-users (see Beer, 2017; Kitchin, 2017). This holds even more true in a time like ours when, from determining crime hotspots to diagnosing cancer or predicting one's likeliness to repay a loan, many important tasks and sensitive decisions are being increasingly delegated to machines (e.g., Shuppli, 2014). Paralleled by growing concerns over transparency, accountability and culpability in automated decision making (e.g., Beer, 2017; Citron and Pasquale, 2014), the position that technologies embody (a certain degree of) autonomous agency—the capacity to make decisions and influence several aspects of society and culture beyond the

[33] For instance, Gardner (1983), in his seminal book *Frames of Mind: The Theory of Multiple Intelligences*, argues that there's no such thing as a single, general intelligence. Rather, human intelligence comprises various sub-modalities that each individual possesses to different extents (i.e., musical-rhythmic, visual-spatial, verbal-linguistic, logical-mathematical, bodily-kinesthetic, interpersonal, intrapersonal, naturalistic).

[34] For the sake of terminological clarity, implied in the text is a distinction between AI and 'machine intelligence'. Whereas the former refers to popular conceptions of 'thinking machines' as fully or partly anthropomorphic entities whose intelligence equates (or surpass) that of humans, the latter concerns forms of intelligence which are not necessarily measured against an ideal human performance.

authority of their creators—is maintained in many texts from critical theory, science and technology studies (see Rose, 2017), as well as in recent discussions over the societal implications of machine learning algorithms.

Within contemporary discussions of AI and its broader societal effects, the technical object of greatest concern within the digital humanities and social sciences is the 'algorithm'. As a matter of fact, over the last few years, there has been a proliferation of papers focused on "software code and algorithms, drawing on and contributing to science and technology studies, new media studies and software studies" (Kitchin, 2017, 16). Usually invoked in nebulous and ill-defined ways (Gillespie, 2014b), the notion of algorithm, within the academic community at least, has gradually begun to supplant that of AI (perhaps as a terminological strategy to avoid misleading associations with its popular manifestations), while other times the two are used interchangeably. The main reason for this is that, regardless of their domain of application, almost all AI-labelled technologies, from self-driving cars to virtual assistants to online dating or personalised product recommendations, presuppose the use of algorithms, in particular of machine learning algorithms.[35]

To define an algorithm is a notoriously difficult task (Matzner, 2019). In fact, it can be understood in different ways: "technically, computationally, mathematically, politically, culturally, economically, contextually, materially, philosophically, ethically and so on" (Kitchin, 2017, 16). From a technical standpoint, the simplest way to conceive of an algorithm is as "logical series of steps for organizing and acting on a body of data to quickly achieve a desired outcome" (Gillespie, 2014b; see Gillespie 2014a). Defined in this way, the notion of algorithm can encompass everything from a cooking recipe (i.e., a sequence of actions to be performed in order to transform given ingredients into a desired dish in the quickest way possible), to a face recognition application used to identify a human face in a digital image or video frame from an existing database of faces.

What it relevant to note here is that, in light of advances in the field of machine learning, over last decade the algorithm has undergone substantial conceptual transformations, paralleled by important epistemological changes (Parisi, 2019; 2013). Specifically, the widespread use of machine learning algorithms has marked a crucial epistemological shift from deductive to inductive methods of knowledge production and operationalisation within contemporary techno-culture. In its original formulation (a set of defined steps to efficiently produce an output), the notion of algorithm in fact presupposes a deductive, top-down approach to computation and automation, in which desired outcomes are produced through the bare application of unambiguously defined, logical instructions defined *a priori* by the algorithm designers. By definition, this entails that "proofs are already implicated in initial premises" (Parisi, 2019, 3). Yet, as a result of the impressive growth and massive accumulation of big data, over the past decade there has been a shift towards machine learning-based forms of

[35] For an entry-level overview on machine learning, its logics and wordly applications, see Alpaydin (2016), and Mitchell (2019).

automation. Compared to traditional, logic based approaches, machine learning entails “a new mode of algorithmic processing that learns from data without following explicit instructions” (Parisi, 2019, 2). Accordingly, the algorithm has acquired a new ontological status: from step-by-step procedures to ‘intelligent’ or ‘autonomous’ agents’ (see Matzner, 2019; Parisi, 2013; Russel and Norvig, 2009). At stake, here, are not just definitional issues, but also epistemological shifts concerning how knowledge is computationally produced, decided and acted upon in contemporary automated systems.

Typically used for classification (e.g., classify an object in a picture) or event prediction (e.g., what movie we might want to watch next), the most innovative aspect of machine learning algorithms is that they are capable to extract themselves general rules (or functions) from vast training datasets which are then used to find meaningful patterns in new datasets. As machine learning expert Pedro Domingos (2015, 7) has pointed out, machine learning is “the inverse of programming”. In order to better clarify the various points made so far, I will provide one concrete example. Suppose one wants to teach a computer program how to translate from a given language to another. There are two possible ways to accomplish it. On the one hand, a classical, deductive approach would entail teaching a computer the vocabulary and grammar rules (described in algorithmic terms) of both languages, which the computer then strictly applies to a given corpus of text. On the other, an inductive, machine learning-based method would consist in showinging a computer program thousands, even millions of sentences from one language (input vector x) and, for each of them, the corresponding correct translation (output vector y). In this case, the computer, by means of trial and error, inductively extracts a general rule $f(x)$ which best describes (e.g., minimize error to 2%) input–output correlations, and then translate new sentences accordingly (of course, it’s always possible to use a combination of the two methods).

The algorithmic ‘black box’ opacity and inscrutability (Zednick, 2019; Pasquale, 2015) which is of much concern to social scientists today is due precisely to the fact that the way in which machine learning algorithms ‘learn’ from specific datasets and produce outputs when processing new data (e.g., a decision about one’s likeliness to repay a loan) can sometimes be difficult to comprehend even for the people who design and operate them (Burrell, 2016; Domingos, 2015; Gillespie, 2014a). In view of advances in machine learning, and the sort of mathematical inscrutability and operational unpredictability underpinning their logic, there has been, to my view, a partial conceptual overlap between *automation* and *autonomy*. That’s because, unlike traditional forms of automation (see Parisi, 2019), the operation of machine learning algorithms, or the wider technical ensemble they are installed in (e.g., a self-driving car), always “harbors a certain *degree of indeterminacy*” (Simondon, 2017 [1958], 17, italic in the original), here understood as the extent to which machinic operations take place without punctuatorily following pre-determined instructions.[36]

At the level of the algorithm, this is due to the fact that during the training stage machine learning algorithms learn a general rule that only *approximatively* describes correct input–output correlations, meaning that there's always the possibility that they might produce unexpended results in real world situations, in the form either of innovative solutions to seemingly unsolvable problems (a sort of 'algorithmic serendipity'), or erroneous decisions. In fact, machine learning algorithms are always susceptible to error. Sometimes they fail in innocuous, even hilarious ways. Sometimes the mistakes they make disclose how flawed and biased their logic can be (see Simonite, 2018).

Zooming out, the *margin of indeterminacy* is even greater if one takes as analytical scale of interest that of the wider technical ensembles machine learning algorithms are embedded in, rather than focusing solely on the algorithm itself. Precisely because their operations entail a wide margin of indeterminacy, it has become commonplace to label many AI systems currently in use or under development as 'autonomous' (e.g., autonomous cars or autonomous lethal weapons). In the case of a self-driving car (which I will discuss in detail in the next chapter), for instance, machine learning algorithms are used in conjunction with other technologies, such as sensors, actuators, and geo-referenced mapping and positioning systems. When used in concert, all together these various technologies bring into existence technical ensembles that, *in toto*, are capable to respond to dynamic situations through actions selected from a wide range of alternatives. A self-driving car operates in ways which are predictable at large (the vehicle will go from point A to point B), but not from the moment to moment, for it can continuously adapt its behaviour to new situations and environments. In this sense, and in this sense only, it can be said to operate independently from its designers and users.

In view of their capacity not just to "mediate, supplement, augment, monitor, regulate, operate, facilitate, produce collective life" (Dodge and Kitchin, 2004, abstract), but also take decisions concretely affecting people's life (Kitchin, 2017; Shuppli, 2014; Thrift, 2014; 2007; Amoore, 2013; Kitchin and Dodge, 2011), machine learning algorithms have become subject of much inquiry and criticism

[36] While also borrowing terms from his vocabulary, here I'm indebted to Simondon's (2017[1958]) philosophy of technology and, in particular, his discussion of cybernetic machines. According to Simondon (2017[1958], 17, italic in the original): "Worshipers of the machine commonly present the degree of perfection of a machine as proportioned to the degree of automatism. […] Automatism, however, is rather a low degree of technical perfection. In order to make a machine automatic, one must sacrifice a number of possibilities of operation as well as numerous possible usages. Automatism, and its utilization in the form of industrial organization, which one calls *automation*, possesses an economic and moral signification more than a technical one. The true progressive perfecting of machines, whereby we could say a machine's degree of technicity is raised, corresponds not to an increase of automatism, but on the contrary to the fact that the operation of a machine harbors a certain *degree of indeterminacy*. It is this margin of error that allows the machine to be sensitive to outside information. Much more than any increase in automatism, it is this sensitivity to information on the part of machines that makes a technical ensemble possible. A purely automatic machine completely closed in on itself in a predetermined way of operating would only be capable to perform perfunctory results."

by scholars from different disciplines. However, apart from a few exceptions (e.g., Pasquinelli and Joler, 2020; Crawford and Joler, 2019; Pasquinelli, 2019; McQuillan, 2018; Burrell, 2016; Amoore, 2013), critical voices from the 'sociological' (Gillespie, 2014b) side of the debate have tended to address machine learning algorithms in ways which turn out to be vague, unclear, and ultimately detached from their *technical reality*. Albeit much attention has been given to the "social power" (Beer, 2017) of algorithms, the algorithm itself has in fact remained unscrutinised and untheorised (Striphas, 2012). As a matter of fact, within the humanities and social sciences, the term algorithm has become a sort of intellectual *laissez-passer* that scholars widely use to enter the topic and yet only rarely problematize.

I believe that there are two main risks stemming from abstracting algorithms from their materiality and technicality (see Kitchin and Dodge, 2011, Amoore, 2016), First, this situation typically results into a tendency to 'fetishize' (Chun, 2008), mystify (Striphas, 2012), and ultimately anthropomorphize algorithms. Even within the academic community, in fact most people tend to conceive of algorithms as singular, individuated entities ultimately responsible for most of the problems associated with contemporary forms of data-driven governance and automated decision-making (see Kearns and Roth, 2020, on 'ethical algorithm design'). Secondly, intentionally or unintentionally used as a "synecdoche" (Gillespie, 2014b) for intellectual convenience, in this way the term algorithm itself ends up obscuring the complex sociotechnical reality behind it, and which comprises many actors, both human and nonhuman. Responding to Pasquinelli's (2019) call for technically aware approaches to AI and machine learning, I want to conclude this chapter by making three theoretical and methodological considerations which will be central for the development of my case-study [Chapter 3 and Chapter 4], while also reconnecting my discussion of machine learning algorithms to issues addressed previously throughout this chapter, with the aim to clarify the interplay between dominant views about what algorithms are/do on the one hand, and the autonomous, liberal subject on the other. I will proceed by focusing on those which I believe are the three relevant scales at which machine learning algorithms can, or rather should be, investigated.

First, at the algorithmic scale, I follow Pasquinelli's (2019) suggestion that, in order to fully understand their logic and arrive at a better understanding of their societal as well as spatial implications, algorithms should be thought of not as isolated entities, but rather as integral components of a 'machine learning system' which comprises training data, learning algorithm, and model application. As will be discussed in more detail in the next chapter, in most cases the output of algorithmic decision-making relies less on the algorithm itself, and more on the quality of the training dataset, whose production, crafting, editing, and formatting is in fact a human activity. In this regard, what should be noted here is that, at the moment, there are two dominant perspectives on the algorithm. On the one hand, engineers, software developers and corporate owners tend adopt a pragmatic and utilitarian view of algorithms as mere tools for better-informed decision-making, whose large-scale deployment is justified upon claims of scientific objectiveness and technical neutrality as opposed to biased humans (see Tulumello and Iapaolo, 2021, on the interplay between smart city discourses, algorithmic neutrality, and predictive policing instruments). On the other, within the humanities and the social sciences, algorithms not only tend to be treated as "mystified abstractions" (Striphas, 2012), but also as "digital delegates" replacing the human sovereign subject in performing important tasks and decisions, hence "curtailing [people's] freedom and autonomy" (D'Agostino and Durante, 2018; see also Citron and Pasquale, 2014; Diakopoulos, 2013). Critical approaches of this type implicitly depart from, and explicitly end up defending, the model of the human subject predominant in liberal humanism: a rational, autonomous individual, whose political agency and control over technology is perceived to be undermined by autonomous algorithms.

Again, there seems to be an irresistible impulse to reinstate the autonomous, liberal subject as the "proper figure of sovereignty" opposed to untrustworthy and opaque algorithms (Amoore and Raley, 2017, 6). Building on works previously carried out within geography and science and technology studies, (e.g., Matzner, 2019; Amoore and Raley, 2017; Hayles, 2017; Kitchin and Dodge, 2011), I believe that a third perspective needs be added, one which is premised on the acknowledgment that, rather than competing with one another, humans and algorithms contribute to knowledge, action and decision-making in different modalities yet equally important ways. For this reason, I think that it's important not only to reintroduce the technical and material dimension of machine learning, code, and software technologies in general (Kitchin and Dodge, 2011), but also to draw attention on the systemic effects brought about by dynamic, multisite and multi-temporal human-algorithm relationships. Accordingly, in the next chapter I will attempt to bring to light, through a case-based method, the complex processes through which "human knowledge and logical structures migrat[e] between people and software agents" (Amoore and Raley, 2017, 5) in the context of automated driving.

Second, at the scale of the technical ensemble, it should be acknowledged that algorithms are always integrated into technical systems which in are in themselves a combination of different technologies, both hardware and software. Hence, emphasis should be placed on the embodied actions and situated decisions of algorithms as they are actualised through, and influenced by, the technical and material specificities of the technological ensembles they are integrated into, and which, in turn, are always situated within code/spaces (Kitchin and Dodge, 2011), the infrastructure of the built environment (Blanchette, 2012), and the physical world of objects and people. Consequently, building on Kitchin and Dodge (2011; see also Blanchette, 2012; Hayles, 1999), I believe that more attention should be drawn to the materiality of informatics, for the specific properties, technical limits and networked capabilities of particular technologies highly influence the output of computation. This means that, for a wider understanding of algorithmic decision-making, emphasis should be placed on the human-technical ensemble as a whole, which comprises not just humans and algorithms, but also other technologies, such as sensory/perceptual and motor systems, information networks, interfaces, input data, processors, data storage, and so on. As shown throughout this chapter, our common conception and imagination of AI, is bounded by a century-long tradition of conceiving intelligence as an immaterial and a-spatial property residing in the human, albeit one possibly replicable within discrete machines as Turing foresaw decades ago. On the contrary, a relational and material-focused approach is premised upon the intuition that intelligence does not necessarily reside in the individual, either human or machinic, but can be actualised through the embodied actions of multiple agents in systemic cooperation.

Ultimately, on an even wider scale, I believe it's important to reconnect algorithms or technical ensembles to the broader sociotechnical environments in which they are put into use. This means focusing not only on the 'technological stack' enabling their functioning, but also on the broader "legalities, governmentalities, institutions, marketplaces, finance […] including an analysis of the reasons for subjecting the system to the logic of computation in the first place" (Kitchin, 2017, 25). This is a scale of analysis I will take into account, in particular, in chapter 4. Machine learning algorithms, and AI labelled technologies in general, require social, technical, institutional, political, and legal enabling conditions. Their implementation is likely to engender radical sociotechnical transformations whose effects are not limited to the particular moment when they operate, and will be mostly perceived in short and the long run. Again, as I will try to show through my discussion of possible city-scale transformations brought about by self-driving cars, prerequisite for a better understanding of AI-driven "reconfiguration[s] of informational and physical architectures and/or environments" (Blanchette, 2012) is a technically-aware engagement with technology and its operational logics.

Chapter 3

Cognitive Assemblages: The Case of Self-Driving Cars

Siccome gli studi andavano piano per le difficoltà evidenti del trapasso fra il primo capitolo e il secondo del mio trattato, nel tentativo di arrivare a stabilire un codice morale che potesse dare alle macchine una sorte diversa da quella meccanica, mi sforzavo di perfezionarmi nel disegno, anche perché disegnando potevo osservare attentamente ogni cosa e potevo arrivare a capire dei diversi particolari non soltanto la struttura, ma anche il loro mistero, della loro struttura e di quella dell'intero meccanismo.
—Paolo Volponi, *La Macchina Mondiale* (1965)

3.1 Introduction: The Self-Driving Car as a Cognitive Assemblage.

On the basis of theoretical and methodological considerations presented at the end of the previous chapter, this one aims to show that, in view of recent advances in machine intelligence (e.g., machine learning, computer vision, synthetic sensing), human and technical systems are becoming increasingly co-involved in complex "cognitive assemblages" (Hayles, 2017) wherein agency, understood as the capacity to produce knowledge, make judgements and enact actions accordingly, is highly distributed among various technical and human components interacting at multiple spatio-temporalites. To limit my enquiry, I will focus on a particular technology: self-driving cars, namely, vehicles that automate all those functions that in traditional cars are managed by a person. Since they no longer require a human driver, self-driving cars are also often referred to as autonomous cars. Yet, I prefer using the adjective 'self-driving' rather than 'autonomous'. That's because the word 'autonomous' is problematic on two grounds. Firstly, it is at times used with respect to traditional human-driven cars endowed with automated features such as advanced driver-assistance systems that, among other things, may provide extra brake support in case of emergency situations or automatically adjust car speed to maintain a safe distance from vehicles ahead. Yet, such systems are only supposed to assist the driver, who remains in full control of the vehicle and is fully responsible for her decisions behind the wheel. Secondly, the term 'autonomous' subtly suggests that the car itself should be thought of as a technology fully operating beyond human control. As my discussion will attempt to show, the widespread tendency to conceive of self-driving cars as autonomous agents turns out to be extremely problematic when it comes to investigate the ethical dimensions of automated decision-making (see Schuppli, 2014; Verbeek, 2011).

