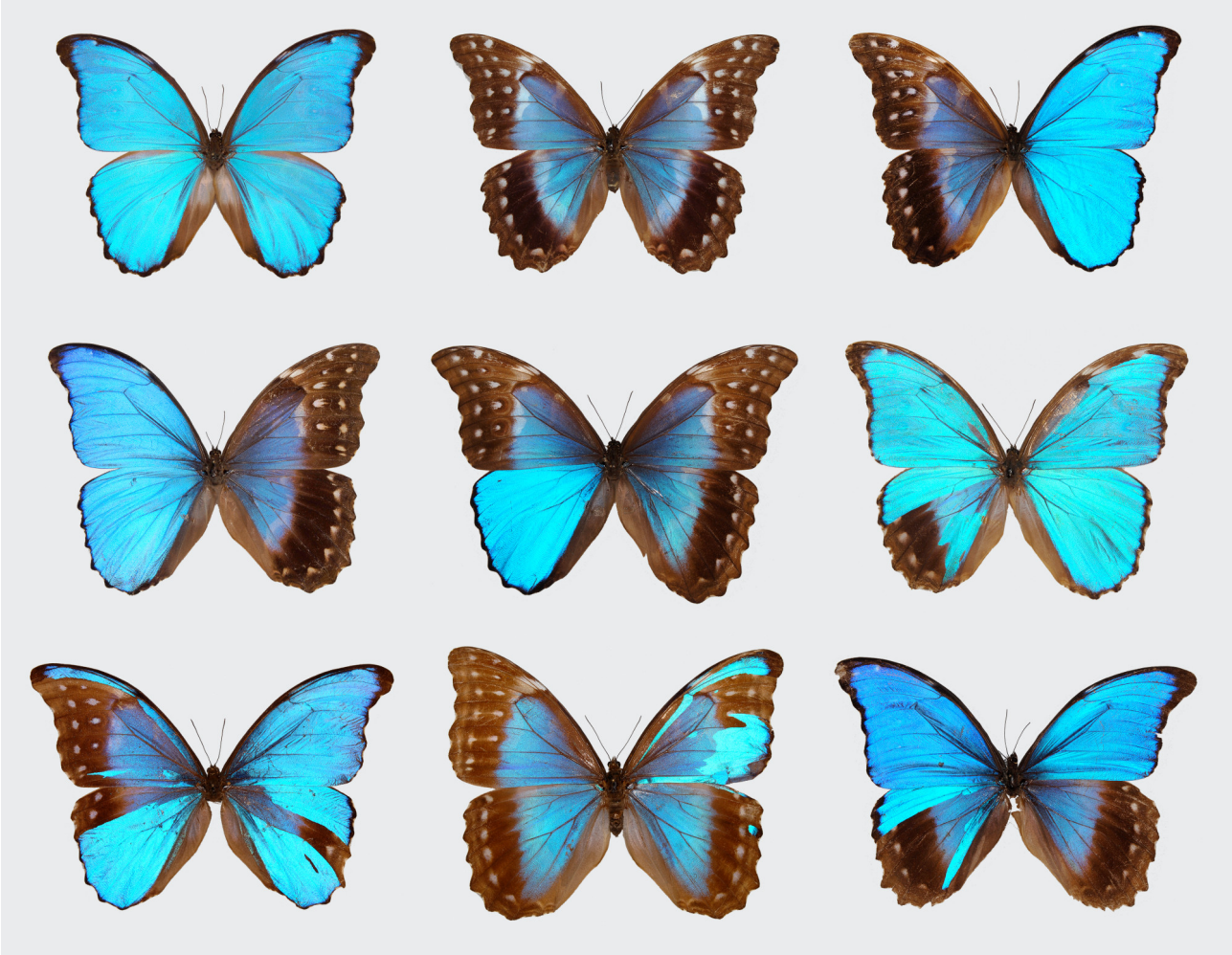
Biased-by-default

position paper AI Culture Lab
Chris Julien



waag
technology & society

colophon

Biased-by-default, position paper AI Culture Lab

Author: Chris Julien, co-authors: Tom Demeyer, Stefano Bocconi

2019, Waag - Amsterdam

Waag | technology & society
Nieuwmarkt 4
1012 CR Amsterdam

waag.org

Cover image:
Top left, a male blue morpho butterfly; top middle, a female. The remainder are gynandromorphic, with both male and female characteristics. - Nipam H. Patel

abstract

In the following, I will argue by means of cultural analysis that a radical shift is necessary in our understanding of bias if we are to develop positive futures with AI. Cultural analysis entails the comparative and interdisciplinary study of technology as a cultural and material practice. The following, as a position paper, describes points of departure for Waag's AI Culture lab, marking territories for further research by raising questions, issues and hypotheses. By exploring the historical context and cultural assumptions of AI, this paper proposes a non-modern framework through which to understand and develop this 'general purpose' technology.

The paper is divided into three parts:
- *chapter one* analyses the current discourse on AI, with its spectacular scenarios;
- *chapter two* looks to a genealogy of AI in modern times, at the ghost in the machine;
- *chapter three* proposes a non-modern notion of bias, as an accounting for ourselves.

introduction

Over the past two years, the phenomenon of Artificial Intelligence has claimed pole position in the rush to be the next technological disruption shaping our near-future. With advances in hardware and data gathering fuelling the ascent of *machine learning* and associated techniques, the promise of Artificial Intelligence has filled our collective consciousness with a weird mix of hope and dread. It calls to mind dire scenarios in which humanity is annihilated by a robot apocalypse, or has merely become redundant in the automation of everyday life. Equally, AI promises a frictionless society of leisure and automated labour, with at its apex our assimilation into a 'singularity' of digitised, transhuman consciousness.

Beyond these spectacular scenarios, technologies associated with artificial intelligence are rapidly being integrated into diverse realms of human activity. The roll-out of ubiquitous computing creates a universal pathway for AI into our lives, ranging from autonomous vehicles to predictive policing and from micro-targeting feeds to urban surveillance networks, and raising complex questions of ethics and control.

hyperbolic phenomena

"Although science fiction may depict AI robots as the bad guys, some tech giants now employ them for security. Companies like Microsoft and Uber use Knightscope K5 robots to patrol parking lots and large outdoor areas to predict and prevent crime. The robots can read license plates, report suspicious activity and collect data to report to their owners." — Gartner Top 10 strategic technology trends for 2019[1]

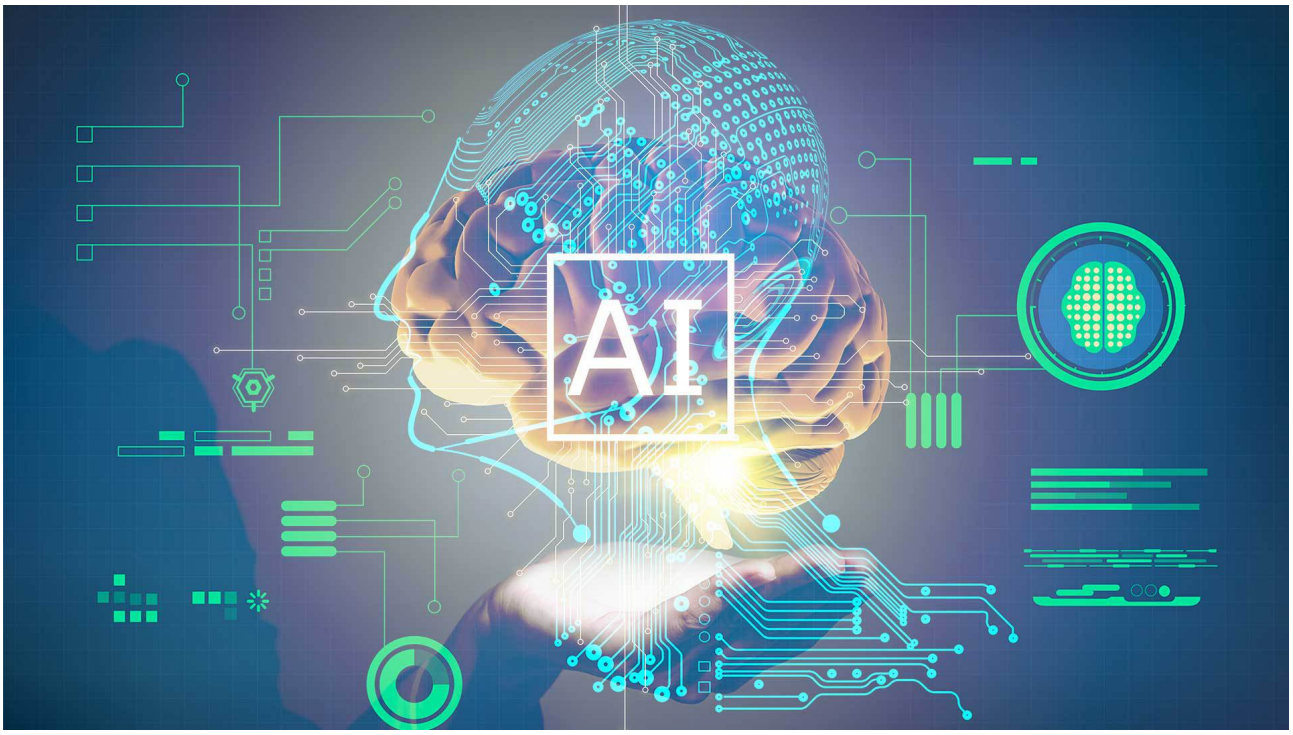
How to make sense of the apparent paradox that AI is predicted to be both our saviour and our undoing? I'll start by taking a closer look at sources of discourse on Artificial Intelligence, focussing on a couple of tropes (both visual and textual) that express our attitudes - and indeed biases (!) - to the phenomenon. These tropes, oft-repeated 'facts' and predictions about AI that appeal to our common sense, seem at times wildly hyperbolic. Yet, by their very superlative character, these hyperboles might help make sense of collective interests in, and expectations of AI technologies as they circulate today[2].