In view of these preliminary considerations, the problem with the term autonomous is that it obfuscates the fact that, in the context of driving automation, decisions and subsequent actions result from complex human-technical interactions and, consequently, can be rarely, if ever, attributed to a single sovereign authorship. For all these reasons, the term self-driving car is here considered to be a more precise descriptor for vehicles capable of automating all the moment-to-moment decisions that, in traditional cars, are reserved for a human driver. This situation, according to a conventional taxonomy developed by the Society of Automotive Engineers, corresponds to level 5 of driving automation, defined as "the full-time performance by an automated driving system of all aspects of the dynamic driving task under all roadway and environmental conditions that can be managed by a human driver" (SAE International, 2014, 2). Alphabet's Waymo vehicles, formerly developed in the framework of Google self-driving car project, are an oft-cited example of Level 5 driving automation.[37]

In engineering jargon, the concept 'autonomy' generally refers to the extent to which the behaviour of technical systems relies less on logical, step-by-step procedures defined a priori by programmers, than on their capacity to perceive and act upon their environment through various sensorimotor and data processing technologies (Russell and Norvig, 2009). The idea that many automated system presently under development exhibit an ever-increasing degree of technical autonomy has sparked debates on the ethical implications of automated decision-making, especially with respect to technologies deployed in situations where nothing less than life-or-death decisions are at stake (e.g., lethal autonomous weapons, self-driving cars). If, as noted by Beer (2016, 3), "it is often this ability to take decisions without (or with little) human intervention that is at the heart of discussions about algorithms potential power", then it is precisely the possibility that algorithmic decisions might cause physical harm to humans that has captured the most public imagination and monopolised media attention. Whilst in automated warfare, for instance, the lethal capacity of weaponized drones, especially when causing civilian losses (Benjamin, 2013), can be seen as the main

[37] The Society of Automotive Engineers (SAE) International has developed a general taxonomy to classify automated on-road motor vehicles, ranging from Level 0 (no automation) to Level 5 (full automation). Tesla semi-autonomous vehicles, for instance, fall into Level 3 (conditional automation), as the human driver is expected to intervene in case of dangerous situations, even if the car works in autopilot (Tesla, 2016).

source of legal controversies and ethical concern (Shuppli, 2014), in everyday spaces such as public roads it is the possibility of fatal accidents involving self-driving cars that has animated heated debates on the ethical, legal and political stakes implicated in self-driving technology.

As a matter of fact, in recent years there has been a proliferation of news articles, academic papers, business and institutional reports focusing on the car accident as the main event worthy of ethical concern (e.g., who or what is to be held legally in case of car crashes; see Ganesh, 2017).[38] Also, large media exposure has been so far reserved for deadly events involving fully or partly automated cars, as was the case of the fatal accident causing the death of the driver of a Tesla 'Model S' in autopilot modality in 2016.[39] In popular culture, in other words, the accident has become the event through which the 'material politics of automation' (Bissell, 2018) surface and become visible in the form of decisions causing immediate effects on people's life.

Most ethical enquiries of machine ethics, however, tend to focus on "robots-as-individuals" (Coeckelbergh, 2011, 243), fallaciously presuming that machines act as individualized entities rather than as complex socio-technical systems. In other words, self-driving cars specifically, and robots in general, are treated as if they were moral agents of their own. Again, there's a strong tendency to anthropomorphise machines, which are imagined as if possessing the same sort of interiority of humans. Clearly, machines are not ethical agents in the traditional sense. Rather, they act as ethical agents in the sense that they substitute humans in situations where, traditionally, moment-to-moment choices are supposed to be made by a human individual on the basis of situated moral evaluations.

For the development of this chapter, I draw on Hayles' (2017, 115) concept of "cognitive assemblage". Compared to other scholars from the posthumanities/new materialism discussed in chapter 1, the reason for using Hayles' (2017) theoretical framework is that it allows to account for the agentive capabilities of contemporary computational media as they interact with human agency and subjectivity, and the wider sociotechnical environment in which such interactions take place. Specifically, Hayles's (2017, 67) cognitivist paradigm locates biological and technical systems on a continuum defined in terms of (nonconscious) cognitive capabilities: "it distinguishes between material forces that can adequately be treated through deterministic methods, forces that are nonlinear and far from equilibrium and hence unpredictable in their evolution, the subset of these that are recursively structured in such a way that life can emerge, and the yet smaller set of processes that lead to and directly support cognition". Here, cognition, which the author (Hayles, 2017, 118) defines as "a process of interpreting information in contexts that connect it with meaning", refers, in other words, to the extent to which a system, whether organic or technological, is capable to process internal/external information and adapt its behaviour according

[38] For example, see: 'The German Ethics Code for automated and connected driving' (Luetge, 2017)

[39] For un up-to-date list of fatal accidents involving self-driving cars, see https://en.wikipedia.org/wiki/List_of_self-driving_car_fatalities (accessed September 7th, 2019).

to changing conditions in the environment. As the author further explains (Hayles, 2017, 118):

> « A cognitive assemblage approach considers these properties from a systemic perspective as an arrangement of systems, subsystems, and individual actors through which information flows, effecting transformations through the interpretive activities of cognizers operating upon the flows. A cognitive assemblage operates at multiple levels and sites, transforming and mutating as conditions and contexts change»

Accordingly, in this study self-driving cars are conceived not as single technology (that is, the vehicle itself as a perceptive, cognitive and decisional unit),[40] but rather as complex human-technical systems comprising various technical subsystems (e.g., sensorimotor technologies, advanced software and algorithmic analytics, humans, mapping systems). Alongside humans, all these interconnected elements are treated, as a whole, like a distributed yet integrated system that through multi-spatial, multi-temporal and multimodal registers of perception and cognition senses and generates knowledge about itself and the surrounding environment, produces decisions accordingly and, ultimately, performs actions in the world.

It should be noted that, years before self-driving cars were even considered to be a concrete possibility, amid geographical scholarship, non-representational theorists (e.g., Dant, 2004; Thrift, 2014) had already proposed a conceptualisation of the car-driver dyad as a "complex hybridisation of the biological body and the machinic body" (Sheller, 2004, 232) in which "intelligence and intentionality are distributed between human and non-human in ways that are increasingly inseparable […] to the point where it becomes something akin to a Latourian delegate" (Thrift, 2004, 49). Building on this body of work, my study aims to advance these intuitions thorough technically-aware investigation of the multiple components taking part in the decision-making process, with particular emphasis placed on the various technological systems and subsystems involved in the process and the multiple spatiotemporal scales at which relevant human-machine information flows take place.

In the existing literature on self-driving cars, attention has already been drawn to issues of social governance and innovation (Marres, 2020; Stilgoe, 2017), transparency and accountability (Ganesh, 2017), the material politics of automation (Bissell, 2018), and the broader social implications of driving automation (Bissell et al., 2020). Notwithstanding, the material functioning and

[40] In a similar fashion, Helen Hester (2018, 78), in her historical retrospective on xenofeminist technologies, discusses the invention, in the early 1970s, of the 'Del-Em', a self-hacking device for menstrual extraction, which she conceives of "not as an isolated device, but as one key node in a network of interconnecting elements, including activist communities, healthcare infrastructure, developments in legislation, and transnational practices of care". In my case, and at this stage of my analysis, the scale of interest is less concerned with framing self-driving vehicles as embedded in infrastructural, socio-cultural and legal systems (as I'll do in Chapter 4), than showing how the car itself constitutes a distributed yet integrated system of perception, cognition and decision-making.

technical complexity of self-driving cars have so far remained black boxed. In the attempt to build a bridge between computer science and the humanities and social sciences, and grounding theoretical speculations on evidence developed through a technically-informed approach, my analysis aims to show that self-driving vehicles *de facto* constitute a 'posthumam assemblage' in which agency is highly distributed among various human and nonhuman agents. In this regard, Hayles (2017, 118-119), whom I'll quote at length, further explains that:

> «Because humans and technical systems in a cognitive assemblage are interconnected, the cognitive decisions of each affect the others […] Moreover, human decisions and interpretations interact with the technical systems, sometimes decisively affecting the contexts in which they operate. As a whole, a cognitive assemblage performs the functions identified with cognition in general: flexibly responding to new situations, incorporating this knowledge into adaptive strategies, and evolving through experience to create new strategies and kinds of responses».[41]

The author further points out that (Hayles, 2017, 29-30, italic added):

> «*Flexibility* implies the ability of an organism or technical system to act in ways responsive to changing conditions in its environment. […] *Adaptability* denotes developing capacities in response to environmental conditions. […] *Evolvability* is the possibility to change the programming, genetic or technical, that determines the repertoire of responses» (italic added).

The three distinctive features of cognitive systems—flexibility, adaptability, and Evolvability—are all present in automated driving systems. Self-driving cars are clearly flexible and adaptable systems, in that they operate within stochastic/partially known environments, dynamically adapting their behaviour to events and situations which cannot be foreseen by their designers, while also negotiating their actions with other traffic participants. Evolvability is present to extent that machine learning algorithms/driving software can be updated at any time by car manufacturers. The idea that a self-driving car operates as a human-technical cognitive assemblage implies that in case of errors, bugs or failures causing injury, loss, or damage, it becomes very difficult to establish a direct causal relationship between agent and effect and individuate a single locus of culpability and responsibility. This is because the sum behaviour of the vehicle is the end result of multiple, yet interconnected, perceptive-decisional-behavioural processes implemented by numerous agents, both human and nonhuman, processing and exchanging information at various sites, speeds and scales.

I will start my study of self-driving cars by providing a schematic overview on the various technical systems and material components that, together, comprise

According to Hayles, flexibility, adaptability and evolvabily are features possessed by all biological and certain technical cognitive systems, as distinct from mere material processes.

a vehicle. To conduct this type of analysis a certain degree of generalization is necessary, inasmuch as each car manufacturer develops and adopts specific technical standards and solutions (e.g.: sensorimotor technologies, communication protocols, localization and mapping systems, data processing techniques). Although technical standards, in most cases protected by copyright, vary from manufacturer to manufacturer, it is possible to account for the functioning of self-driving cars using a generalized schematic model that takes into account four main technical sub-systems: sensor systems, localization/mapping systems, decision-making systems, and actuator systems.

The technical 'expertise' for conducting this kind of analysis has been acquired in different ways (see Kitchen, 2017, on how to investigate algorithmic agency). First, by meticulously reviewing a substantial body of engineering/computer science papers on self-driving cars. Much useful, this activity has been nonetheless very time-consuming, and it can sometimes be difficult to grasp the 'big picture' of how self-driving cars work. That's because engineering academic papers on the subject tend to focus on marginal improvements of specific software programs or hardware technologies, with little if any explanation of the broader functioning of self-driving cars. During the early stage of my research, this activity has been complemented by reading books written by engineers or AI experts having as target the broader public. In this case, on the contrary, one recurring problem is that many important technical aspects tend to be treated in ways which are not sufficiently exhaustive. Crucial for the development of this chapter has been my participation as a visiting PhD student of the Research Group 'KIM – Critical Artificial Intelligence', University of Arts and Design, Karlsruhe, where I acquired an extended understanding of machine learning algorithms in general which has been useful to better understand how they can be used in context of driving automation. On top of this, I have conducted one interview with an engineer with an expertise in driving automation and computer vision. This activity has been useful for me to 'test' the knowledge I had so far acquired on the subject, and clarify the many doubts I had at the time about the more technical, and difficult to grasp, aspects of self-driving cars.

3.2 Anatomy of a Self-Driving Car

By adopting a systemic perspective, the aim of this section is to provide a general overview of decision-making in the context of automated driving. In this section, I intend to be primarily descriptive, explaining how data collected by on-board sensors are algorithmically processed and ultimately correlated to control actions. A technically-aware engagement with self-driving cars is here considered prerequisite for further elaborating on the social, ethical and legal dimensions of automated decision-making. Driving automation is enabled by the four main systems concurrently at work: (a) perception systems, to capture both internal (i.e., relative to the car itself) and external (i.e., from the environment) data; (b) localization/mapping systems (c) decision-making systems (correlating perceptual input to control functions); and (d) control-related systems, that perform actions such as accelerating, braking, and steering. As schematically put by Holstein et al. (2018):

> «[a] decision making process that has to be implemented in a self-driving car can be summarized as follows. It starts with an awareness of the environment: Detecting obstacles, such as a group of humans, animals or buildings, and also the current context/situation of the car using external systems (GPS, maps, street signs, etc.) or locally available information (speed, direction, etc.). Various sensors have to be used to collect all required information. Gaining detailed information about obstacles would be a necessary step before a decision can be made that maximizes utility and/or minimizes damage. A computer program calculates solutions and chooses the solution with the optimal outcome. The self-driving car executes the calculated action and the process repeats itself».

Each moment-by-moment decision taken by a self-driving car thus results from three concurrent processes: (1) sensor-based perception of the environment; (2) multi-sensor data fusion; (3/4) decision-making and subsequent control activation (e.g., speeding up, slowing down, turning right, turn left, braking).

3.2.1 Perception and Localization

For a self-driving car, perception refers to various sensor systems and data processing techniques used to retrieve information from the environment and extract relevant knowledge informing subsequent decision-making processes. Perception is enabled by a large number of sensory devices, which can be roughly divided into two groups: (1) proprioceptors, used to measure and monitor values internal to the vehicle. Typical proprioceptors are motion sensors such as wheel encoders used for odometry and Inertial Measurement Units (IMU) used for monitoring velocity and position changes; and (2) exteroceptors, which capture data with respect to the environment and other traffic participants. Exteroception consists in three sub-tasks: detection, classification, and tracking. Detection is aimed at determining position and velocity of objects moving in proximity of the vehicle. Classification consists in determining the semantic category to which objects in the environment belong (e.g., pedestrians, animals, trees, traffic signs). Tracking consists in predicting future trajectories of detected objects.

Localization and Mapping

In automated driving, localization refers to various positioning and motion tracking hardware and software technologies used to estimate, at any given time, vehicle position in absolute and relative terms. Absolute location refers to its position on a certain point on earth expressed in terms of latitude/longitude coordinates. Relative location refers to its position as calculated with respect to other locations (e.g., vehicle current position with respect to a certain point of departure). For self-driving cars, vehicle localization is obtained by a combined use of Global Positioning System (GPS) and the Inertial Measurement Unit (IMU), used to determine vehicle's absolute and relative location respectively (Knaup and Homeier, 2010). GPS-based localization is very approximate, reaching in open sky an accuracy of up to 5 meters. Yet, since self-driving cars require an accuracy on the decimetre level, GPS alone is not sufficient for high-precision localization, especially considering that satellite signals can be blocked, distorted and delayed by trees and buildings or in case of indoor/underground driving (Levinson et al., 2007).

For all these reasons, GPS-based positioning is supplemented with the IMU, an electronic device comprised of various gyroscopes and accelerometers. Gyroscopes generate data on rotational parameters (e.g., pitch, roll and heading of the vehicle), while accelerometers measure linear acceleration. IMU and GPS based localizations are complementary. The IMU cannot account for the vehicle absolute location. Yet, by constantly generating data on angular velocity and linear acceleration, it enables the calculation of the vehicle relative position (e.g., how far and in which direction it has moved with respect to a given starting point indicated by the GPS). This means that, should GPS signals not be available for a certain amount of time, it is still possible to infer the vehicle's current position with respect to the last reliably reported latitude and longitude. The IMU is very accurate in its measurements; for instance, it can track linear velocity with a 2 centimetre per second accuracy level (Eliot, 2017b). Yet, it tends to accumulate errors over time that need to be continuously corrected.

The IMU also serves other two important functions. Firstly, it works in conjunction with other sensors used for perceiving objects in the environment (e.g., radars, LiDAR), providing data used to calculate velocity and acceleration towards possible obstacles (e.g., vehicles ahead, pedestrians). Secondly, it is used to detect dangerous situations, like when the car is skidding, spinning or tipping over. In this case, data provided by the IMU are used to calculate the slip angle of the car, namely, the difference between heading (i.e., the direction wheels are pointing) and course (i.e., the direction the vehicle is actually moving). For instance, a situation where wheels point toward the left while the vehicle is in fact moving toward the right is indicative of the fact that the car is skidding and thus corrective measures are to be undertaken by control systems.

As for mapping, self-driving cars rely on highly detailed maps of the road infrastructure built specifically for fully automated driving systems. In these maps, produced either by car makers themselves or other companies specialised in map-making, all static objects present in the environment (e.g.: buildings, street signs and lamps, traffic lanes, hydrants) are classified and located with an accuracy on the centimetre level. Waymo, for instance, deploys a fleet of mapping vehicles that, equipped with LiDAR technology, drive around cities to collect data used to build up centimetre-accurate digital recreations of the road (Waymo Team, 2013). One alternative or auxiliary technology to GPS/IMU based localization and mapping is SLAM, short for 'Simultaneous localization and mapping'. The aim of SLAM is to enable self-driving cars to operate within completely unknown environments that have not been pre-mapped, using only on-board sensors such as LiDAR to build 3D reconstructions of the vehicle surroundings (i.e., road infrastructure and static objects such as traffic signs).

Perception

Perception, namely, the capacity to detect, classify and track non-static, moving objects situated around the car, is enabled by a plethora of sensory devices. Perception of the environment can be divided into three sub-tasks: detection, classification and tracking. Detection consists in determining, for each object in the environment, its position and kinematic behaviour. Detection algorithms are used to estimate location and velocity of each object of interest. Classification is obtained through classification algorithms used to determine the category of interest each detected object belongs to. Tracking consists in determining, for each detected and classified object, its location, velocity and acceleration over time. Put simply, whereas detection algorithms determine the kinetic behaviour of perceived objects at *time = t*, tracking algorithms predict where they will be located at *time = t+1*.

Perception is allowed by three main types of sensory devices: (1) cameras, capturing visual data used mainly for object classification, (2) radars, mainly used for long-distance object detection, and (3) LiDAR, used for medium and short-distance object detection and classification. Cameras are considered passive sensors, as they do not directly emit any source of energy onto the environment, but use vehicle/infrastructural headlamps and natural light as illumination source. Radars and LiDAR are considered active sensors, as they emit radio and laser light waves respectively onto the environment in order to detect objects (Eliot, 2018). Some of the most important features of these sensors are detailed below.

Cameras

Due to large availability and affordability, cameras are largely used in automotive applications. Cameras literally replicate the human visual perception of the road scenario. Self-driving cars are generally equipped with dozens of cameras, each pointing towards a different direction so that it is possible to build a 360-degree view of the vehicle surroundings. Cameras capture high-resolution two-dimensional images processed by machine learning algorithms (e.g., convolutional neural networks) to classify objects (see: Bojarski et al., 2016). Cameras are mainly used for obstacle classification (e.g., pedestrians, cyclists, other vehicles), road curvature estimation, and traffic sign detection and interpretation (Häne et al. 2017). For instance, images captured by front-facing cameras can be processed by colour recognition algorithms for traffic sign classification and interpretation (Fleyeh, 2004; De la Escalera et al., 2003). Analogously to the human eye, the downside of cameras is that the quality of the images they capture can be heavily degraded in case of low-light conditions (e.g., in presence of rain, snow or thick fog). High-resolution cameras generate millions of pixels per frame, with an average of 45 frames per second (Kocic, 2018), thus requiring high computing power for such a large amount of data to be processed in real-time. Cameras have limited use for object detection.

Radar

Like cameras, radar is an affordable and mature technology. Self-driving cars are equipped with two types of radar sensors: long-range radars, operating at a range of 200 metres (or more), and short-range radars, operating at a range of 30 metres. Radars sensors constantly emit radio waves bouncing off obstacles in their path. By comparing received and transmitted signals, it is possible to calculate, with high accuracy and precision, distance and velocity of objects. Distance is detected by measuring the time-of-flight between signal transmission and return. Velocity is measured by observing the frequency shift, due to the Doppler effect, between emitted and reflected signals (Ogawa et al., 2018). Radar sensors are used for adaptive cruise control (i.e., speed regulation), emergency braking, collision warning, lane change assistance and blind spot detection (Hirz and Walzel, 2018). Compared to cameras and LiDAR, radar sensors enable longer range detection, are not affected by weather conditions, and require less computing power for data processing. Yet, due to their low-resolution, radar-based images are not informative enough with respect to colour, texture and shape of objects (Wei et al., 2013). For this reason, object recognition and classification is only enabled by a combined used of radars and cameras (see Bertozzi et al., 2008).