Right off the bat, the term 'Artificial Intelligence' *itself* gives rise to hyperbole, as this word-pair conflates the intelligence of the human mind and the 'intelligence' expressed through algorithms. In linguistics this comparison would properly be called a simile, where two essentially unlike things are compared. Such a comparison is signalled by the interjection of 'as' or 'like'. In the case of AI then, we ought to speak of 'intelligent-like' qualities, an essential nuance which is often lacking in contemporary discourse. To put it in more material terms, the human brain and the neural net are conflated. From a material, i.e. physical perspective it is obvious that our brain (sitting right *here* in your head) and its complex cultural manifestations is really, really different from a coded neural net spinning on a server farm *somewhere* (operated by *someone*), communicating in binary with a server farm (or your phone!) somewhere else. Yet, in our imaginary, an abstract impression of sameness persists, irresistibly portraying these two things in overlay.

---

[1] https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2019/

[2] The style of delivery of these 'facts' is captured brilliantly by Jake Elwes and his 'Dada Da Ta' video work, available via https://www.jakeelwes.com/project-DaDaTa.html



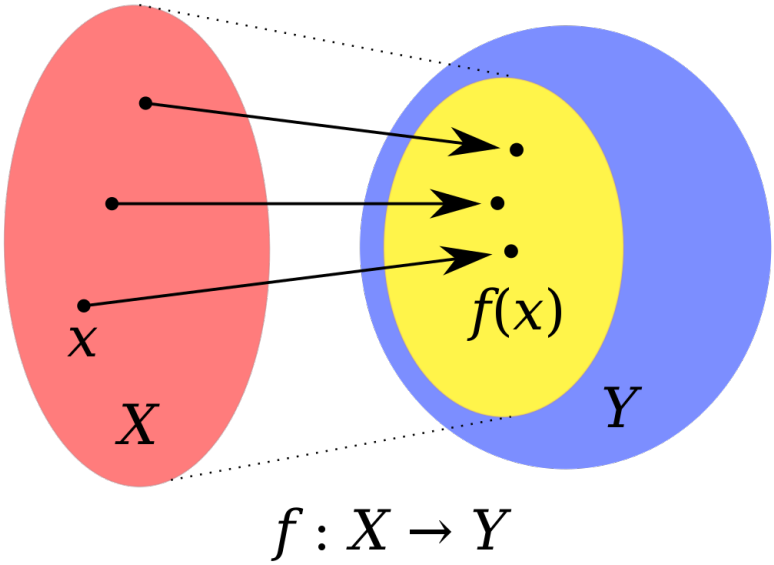
[image 1]

Images such as this seduce us to see the digitally rendered pattern of the neural net in equal relation to its biological, physical source. We see neural nets and smooth brains, floating convivially in a cybernetic and self-referential cloud of data. These images, a staple in reporting on AI, evoke a sense of meta-ness and control: abstracted from their physical bodies and material connections, they are designed to suggest phenomena we can interface with directly and modulate according to our will.

equation

Both physically and functionally, the relationship between the neural net and the brain is not one of equality. The neural net was originally *derived* from certain characteristics of the brain's synaptic function[3], yet in its development and implementation through algorithms in computational systems it has become a distinct and non-equitable phenomenon. What explains the persistent equation of human intelligence and the capacities of AI, even though the original 'neural' connection is tenuous? Referencing mathematics, the relationship between a *derivative* and its *domain* might help to clarify, as illustrated in the diagram below:

[3] The foundations were laid in 1943 by by Warren McCulloch and Walter Pitts in their 'A Logical Calculus of Ideas Immanent in Nervous Activity'.



[image 2]

Take a moment to observe the situation described here. We see a *domain* 'X' (red), which I take to stand for the cognitive capacities of the human brain as we understand them today. This domain is related to a *function* (f(x)), to be understood for present purposes as an implemented neural net. Their relationship is manifested by 'x', or 'argument', which is an extracted member of the domain, and serves as the input for the function, which in its turn creates an output value. This relationship helps to grasp the non-equivalence between biological neurology and computational neural networks. The neural network is indeed a derivative of the human brain; it's an extraction of certain characteristics of the neuron (or 'x'), as observed within the overall set of brain functions, and their subsequent harnessing in a specific, computational function. In terms of human intelligence, this entails an extraction and a morphing of that overall capacity, as a *partial function* of the brain. This suggests the 'intelligence-like' characteristics of machines are a functionally limited, algorithmically rendered version of a certain feature of human intelligence[4]. It is a bit 'like' the brain, and as a technological system certainly has capacities that the human brain does not, but that does not mean they are equal or comparable.

The fact that AI functions were originally derived from certain functions of the human brain, yet are not functionally equal, is important in understanding the hyperbolic

[4] For further reference to the mathematical underpinnings to this image, see: https://en.wikipedia.org/wiki/Domain_of_a_function and https://en.wikipedia.org/wiki/Partial_function - in particular as the domain-function relationship bears on morphism in category functions and the subset-superset interactions of set-theory.

projection of the potential applications of AI. In the diagram above, these relate to the blue 'Y' field, which indicates the *co-domain, or target set* of the function, within which the yellow region is known as the *image* of the function. This image extrapolates the potential of the function, and can be equal or smaller than the co-domain. How we understand the value of AI, expressed in its image and target-set, goes to the heart of our overall understanding - tropes included - of the phenomenon 'artificial intelligence'. If we accept that the function of AI is not a fully *reproduced extension* of human intelligence by technology, but rather a *reductive extraction* of human intelligence into technology with its own distinct functioning, then the range of outputs and target set are not equal to the domain from which it derives. This means that - however we understand the function, image and target set of AI to manifest based on their technological capacities - to assume their equivalence to human intelligence and extrapolate from there seems to be a category mistake. Put simply, what we can reasonably expect AI to do, depends on what we reasonably accept AI to be.

googlespeak

The non-equal relationship between the domain of human intelligence (X), its functional derivative in AI systems (f(x)) and the resultant target set (Y) helps to clarify another mainstay of hyperbole on AI. This involves the projection and extrapolation of the 'intelligent' qualities of AI unto the future. Speculative projection is the staple of futurists, but in the case of AI many serious engineers and academics engage in such flights of fancy. They do so in a genre we might describe as so-called 'non-fiction about the future' (or 'googlespeak'). Statements of this kind can be identified by their equally spectacular and unverifiable claims, for example this classic:

"*Human level AI will be passed in the mid 2020's, though many people won't accept that this has happened.*"

— Shane Legg, Chief Scientist at Google Deep Mind[5]

Here, the non-linear quality of the hyperbole of extrapolation and projection of Artificial Intelligence is at full play as 'human level AI'. It's a movement that we can now break down into 1) a huge potential range of AI is projected, founded on 2) an imprecise

[5] quoted in: https://slatestarcodex.com/2015/05/22/ai-researchers-on-ai-risk/

understanding of its function, which stems from 3) the incorrect equation of the function to its domain. Of course, we cannot fault anyone simply for enthusiasm - indeed, AI technologies clearly show significant potential applications in automating cognitive tasks. Honest confusion on the relationship between the derivative and its range aside, the point is that there are obvious material interests in presenting hyperbolic futures for AI.

commodity

The greater the confusion about the actual function of AI, the larger the spectacular image of outputs that can be convincingly projected onto the future. And with bigger expectations of the technology come larger investments, higher salaries, greater regulatory leniency and policy support. In particular, big tech, driven by a venture capital funding model, requires such promises to accumulate capital and adjust (economic) activities. From SoftBank to Google, the corporate interests for inflating the promise/prowess of AI for data-driven tech are evident - it's their product, after all. Many academic partners to such companies worryingly parrot these hyperboles, while adding some of their own:

"Artificial Intelligence (AI) will, as a key technology, change the world as radically as the industrial revolution did in the 18th and 19th century... On this topic, Sundar Pichai, CEO of Google, commented in February 2018 that: 'AI is one of the most important things that humanity is working on and will have more impact than electricity or fire.'"

— AI voor Nederland rapport 2018, pp 1[6]

Remarkably, this NWO-published report, presented to the Dutch government by a group of leading academics in collaboration with the employer's organisation VNO-NCW and Boston Consulting Group, equates in no uncertain terms AI's impact to that of the Industrial Revolution, only to trump their own claim by way Google's Pichai, who promises it to be of greater consequence than the invention fire and electricity. Given that we are only today beginning to grasp the full consequences of the industrial revolution on our planet, and that these consequences have yet to play out over decades, centuries and perhaps millennia, the confidence with which this scientific

[6] https://www.nwo.nl/documents/enw/rapport-ai-voor-nederland-vergroten-versnellen-en-verbinden

commission heralds the future impact of AI is quite astonishing. Pichai's prediction that AI in and of itself will exceed the impacts of *fire* and *electricity,* is particularly significant in its impossibility, as such factual claims about the future by definition cannot be falsified, and more pertinently AI obviously cannot exist without electricity.

One of the reasons that such wild claims can be suggested, is another trope contained in these images and statements. This is the trope of the 'brain in the jar', where AI is presented as a stand-alone technology, calling to mind the images of disembodied, floating brains to which AI is equated. Rather than seeing AI as part of an *ensemble* of technologies, implemented in real-time, living societies, AI is presented as a stand-alone *commodity* that can be exponentially developed and as such promises to be highly profitable - irrespective of its contexts or history. Under the pressure of intense global competition and the thinly veiled threat of opportunity costs that will have to be paid for missing the boat, it is advised that this commodification should proceed in all haste, with minimal 'friction' from considerations for externalities and ethics.