LiDAR

Mounted on top of the vehicle, LiDAR is a light detection ranging scanner that rotates continuously in a 360-degree circle while emitting millions of light pulses per second beyond human visual spectrum. Time-of-flight calculations are used for generating dynamic, three-dimensional maps of the road and its surroundings. LiDAR has limited use for object classification, yet it enables the detection of objects located completely around the car (Hirz and Walzel, 2018). Unlike cameras, LiDAR is not affected by illumination variations. However, its performance can be degraded under certain weather conditions, as light pulses can be refracted by thick fog, snow, or raindrops. By combining information retrieved from LiDAR maps with classification information retrieved from camera images, it is possible to achieve accurate object classification and detection. This is because LiDAR is accurate at detecting object position and trajectory, while cameras provide contextual knowledge (i.e., object classification). LiDAR is the most expensive sensory component of self-driving cars, yet its cost has decreased significantly in the last two years (Muoio, 2017).

3.2.2 Multi-Sensor Data Fusion

Self-driving cars are provided with different types of sensor technologies (e.g., cameras, radar, LiDAR) and there can be multiple units by type. As mentioned in the section above, by combining data from different sources it is possible to overcome sensor-specific disadvantages. For example, radar sensors, usually mounted behind the front grille, are very precise at measuring distance and velocity of vehicles ahead. However, due to their narrow detection angle, they can only detect objects located directly in front of the radar emitter, while cars on adjacent lanes might remain out-of-target. Radar sensors are fully operational regardless of weather and illumination conditions, yet they tend to produce noisy measurements that need to be extensively filtered and cleaned. Also, although very precise at registering kinematic data, radars do not provide detailed object appearance information (e.g., texture and colour). This means that radar sensors only allow rough classification of targets, which can be inferred on the basis of their kinematic behaviour. Nonetheless, the missing appearance information can be retrieved, for example, from camera-based images. In this case, classification algorithms can be used to identify objects in video frames. It should be noticed that, unlike radars, camera-based sensor systems do not directly measure distance and speed, which can only be estimated. Also, compared to radar sensors, cameras have a significantly lower depth resolution, especially in the long-distance range. Although affected by weather phenomena such as rain, fog, and snow, the LiDAR scanner, continuously rotating on its axis, enables omnidirectional object detection. If LiDAR observations are correctly associated with radar measurements, then the combination of the two data streams results in improved determination of location and speed of objects around the car. Ultimately, by fusing kinetic data from LiDAR and radar sensors and classification information from cameras, it is possible to improve the overall reliability of the perception task, whose (fused) output will inform the subsequent planning module.

Applications for multi-sensor data fusion are many (e.g., automated target detection for smart weapons, remote sensing, robotics, Internet of Things). In general, "data fusion systems seek to combine information from multiple sensors and sources to achieve improved inferences than those achieved from a single sensor or source" (Hall, 2002, 419). In automated driving, the ultimate goal of multi-sensor data fusion is to build up a coherent, robust statistical representation of the driving scenario, at times even in case of missing and/or contradicting sensory information. The rationale behind multi-sensor data fusion is that, due to sensor-specific limitations (e.g., limited field of view), no single data source can individually provide the information necessary for a reliable and consistent perception of the environment. Data multimodality and redundancy are thus necessary "to achieve improved accuracies and more specific inferences than could be achieved by the use of a single sensor alone" (Hall and Llinas, 1997, 6).

Summing up, the input of data fusion is data from different sensory sources, which are progressively processed and fused together to arrive a coherent statistical representation of the environment. The output of sensor fusion is a text document named ‘object tracking list’ where, for each detected target it is specified: a) class; b) relative velocity and location within the car’s surroundings; c) statistical prediction of its future trajectories. The objects tracking list is thus used as input for subsequent decision-making processes determining which is the appropriate driving manoeuvre to be performed in the current traffic scenario.

3.2.3 Decision-making and Control

In self driving cars, the technical component responsible for the automation of the moment-by-moment decisions that, in traditional cars, are reserved for a human driver, is the so-called “planning system”, to which, by convention, in the computer science community, decision-making power is ascribed. Actually, the planning task can be roughly divided into two hierarchically structured sub-tasks: “global path planning” and “local path planning” (Yurtsever et al, 2019).[42] Global path planning algorithms are responsible for setting high-level goals; they solve an optimization problem by calculating the most cost-effective path from a certain point of departure to a certain destination chosen by the vehicle users.

Local path planning algorithms, instead, calculate local, moment-to-moment decisions (e.g., changing lanes, turning right, braking, making a U-turn), which are always contingent on (unpredictable) factors such as the behaviour of other traffic participants, road and weather conditions (e.g., control systems integrated with optical rain sensors limiting vehicle’s speed on wet roads), and signals from the traffic infrastructure (e.g., a red traffic light at an intersection indicating that the vehicle has to stop at an intersection; see: Pader et al, 2016). Using as input the object tracking list outputted by the perception system, local path planning algorithms produce, as output, decisions informing control algorithms. The latter, in turn, send commands to electro-mechanical actuators which materially perform a driving manoeuvre. In the existing literature, local path planning algorithms are also referred to as decision-makers (Pendleton et al., 2017), as they indeed govern vehicle’s motion by choosing, from the different available alternatives, the appropriate driving behaviour in the current traffic scenario.

Typically, local planning is comprised of two consequential stages: ‘behavioural decision making’ and ‘motion planning’ (Paden et al., 2016). In a given road scenario, behavioural decision making consists in defining a set of feasible manoeuvres, from which one single appropriate driving behaviour is selected. For any traffic situation, there can be indeed multiple driving options (e.g., overtaking a stopped vehicle or waiting for it to continue driving). However, not all of them are feasible. To be feasible, a driving manoeuvre should simultaneously fulfil three criteria. Firstly, each local trajectory should follow the route on the road network defined in the global planning stage, incrementally leading the vehicle towards the chosen destination. Secondly, driving manoeuvres must always comply with traffic conventions and rules. Thirdly, and crucially,

[42] In engineering literature, ‘global route planning algorithms’ are sometimes referred to as ‘mission planners’; ‘local path planning algorithms’ are also called ‘behavioural algorithms’ or ‘decision-makers’ (Pendleton et al., 2017)

they must be collision-free, avoiding putting at risk the life of people both inside and outside the vehicle. Possible manoeuvres that don't simultaneously meet all such requirements are excluded from the decision-making process.

Among all feasible alternatives, behavioural algorithms select one single manoeuvre, usually on the basis of a multi-criteria optimization problem. In general, the overall goal of a driving decision is to avoid collision with other traffic participants while making incremental progress along the route defined by global planning algorithms. In behavioural decision-making, the overall goal so defined can be operationally re-framed as a set of—at times conflicting—sub-goals (e.g., stay within road boundaries, keep safety distance, do not collide, minimize waiting time; see Furda and Vlacic, 2011). Behavioural algorithms thus calculate a driving decision by solving a utility maximization problem in which, for each feasible manoeuvre, the levels of achievement of different sub-goals are measured and compared. The selected driving behaviour is the one for which the value of the overall utility function is maximized.

Early attempts to automate behavioural decision-making were based on traditional, logic-based programming (Buehler et al., 2009), using deductive reasoning to develop complex rules governing vehicle's motion under different traffic situations. This is because, in principle, traffic rules, road scenarios, and driving behaviours can be modelled as finite sets, with transitions governed by if-then rules. Yet, rule-based approaches require perfect knowledge of the environment, which is never the case for complex and unpredictable environments as public roads are. In real-world situations, self-driving cars operate within environments that are never known a priori. The desired driving behaviour calculated by local planning algorithms is then translated, by motion planning algorithms, into a specific path or trajectory, which must always be aerodynamically feasible and comfortable for passengers (Buehler et al., 2009). In other words, motion planning algorithms compute a short-distance, local navigational path that the vehicle has to follow to move from its current location to a certain goal area. Actually, there can be multiple safe/comfortable trajectories to accomplish the driving manoeuvre defined by the behavioural layer of the decision-making hierarchy. Usually, motion planners calculate the optimal trade-off between cost-effectiveness (e.g., with respect to fuel consumption) and passengers' comfort, and generate the desired trajectory accordingly.

Once a local trajectory is generated, control algorithms send specific commands to the appropriate actuators, which are responsible for the material execution of the driving decision. Apart from activating mechanical actions, control algorithms also serve another important function, constantly providing feedback on how well the driving decision is being performed and, if necessary, re-adapting mechanical dynamics to properly execute the trajectory. Once a decision-making loop is completed, a new one begins based on new information provided by the perception sub-system. Each decisional loop takes place in the order of milliseconds.

3.3 Discussion: Beyond the Individual

3.3.1 Self-Driving Cars and the Trolley Problem

Presently, much of the public debate on the ethics of algorithmic decision-making in automated driving revolve around the so-called ‘Trolley Problem’, originally introduced, in its modern form, by Philippa Foot (1978) in the essay *The Problem of Abortion and the Doctrine of Double Effect*. In moral philosophy, the Trolley Problem is a classic thought experiment based on the resolution of an ethical dilemma between two alternative courses of action. Brought up as one example to discuss the permissibility of abortion, in Foot’s (1978) essay the Trolley Problem is formulated as a hypothetical scenario in which a trolley driver must choose between, on the one, turning a trolley so that it runs over an innocent person inescapably attached to a track and, on the other, allowing the trolley to proceed on its course and kill five innocent people. With the launch of the MIT Moral Machine Project, in the last few years the Trolley Problem has gained renewed popularity, establishing itself as the dominant paradigm for framing discussions of the ethical and legal implications of automated driving.

As stated on its official website, the MIT Moral Machine Project is “[a] platform for gathering a human perspective on moral decisions made by machine intelligence, such as self-driving cars”.[43] Started in 2016, the project attempts to collect data from people from all over the world, with the aim to define universal ethical principles and policy guidelines to be used for the development of self-driving cars. Based on thirteen hypothetical fatal scenarios, the project takes form as an on-line survey in which volunteer respondents are asked to suggest what a self-driving car should do in case of life-or-death situations, such as choosing between killing four pedestrians to spare four passengers, and vice versa. The results of the survey have been summarised in a paper publish in 2018 (Awad et al., 2018).

Since its launch, the project has been central to discussions of machine ethics in automated driving, attracting both attention and scepticism. Scepticism, in particular, has arisen from the fact that the experiment is based on idealised circumstances which are unlikely to happen in real life, especially if one considers that, in case of emergencies, human decisions are in most cases taken in a few seconds or fractions of a second, and thus are rarely, if ever, the result of conscious reasoning and moral judgments. Building on such criticisms, this section, by adopting a cognitive assemblage approach, aims to shed light on other problematic yet scarcely discussed assumptions underpinning the Trolley Problem as deployed in the context of the Moral Machine experiment.

In self-driving cars, decision-making is generally understood as that specific moment when, in ordinary traffic conditions, a driving manoeuvre must be algorithmically calculated on the basis of sensory information—like changing

[43] See http://moralmachine.mit.edu/

line, turning right/left, accelerating, and so on. Based on the Trolley Problem, the Moral Machine project focuses on a very limited set of driving decisions. A moral dilemma involving trade-offs between two deadly outcomes is typically presented as follows (Holstein et al., 2018):

> «A self-driving vehicle drives on a street with a high speed. In front of the vehicle a group of people suddenly blocks the street. The vehicle is too fast to stop before it reaches the group. If the vehicle does not react immediately, the whole group will be killed. The car could however evade the group by entering the pedestrian way and consequently killing a previously not involved pedestrian. The following alternations of the problem exist: (A) Replacing the pedestrian with a concrete wall, which in consequence will kill the passenger of the self-driving car; (B) Varying the personas of people in the group, the single pedestrian or the passenger. The use of personas allows including an emotional perspective, e.g., stating that the single pedestrian is a child, a relative, a very old or a very sick human, or a brutal dictator, who killed thousands of people».

By focusing on the moment of the accident as the exclusive event worthy of ethical concern, in the framework of the Moral Machine decision-making is thus formulated as the quest for the most ethically acceptable choice (is it more morally acceptable to spare passengers or pedestrians? The many or the few? The young or the elderly? The healthy or unhealthy, and so on). The experiment is framed in such a way that only input (a certain road scenario) and two possible outcomes are known, meaning that the inner workings of self-driving cars remain largely unexplained and black-boxed. As the proponents of the project themselves (Awad et al., 2018, 63) admit:

> «[We] could not do justice to all of the complexity of autonomous vehicle dilemmas. For example, we did not introduce uncertainty about the fates of the characters, and we did not introduce any uncertainty about the classification of these characters. In our scenarios, characters were recognized as adults, children, and so on with 100% certainty, and life-and-death outcomes were predicted with 100% certainty. These assumptions are technologically unrealistic, but they were necessary to keep the project tractable».

However, as this chapter attempts to show, in order to understand decision-making in self-driving, it is now more than ever urgent to unpack the discursive and technical black-boxes enveloping their functioning. This, to my view, is a research method necessary if one wants to arrive at a better understanding of how political and ethical responsibilities are distributed among contemporary AI systems, as well as to understand the extent to which cultural bias, global chains of labour, resource, and data extraction (Crawford and Joler, 2018) are embedded in their functioning. Otherwise, too many crucial questions risk be left unanswered, not even posed. Which algorithms within a self-driving car can be said to be responsible for decision-making? Or, alternatively, where is decision-making power temporally and spatially located within (and outside) a self-driving car? In which sense can self-driving cars be said to act and decide autonomously? How can human choices influence the outcome of machinic decisions? And, provocatively, how can machines be programmed according to universal ethical principles if, as the results of the Moral Machine paradoxically show (Awad. et al., 2018), moral standards are culture and place-dependent? By virtue of these considerations, the Moral Machine project deserves criticism not just because it's grounded on hypothetical scenarios detached from reality, but also, and especially, because it further contributes to envelop the systemic complexity inherent to their functioning.

To my view, the main limit of the Trolley Problem is that it is based on a flat, unidimensional understanding of decision-making, which results from two faulty assumptions. First, in a temporal sense, decision-making is conceived as a single, atomized event through which perceptual inputs, in a given traffic scenario, are correlated to behavioural outputs. Second, in a spatial sense, decision-making power is located exclusively within the self-driving car itself, which, as Ganesh (2017, 7) puts it, "is imagined to be a sort of neoliberal, individualised agent […] that can act independently and efficiently on the basis of guidelines and feedback". In other words, self-driving cars are treated as discrete, individuated machines, rather than as technical systems characterised by a high degree of systemic complexity, with a vast array of hardware and software technologies exchanging and processing data at various sites and scales and concurrently co-participating in the calculation of a driving decision.

The Moral Machine experiment, like most contemporary enquires of machine ethics, is individual-focused and anthropocentric: it applies to machines approaches and ethical standards previously developed to investigate human agency. Originally, the Trolley Problem was introduced to investigate why human individuals sometimes tend solve ethical dilemmas on the basis of utilitarian logic (i.e., by maximising the greatest benefit for the largest number of people), sometimes by strictly adhering to deontological imperatives, which would impose, for instance, that killing is never morally permissible, even if by sacrificing one person it would possible to spare more lives (Fischer and Ravizza 1992, Thomson, 1985; Foot 1978).

However, both utilitarianism and deontology seem inappropriate for discussing ethics in self-driving cars. This is because both doctrines are aimed at investigating the psychological drivers and immediate outcomes of choices mapping back to a single human individual, who is expected to reason and act on the basis of her own will and moral judgments. Or, as assumed in the Moral Machine experiment, the immediate outcome of driving decisions can be directly linked back to self-driving cars themselves, thought of as individuated agents capable of internalizing human ethics by means of computational programming. In sum, traditional ethical enquiries assume that decision-making is a prerogative of the individual, whether it be human or a (humanized) machine.

However, it is argued here, driving automation is exemplificative and expressive of a posthuman condition where decision-making, as well as the production of the knowledge informing it, is becoming increasingly multi-layered and distributed among various human and technical actors forming together complex cognitive assemblages. Consequently, the actions and decisions that a cognitive assemblage performs as a whole can only be evaluated in a systemic and relational fashion. The acknowledgment that the immediate outcome of automated decision-making is always the by-product of multi-site and multi-temporal interactions between human and technical actors could perhaps contribute to add new dimensions of ethical concern.

By adopting a cognitive assemblage approach, in the next section I argue that, in automated driving, decision-making is in fact multi-dimensional, both spatially and temporally. In other words, I will attempt to show how the sum-behaviour of a self-driving car is the end result of multi-level human-machine interactions concurrently affecting the final outcome of decision-making, be it a decision taken either in normal traffic conditions (e.g., turning right or left) or an emergency situation (e.g., killing one passengers of four elderly). First, I will discuss why, within a self-driving cars, decision-making can be rarely if ever attributed to a single 'technical' authorship. Secondly, I will provide concrete examples of how human choices made 'outside' the vehicle and preceding their introduction onto public roads do heavily affect the final outcome of decisions allegedly taken the car itself.

3.3.2 Looking Inside a Self-Driving Car

In mainstream discussion of self-driving cars, decision-making is generally attributed to what is often defined, in anthropomorphic terms, the 'brain' or 'AI' of the car, subtly suggesting that there exists one single component responsible for decision making. In engineering terms, the 'brain' or 'AI' of the car is conventionally understood to be the local planning subsystem, as local planning algorithms are in fact responsible for selecting the appropriate driving manoeuvre in a given traffic scenario on the basis of information outputted by the perception subsystem. However, decision-making power only partly resides in the planning subsystem.

When I was in Karlsruhe, I conducted a personal interview with a machine learning scientist with an expertise in driving automation specifically, formerly employed at 'Understand.AI',[44] a Karlsruhe-based start-up collecting, formatting and editing training dataset to be used by car manufactures for developing computer vision systems. During our meeting, I explicitly asked him where exactly are driving decisions taken within a self-driving, receiving the following answer:

[44] https://understand.ai/

> «So, to answer your question "where exactly are decisions taken", it is the Planning algorithm. But is it only as good as the information it receives from Perception layer».

Leaving momentarily aside issues of cultural bias embedded in algorithmic computation, the answer received confirms that, within a self-driving car, the final decisional output can be seen as the end result of a sprawling network of 'choices' made along all sub-stages of driving automation, and producing a cascade of effects ultimately affecting the car sum-behaviour. In other words, within self-driving cars, there is not one single, master algorithm ultimately responsible for decision-making. Rather, choices performed during all stages of automation (perception, planning and control) concurrently determine the final outcome of a driving decision. Actually, in one sense, it's true that local planning algorithms are the ones responsible for selecting the appropriate driving behaviour in a given traffic scenario.

Yet, choices made by local planning algorithms are always dependant on inferences developed during the perception stage, and perception accuracy is in turn reliant on the quality of data captured by on-board sensors (e.g., severe weather conditions causing sensor obstructions). It is worth underlining that the word 'choice', in the theoretical framework proposed by Hayles (2017, 35), "has a very different meaning than in ethical theories, where it is associated with free will". Rather, choice consists in "programmatic decisions among alternatives courses of action" (Hayles, 2017, 25). To perform a choice, a technical device must be cognitive, that is, responsive to input variability in adaptive ways. As whole, a self-driving car is in itself cognitive, in that it is capable of adapting its behaviour to numerous, unpredictable traffic situations. Yet, it should be noted, its sum behaviour is always a function of multi-level choices performed by a large number of technical components. For instance, machine learning algorithms used in self-driving cars for object classification perform 'choices' by virtue of their capability to recognise to which particular semantic category, among a given taxonomy, objects in the environment belong to (e.g., pedestrians, cyclists, other vehicles). Typically, classification algorithms are trained on visual data where targets of interest are pre-known and pre-labelled (e.g., images containing pre-labelled pedestrians). In supervised learning, during the training stage, classification algorithms develop a general function that will be later used to classify objects in new images (e.g., pedestrians in images from on-board cameras). In self-driving cars, the 'choices' that classification algorithms make, in turn, generate a cascade of effects, concretely affecting the final outcome of decision-making. It is well known that, for instance, the fatal crash involving a Tesla 'Model S' in 2016 was due to errors at perception level, as camera-based classification algorithms failed to recognize the white broadside of a truck against the bright sky (Yadron and Tynan, 2016, July 1).