Beyond the financial drivers for hyperbole, which are a common characteristic in capitalist markets, propagating such future images of AI comes with societal and cultural consequences. Not only are we being served up weird futures based on a simplistic, quantified world view that's driven by profit-seeking agents, but we're also being distracted from matters of urgent concern today. The hyperboles of AI produce images of spectacular futures that obscure concrete implementations of the tech right now, as well as the economic and political interests behind them. Pedro Domingos captures the intentional paradox of this situation well with his much-quoted reverse-trope: "people worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world."[7]

spectacle

Indeed, the current performance of AI systems seems to structurally go astray as soon as implementation shifts from well-bounded games such as chess or go[8] to real-world

[7] Pedro Domingos, The Master Algorithm (2016)

[8] See AlphaGo - the movie for an exposé of the labour and reasoning behind machine learning besting human players at abstract games — https://www.alphagomovie.com/

issues such as criminality or employability[9]. Presenting a spectacular future while obscuring the movements of power and its societal consequences in the present was analysed well by Guy Debord in his seminal 'Society of the Spectacle', where he states:

"*In form as in content, the spectacle serves as a total justification of the existing system… The spectacle divides the world into two parts, one of which is held up as a self-representation of the world, and is superior to the world… The spectacle is* capital *accumulated to the point where it becomes image*."

— Guy Debord, Society of the Spectacle (1967)

The hyperboles of the AI phenomenon, from its naming, to its visual representations and its future promise, can thus be best understood from a systemic perspective as spectacles. With these spectacles, the real-time implementation of AI systems is *intentionally* obscured, as are the interests driving them, and their consequences in our societies. For us, as part of a technological development trajectory rushing towards societies where human judgement can be automated and creative human behaviour is foreclosed[10], a critical attitude towards the current discourse on AI is as urgent as another, more productive understanding of the the technology's uses and ends.

[9] The practical failure of AI systems to implement 'effective computation' in relation to complex problems has been well documented. See for example 'Artificial Unintelligence' by Meredith Broussard (2018) for a contemporary account of the structural failure of information technologies to deliver us the windows promised by what she calls '*technochauvinism*'.

[10] This crucial argument of foreclosure of future choice as a systemic feature of current AI development is developed by Shoshana Zuboff in 'The Age of Surveillance Capitalism' (2018)

images of neutrality

"*We thought there were windows but actually they're mirrors*."

— Julieta Aranda, Brian Kuan Wood, Anton Vidokle,
The Internet Does Not Exist (2015)

In assessing the sprawling logic of the internet, Aranda, Wood and Vidokle note the impossibility of getting *inside* of an information network. This leads them to conclude that the internet held a promise of windows, which have turned out to be mirrors. With their analysis in the introduction of 'The Internet Does Not Exist', they suggest a genealogy that 'lifts AI out of the jar' and embeds it in larger historical trajectories of information technologies, questioning why the technological windows that promise views of the world structurally turn into mirrors, in which we only see ourselves.

end of history

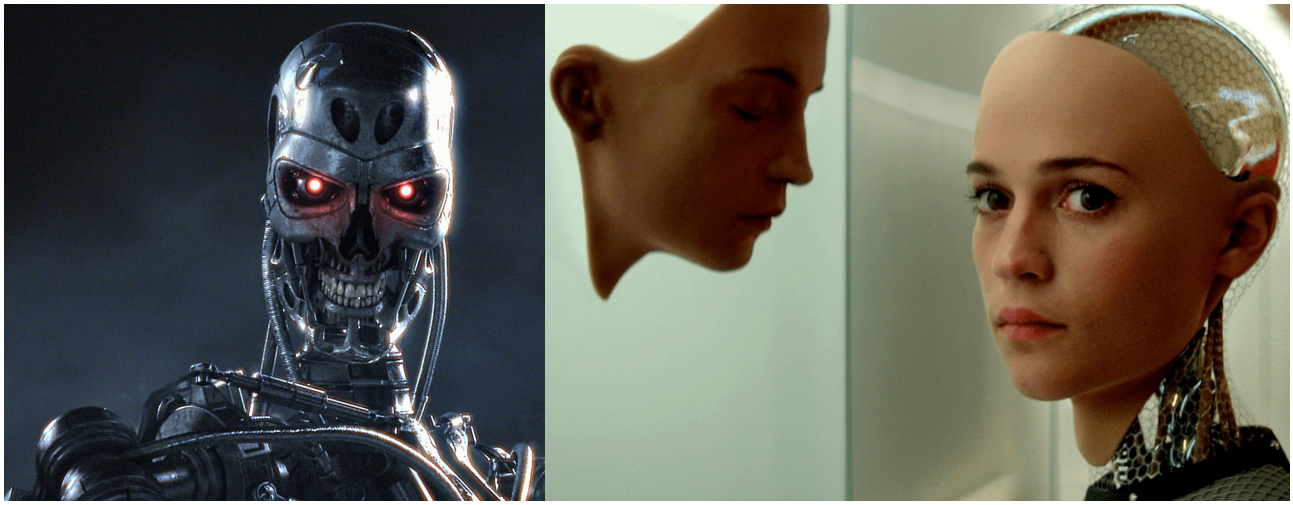
Aranda, Wood and Vidokle critically assess the post-cold war development of the internet, shifting towards data-driven and attention-oriented economic expansion, whose "information networks actually do have an ideological structure". It's a structure styled after Fukuyama's *End of History* thesis, where the free interplay of markets and representative democracy became the post-ideological norm. This post-ideological attitude drove the realisation of the "commercial and economic potential of information exchange, placing it at the centre of the era of globalisation and acceleration in the financial sector", leading the internet down the path of privatisation and personalised exploitation we have before us today. Although Fukuyama's thesis has been widely derided for its arrogant assertion of Western values, its reflection of the Zeitgeist of the 1990s remains pertinent in understanding the re-assertion of Western hegemony after the Cold War. By presenting this victory of the West as a natural and necessary outcome, its mode of liberal democracy and associated 'modern' world view became the default global perspective.