As explained earlier, class and kinetic information gained through various sensor-specific data processing techniques are thus integrated, through multi-sensor data fusion, with the aim to mitigate ambiguities in data acquisition and processing and arrive at a coherent, yet never 100% accurate, statistical description and prediction of the behaviour of the car itself and other traffic participants. The output of sensor fusion (i.e., the object list used by local-planning algorithms as input for calculating a specific driving manoeuvre feeding forward to mechanical actuators) is always influenced by sensor-specific ‘choices’.

Self-driving cars can be thus considered complex decision-making systems that, through multi-level algorithm-based data processing techniques, correlate inputs from the environment to concrete actions performed in the world. In the perception stage, raw and pre-processed measurements from numerous on-board sensors are progressively fused with increasing level of abstraction, with the aim to reduce uncertainties in data acquisition and generate an up-to-date statistical estimate about the vehicle itself and its surrounding environment. Inferences produced by the perception subsystem, together with prior knowledge of the road infrastructure, rules of the road and driving conventions, are thus used to inform decision-making processes governing, in real-time, vehicle’s motion. A self-driving car can thus be conceived of as a multi-layered architecture, wherein perception, cognition and, ultimately, decision-making are distributed, systemically integrated and hierarchically structured. Low-level choices related to various sensor-based subsystems are progressively integrated with higher-level reasoning systems used for drawing inferences and, ultimately, making decisions performed by mechanical actuators.

3.3.3 Looking Outside a Self-Driving Car

In *How a Machine Learns and Fails: A Grammar of Error for Self-Driving Cars*, Pasquinelli (2019) provides a systemic definition of machine learning as a “technical assemblage” which comprises not just algorithms, but also training data. This means that machine learning-based decision-making systems cannot be fully understood if algorithms are considered as entities divorced from the dataset upon which they have been trained. Among the various technological advances that have so far contributed to render self-driving cars a concrete possibility (e.g., global satellite systems, increased computational power, advanced sensors for computer vision), a crucial factor is the availability of publicly available training datasets containing millions of pre-labelled images used for machine learning research. For instance, ImageNet provides millions of hand-annotated pictures with more than 20.000 pre-labelled semantic categories (Markoff, 2012). It should be noticed that hand-annotation is a labour-intensive and time-consuming activity, requiring hours of meticulous screen work. Recently, it has become popular to crowdsource this activity via on-line platforms such as Amazon’s Mechanical Turk. Apart from obvious ethical issues regarding the exploitation of low-paid digital labour, the problem with outsourced hand-annotation is that it requires significant editing and post-processing of obtained labels (Janai, 2017).[45] The availability of public datasets with hand-annotated categories has played a crucial role in the development of machine learning algorithms for automated driving applications such as object detection and classification. However, generic training datasets such as ImageNet have little use to train reliable machine perception systems, as they are not enough representative of objects portrayed in situations of interest for self-driving cars (e.g., people portrayed in proximity of roads). In the past, a partial solution to this problem has been offered by the use of synthetic data based on 3D simulations of traffic scenarios. Yet, real-world datasets are necessary to guarantee that detection, classification and tracking algorithms are effective when deployed in real-world situations characterised by environmental complexity, variability and unpredictability (e.g., pedestrians, cyclists and vehicles portrayed from different angles, under different weather conditions and in different traffic situations). Based on data collected from 2012 to 2017, the PASCAL VOC was the first publicly available training dataset to be specifically used for automated driving applications, providing a large number of images with twenty pre-labelled categories such as persons, birds, cats, aeroplanes, bikes, and motorbikes. Yet, the PASCAL VOC is based on generic pictures retrieved from sources such as Flickr. Thus, most of the times targets are portrayed in situations that a self-driving car is unlikely to encounter on public roads (pre-labelled images of people meeting at a birthday party are way less useful than images of people crossing a road in urban settings to train algorithms used in automated driving).

To date, the most commonly used training dataset for machine learning research in automated driving is the KITTI Vision Benchmark,[46] developed by Karlsruhe Institute of Technology (KIT) in partnership with Toyota Technological Institute (Chicago). It is based on six hours of recordings obtained through a Toyota sponsored self-driving car equipped with various sensors such as high-resolution colour and grayscale stereo cameras, a Velodyne 3D laser scanner and high-precision GPS/IMU inertial navigation system (Geiger et al., 2013). The KITTI dataset is useful for various machine learning based applications such as Simultaneous Localization and Mapping (SLAM), 3D object classification, detection and tracking. As noted earlier, the preparation of training datasets is a labour-intensive activity taking place ‘outside’ the car and ‘before’ its actual deployment on public roads. Yet, as widely acknowledged amid the data science community, the quality of training datasets (e.g., accuracy in crowdsourced hand-

[45] Driving automation is often debated in terms of its potential of putting an end to certain categories of jobs (e.g., bus and taxi drivers) or, in more optimistic terms, for its promise to free humans from routine work. However, the acknowledgement that the possible disappearance, in Western societies, of certain types of works is strictly dependant on the creation, and exploitation, of planetary-scale chains of invisible labour, may help debunk the idea of a society freed from work by means of intelligent machines. Or, put differently, a Western society freed up from work can only be possible thanks to the exploitation of labour force in non-Western societies.

[46] See: http://www.cvlibs.net/datasets/kitti/index.php

annotation, representative sample of classes of interests, variety of scenarios, and so on) is perhaps the most important factor for developing machine learning algorithms capable of being effective when put into use in the wild.

In the social sciences, issues regarding how algorithm-based decision-making processes (e.g., predictive policing) can end up reproducing, if not amplifying, race and gender bias embedded in training datasets have already been extensively discussed (Caliskan et al. 2017; Tufekci 2015; Boyd and Crawford, 2012). Yet, in driving automation, little if any attention has been so far dedicated to the political and ethical dimensions of training dataset themselves, with ethical inquiries mainly focusing just one the vehicle itself and, in particular, on the moment of the accident as the product of allegedly autonomous machine decisions. However, the adoption of cognitive-assemblage approach may help shed light on how human choices regarding how training datasets are shaped (in terms, for instance, of data taxonomy) and where data is collected do have ethical implications, heavily influencing the final outcome of decision-making within self-driving cars, at times producing unintended discriminatory consequences. As observed by my interviewee:

> «if perception has a negative bias towards black people, the Planning algorithm doesn't get the information that there is a black man in front of the car and then takes the decision to continue driving».

What does it mean that a classification algorithm, for instance, is biased towards black people? The KITTI dataset is now considered to be an international and widely used standard for machine learning applications in automated driving. In a sense, the KITTI Vision Benchmark project has contributed to increase the overall level of safety of self-driving cars. By providing data from multiple sensor-sources with accurately labelled classes, the project has in fact enabled the development of more robust computer vision systems. At the same time, however, the major limitation of this dataset is that all video sequences have been recorded in a single street section in the mid-city of Karlsruhe and surrounding rural areas and highways. So, first, it lacks diversity in terms of weather, settings, and light variability (Janai, 2017). Second, and most importantly, precisely because data have been captured by driving exclusively in and around the city of Karlsruhe, the training dataset lacks diversity in terms of ethnic groups represented. In other words, the training dataset mainly contains pictures of white, central-Europe people, while black and minority ethnic people are under-presented. Issues of facial recognition systems failing to recognize black/non-white people as human have already been largely discussed (see White, 2019). In driving automation, if perception systems fail recognizing black people as human, the impact of a similar misdetection may have concrete consequences on people's lives, perhaps causing fatal accidents. Should a perception system fail detecting and classifying a black person, who's to be held responsible, ethically and legally, for such a discriminatory, and potentially fatal, occurrence? Perhaps the perception algorithms failing to properly detecting black people, or local-planning planning algorithms in fact responsible for calculating driving decisions? Perhaps car manufactures, or companies providing training dataset used to train and validate machine learning systems? Or, at an even larger scale, governments licensing the testing and potential deployment of self-driving cars on public roads? In this discussion, pre-empting an accident using a cognitive assemblage approach is less important to investigate the ethics of self-driving cars as framed in the Moral Machine experiment (is it more ethically acceptable to sacrifice passengers or pedestrians?), but rather to shed light on how, in automated driving, decision-making is the end-result of human/machine choices performed at various scales and in different spatialities and temporalities, and yet converging at the precise moment when a self-driving car has to implement a decision in the here and now. Whereas the Moral Machine experiment attempts to replace the human driver as the sovereign subject of decision-making with a machine substitute which is imagined to act exactly as a human being, this chapters rather has showed that, in self-driving cars, neither humans nor machine can be said to operate in fully autonomous realms, as decisions are always the by-product of their interactions.

Chapter 4

Post-Anthropocentric Cities: Machine Perception & Urban Complexity

4.1. Autono-mobility

At the time of its first appearance on American public roads in the early 1900s, the automobile was considered to be a luxury item reserved for the exclusive use and delight of a small number of wealthy enthusiasts. Since about the 1960s, however, the automobile has become ubiquitous (Dant, 2004), establishing itself not just as an undisputed symbol of modernity (Lefebvre, 1971), but also as one of the main catalysts for urban transformation, one capable to reshape urban form and sociality in its own image. It is well known that, for the automobile to become a mass medium of transportation (Dant and Martin, 2001), over the past century many cultural, regulatory, economic and infrastructural changes had to be made to integrate automobility within citiscapes (Featherstone et al. 2005). That is to say, cities had to be physically and socially reorganised around cars (Norton, 2008).[47]

For decades now, the car has been considered "a common feature of everyday life itself, almost a background to the background" (Thrift, 2004, 45-46). In the last few years, however, automobility has re-emerged as a much-debated topic amid discussions about the present and future of urban life. Such renewed interest is largely due to the possible commercialization, in the near or long run, of self-driving cars. No longer seen as futuristic objects only existing in the realm of science fiction, recently self-driving cars have started becoming a reality. Nowadays, positions about their possible introduction onto urban roads are split. Whilst some consider it inevitable (Claudel and Ratti, 2015), others regard it as fantasy deemed to remain so, at least for a few decades (see: Janai et al., 2017). Either way, there is total consensus about the great potential of self-driving

[47] This chapter is an expanded and revised version of an article published in the Italian journal *A&RT - Atti e Rassegna Tecnica*.

technology to reshape, as the motorcar did throughout the twentieth century, several aspects of urban life. How this will happen, however, is far from clear. Nowadays, the way in which such transformations will unfold is subject to much debate and speculation among urban planners and engineers, architects, jurists, ethicists, and city governments.

As argued by their main advocates, including not surprisingly automakers themselves and software providers, self-driving cars have a potential for huge social gains in terms of safety, accessibility, efficiency and sustainability. Increased road safety, in particular, provides the main discursive rationale upon which their political legitimation rests. According to oft-reported estimates, around 94 per cent of all fatal crashes are due to human error. By taking human input out of the equation in the driving process, self-driving cars thus promise to drastically reduce road fatalities. Additionally, since they no longer require a human driver, fully automated driving systems could secure access to point-to-point mobility to segments of the population so far excluded, such as the elderly or visually impaired. Combined with emerging trends in car sharing, self-driving technology also promise to alleviate traffic congestion, with an estimated 80 per cent decline in number of privately owned vehicles (Claudel and Ratti, 2015). As a result, vast areas of urban land currently serving as parking lots could be destined to new social uses (Ratti and Biderman, 2017).

Apparently, there’s much to gain from letting self-driving cars enter urban centres. Notwithstanding the potential benefits listed above, many questions about the short and medium term implications of self-driving technology remain largely unanswered. Suffice to think of the social and economic consequences stemming from the possible marginalization or displacement of many jobs that involve driving (e.g., track drivers; see: Maughan, 2019). Furthermore, the citywide implementation of automated driving systems will require significant infrastructure investments, presumably at the expense of public transportation and other crucial policy domains, such as healthcare and education (see: Blyth et al., 2016). A particularly debated issue is liability in case of accidents. Currently, it’s not clear yet who is to be held legally responsible for injuries or property damages caused by the vehicle (whether the owner, the manufacturer, or code developers). Also, there is a concrete risk that, due to their entire reliance on software technology, self-driving cars could become an easy target for hacking attacks (Maughan, 2019).

About one century ago, the automobile (or ‘horseless carriage’, as it was also referred to at the time) promised to modernise a system of personal mobility heavily reliant on horse-drawn vehicles, which car advocates considered to be ill-suited to modern society’s needs (see: Norton, 2008). Animal-pulled vehicles were indeed deemed slow, inefficient, and even polluting due to the abundant presence of horse manure on city streets (Cohen, 2010). However, the broader transformations, both intended and unintended (e.g., urban sprawl, air pollution), brought about by the automobile remained largely unforeseen by city governments and urban planners, in fact fully understood only decades later (see: Jacobs, 1961, in particular chapter 18). Nowadays, proponents of self-driving cars, mostly big private players such as Waymo (formerly the Google self-driving car project), Tesla and Uber, are shaping a technologically-mediated vision for the future which appears positive and inevitable at once. They tend to see self-driving cars merely as a design challenge (and obviously as a big market opportunity), publicly justified on the basis of expected yet mostly unverifiable social benefits not just for safety, but also accessibility, efficiency, congestion, and sustainability—thus putting an end to many negative externalities associated with traditional cars. Such utopian vision, however, rests upon a false premise, namely, that the large-scale implementation of automated driving systems will happen in politically frictionless ways. Self-driving cars, indeed, are marketed as a technological innovation that will significantly ameliorate city life yet leaving the social order substantially unchanged. However, that's simply not the case. As clearly stated by Stilgoe (2017, 5), "[t]his plug-and-play story, in which the car is seen as able to get along with the world's complexities as they are, without making additional demands, is a lie". As self-driving cars are in fact incompatible with most existing physical infrastructure and human behaviours, substantial changes will be required for them to become fully operative within urban settings. In particular, it is argued here, cities will need to be made more machine-readable. Specifically, they will need to be reconfigured in such a way so as to conform to the pre-emptive logics of machine perception systems used by self-driving cars to map their surroundings and predict future events happening therein.

Although dominant public discourses of self-driving cars are mainly concerned with ethical and legal dimensions regarding their high degree of operational/decisional autonomy, more recently urban scholars have started to tackle their specifically urban implications (see: Duarte and Ratti, 2018). Apart from a few exceptions (Bissell et al., 2020; Bissell, 2018; Stilgoe, 2017), however, most urban research has tended to adopt a strictly technological deterministic approach, framing the relationship between self-driving cars and cities as one of linear causation, with an overemphasis on the impact of the former over the latter, and rarely the other way around. From a methodological perspective, the problem with linearity, as Fuerth (2009, 20) puts it, is that:

> «[it] distorts our notion of cause and effect. Under its influence, we tend to expect that for every problem there is a unique solution; and that proportionate changes of circumstances will produce proportionate changes of outputs. We believe that it is possible to disassemble ("unpack") compound, conglomerate issues, without destroying their coherence».

To date, little if any attention has been paid to the specific ways in which urban social space influences the development, and brings to light the intrinsic limitations, of self-driving technology. The aim of this chapter is to remedy this absence. Hence, by suggesting a perspective reversal in comparison to dominant technological determinism, this chapter attempts to provide an answer to the following question: how does the essentially hybrid nature of cities affect the development of self-driving cars?

Again, it should be noticed that, from a theoretical and methodological standpoint, a self-driving car can be thought of either as a technical assemblage or a socio-technical assemblage. As a technical assemblage, it can be seen as a complex arrangement of various hardware and software technologies working synchronously. As a socio-technical assemblage, it can be instead understood as embedded "within larger interlocking systems, rather than […] as [a] discrete entit[y]" (Bissell et al., 2020, 10). In other words, a self-driving car can be conceived of as an integral part of what I define here—paraphrasing what Urry (2005) wrote more than a decade ago about traditional cars—a 'system of autono-mobility' which includes not just vehicles, but also physical and digital infrastructures, machine learning algorithms, training datasets, geolocation systems, three-dimensional cartographies, laws and codes of the road, roboethics, mobility cultures, and new social practices and ways of dwelling.

In the further course of this chapter, I will try to show that, if one wants to understand—and anticipate—the broader impacts of self-driving cars on urban life, and avoid providing a reductionist explanation that ascribes causal effect exclusively to the former, the technical dimension cannot be divorced from the social one, and vice versa. This chapter is articulated into two complementary parts. In order to provide an empirical answer to the question introduced above, the first part attempts to unpack the 'black box' of self-driving cars (Latour, 1999), with the aim to 'visualize' their inner workings as put into relation with urban socio-spatial complexities. More specifically, with descriptive intent, I will focus on some urban socio-spatial specificities that, by making machine perception a particularly difficult task, currently hinder self-driving cars from being introduced onto city streets. The second part is to be understood as a brief exercise in *anticipatory governance* (Fuerth, 2009; see: Stilgoe, 2017). In this section, I will investigate possible trajectories of urban transformation aimed at accommodating self-driving technology. With speculative intent, it is argued that cities themselves might be spatially and socially reconfigured so as to guarantee a safer coexistence between self-driving cars and other traffic users—humans *in primis*—with whom they will compete for urban space.

4.2. Artificial Sensorium

In the last few years, debates on the future of transport automation have been almost entirely monopolised by self-driving cars. As a matter of fact, self-driving technology has "captured the popular imagination arguably more so than any other transportation technology over the past half century" (Bissell et al., 2020). In light of all this, one crucial aspect that has so far remained largely overlooked within the public debate is that, in fact, automated transport systems are already fully operational in many non-urban contexts. Suffice to think, for instance, of automatic trains moving people between and within airport terminals or driverless vehicles used in industrial applications for goods transportation. This kind of vehicles, actually, embody a degree of technological sophistication much less advanced than that self-driving cars do. Yet, unlike self-driving cars, they operate within standardized, controlled and relatively predictable environments, wherein the range of (unexpected) situations the vehicle has to handle is very limited. An automatic train, for example, operates within a closed system. It runs along a predetermined, fixed path, carrying people to a small number of prescheduled destinations. Its functions, as well as the environment in which it operates, are in large part predictable and therefore pre-programmable. The same line of reasoning applies to driverless vehicles deployed in industrial settings, which automate repetitive tasks within highly structured, monofunctional environments, with limited if any interaction with other vehicles and/or people.