With this history, the authors situate Artificial Intelligence within a political and economic trajectory of information technologies, which matured and spread in the

internet during a recent, post-ideological period. This suggests that the spectacular images of AI we've studied are part of a longer history, moreover, a history where big promises have been made ánd broken before! Back then, the promised vistas were to be enjoyed through the window that the internet promised to be. Now, we've come to understand that these vistas become worryingly distorted by cultural and indeed spectacular reflections. Nonetheless, with AI, the revolutionary promise of machines has returned. Apparently, such promises appeal to deep-seated cultural assumptions, which might be better understood by tracing a longer legacy of windows and mirrors.

The human-machine relationship, with its master-slave dynamics was a topic of debate and speculation throughout much of the 20th century. This complex relationship was signalled by philosopher Martin Heidegger in terms of our extension of agency through tools, which in their use come to determine us[11]. Indeed, since early modern times technologies have offered us scientific apparatuses of measurement and automation, allowing us to pull the curtain back ever further, to see and manipulate our 'domain' much as we would through a window. Yet as our apparatuses grow more complex, tools increasingly give rise to tools, heightening the mirroring effects of technology *on* humanity. Beyond systemic, instrumental reasons for the hyperbolic images of AI, the 'mirror' effects of AI seem to have cultural and epistemological significance regarding our relationship with technology that is more deeply ingrained in our understanding of the world, and therefore both harder to pin down, and more important to understand.

ghost in the machine



[11] Heidegger states: "So long as we represent technology as an instrument, we remain held fast in the will to master it." — Martin Heidegger, The Question Concerning Technology (1954)

[image 3]

For almost a century, the spectacles of ‘man-machine’ futures have been projected onto the silver screen. Cinema, and blockbuster movies in particular, are often understood to provide a cathartic function, articulating our collective anxieties within the safe zone of fiction and entertainment. As such, cinema provides us with a lineage of images that precede our current fascination with AI. Anyone with an affinity for cinema can trace a line of fantastic stories about our future relationship with our techno-spawn, drawing a timeline of how we’ve projected our anxieties and aspirations onto technology over the past century. Such a list of movies on the becoming-human of technology might include such memorable embodiments as Frankenstein’s ghoulish monster (1910, 1931), the genocidal Terminator (1984), a lonely Blade Runner (1982), and recently the ephemeral Ex-Machina (2014). Remarkably, in each of these cases the projection of the human on technology heralds the potential end of that very humanity.

The narrative arc common to these movies shows us information networks that catalyse the autonomous development of machines as a precursor to violent confrontation with their creators. Usually, this narrative is understood to be an expression of the hopes and fears of a humanity staring across the uncanny valley at their own creations - staring in the mirror, if you will. In this reading, our inner workings are projected on the more or less blank slate of technologies-to-come. Certainly, this aesthetic-reflective line of reasoning involves a lot of mirroring. Yet the level of truth that is claimed for the tropes surrounding AI, suggest that these images speak in a more existential way to our relationship with technology, than merely expressing unreasonable yet understandable anxieties about selfhood, etc. Why has European culture for over a century fantasised time and again that the end days will come with technological transcendence? Do the riders of the apocalypse indeed wear metal skins?

progress

Such paradoxically seductive fantasies show us how in modern times, our belief in windows, with their promise of agency and even liberation, cannot be separated from the threat of the mirror and the 'ghost in the machine’. What cultural assumptions lead us to associate a glimpse of ourselves in technology with our downfall? Such narratives are wholly distinct for example from the animistic perspectives familiar to many non-

modern cultures. There, the movements between the human and non-human are often as spiritual as they are mundane, and certainly do not automatically signify impending doom[12]. If it is indeed a typically modern, European view to see the human as set apart from the world, then what do these fantasies tell us about modernity? It seems to me that what these images of apocalyptic transcendence represent is a release from that very modernity. From our perspective, becoming machine entails our liberation from the ever onward march of linear, modern history; a march that we otherwise must trudge along with until the ends of time. It seems that the modern trajectory 'terminates' *only when we* are *subsumed in our own progress*[13]. On the silver screen we encounter the spectacular image of our modern destiny: to become ourselves automated, completing our capitalist teleology of general equivalence by becoming commodities. As Guy Debord envisaged: "the spectacle is the bad dream of modern society in chains, expressing nothing more than its wish for sleep. The spectacle is the guardian of that sleep."[14]

Upon closer inspection then, our hyperbolic, spectacular images of AI appear within a modern frame that fundamentally offers us a choice between two dreams: to progress-without-end... or to end-in-progress. Conceiving of different fates with technology, seems then to entail taking leave of some specifically modern assumptions about the relationship between ourselves and the world. What critical element of our modern frame forces us to choose between being 'only human' and a 'ghost in the machine'? Perhaps what's particular about this modern frame can be discerned by continuing to investigate the legacy of what's in the frame, those windows and mirrors. With their capacity for transparency and reflection, windows and mirrors share a common quality: allowing light to pass or reflect without distortion. This shared lack of distortion means both are *neutral media*. Is neutrality the general quality of the modern frame, and the quality that inhibits more positive ends to be conceived of for AI?

wobble

---

[12] As George Manuel notes in his 'Fourth World: An Indian Reality' (1974): "Spiritual and material power have never been wholly separated in the Indian world as they seem to have been elsewhere." pp 43

[13] This paradox of recurrence-or-nothingness in our modern destiny was announced by Nietsche with his 'eternal return of the same' in The Will to Power (1887): "Let us think this thought in its most terrible form: existence as it is, without meaning or aim, yet recurring inevitably without any finale of nothingness: '*the eternal recurrence*.' This is the most extreme form of nihilism: the nothing (the 'meaningless'), eternally!"

[14] Guy Debord, 'Society of the Spectacle' (1967) pp 18

After all, the techno-mirrors we've been discussing do not show us direct reflections of ourselves, nor do they show simply the clear blue skies outside. Rather, both the internet and AI display weirdly distorted, partial images mediated through dizzying arrays of functions and circuits. This distortion occurs irrespective of whether these arrays are intended for transparent or reflective functions, pointing to a shared underlying process. It seems that gazing into the machine - into the black box - shows us weird, wobbly versions of ourselves *and* the world, all mixed up and spliced. Instead of a smooth passage, any information crossing technological surfaces becomes weird and ghostly; *diffracted*. Under such conditions[15], can we even distinguish the windows from the mirrors? If this 'wobble' is indeed a shared quality of windows and mirrors, inevitable when we gaze into black boxes, then our view both of the outside and of ourselves becomes questionable. Even though easily discernible, this disturbs and confuses us, causing self-inflicted myopia when it comes to 'the wobble'. We'd rather see it as an uninvited guest, a ghostly haunting, or an attack surface; something to design *against* or weaponise. It makes sense that wobble 'looks like a nail to people holding hammers', which is why engineers respond by promising a fix for the next version. Yet, in the context of ubiquitous computing and connectivity, such incremental updates turn into a perpetual state of deferral, which is usually known as 'solutionism'[16]. Instead of suppressing 'those little things', we should study this wobble, which seems to be giving off some crucial information about modernity - not to mention on why computation never delivers on its promise of efficient objectification and accurate prediction.

What is this wobble, shared by windows and mirrors alike telling us? Might it be reasonable to assume that this wobble is not a weird deficiency of technology, but that it is caused by our hand holding the camera as we try to take a selfie, the hand that built the window? In other words, this wobble seems to raise a critical issue of measurement, which is whether it is possible at all to subtract ourselves from our view of the world. Although our myopic view has hindered widespread acknowledgement of this fact, we actually have strong experimental evidence confirming the wobble. Since the development of quantum physics in the 1930s, experimental studies have

---

[15] Karen Barad, in 'Meeting the Universe Halfway' develop Donna Haraway's *diffractive method* in terms of a new materialist notion of agency based on quantum field theory, which will inform further research of the AI Culture lab into measurement and data.

[16] For a forceful critique of this notion, see Evgeny Morozov: 'To Save Everything, Click Here: The Folly of Technological Solutionism' (2013)

confirmed that our measurement apparatuses, by their mere presence in a situation, i.e. without actively carrying out a procedure, *change the possibility field of the observed phenomena* at the quantum level.[17] In other words: any black box we build physically influences what can be subsequently measure, i.e. perceive. This crucial insight on material interaction at the quantum level tells us that 'the wobble is real', and it's known as a phenomenon Einstein referred to as 'spooky action at a distance'[18]. If it is not possible to separate the act of measuring from the phenomenon being measured, then windows and mirrors collapse into each other. This means that the modern frame, with its promises of clear windows and mirrors, can only promise clarity by obscuring the human element. This element is subtracted both from the act of observation and from the material build of the black box, even though the wobble forces us to recognise how we influence both and thereby any situation we seek to access through technology.

In an unexpected way, this wobble caused by black boxes of AI throws into question our whole idea of a neutral perspective. Suddenly a Starbucks cup appears in the frame of the spectacle, breaking the fourth wall of representation and pointing our perceived 'natural' reality to be a man-made construct. From a modern standpoint, the absence of neutrality is an awkward thing to find out. Indeed, the absence of neutrality in technological systems of measurement connects the history of AI and the images it brings forth in our present 'post-truth moment' with a history stretching back to the Renaissance.