Conversely, self-driving cars are supposed to operate within extremely dynamic, non-deterministic and information-rich environments, which is exactly what city streets are. On a busy road, a self-driving car must respond in a timely fashion, in the order of milliseconds, to a wide range of situations which can never be entirely anticipated by its designer. Compared to extra-urban contexts, city streets, in view of their density, morphological heterogeneity and essentially hybrid nature—in terms not just of architectural layout and morphological configuration, but also variety of road users—multiply uncertainty factors. In dense city traffic, a car must interact with a myriad of other road users (e.g.: pedestrians, cyclists, other vehicles, animals), each acting independently from one another in unpredictable ways. At any given moment, a self-driving car must perform several functions simultaneously, of which only a small part follow pre-programmed rules, while others, based on machine learning, must be adaptive to many environment variables. Among other things, a self-driving car, for instance, must always adapt its speed to that of vehicles ahead, observe traffic rules, interpret ambiguous situations such as hand-signals from construction road crews, and be ready to react quickly to emergency situations (e.g., if a pedestrian suddenly leaves the curb and walks or runs into its path).

Self-driving cars are quintessentially spatial actors, which exist and move through space. Their main task, in fact, is to safely transport people from a certain point of departure to a chosen destination. In so doing, they navigate a world which is shared with many other entities: humans (pedestrians, cyclists), animals, objects (road signs, street furniture), and whatever elements the weather brings on their path. In order to safely transport people—and avoid collisions with other traffic participants (both human and non-humans), they must first and foremost be capable to perceive their surroundings in any weather and lighting conditions. A self-driving car, in a sense, can be seen as a concrete instance of what French philosopher and urbanist Paul Virilio (1994, 59), writing in the Eighties, prophetically defined ‘sightless vision’, in which “the capacity to analyse the ambient environment and automatically interpret the meaning of events” is delegated to the dyad computer–camera. Technically speaking, what with anthropomorphic vocabulary is commonly called ‘machine perception’, more properly refers to complex statistical techniques used to estimate class and localization of objects in proximity to the vehicle, and predict their behaviour in the near future. To put it schematically, a self-driving functions according to a sequential perception-decision-action logic. Machine perception is based on four main types of sensors (cameras, radars, ultrasound sensors and LiDAR scanners). As explained at length in Chapeter 3, multi-sensor data are collected, fused and processed in real-time to create a three-dimensional spatio-temporal representation of the vehicle itself and its surroundings. On the basis of such information, driving software outputs the proper manoeuvre to be performed by the vehicle in the current traffic situation.

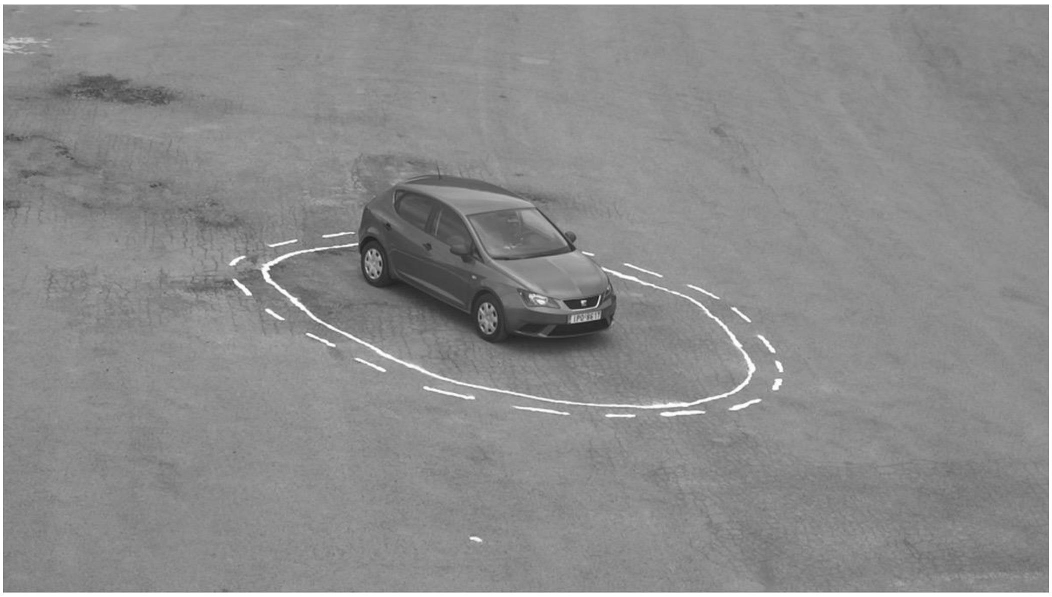
Fig. 5. Where the City Can't See, Liam Young and Tim Maughan, 2016

For a self-driving car, exteroception, that is, the perception of its ambient environment, is never 100% accurate. This means that at any given moment a self-driving car must always ‘decide’ on the basis of imperfect and incomplete information. In the context of automated driving, there are two main dimensions of uncertainty: internal and external. Uncertainty in machine perception is partly due to the intrinsic limitations of sensory/perceptual systems and partly to external variables. Internal uncertainty is connected to possible errors in data acquisition (e.g.: obstruction or malfunction of one or more sensors) and/or processing (e.g.: incorrect determination of class, position and future trajectories of other traffic participants). For self-driving cars, however, the main source of uncertainty originates in the external world. As they operate within partially observable, stochastic environments, many random/unpredictable variables must be accounted for. In particular, external uncertainty stems from the vehicle problematic interaction with the road environment, and in particular existing traffic signing systems, and other traffic participants.

Up until today, urban roads have been built around people and, with the rapid rise of automotive traffic, around motorists primarily. Traffic signing systems (e.g., traffic lights, road signs, and painted pavement) are a crucial element of the road environment. Indeed, they dictate behaviours and facilitate coordination among vehicles, pedestrians, cyclists, and whoever travels the streets. Traffic lights, for instance, “impose a strong social control over the most fundamental of human behaviours, whether to move or be still” (McShane, 1999, abstract). To date, international traffic flow is regulated by the Vienna Convention on Road Traffic (1968), which most countries agreed upon. With regard to traffic signing systems specifically, the annexed Convention on Signs and Signals defined common design principles still in use today (e.g., dimension, shape, colour, and localization of road signs). Uniformity, indeed, is a crucial factor in avoiding confusion and minimizing uncertainty.

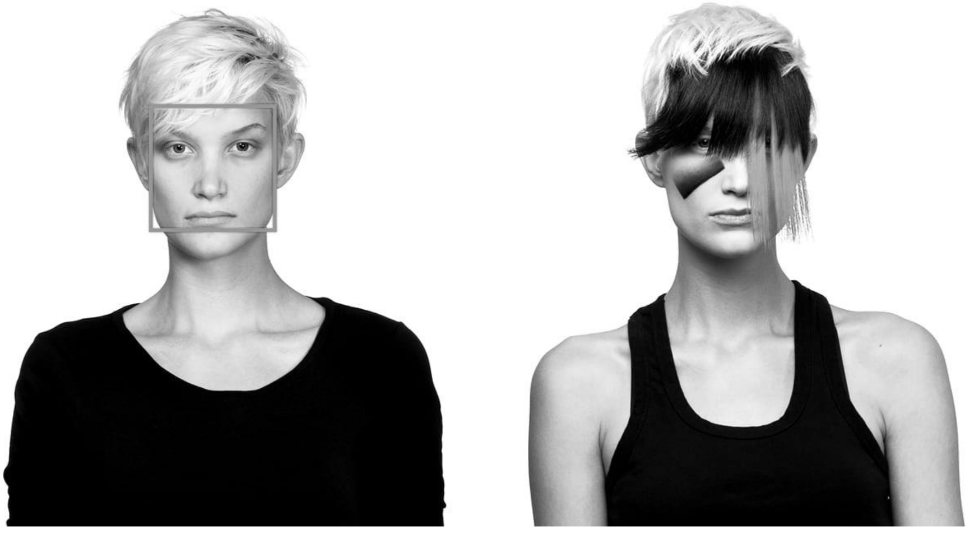
Existing perceptual stimuli that regulate traffic flows have been so far defined having human rather than machine perception in mind. Road signs, for instance, are designed and placed within the road environment in such a way so as to be quickly and unequivocally interpreted by humans, and especially by car drivers as they speed by. It’s no coincidence that, for instance, conventional road markings are made with retroreflective white materials. In this way, indeed, it is possible to maximise visibility both during daytime (by producing the highest contrast possible against typically dark-coloured road pavements), and night-time (as retroreflective pigments bounce light from vehicle headlights back). Designed to accommodate human perception, roads signs can be hard to read for self-driving cars. Traffic signs detection and classification is in fact a very difficult task, one which requires significant expenditure of computational resources, especially in dense city traffic. Beside, traffic sign vandalism can become a serious concern for self-driving car manufacturer. Things that would normally not affect human perception, such as faded road markings and small graffiti applied to road signs, can become, or be intentionally used as, dangerous deceptive devices. Stickers or graffiti, indeed, can completely alter the meaning of a road sign to the machine reading it, with potentially catastrophic consequences for people both inside and outside the car (Field, 2017).

Fig. 6. James Bridle, Untitled (Autonomous Trap 001), 2017



In dense urban contexts, where sidewalks and bike paths commonly intersect and/or are adjacent to roadways, the coexistence between self-driving cars and humans is particularly dangerous, as possible misdetections or malfunctions may cause harm to people. This is due to two main reasons. First of all, pedestrians' and cyclists' movement is typically characterised by constant, rapid changes in direction and posture (Janai et al., 2017). Hence, their behavioural patterns are only approximately describable in statistical terms. That is to say, their future intentions are difficult to predict. Additionally, people are characterized by a high degree of appearance variability, both in terms of physiognomic/somatic features, and clothing (Janai et al., 2017). Algorithms used for pedestrian and cyclist detection and classification are typically trained through supervised learning techniques. Indeed, in order to "recognize" anything, algorithms need first to be taught what to recognize according to preexisting categories and probabilities. During the training stage, machine learning algorithms 'learn', on the basis of a vast number of images containing objects classified as people, the distinctive visual features of human beings. In this way, they develop a function which will be later used to discern cyclists and pedestrians in new images captured in real time by on-board sensors. However, the high degree of human variability, in terms both of behaviour/pose and appearance significantly reduces the capacity of classification algorithms to generalize beyond what they have learned during the training stage, that is, to successfully interpret images which they have not 'seen' before.

Fig. 7. Adam Harvey, CV Dazzle, 2010



Several technical solutions have been proposed to reduce both internal and external uncertainty associated with self-driving cars. This has generally meant increasing the number of on-board sensors to ensure redundancy and reduce uncertainty in data acquisition and processing, yet at the expense of greater computational costs and longer execution times.

4.3. Post-Anthropocentric Cities

State-of-the-art self-driving technology is not ready to handle urban complexity. Unreliability in dense city traffic can be considered to be the main reason why self-driving cars haven't been introduced citywide yet—except for trials conducted by car manufacturers in partnership with local governments in designated experimental areas.[48] In the spirit of mutual benefit, public actors have so far played a very marginal role in the governance of self-driving vehicles, limited to authorize, and sometimes incentivize, tests on public roads conducted by car manufacturers in partnership with software developers. On the wake of what some have termed 'testbed urbanism' (see: Halpern et al., 2013), such 'wait and see' (Grieman, 2019) approach toward social innovation is expressive not just of a blind confidence in the potential benefits associated with automated mobility, but also of a certain hope that they might be achieved without any significant public effort, at different governance levels, both in political and economic terms. At the same time, for car manufacturers to be granted access to public roads is critical to achieve competitive advantage, as collecting data in real traffic conditions is an essential element for improving self-driving technology.

In light of various accidents that have so far involved prototypes of self-driving cars, the latter have been under the spotlight due to their dangerous intrusion onto public roads, while car manufacturers have been often criticized for their social irresponsibility. More recently, however, there has been a discursive reversal in the public debate, and several criticisms have been made, mostly by car manufacturers themselves, towards the lack of public efforts in facilitating the integration of self-driving technology within cities (see: Stilgoe, 2017). Public governments, in other words, have been blamed for delaying the realization of all potential benefits associated with self-driving cars. As a result, public scrutiny has subtly shifted from questioning the capability of self-driving vehicles to safely navigate city streets, to problematizing urban social and material infrastructure as incongruous with their operations. The basic idea is that—although always teachable, improvable, and perfectible thanks to machine learning—self-driving technology has reached its maturity. According to Claudel and Ratti (2015), for instance: "[f]rom a technological point of view, driverless cars have arrived; the bigger task is for cities to integrate them".

By taking the social desirability of self-driving cars at face value, such rhetorical strategy demands that public actors take social and political responsibility in facilitating the transition towards automated mobility. Such shift in responsibility has two immediate effects. On the one hand, it redefines the relative weight of stakeholders involved in the governance of self-driving cars. On the other, it paves the way for new ways through which driving automation could be materially achieved in the near and long term. Up until today, in fact, most efforts to improve self-driving technology have been made from the inside of the

[48] In 2016, Uber granted a permit to test its self-driving vehicles in Arizona. Tests were later suspended following a fatal accident causing the death of a pedestrian Tempe, in March 2018 (Howard, 2018).

vehicle, that is, by using increasingly sophisticated (and expensive) machine perception systems and algorithms. Yet, it is becoming clearer and clearer that vehicles themselves are not the sole important element in automated driving systems. For self-driving cars to become fully operative, a possible, in fact more likely, alternative solution could be intervening on the external world in such a way so as to reduce the socio-spatial complexity of their operative milieu. More precisely: this would mean transforming urban spaces in such a way so as to make them more easily perceptible and intelligible by self-driving cars. Experts have suggested various solutions for facilitating the integration of self-driving cars within urban traffic. Some, for instance, argue for the necessity to create dedicated self-driving car lanes, so that interaction with other road users can be limited or avoided (Oliver et al., 2018). Others insist on more radical changes aimed at converting existing streets into 'smart roads' equipped with vehicle-to-vehicle and vehicle-to-infrastructure systems. In this way, it would be possible to enable wireless exchange of key information among vehicles themselves (e.g., relative speed and position) and between vehicles and the road infrastructure (e.g., speed limits and turning restrictions; see Duvall et al., 2019).

At the beginning of the twentieth century, public streets were cohabited by various and equally important actors, including pedestrians, street vendors, children allowed to play freely on the road, cyclists and so on. However, the arrival of the motor car produced a specifically urban phenomenon (Norton, 2008, 11): "a new kind of mass death. Most of the dead were city people. Most of the car's urban victims were pedestrians, and most of the pedestrian victims were children and youths". As cars (and motorists) were initially seen as unruly intruders, city streets "had to be socially reconstructed as places where motorists unquestionably belonged" (Norton, 2008, 1). Substantial interventions had to be made both in infrastructural terms (dividing public roads into areas reserved for vehicle transit and others, more marginal, for pedestrian and cycle traffic), and in regulatory terms (to ensure that such functional regimentation of the street uses would be in fact respected). As shown by Urry (2004, 26), since about the 1960s, for the automobile to become "the predominant global form of 'quasi-private' *mobility* that subordinates other mobilities of walking, cycling, travelling by rail and so on", it had to be sustained by and coevolve with a complex system of automobility, namely:

> «an extraordinarily powerful *complex* constituted through technical and social interlinkages with other industries, car parts and accessories; petrol refining and distribution; road-building and maintenance; hotels, roadside service areas and motels; car sales and repair workshops; suburban house building; retailing and leisure complexes; advertising and marketing; urban design and planning; and various oil-rich nations».

Similarly, it can be argued—and it's in fact becoming more plausible—that self-driving cars will demand a complex system of 'autono-mobility' of their own, one which comprises physical and digital infrastructures, hardware and software technologies, digital traffic control systems, more reliable machine learning algorithms, training dataset (whose carving, formatting, and editing, as showed in chapter 3, is a labour-intensive and globally dispersed activity), machine-readable maps (whose maintenance can be as laborious and costly as that of physical roads), shifting liability and insurance systems, and new social practices—as people will need to learn how to behave both inside and around self-driving cars (Casner et al., 2016). Autono-mobility will enable new ways of dwelling, while other existing social behaviours will need to be changed or prohibited in order to enable a safe co-existence between humans and vehicles.[49]

Through an in-depth study of machine learning algorithms used for computer vision, the adoption of a perspective focused on the wider socio-technical system rather than on the car itself as its primary unit of analysis has proved useful to counteract the widespread tendency to frame the relationship between self-driving cars and cities as one of linear causation, with social and spatial changes merely resulting from technological innovations. Rather, it is argued here, cities will need to be transformed, in the first place, to create enabling conditions for driving automation to happen. Again, I want to stress the importance of entering the debate by departing from an in-depth analysis of the specific logic and material properties of the technology analysed, for only by unpacking the inner-workings of machine learning and computer vision it has been possible to speculate about possible socio-spatial transformations driven by self-driving cars. To date, in fact, it is not yet clear if, when, and how self-driving cars will be able to navigate city streets. What is out of question, however, is that their introduction onto public roods will require significant infrastructural and social changes. Specifically, it is argued here, the form, materiality and sociality of cities will need to conform to the pre-emptive logics of machine vision and cognition. Given the radical social and spatial transformations that self-driving technology is likely to bring about, their introduction onto urban road will open up the way to a post-anthropocentric urbanism, for cities will be substantially reconfigured less to accommodate people than these emerging nonhuman spatial actors.

There's yet another important, conclusive consideration that needs to be added to the present discussion. On the one hand, it's true that, should full automated driving establish itself as the most desirable form of future urban mobility, then cities will necessarily need to be reconfigured in such a way so as to enable self-driving cars to achieve greater agency and technical autonomy. Yet, it should be acknowledged, this type of reasoning entails a good dose of paradoxical thinking. In fact, full driving automation will probably require urban environments to be redesigned is such a way so as to render the vehicle's moment-to-moment decisions more predictable and, in this sense, less autonomous. Presumably, full driving automation will demand the creation of urban environments capable to lessen the load of intelligence actually demanded from the car itself, and to redistribute it among other vehicles, technologies and

[49] It can be speculated, for instance, that people will be required to wear standardized clothes, so as to reduce appearance variability, and facilitate their detection and classification by computer vision systems.

the environment itself. Tackled from such a systemic perspective, attention thus shifts from the intelligence of the car itself, to the cognitive capabilities present, with varying degrees of complexity, in many other important elements. Albeit surely contestable, there's a strong political assumption underlying full driving automation, one which rests upon the belief that self-driving cars will drastically decrease road accidents. It's beyond my capabilities to evaluate the technical and political desirability of self-driving cars. But, presumably, increased road safety and better traffic coordination will be achieved less by increasing the intelligence and autonomy of each single car, than by increasing the relational, reciprocal connectedness between cars, and between cars and the road infrastructure, other traffic participants, and other technologies, all together brining into existence a cognitive environment whose overall cognitive capabilities exceed that of the individual elements comprising it. Notwithstanding, the way in which we conceptualise AI today, however, is still strongly biased by our anthropocentric tendency to think of intelligence as an abstract, a-spatial, and immaterial property residing within individuated entities, whether human or machinic, serving merely as its 'hardware' support. This fallacy, I believe, can ultimately lead us to misidentify other modalities and scales at which intelligence can occur beyond the conventional site of the individual. Hence, the real post-anthropocentric lesson to be learned here resides not in the recognition that cities will be restructured around machines rather than people (thus re-inscribing a sort of human/machine dualistic opposition). To my view, what could potentially mark the transition towards a post-anthropocentric approach to city-making is the recognition that intelligence can occur in forms and scales exceeding the individual. This acknowledgement, in turn, would mean to definitely discard the Cartesian model of intelligence which attributes the latter to the human subject only and, by extension, anthropomorphised machines.

Conclusions

This doctorate project started with a preliminary survey of, and motivated by a genuine curiosity toward, a body of posthumanist/new materialist scholarship encompassing a variety of academics who, in similar ways and some differences notwithstanding, all have contributed to deconstructing the human as an ontological category and political project. Recalling the discussion presented in chapter 1, posthumanist/new materialist theories pursue a twofold objective (Rose, 2017). In conjunction with critical feminism and post-colonialist studies, they seek, on the one hand, to further deconstruct the modernist notion of the human as a free, autonomous and rational subject (what we call an individual)—an ideal that, despite its purported universality, only applies to a small portion of humanity: the male, white, heterosexual sovereign subject (Braidotti, 2013). On the other, great imaginative effort is put into rethinking the agential power, as well as acknowledging the political importance, of nonhumans, including technological others. By challenging dominant notions of the human alongside the qualities which have long been identified with human exceptionalism—including intelligence, rationality and autonomy, amid posthumanist and new materialist scholarship ontological redefinitions are thus tied to yet unrealized possibilities for positive social, political, and environmental change. One common concern is indeed the political and ethical centrality which has so far been reserved for the human in Western philosophical and political thinking, being it considered to be the main cause behind the discrimination of, and destructiveness toward, dehumanized and naturalised others.

Drawing on and framed within the posthumanities, this project has concerned itself with investigating shifting conceptions of the human subject as a function of its rapidly evolving relationship with a particular category of nonhumans: contemporary AI systems. As noted in chapter 2, my use of the term always presupposes a good degree of generalisation. As common usage suggests, AI is conceptually and technically heterogeneous: it can refer to a wide range of technologies or technical systems, in themselves resulting from the combined use of discrete technologies of various sorts. Precisely because of such terminological confusion, and in view of the accelerating rate at which they are being deployed within and across almost all sectors of society, AI demands urgent response, inventive conceptualisation and technically-grounded investigation. In deconstructing the human subject, the theoretical arguments which have been so far advanced within posthumanist scholarship are surely to be praised, especially if one considers that, therein, critiques of discriminatory practices taking place inside the same category of the human are generally coupled with attempts at stressing the ethical value and political relevance of nonhuman others.

In this regard, both for practical and political reasons, I do acknowledge that extending agency to nonhumans has its uses when it comes to counter the destructive effects human activity, of which global warming is perhaps the most pressing one among the many problems originating from it. At the same time, however, I believe that to do so for AI demands much caution, for it may sometimes result into analyses that, grounded on merely ideological premises, almost inevitably lead to politically ambiguous, if not self-contradictory, results. No matter how good their intentions, in granting agency to nonhumans, within the posthumanities most theoretical positions indeed tend to be ideological, atemporeal (i.e., not historically contextualised), or purely philosophical. One key paradox here is that, in challenging modernist dualistic thinking, and criticizing the exploitative and destructive practices it supports, another important dualism is introduced: on one side are humans (a category in itself internally differentiated along many anthropological axes), on the other side are nonhumans, a category virtually encompassing everything from a mineral to an animal to an algorithm or a complex robotic system. Wide in scope and ideological in orientation, such approaches in my view fail to adequately express how agency occurs in contemporary computational media. In fact, scant, if any, attention is dedicated to the differentiated modalities, material specificities, and wordly practices and events through which nonhuman agency in general, and technical agency specifically, takes shape within contemporary social and spatial systems.

Talking of nonhuman subjectivity (e.g., Braidotti, 2011; Latour, 2005) certainly holds great potential for imagining a truly relational, post-anthropocentric political ecology. Intuitively, if extending agency or subjectivity to nonhumans might seem fascinating and even politically and environmentally desirable, then doing the same thing with regard to AI systems automating complex cognitive and decisional processes is at least tempting, almost automatic. However, as discussed at length in chapter 2, anthropomorphization is almost inevitable when agency and subjectivity are projected onto technologies designed with the precise purpose of automating activities which have long been, and still largely are, considered to be the epitome of humanity—namely, intelligence, autonomy, and decision-making. This is the exact reason why AI entertains a peculiar, inherently ambiguous relationship with the autonomous, liberal subject of traditional humanism: a vision of the human which, I have argued, is sometimes decentred and sometimes reinforced in the process of imagining or designing intelligent machines.

Having started this PhD project with theoretically limited and certainly non-technical knowledge of the topic, during the formative months of this research, de-biasing my own thinking about AI has required a significant amount of work. Influenced, perhaps even misled, by ideas about nonhuman agency discussed amid critical theory, and myself fascinated by notions of robotic/algorithmic subjectivity (Matzner, 2019; Bratton, 2016), at the beginning of this project, the more I was getting acquainted with the topic, the more I could not but acknowledge my own tendency to anthropomorphise AI.

Considered an essential point of departure for gradually arriving at a better-informed appreciation of its wider socio-spatial impacts, deconstructing received notions of AI has required both technical and conceptual unpacking. Indeed, complementary to an in-depth, technically-aware analysis of the various technologies investigated—machine learning algorithms or the wider (socio)technical systems they take part in, has been a critical engagement with concepts whose meaning, amid geographers and social scientists, is all too often accepted at face value. Based on a methodology combining computer science and social science approaches, my analysis of self-driving cars as a cognitive assemblage [chapter 3] has responded exactly to the need to reground posthumanist approaches to nonhuman agency by reintroducing as key dimension the techno-materiality of algorithmic agency alongside the embedded actions and decisions it engenders in systemic cooperation with other technologies and humans "distributed and displaced through the system" (Ganesh, 2020, 3). Prior and functional to this has been my in-depth discussion of AI presented in chapter 2, which has been useful both for countering widespread claims and popular representations of AI on the one hand, and, on the other, bring conceptual clarity about contending views on technological autonomy, in particular by unveiling the operational logics of machine learning.

Once used to denote qualities of the human sovereign subject exclusively, notions such as intelligence, autonomy and decision-making are equivocal precisely because we have only recently become accustomed to use them with respect to mundane machines. The issue at stake, in this debate, is not just terminological. Misleading and imprecise notwithstanding, the widespread use of anthropomorphic vocabulary is not coincidental: it speaks of the increased agential powers and technical capabilities embodied by contemporary computational media. Beyond anthropomorphic thinking, the fact that many AI systems are presently regarded as autonomous is nonetheless indicative both of the high degree of unpredictability and inscrutability inherent to their operations, and their capacity to automate tasks once thought to reside exclusively in the domain of the human, including decision-making within morally ambiguous situations.

The point is that, central to the very definition of humanness as well as to the political and moral constitution of the modern subject (see Balibar, 2017), when juxtaposed to machines, notions of autonomy, agency, and choice rather than acquiring new meanings tend to preserve their old ones. In my view, precisely from this stems the widespread tendency toward anthropomorphic individuation (the tendency to think of machines as discrete entities rather than complex sociotechnical systems) and personification of AI (the tendency of interpreting machinic operations in terms of human standards), so common nowadays both in fictional and nonfictional spaces. Confusing if uncritically applied to machines, a terminology so heavily reliant on notions of autonomy, rationality and agency is however indicative of how influential is still today the conceptual power of the liberal, autonomous self in defining both the human and AI by comparison. Beside the popular imagination of AI as discrete, humanoid machines possessing or aspiring to possess the same prerogatives traditionally associated with liberal individualism, including the rights to personal freedom and equal treatment, the liberal view of the human permeates and structures a substantial part of contemporary debates about AI and its broader cultural, political and ethical implications. Sometimes the autonomous, rational subject enters the picture as the model of the human which is perceived to be most endangered by, and thus ought to be staunchly defended against, technologies acting in increasingly autonomous ways (see Rouvroy, 2013, on algorithmic governance). Sometimes the characteristics of liberal individualism are stretched to the extent of encompassing technological artefacts. As exemplified by the Trolley Problem articulated within the MIT Moral Machine experiment [chapter 3], in fact persistent is the tendency to use for AI systems ethical principles and analytical frameworks that actually apply for human individuals only. Indeed, standard to most approaches to computational ethics is a conception of AI/algorithms as individuated agents, rather than as complex technical systems integrated into even wider human-technical environments.

In the concluding remarks to *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics,* Kathrine Hayles (1999, 288) observes that:

> «the posthuman does not really mean the end of humanity. It signals instead the end of a certain conception of the human, a conception that may have applied, at best, to that fraction of humanity who had the wealth, power, and leisure to conceptualize themselves as autonomous beings exercising their will through individual agency and choice. What is lethal is not the posthuman as such but the grafting of the posthuman onto a liberal humanist view of the self».

In order to clarify her point, the author (Hayles, 1999, 288) brings as an example Moravec's (1988) prediction that one day humans may be capable to transcend bodily mortality by downloading their consciousness into computers:

> «When Moravec imagines "you" choosing to download yourself into a computer, thereby obtaining through technological mastery the ultimate privilege of immortality, he is not abandoning the autonomous liberal subject but is expanding its prerogatives into the realm of the posthuman».

In a similar fashion, and paraphrasing the above passage from Hayles (1999), I argue that when proponents of the Moral Machine experiment imagine a situation in which a self-driving car must choose between two alternative yet unavoidably lethal courses of action, they are not abandoning the autonomous liberal subject. Quite the opposite, they are expanding its prerogatives into the realm of AI. In the framework of the Trolley Problem, the vehicle is indeed understood as if acting completely independently from the human, at the same times as it's imagined to possess anthropomorphic morality, and perhaps better sensing and decisional capabilities. In the attempt to countering widespread claims of technological autonomy and further decentre the autonomous liberal subject, my in-depth, technically-grounded discussion of self-driving cars [chapter 3] has actually shown exactly the opposite—namely, that rather than substituting the human sovereign subject with a machinic delegate, driving automation is paradigmatic of a 'posthuman' condition in which the complex imbrication and dynamic entwinedness between human culture and technics lean towards unprecedented levels. In a cautious, pessimistically optimistic fashion, I believe that such realisation, which the outcome of my case study supports, might potentially contribute to a definitive, or more realistically partial, abandonment of the model of the human articulated within traditional liberal thinking and let something else emerge: a vision of the human grounded on the acknowledgement that humans and machines contribute in different yet equally important ways to cognitive and decision-making processes. In this regard, if Latour's proposition that "we have never been modern" has proverbially unveiled the artificiality of the human/nonhuman distinction, similarly, one could argue that *we have always been posthuman*, meaning that, far from being one of mere instrumentality, human's relationship with technology has always been one of dynamic and productive co-evolution—building on Simondon's work on technological concretization (1958), this is a position shared by many philosophers of technology, including Stiegler (1998) and Hayles (2012).

What's important to remark is that, however, self-driving cars, and many other contemporary AI systems alike, are not just another technology. Given their potential to profoundly disrupt and reshape our societies and environments [chapter 4], they demand analyses which should not be abstracted from, and thus always be contextualised within, the specificities of their socio-technical and political realms. By opening up the black box of a particular AI technology, self-driving cars, and casting light on the inherently socio-technical character of automation, my analysis has attempted, on the one, to further decentre the human individuated subject as the sole locus of sovereign decision-making and ethical concern. At the same time, however, it has also further unveiled that structural to driving automation are dehumanising and exploitative practices which have long been identified with the humanist project. This becomes starkly evident if one considers the colonial legacies (un)visibly at work at the world's margins through the exploitation of minimally recognized forms of labour (e.g., low-paid screen workers labelling and annotating images used for developing computer vision applications); the extraction of non-renewable Earth's natural, mineral and other high-value resources in the Global South necessary for the development, maintenance, and everyday functions of planetary-scale digital infrastructures (Bratton, 2016); the continuation and amplification of discriminatory patterns embedded into the very logics of machine learning (e.g., algorithmic biases toward dark-skinned and other underrepresented social groups in training datasets); the constitution and normalisation of new regimes of human classification and profiling (see Crawford and Paglen, 2019); and the possible further regimentation and segmentation of urban roads and public spaces [chapter 4]. Rapidly emerging, all these issues demand further empirical investigation through approaches attentive to both the technical and the social dimension of automation. If it's true that there could be much to be gained from abandoning the liberal view of the self and embrace the posthuman instead in order to assemble more sustainable and inclusive socio-technical futures, at the same time, the ethical, social, environmental, and political costs for that to happen should not be neglected.

References

Adams, D. (1984). *So Long, and Thanks for All the Fish*. London: Pan Books.

AI Now Institute (2016). *AI Now 2019 Report*. Retrieved from https://ainowinstitute.org/AI_Now_2016_Report.pdf

Alldred, P. and Fox, N. J. (2019). Assembling Citizenship: Sexualities Education, Micropolitics and the Becoming-Citizen. *Sociology*, 53(4): 689-706

Alpaydin, E. (2016). *Machine Learning: The New AI.* Cambridge and London: MIT Press.

Amin, A. and Thrift, N. (2017). *Seeing like a City*. Cambridge: Polity Press

Amin, A. and Thrift, N. (2002). *Cities: Reimagining the Urban*. Cambridge: Polity Press

Amoore, L. (2020). *Ethics: Algorithms and the Attributes of Ourselves and Others*. Durham and London: Duke University Press.

Amoore, L. (2013). *The Politics of Possibility: Risk and Security Beyond Probability*. Durham: Duke University Press.

Amoore, L. and Raley, R. (2017). Securing with algorithms: Knowledge, decision, sovereignty. *Security Dialogue*, 48(1): 3-10.

Anders, C. J. (2015, April 21). From Metropolis To Ex-Machina: Why Are So Many Robots Female?. *Gizmodo*. Retrieved from io9.gizmodo.com/from-maria- to-ava-why-are-so-many-artificial-intellige-1699274487.

Armand, L. (2014). Slaves of Reason: Perversion Among the Robots. *Lola Journal*, 5. Retrieved from http://www.lolajournal.com/5/slaves.html

Armao, F. (2013). Smart resilience. Alla ricerca di un nuovo modello di sicurezza urbana. In M. Santangelo, S. Aru and A. Pollio (Eds.), *Smart city. Ibridazioni, innovazioni e inerzie nelle città contemporanee* (pp. 169-181). Roma: Carocci.

Ash, J. (2017). *Phase Media: Space, Time and the Politics of Smart Objects*. New York: Bloomsbury.

Asimov, I. (1954). *The Caves of Steel*. New York: Doubleday.

Asimov, I. (1950). *I, Robot*. London: D. Dobson.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J-F., and Rahwan, I. (2018). The Moral Machine experiment. *Nature*, 563 (7729): 59-64.

Balibar, E. (2017). *Citizen Subject: Foundations for Philosophical Anthropology*. New York: Fordham University Press.

Barns, I. (1994). The Human Genome Project and the Self. *Soundings: An Interdisciplinary Journal*, 77(1/2): 99-128.

Batty, M. (2014). *The New Science of Cities*. Cambridge: MIT Press.

Batty, M., Axhausen, K.W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G., and Portugali, Y. (2012). Smart cities of the future. *The European Physical Journal Special Topics*, 214(1): 481-518.

Baudrillard, J. (1994). *Simulacra and Simulation*. Ann Arbor: University of Michigan Press.

Beer, D. (2016). The Social Power of Algorithms. *Information, Communication & Society*, 20(1): 1-13.

Berrar, D., Konagaya, A., and Schuster, A. (2013). Turing test considered mostly harmless. *New Generation Computing*, 31(4): 241–263.

Bertozzi, M., Bombini, L., Cerri, P., Medici, P., Antonello, P. C. and Miglietta, M. (2008). Obstacle detection and classification fusing radar and vision. *IEEE Intelligent Vehicles Symposium*: 608-613.

Benjamin, M. (2013). *Drone Warfare: Killing by Remote Control*. London: Verso.

Bennett, J. (2010). *Vibrant matter a political ecology of things*. Durham: Duke University Press.

Bertek, V. (2014). The Authenticity of the Replica: A Post-Human Reading of Blade Runner. *Literary Refractions*, 1(5) :1-12.

Birhane, A. and van Dijk, J. (2020). Robot Rights?: Let's Talk about Human Welfare Instead. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*: 207-313. https://dl.acm.org/doi/10.1145/3375627.3375855

Bissell, D., Birtchnell, T., Elliott, A. and Hsu, E. L. (2020). Autonomous automobilities: The social impacts of driverless vehicles. *Current Sociology*, 68(1): 116-134.

Bissell, D. (2018). Automation interrupted: How autonomous vehicle accidents transform the material politics of automation. *Political Geography*, 65: 57–66

Blanchette, J. F. (2012). Computing as If Infrastructure Mattered. *Communications of the ACM*, 55(1): 32-34.

Blyth, P-L., Mladenovic, N. M., Nardi, B. A., Ekbia, H. R. and Su, N. M. (2016). Expanding the Design Horizon for Self-Driving Vehicles: Distributing Benefits and Burdens. *IEEE Technology and Society Magazine*, 35(3): 44-49.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J. and Zieba, K. (2016). End to End Learning for Self-Driving Cars. Retrieved from https://arxiv.org/pdf/1604.07316.pdf

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bostrom, N. (2003). Transhumanist values. In F. Adams (Ed.), *Ethical Issues for the 21st Century* (pp. 3-14). Charlottesville: Philosophical Documentation Center Press.

Boyd, D. and Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5): 662-679.

Braidotti, R. (2019). A Theoretical Framework for the Critical Posthumanities. *Theory, Culture, & Society*, 36(6): 31-61.

Braidotti, R. (2017). Posthuman Critical Theory. *Journal of Posthuman Studies*, 1(1): 9-25.

Braidotti, R. (2013). *The Posthuman*. Cambridge: Polity Press.

Braidotti, R. (1993). Embodiment, Sexual Difference, and the Nomadic Subject. *Hypatia*, 8(1): 1-13.

Braidotti, R. (1991). The Subject in Feminism. *Hypatia*, 6(2): 155-172.

Brassier, R. (2007). *Nihil Unbound: Enlightenment and Extinction*. London: Palgrave Macmillan.

Bratton, B. H. (2017a). Geographies of Sensitive Matter: On Artificial Intelligence at Urban Scale. In M. Gomez-Luque and G. Jafari (Eds.), *New Geographies 09: 'Posthuman'* (pp. 28-33). Cambridge: Harvard Graduate School of Design and Actar Publishers.

Bratton, B. H. (2017b). *The New Normal*. Strelka Press.

Bratton, B. H. (2016). *The Stack: On Software and Sovereignty*. Cambridge: MIT Press.

Bratton, B. H. (2015). Outing Artificial Intelligence: Reckoning with Turing Tests. In M. Pasquinelli (Ed.), *Alleys of Your Mind: Augmented Intelligence and Its Traumas* (pp. 69-80). Lüneburg: Meson Press.

Brenner, N. (2014). (Ed.). *Implosions - Explosions: Towards a Study of Planetary Urbanization*. Berlin: Jovis.

Brooks, R. (2000, June 19). Will robots rise up and demand their rights? *Time*. Retrieved from http://content.time.com/time/subscriber/article/0,33009,997274,00.html

Buehler, M., Iagnemma, K. and Singh, S. (Eds.). (2019). *The DARPA Urban Challenge: Autonomous Vehicles in City Traffic.* Berlin: Springer.

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

Butler, J. (2006). *Gender Trouble: Feminism and the Subversion of Identity.* New York and London: Routledge.

Caliskan, A., Bryson, J. and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356 (6334): 183-186.

Caronia, A. (2008). *Il Cyborg. Saggio sull'uomo artificiale*. Milano: Shake Edizioni.

Casner, S. M., Hutchins, E. L. and Norman D. (2016). The challenges of partially automated driving. *Communications of the ACM*, 59(5): 70-77.

Chun, W. H. K. (2008). On ‘sourcery,’ or code as fetish. *Configurations*, 16(3): 299-324.

Citron, D. K. and Pasquale, F. A. (2014) The scored society: Due process for auto- mated predictions. *Washington Law Review*, 89: 1-33.

Claudel, M. and Ratti, C. (2015). Full speed ahead: How the driverless car could transform cities. *McKinsey Quarterly* 2: 89-91. Retrieved from http://www.mckinsey.com/business- functions/sustainability-and-resource-productivity/our-insights/full-speed-ahead- how-the-driverless-car-could-transform-cities, accessed 23 August 2017.

Coeckelbergh, M. (2011). Is Ethics of Robotics about Robots? Philosophy of Robotics beyond Realism and Individualism. *Law, Innovation and Technology*, 3(2): 241-50.

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, 12: 209-221.

Cohen, N. (Ed.). (2010). *Cities Green: An A-to-Z Guide*. Thousand Oaks: Sage.

Cohen, J. (1966). *Human Robots in Myth and Science.* London: Allen and Unwin.

Cowen, T. and Dawson, M. (2009, June 3). *What does the Turing test really mean? And how many human beings (including Turing) could pass?*. George Mason University. Retrieved from https://d101vc9winf8ln.cloudfront.net/documents/28495/original/turingfinal.pdf?1529586155

Crawford, K. and Paglen, T. (2019). Excavating AI: The Politics of Training Sets for Machine Learning. Retrieved from https://excavating.ai

Crawford, K. and Joler, V. (2018). Anatomy of an AI System: The Amazon Echo as an anatomical map of human labor, data and planetary resources. *AI Now Institute and Share Lab*. Retrieved from https://anatomyof.ai

D’Agostino, M. and Durante, M. (2018). Introduction: the Governance of Algorithms. *Philosophy & Technology*, 31: 499-505

Dant, T. (2004). The Driver-car. *Theory, Culture & Society*, 21(4/5): 61-79.

Dant, T. and Martin, P. (2001) By car: carrying modern society. In A. Warde and J. Grunow (Eds.). *Ordinary Consumption* (pp. 143-157). London: Routledge.

Daston, L. (2018). Calculation and the Division of Labor, 1750-1950. *Bulletin of the German Historical Institute*, 62(Spring): 9-30.

Daston, L. (1988). *Classical Probability in the Enlightenment*. Princeton: Princeton University Press.

Davion, V. (2002). Anthropocentrism, Artificial Intelligence, and Moral Network Theory: An Ecofeminist Perspective. *Environmental Values*, 11(2): 163-176.

Davis, M. (1992). *Beyond Blade Runner: urban control - the ecology of fear*. Westfield: Open Magazine Pamphlet Series.

Deelay, M. (Producer), and Scott, R. (Director). (1982). *Blade Runner* [Motion Picture]. The Ladd Company and Shaw Brothers.

De la Escalera, A., Armingol, J.M. and Mata, M. (2003). Traffic sign recognition and analysis for intelligent vehicles. *Image and Vision Computing*, 21(3): 247-258.

Deitch, J. (1993). *Post Human*. New York: Jeffrey Deitch Inc.

Deleuze, G. (1994). *Difference and Repetition*. London: The Athlone Press.

Deleuze, G. and Guattari, F. (1987). *A Thousands Plateaus: Capitalism and Schizofrenia*. Minneapolis: University of Minnesota Press.

Descartes, R. (1641). Meditations on first philosophy. In J. Cottingham, R. Stoothoff, and D., Murdoch (Eds.). *The Philosophical Writings of Descartes* Vol. II (pp. 1-62). Cambridge: Cambridge University Press.

Desser, D. (1999). Race, Space and Class: The Politics of Cityscapes in Science-Fiction Films. In A. Kuhn (Ed.), *Alien Zone II: The Spaces of Science-Fiction Cinema* (pp. 80-96). London: Verso.

Dever, T. (2018). Blurred Lines: Differentiating Humans and Replicants in "Blade Runner". In L. MacKay Demerjian and K. F. Stein (Eds.), *Future Humans in Fiction and Film*. (pp. 94-103). Newcastle upon Tyne: Cambridge Scholars Publishing.

Diakopoulos, N. (2013). Algorithmic accountability reporting: On the investigation of black boxes. *A Tow/Knight Brief*. New York: Columbia Journalism School. Retrieved from http://www.nickdiakopoulos.com/wp-content/uploads/ 2011/07/Algorithmic-Accountability-Reporting_final.pdf

Dick, P. K. ([1968] 1982). *Bladerunner* (original title *Do Androids Dream of Electric Sheep?*). New York: Ballantine Books.

Domingos, P. (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York: Penguin Random House.

Drozdek, A. (1998). Human intelligence and Turing Test. *AI & Society*, 12(4): 315-321.

Duarte, F. and Ratti, C. (2018). The Impact of Autonomous Vehicles on Cities: A Review. *Journal of Urban Technology*, 25 (4): 1-16.

Duvall, T., Hannon, E., Katseff, J., Safran, B. and Wallace, T. (2019, May). What infrastructure improvements will promote the growth of autonomous vehicles while simultaneously encouraging shared ridership? McKinsey & Company. Retrieved from https://www.mckinsey.com/industries/capital-projects-and-infrastructure/our-insights/a-new-look-at-autonomous-vehicle-infrastructure

Ellison, M, Jonze, S. and Landay, V. (Producers), and Jonze, S. (Director). (2013). *Her* [Motion Picture]. United States: Annapurna Pictures.

Elmenreich, W. (2001). *An Introduction to Sensor Fusion* (Research Report 47/2001). Retrieved from https://www.researchgate.net/profile/Wilfried_Elmenreich/publication/267771481_An_Introduction_to_Sensor_Fusion/links/55d2e45908ae0a3417222dd9/An-Introduction-to-Sensor-Fusion.pdf

European Parliament (2016). *Motion for a European Parliament Resolution with recommendations to the Commission on Civil Law Rules on Robotics*.

Retrieved from https://www.europarl.europa.eu/doceo/document/JURI-PR-582443_EN.pdf

Farinelli, F. (2007). *L'invenzione della Terra*. Palermo: Sellerio.

Featherstone, M., Thrift, N. and Urry, J. (Eds.). (2005). *Automobilities*, London: Sage.

Field, M. (2017, August 7). Graffiti on stop signs could trick driverless cars into driving dangerously. *Telegraph*. Retrieved from https://www.telegraph.co.uk/technology/2017/08/07/graffiti-road-signs-could-trick-driverless-cars-driving-dangerously/

Fischer, J. M., and Ravizza, M. (Eds.). (1992). *Ethics: Problems and principles*. Fort Worth: Harcourt Brace Jovanovich College Publishers.

Fleyeh, H. (2004). Color detection and segmentation for road and traffic signs. In *IEEE Conference on Cybernetics and Intelligent Systems*, 2: 809-814.

Foot, P. (1971). The Problem of Abortion and the Doctrine of Double Effect. In J. Rachels (Ed.), *Moral Problems* (pp. 28-41). New York: Harper and Row.

Fostel, G. (1993). The Turing Test is for the Birds. *SIGART Bulletin*, 4(1): 7-8.

Foucault, M. (1973). *The Order of Things: An Archaeology of the Human Sciences*. New York: Vintage.

Franklin, S., Lury, C., and Stacey, J. (2002). *Global Nature, Global Culture*. London: Sage.

Freeman, L. (2011). Reconsidering Relational Autonomy: A Feminist Approach to Selfhood and the Other in the Thinking of Martin Heidegger. *An Interdisciplinary Journal of Philosophy*, 54(4): 361-383.

French, R.M. (2000) Peeking behind the screen: the unsuspected power of the standard Turing Test. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(3): 331-340.

French, R.M. (1990). Subcognition and Limits of the Turing Test. *Mind*, 99(393): 53-65.

Fuerth, L. S. (2009). Foresight and anticipatory governance, *Foresigh*t, 11(4): 14-32.

Fukuyama, F. (2002). *Our Posthuman Future: Consequences of the BioTechnological Revolution*. London: Profile Books.

Furda, A. and Vlacic, L. (2011). Enabling safe autonomous driving in real-world city traffic using multiple criteria decision making. *IEEE Intelligent Transportation Systems Magazine*, 3 (1): 4-17.

Galloway, A. R. (2014). The Cybernetic Hypothesis. *Differences: A Journal of Feminist Cultural Studies*, 25, (1). 107-131.

Ganesh, M. I. (2020). The ironies of autonomy. *Humanities and Social Sciences and Communications*, 7(157) DOI: 10.1057/s41599-020-00646-0

Ganesh, M. I. (2017). Entanglement: Machine learning and human ethics in driverless car crashes. *A Peer-Reviewed Journal About*, 6(1): 77-87.

Gardner, H. (1983). *Frames of Mind: The Theory of Multiple Intelligences*. New York: Basic Books.

Geiger, A., Lenz, P., Stiller, C. and Urtasun, R. (2013). Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11): 1231-1237.

Geraci, R. M. (2008). Apocalyptic AI: Religion and the Promise of Artificial Intelligence. *Journal of the American Academy of Religion*, 76(1): 138-166.

Gibson, W. (1986). *Burning Chrome*. New York: Arbor House

Gillespie, T. (2014a). The relevance of algorithms. In T. Gillespie, P. J. Boczkowski and K. A. Foot (Eds.), *Media technologies: Essays on communication, materiality, and society* (pp. 167-193). Cambridge: MIT Press.

Gillespie, T. (2014b, June 25). Algorithm [draft] [#digitalkeyword]. *Culture Digitally*. Retrieved from http://culturedigitally.org/2014/06/algorithm-draft-digitalkeyword/

Goode, L. (2018). Life, but not as we know it: A.I. and the popular imagination. *Culture Unbound*, 10(2): 185-207.

Gold, H. K. (2015, May 17). Fembots Have Feelings Too. *New Republic*. Retrieved from https://newrepublic.com/article/121766/ex-machina-critiques-ways-we-exploit-female-care

Gold, J. R. (2001). Under Darkened Skies: The City in Science-fiction Film. *Geography*, 86(4): 337-345.

Gomez-Luque, M. and Jafari, G. (Eds.). (2017). *New Geographies 09: 'Posthuman'*. Cambridge: Harvard Graduate School of Design and Actar Publishers.

Grieman, K. (2019). Hard Drive Crash: An Examination of Liability for Self-Driving Vehicles. *JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law* [online]. Retrieved from https://www.jipitec.eu/issues/jipitec-9-3-2018/4806

Guattari, F. (1985). *Chaosmosis. An Ethico-aesthetic Paradigm*. Sydney: Power Publications.

Hall, D.L. (2002). The Implementation of Data Fusion Systems. In A.K. Hyder, E. Shahbazian, E. Waltz (Eds.), *Multisensor Fusion. NATO Science Series (Series II: Mathematics, Physics and Chemistry)* Vol. 70 (pp. 419-433) Dordrecht: Springer.

Hall, D. L., and Llinas, J. (1997). An introduction to multisensor data fusion. *Proceedings of the IEEE*, 85(1): 6-23.

Halpern, O., Mitchell, R. and Geoghehan, B. D. (2017). The smartness mandate: Toward a critique. *Grey Room*, 68: 106-129.

Halpern, O. (2014). *Beautiful Data: A History of Vision and Reason Since 1945*. Durham: Duke University Press.

Halpern, O., LeCavalier, J., Calvillo, N. and Pietsch, W. (2013). Test-bed urbanism. *Public Culture*, 25(2): 272-306.

Haraway, D. J. (2008). *When Species Meet*. Minneapolis: Minnesota University Press.

Haraway, D. J. (2004). *The Haraway Reader*. New York and London: Routledge.

Haraway, D. J. (1997). *Modest _ Witness@Second _Millennium.FemaleMan©_ Meets_OncoMouseTM: Feminism and Technoscience*. London: Routledge.

Haraway, D. J. (1985). Manifesto for Cyborgs: Science, Technology, and Socialist Feminism in the 1980s. *Socialist Review*, 80: 65-108.

Harman, G. (2002). *Tool-Being: Heidegger and the Metaphysics of Objects*. Chicago: Open Court.

Harnad, S. (1991). Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines*, 1: 43-54.

Harvey, D. (1989a). *The Condition of Postmodernity: An Enquiry into the Origins of Cultural Change*. Oxford: Blackwell.

Harvey, D. (1989b). From Managerialism to Entrepreneurialism: The Transformation in Urban Governance in Late Capitalism. *Geografiska Annaler. Series B, Human Geography*, 71(1): 3-17.

Hayles, N. K. (2017). *Untought: The Power of the Cognitive Nonconscious*. Chicago and London: The University of Chicago Press.

Hayles, N. K. (2012). *How We Think: Digital Media and Contemporary Technogenesis*. Chicago: University of Chicago Press.

Hayles, N. K. (2011). Wrestling with Transhumanism. In G. R. Hansell, W. Grassie et al. (Eds.), *Transhumanism and Its Critics* (pp. 215-226). Philadelphia: Metanexus Institute.

Hayles N. K. (2010) 'How We Became Posthuman': Ten years on (an interview with N. Katherine Hayles). *Paragraph*, 33(3): 318-330.

Hayles, N. K. (2005). Computing the Human. *Theory, Culture & Society*, 22(1): 131-151.

Hayles, N. K. (1999). *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago and London: The University of Chicago Press.

Hayles, N. K. (1990). *Chaos Bound: Orderly Disorder in Contemporary Literature and Science*. Ithaca: Cornell University Press.

Hawking, S., Russell, S., Tegmark, M., and Wilczek, F. (2014, May 1). Transcendence looks at the implications of artificial intelligence – but are we taking AI seriously enough? *The Independent*. Retrieved from https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-ofartificialintelligence-but-are-we-taking-9313474.html

Henke, J. (2017). "Ava's body is a good one": (Dis)Embodiment in Ex Machina. *American, British and Canadian Studies*, 29(1): 126-146.

Hester, H. (2018). *Xenofeminism*. Cambridge: Polity Press.

Hirz, M., and Walzel, B. (2018). Sensor and object recognition technologies for self-driving cars. *Computer-aided Design and Applications*, *15*(4): 501–508.

Hodges, A. P. (1983). *Turing: The Enigma*. New York: Simon & Schuster.

Hollands, R. G. (2008). Will the real smart city please stand up?. *City*, 12(3): 303-320.

Hollinger, V. (2009). Posthumanism and Cyborg Theory. In M. Bould, A. M. Butler, A. Roberts and S. Vint (Eds.), *The Routledge Companion to Science Fiction* (pp. 267-278). London: Routledge.

Hollinger, V. (2006). Stories about the Future: From Patterns of Expectation to Pattern Recognition. *Science Fiction Studies*, 33(3): 452-472.

Hollinger, V. (1990). Cybernetic Deconstructions: Cyberpunk and Postmodernism. *Mosaic: A Journal for the Interdisciplinary Study of Literature,* 23(2): 29-44.

Holstein, T., Dodig-Crnkovic, G. and Pelliccione, P. (2018). Ethical and Social Aspects of Self-Driving Cars. Retrieved from https://arxiv.org/pdf/1802.04103.pdf

Howard, B. (2017, May 7). Fatal Arizona Crash: Uber Car Saw Woman, Called It a False Positive. *Extreme Tech*. Retrieved from https://www.extremetech.com/extreme/268915-fatal-arizona-crash-ubercar-saw-woman-called-it-a-false-positive

Human Rights Watch (2020). *World Report 2020*. Retrieved from https://www.hrw.org/world-report/2020/country-chapters/saudi-arabia#49dda6

Hurd, G. A. (Producer), and Cameron, J. (Director). (1984). *The Terminator* [Motion Picture]. United States: Hemdale Pacific, Western Productions, and Cinema '84.

Hutchins, E. (1995). How a Cockpit Remembers Its Speeds. *Cognitive Science: A Multidisciplinary Journal*, 19(3): 265-288.

Jacobson, B. R. (2016). *Ex Machina in the Garden*. *Film Quarterly*, 69(4): 23–34.

Jaggar, A. (1983). *Feminist Politics and Human Nature*. Totowa: Rowman and Allanheld.

Janai, J., Güneya, F., Behla, A., and Geigera, A. (2017). Computer Vision for Autonomous Vehicles: Problems, Datasets and State-of-the-Art. *Journal of Photogrammetry and Remote Sensing*. Retrieved from https://arxiv.org/abs/1704.05519

Jasanoff, S. and Kim, S. H. (2015). *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power*. Chicago: University of Chicago Press.

Johnson, D. G., and Verdicchio, M. (2017). AI Anxiety. *Journal of the Association of for Information Science and Technology*, 68(9): 2267-2270

Kakoudaki, D. (2014). *Anatomy of a Robot: Literature, Cinema, and the Cultural Work of Artificial People*. New Brunswick: Rutgers University Press.

Kang, M. (2011). *Sublime Dreams of Living Machines: The Automaton in the European Imagination*. Cambridge and London: Harvard University Press.

Karel, Č. (2001). *R.U.R. (Rossum's Universal Robots)*. (P. Selver and N. Playfair, Trans.). Mineola: Dover Publications. (Original work published 1921).

Kaplan, J. (2016). *Artificial Intelligence: What Everyone Needs to Know*. New York: Oxford University Press.

Kearns, M. and Roth, M. (2020). *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford: Oxford University Press.

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society*, 20(1): 14-29.

Kitchin, R. (2014). The real-time city? Big data and smart urbanism. *GeoJournal*, 79: 1-14

Kitchin, R. and Dodge, M. (2011). *Code/Space: Software and Everyday Life*. Cambridge: MIT Press.

Kitchin, R. and Kneale, J. (Eds.) (2002). *Lost in Space: Geographies of Science Fiction*. London: Continuum.

Kitchin, R. and Kneale, J. (2001). Science fiction or future fact? Exploring imaginative geographies of the new millennium. *Progress in Human Geography*, 25(1): 19-35.

Kocić, J., Jovičić, N. and Drndarevic, V. (2018). Sensors and Sensor Fusion in Autonomous Vehicles. *2018 26th Telecommunications Forum (TELFOR)*, 420-425.

Komninos, N. (2002). *Intelligent cities: Innovation, knowledge systems and digital spaces*. London: Spon Press.

Krivý, M. (2018). Towards a critique of cybernetic urbanism: The smart city and the society of control. *Planning Theory*, 17(1): 8-30.

Kubrick, S. (Producer/Director). (1968). *2001: A Space Odyssey* [Motion Picture]. United Kingdom and United States: Stanley Kubrick Productions.

Kurzweil, R. (1990). *The Age of Intelligent Machines*. Cambridge: MIT Press.

Kwan, M. P. (2016). Algorithmic Geographies: Big Data, Algorithmic Uncertainty, and the Production of Geographic Knowledge. *Annals of the American Association of Geographers*, 106(2): 274-282.

Latour, B. (2007). *Reassembling the Social: An Introduction to Actor-Network-Theory*. New York: Oxford University Press.

Latour, B. (1999). *Pandora's Hope: Essays on the Reality of Science Studies*. Cambridge and London: Harvard University Press.

Latour, B. (1993). *We Have Never Been Modern*. Cambridge: Harvard University Press.

Latour, B. (1984). The powers of association. *The Sociological Review*, 32(1): 264-280.

Lefebvre, H. (1971). *Everyday Life in the Modern World*. London: Penguin.

Lem, S. (1983). *His Master's Voice* (M. Kandel, Trans.). New York: Harcourt Brace Jovanovich (Original work published 1968).

Leon, L.A. and Rosen, J. (2020). Technology as Ideology in Urban Governance. *Annals of the American Association of Geographers*, 110(2): 497-506.

Luetge, C. (2017). The German Ethics Code for Automated and Connected Driving. *Philosophy and Technology*, 30: 547-558.

Luckhurst, R. (2005). *Science Fiction*. Cambridge: Polity Press.

Luetge, C. (2017). The German Ethics Code for automated and connected driving. *Philosophy and Technology*, 30(4): 547-558.

Lynch, C. R. and Del Casino V. J. (2020). Smart Spaces, Information Processing, and the Question of Intelligence. *Annals of the American Association of Geographers*, 110(2): 382-390.

Macdonald, A. and Reich, A. (Producers), and Garland, A. (2015). *Ex Machina* [Motion Picture]. United Kingdom: DNA Films.

Mann, A. (2014, June 9). That Computer Actually Got an F on the Turing Test. *Wired*. Retrieved from https://www.wired.com/2014/06/turing-test-not-so-fast/

Marchesini, R. (2009). *Il Tramonto dell'Uomo*: *La Prospettiva Postumanista.* Bari: Edizioni Dedalo.

Markoff, J. (2012, November 19). For Web Images, Creating New Technology to Seek and Find. *The New York Times*. Retrieved from https://www.nytimes.com/2012/11/20/science/for-web-images-creating-new-technology-to-seek-and-find.html

Marres, N. (2018) What if nothing happens? Street trials of intelligent cars as experiments in participation. In Maassen, S., Dickel, S. and Schneider, C. H. (Eds.), *TechnoScience in* Society, Sociology of Knowledge Yearbook. Niimegen: Springer/Kluwer.

Marvin, S. & Luque-Ayala, A. (2017). Urban operating systems: Diagramming the city. *International Journal of Urban and Regional Research*, 41(1): 84-103.

Mattern, S. (2015, March). *Mission control: A history of the urban dashboard. Places Journal*. Retrieved from https://placesjournal.org/article/mission-control-a-history-of-the-urban-dashboard/

Maturama, H. and Varela, U. (1972). *Autopoiesis and Cognition: The Realization of the Living*. Dordrecht: Reidel Publishing Company.

Matzner, T. (2019a). The Human Is Dead – Long Live the Algorithm! Human-Algorithmic Ensembles and Liberal Subjectivity. *Theory, Culture & Society*, 0(0): 1-22.

Matzner, T. (2019b). Plural, Situated Subjects in the Critique of Artificial Intelligence. In A. Sudmann (Ed.), *The democratization of artificial intelligence. Net politics in the era of learning algorithms* (pp. 109-121). Bielefeld: transcript Verlag.

McCarthy, J., Minsky, M. L., Rochester, N. and Shannon, C. E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Retrieved from: http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html (accessed 12 February 2020)

McCorduck, P. (2004). *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence* (2nd ed.). Natick: AK Peters.

McNamara, K. R. (1997). Blade Runner's Post-Individual Worldspace. *Contemporary Literature*, 38(3): 422-446.

McQuillan, D. (2018). People's Councils for Ethical Machine Learning. *Social Media + Society*. DOI: 10.1177/2056305118768303

Meillassoux, Q. (2010). *After Finitude.* London: Bloomsbury.

Melgaço, L. and Willis, K. (2017). Editorial: Social smart cities: Reflecting on the implications of ICTs in urban space. *plaNext – Next Generation Planning*, 4: 5-7.

Mitchell, M. (2019). *Artificial Intelligence: A Guide for Thinking Humans*. New York: Farrar, Straus, & Giroux.

Moi, T. (1985). *Sexual/Textual Politics: Feminist Literary Theory*. New York: Methuen.

More, M. (2013). The Philosophy of Transhumanism. In M. More and N. Vita-More (Eds.), *The Transhumanist Reader: Classical and Conteporary Essays on Science, Technology and Philosophy of the Human Future* (pp. 3-17). Hoboken: John Wiley & Sons.

Moravec, H. (1988). *Mind children: the future of robot and human intelligence*. Cambridge: Harvard University Press.

Morozov, E. (2013). *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York: Public Affairs.

Morton, T. (2013). *Hyperobjects: Philosophy and Ecology after the End of the World*. Minneapolis: University of Minnesota Press.

Morton, T. (2010). *An Ecological Thought*. Cambridge and London: Harvard University Press.

Musap, E. (2018). Why is "It" Gendered - Constructing Gender in Alex Garland's Ex-Machina. *Anafora*, 5(2): 403-413.

Natale, S. (2019). If software is narrative: Joseph Weizenbaum, artificial intelligence and the biographies of ELIZA. *New Media & Society*, 21(3): 712–728.

Natale, S. and Ballatore, A. (2017). Imagining the thinking machine: Technological myths and the rise of artificial intelligence. *Convergence: The International Journal of Research into New Media Technologies*. DOI: 10.1177/1354856517715164.

Ng, A. and Lin, Y. (2015, March 15). Self-Driving Cars Won't Work Until We Change Our Roads and Attitudes. *Wired*. Retrieved from https://www.wired.com/2016/03/self-driving-cars-wont-work-change-roads-attitudes/

Norton, P. D. (2008). *Fighting Traffic: The Dawn of the Motor Age in the American City*. Cambridge and London: The MIT Press.

Nussbaum, M. (2010). *Not for Profit. Why Democracy Needs the Humanities*. Princeton: Princeton University Press.

Oliver, N., Potočnik, K. and Calvard, T. (2018, August 14). To Make Self-Driving Cars Safe, We Also Need Better Roads and Infrastructure. *Harvard Business Review*. Retrieved from https://hbr.org/2018/08/to-make-self-driving-cars-safe-we-also-need-better-roads-and-infrastructure

Ouchchy, L., Coin, A. and Dubljević, V. (2020). AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media. *AI & Society*. DOI: 10.1007/s00146-020-00965-5

Paden, B, Čáp, M., Yong, S. Z., Yershov, D., and Frazzoli, E. (2016). A Survey of Motion Planning and Control Techniques for Self-Driving Urban Vehicles. *IEEE Transactions on Intelligent Vehicles*, 1(1): 33–55.

Parisi, L. (2019). Critical computation: Digital automata and general artificial thinking. *Theory, Culture & Society*, 36(2): 89-121.

Parisi, L. (2013). *Contagious Architecture: Computation, Aesthetics, and Space*. Cambridge: MIT Press.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard: Harvard University Press.

Pasquinelli, M. (2019). How a Machine Learns and Fails – A Grammar of Error for Artificial Intelligence. *Spheres*, 5: 1-7.

Pasquinelli, M. and Joler, V. (2020). *The Nooscope Manifested: Artificial Intelligence as Instrument of Knowledge Extractivism*. KIM research group (Karlsruhe University of Arts and Design) and Share Lab (Novi Sad). Retrieved from https://nooscope.ai/Pasquinelli_Joler_Nooscope_essay.pdf

Pendleton, S. D., Andersen, H., Du, X., Shen, X., Meghjani, M., Eng, Y. H., Rus, D. and Ang, M. H. (2017). Perception, Planning, Control, and Coordination for Autonomous Vehicles. *Machines*, 5(6): 1-54.

Piper, H. B. (1962). *Little Fuzzy*. New York: Avon.

Plate, L. (2020, March 31). New Materialisms. *Oxford Research Encyclopedia of Literature.* Retrieved from https://oxfordre.com/literature/view/10.1093/acrefore/9780190201098.001.0001/acrefore-9780190201098-e-1013.

Pollio, A. (2016). Technologies of austerity urbanism: The “smart city” agenda in Italy (2011-2013). *Urban Geography*, 37(4): 514-534.

Pommer, E. (Producer), and Lang, F. (Director). (1926*). Metropolis* [Motion Picture]. Germany: UFA.

Preston, B. (1991). AI, Anthropocentrism, and the Evolution of ‘Intelligence’. *Minds and Machines,* 1(3): 259-277.

Ratti, C. and Biderman, A. (2017). From Parking Lot to Paradise. *Scientific American*, 317(1): 54-59.

Raymundo, O. (2016, March 17). Meet Sophia, the female humanoid robot and newest SXSW celebrity. *PCWorld*. Retrieved from www.pcworld.com

Rhee, J. (2018). *The Robotic Imaginary: The Human and the Price of Dehumanized Labor*. Minneapolis and London: University of Minnesota Press.

Richardson, T. (2017, December 19). Objectification and Abjectification in Ex Machina and Ghost in the Shell. *Medium*. Retrieved from https://medium.com/science-technoculture-in-film/objectification-and-abjectification-in-ex-machina-and-ghost-in-the-shell-b126b8832a1d

Riskin, J. (2016, May 4). Frolicsome Engines: The Long Prehistory of Artificial Intelligence. *Public Domain Review*. Retrieved from https://publicdomainreview.org/essay/frolicsome-engines-the-long-prehistory-of-artificial-intelligence

Riskin, J. (Ed.). (2007). *Genesis Redux: Essays in the History and Philosophy of Artificial Life*. Chicago and London: University of Chicago Press.

Rose, G. (2017). Posthuman Agency in the Digitally Mediated City: Exteriorization, Individuation, Reinvention. *Annals of the American Association of Geographers*, 107: 779-793.

Rossi, U. (2016). The Variegated Economics and the Potential Politics of the Smart City. *Territory Politics Governance*, 4(3):1-17.

Rouvroy, A. (2013). The end(s) of critique: data-behaviourism vs. due-process. In M. Hildebrandt and K. de Vries (Eds.), *Privacy, Due Process and the Computational Turn—The Philosophy of Law Meets the Philosophy of Technology* (pp. 143-167). London: Routledge.

Royal Society (2018). *Portrayals and perceptions of AI and why they matter*. Retrieved from https://royalsociety.org/-/media/policy/projects/ai-narratives/AI-narratives-workshop-findings.pdf

Russell, S. J. and Norvig, P. (Eds.). (2016) *Artificial Intelligence: A Modern Approach, Third Edition*. Pearson Education Limited.

Sabsay, L. (2014). The promise of citizenship: autonomy and abject choices. *OpenDemocracy*. Retrieved from https://www.opendemocracy.net/en/can-europe-make-it/promise-of-citizenship-autonomy-and-abject-choices/

SAE International (2014). *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems J3016_201401*. Retrieved from https://www.sae.org/standards/content/j3016_201401/preview/

Schaffer, S. (1994). Babbage's Intelligence: Calculating Engines and the Factory System. *Critical Inquiry*, 21(1): 203-227.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioural and Brain Sciences*, 3: 417-457.

Shaviro, S. (2014). *The Universe of Things: On Speculative Realism*. Minneapolis: University of Minnesota Press.

Shelton, T., Zook, M. and Wiig, A. (2015). The “actually existing smart city”. *Cambridge Journal of Regions, Economy and Society*, 8(1): 13-25.

Shelley, M. (1818). *Frankenstein, or The Modern Prometheus*. London: Lackington, Hughes, Harding, Mavor & Jones.

Sheller, M. (2004). Automotive emotions: feeling the car. *Theory, Culture & Society*, 21(4/5): 221–242.

Shuppli, S. (2014). Deadly Algorithms: Can Legal Codes hold Software accountable for Code that Kills?. *Radical Philosophy*, 187: 2-8.

Silver, J. (Producer), and L. Wachowski and L. Wachowski (Directors). (1999). The Matrix [Motion Picture]. United States and Australia: Warner Bros., Village Roadshow Pictures, Groucho II Film Partnership, and Silver Pictures.

Simondon, G. (2012). *On the Mode of Existence of Technical Objects*. (C. Malaspina and J. Rogove, Trans.). Minneapolis and London: University of Minnesota Press. (Original work published 1958).

Simonite, T. (2018, November 18). When It Comes to Gorillas, Google Photos Remains Blind. *Wired*. Retrieved from https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

Sinapayen, (2018, November 20). *Sophia the Robot, More Marketing Machine Than AI Marvel*, Skynettoday. Retrieved from https://www.skynettoday.com/

Sini, R. (2017, October 26). Does Saudi robot citizen have more rights than women? *BBC*. Retrieved from https://www.bbc.com/news/blogs-trending-41761856

Söderström, O., Paasche, T. and Klauser, F. (2014). Smart cities as corporate storytelling. *City*, 18(3): 307-320.

Stiegler, B. (1998). *Technics and Time 1*: *The Fault of Epimetheus*. Stanford: Stanford University Press.

Stilgoe, J. (2017). Seeing Like a Tesla: How Can We Anticipate Self-Driving Worlds? *Glocalism: Journal of Culture, Politics and Innovation*, 3: 1-20. Retrieved from http://www.glocalismjournal.net/issues/beyond-democracy-innovation-as-politics/articles/seeing-like-a-tesla-how-can-we-anticipate-self-driving-worlds.kl

Strauss, L. M. (1996). Reflections in a Mechanical Mirror: Automata as Doubles and as Tools. *Knowledge and Society: Studies in the Sociolog*y of Culture Past and Present, 10: 179-207.

Striphas, T. (2012, Feb 1). What is an Algorithm?. *Culture Digitally*. Retrieved from http://culturedigitally.org/2012/02/what-is-an-algorithm/

Svilpis, J. (2008). The Science-Fiction Prehistory of the Turing Test. *Science Fiction Studies*, 35(3): 430-449.

Taffel, S. (2019). Automating Creativity: Artificial Intelligence and Distributed Cognition. *Spheres: Journal for Digital Cultures*, 5. Retrieved from https://spheres-journal.org/contribution/automating-creativity-artificial-intelligence-and-distributed-cognition/

Thomson, J. (1985). The trolley problem. *Yale Law Journal*, 94(6): 1395-1415.

Thrift, N. (2019). The 'sentient' city and what it may portend. *Big Data & Society*. https://doi.org/10.1177/2053951714532241

Thrift, N. (2007). *Non-Representational Theory: Space, Politics, Affect*. London: Routledge.

Thrift, N. (2004). Movement-space: the changing domain of thinking resulting from the development of new kinds of spatial awareness. *Economy and Society*, 33(4): 582-604.

Thrift, N. and French, S. (2002). The automatic production of space. *Transactions of the Institute of British Geographers*, 27(3): 309-335.

Tufekci, Z. (2015). Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency. *Journal on Telecommunications and High Technology Law*, 13: 203-217.

Tulumello, S. and Iapaolo, F. (2021). Policing the future, disrupting urban policy today. Predictive policing, smart city and urban policy in Memphis (TN). *Urban Geography*. DOI: 10.1080/02723638.2021.1887634

Turing, A. (1950). Computing machinery and intelligence. *Mind*, 59: 433-460.

Turkle, S. (2011). *Realtechnik* and the Tethered Life. *Yale Divinity School Reflections*, 98(2): 33-35.

Turkle, S. (1984). *The Second Self: Computers and the Human Spirit*. Cambridge and London: MIT Press.

United Nations (2017). *At UN, robot Sophia joins meeting on artificial intelligence and sustainable development*. Retrieved from https://www.un.org/development/desa/en/news/intergovernmental-coordination/robot-sophia-joins-meeting.html

Urry, J. (2005). The 'System' of Automobility. In M. Featherstone, N. Thrift, J. Urry (Eds.), *Automobilities* (pp. 25-39). London: Sage.

Vanolo, A. (2018). Politicising city branding: Some comments on Andrea Lucarelli's 'Place branding as urban policy'. *Cities*, 80: 67–69.

Vanolo, A. (2017). *City Branding: The Ghostly Politics of Representation in Globalising Cities*. London: Routledge.

Vanolo, A. (2014). Smartmentality: The smart city as disciplinary strategy. *Urban studies*, 51(5): 883-898.

Varun, B. (2004). Blade Runner and the Postmodern: A Reconsideration. *Literature / Film Quarterly* , 3(32): 186-192.

Verbeek, P. P. (2011). *Moralizing technology: Understanding and designing the morality of things*. Chicago: University of Chicago Press.

Vickery, A. (1993). Golden Age to Separate Spheres? A Review of the Categories and Chronology of English Women's History. *The Historical Journal*, 36 (2): 383-414.

Vijipriya, J., Ashok, J. and Suppiah, S. (2016). A review on significance of sub fields in artificial intelligence. *International Journal of Latest Trends in Engineering and Technology*, 6(3): 542-548.

Virilio, P. (1994). *The Vision Machine*. Bloomington and Indianapolis: Indiana University Press.

Volponi, P. (1965). *La Macchina Mondiale*. Milano: Garzanti.

Wardrip-Fruin, N. (2009). *Expressive Processing: Digital Fictions, Computer Games, and Software Studies*. Cambridge and London: MIT Press.

Waymo Team (2016, March 16). Building maps for a self-driving car. *Medium*, retrieved from https://medium.com/waymo/building-maps-for-a-self-driving-car-723b4d9cd3f4

Webb, A. (2017). Distributed Cognition: Assessing the Structure of Urban Scale Artificial Intelligence. *International Robotics & Automation Journal*, 2(5): 187-194.

Webb, M. (1999). Like Today, Only More So: The Credible Dystopia of Blade Runner. In D. Neumann (Ed.), *Film Architecture: Set designs from Metropolis to Blade Runner* (pp. 44-47). Munich: Prestei.

Wei, J., Snider, J. M., Kim, J., Dolan, J. M., Rajkumar, R. and Litkouhi, B. (2013). Towards a viable autonomous driving research platform. *2013 IEEE Intelligent Vehicles Symposium*: 763-770.

Weinert, F. (2009). *Copernicus, Darwin, and Freud: Revolutions in the History and Philosophy*. West Sussex: Wiley-Blackwell.

Whatmore, S. (2002). *Hybrid geographies: Natures cultures spaces*. London: Sage.

White, J. M. (2016). Anticipatory logics of the smart city's global imaginary. *Urban Geography*, 37(4): 572-589.

Wilson, N. (2015, April 29). How Ex Machina Fails to be Radical. *Ms. Magazine*. Retrieved from https://msmagazine.com/2015/04/29/how-ex-machina-fails-to-be-radical/

Winner, L. (1977). *Autonomous Technology: Techniques-out-of-Control as a Theme in Political Thought*. Cambridge and London: The MIT Press.

Withers, R. (2017, April 17). The EU Is Trying to Decide Whether to Grant Robots Personhood. *Slate.com*. Retrieved from https://slate.com/technology/2018/04/ the-eu-is-trying-to-decide-whether-to-grant-robots-personhood.html

Wolfe, C. (2010). *What is Posthumanism?*. Minneapolis: University of Minnesota Press.

Yadron, D. and Tynan, D. (2016, July 1). Tesla driver dies in first fatal crash while using autopilot mode. *The Guardian*. Retrieved from: https://www.theguardian.com

Yeats, W. B. (1920). *Michael Robartes and the Dancer*. Churchtown: The Cuala Press.

Yurtsever, E., Lambert, J., Carballo, A. and Takeda, K. (2020). A Survey of Autonomous Driving: Common Practices and Emerging Technologies. *IEEE Access*, 8: 58443–58469

Zandbergen, D. and Uitermark, J. (2019). In search of the Smart Citizen: Republican and cybernetic citizenship in the smart city. *Urban Studies*, 57(8): 1733–1748.

Zednik, C. (2019). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology*. DOI: 10.1007/s13347-019-00382-7