---

[17] See Karen Barad, chapter 3 of Meeting the Universe Halfway, where they discuss research on 'which path' experiments: "In other words, *all that is required to degrade the interference pattern is the possibility of distinguishing paths*.... Additional theoretical analysis... confirms this remarkable finding: "It is also important to emphasise that the quantity D(P) is *distinguishability*, and the suffix "ability," connoting physical possibility, is crucial. The limitations upon fringe visibility... are not imposed by the actual information that the observer has extracted concerning the particles of interest, but the information that could *in principle* be extracted within the constraints established by the preparation." (Jaeger et al. 1995, 51; italics mine)"

[18] On the consequences for classical, mechanistic notions of causality that are challenged by these insights, see https://aeon.co/essays/can-retrocausality-solve-the-puzzle-of-action-at-a-distance

neutrality

[image 4]

The hyperboles of AI seem to be inscribed in a fundamental modern assumption about our relationship with the world, which dictates that we need a neutral position to access reality. The wobble we mistook for the 'Ghost in the machine' is challenging this assumption, showing us that every window is also a mirror, which effectively precludes the possibility of a neutral view of the world. In particular when it comes to technologies founded on probabilistic science, the absence of a neutral observer-standpoint spells trouble, as it throws off the relationship between material measurement and mathematical prediction. How did this notion of a neutral standpoint come to stand between us and the world, and a deeply rooted, commonsensical part of our cultural attitudes to technology?

Taking a step back from that common sense, I think it's good to ask ourselves what it actually means to be neutral. Being neutral seems to be an acknowledgement and a disavowal at the same time. Your position is that you have no position. It is in fact, a paradoxical term without clear definition, in many cases referring back to itself, such as in the Mirriam Webster Dictionary, where neutral is defined as: "'one that is neutral"; or to an absence of characteristics, such as in the Oxford Dictionary definition: "Having no strongly marked or positive characteristics or features."[19] In the context of gender studies and decolonisation, it has become highly suspect to claim a lack of positive

[19] Remarkably, in mathematics a neutral element is also known as an 'identity element', or simply 'identity' https://en.wikipedia.org/wiki/Identity_element

features. Such claims are also known as 'unmarked categories'[20], which generally indicate a disavowal of (power) relations. Tellingly, Wikipedia's Wiktionary features as one of its definitions of neutrality: "An individual or entity serving as an arbitrator or adjudicator"[21]. It makes sense that the notion of such a cool, distanced position we attribute to AI today is actually part of a larger modern fantasy, one that goes back centuries. This neutral world view was once conjured up by European men during their ongoing 'discovery' and conquest of the globe. This globe had, before their coming, been properly 'pre-historical', not partaking in the neutral observer position. To those dapper humanistic scientists and explorers, it made complete sense that their perspective, rather than that of an almighty and all-seeing God, should be the pivot of reality, the adjudicator of the real.

However, it seems that there is ample evidence from both the natural and human sciences to question the pivotal assumption of modernity, that of the neutral observer, which turns out to radically separate humanity from reality, opening up an unbridgeable quantum chasm of probability[22] and representationality. Looking closely at our windows confronts us with the appearance of the ghost in the machine[23], the very concept first dreamt up by Renée Descartes when he separated mind from body in his meditations (1637). Inventing *cogito ergo sum,* he birthed the Terminator.

Our images of AI seem to preclude futures that meaningfully deviate from the cybernetic automation of society because of the fundamental modern separation of mind and body, of human and non-human. The representation of an act tells us more

---

[20] As Bruno Latour explains in Facing Gaia (2015) in relation to man/woman (pp 15-16): "this means that the term "man" is an *un*marked category: it poses no problem and attracts no attention. When the term "woman" is used, attention is drawn to a specific feature, namely, her sex; this makes the category *marked* and thus detached from the unmarked category that serves as its background."

[21] If we would like to extend our analysis back to classical antiquity, Giorgio Agamben points out in 'The Adventure' (2018) that before becoming the Goddess of retributive justice, *Nemesis* was the Deity of impartiality among ancient Greeks and Egyptians, the term coming from *nemein*, meaning 'to assign' or 'the one who assigns'. (pp 16-17); in our times, *Nemisis* is more generally understood as that which we cannot conquer or will defeat us.

[22] In 'What is Real?' (2018) Georgio Agamben identifies: "the principle that probability does not concern a real given event but only the tendency to infinity of the number of examined samples… Those who act with probability in mind abide by this superimposition and are compelled to acknowledge, more or less tacitly, that although it never determines an individual case, it can nonetheless influence to some extent their decisions with respect to reality, in spite of the evident paralogism." pp 32

[23] The term was in fact coined as a critique of Cartesian reason by Gilbert Ryle in The Concept of Mind (1949) as *the dogma of the Ghost in the machine*: "with the doubtful exceptions of the mentally-incompetent and infants-in-arms, every human being has both a body and a mind… In consciousness, self-consciousness and introspection, he is directly and authentically apprised of the present states of operation of the mind… I hope to prove that it is entirely false, and false not in detail but in principle. It is not merely an assemblage of particular mistakes. It is one big mistake and a mistake of a special kind. It is, namely, a category mistake."

about the hand of the painter than about its material history, and probability refers more to its mathematical conditions than to material reality. Renouncing neutrality indeed seems to lead us to an impasse, threatening to lock us in to a self-referential loop that precludes direct access to the world. How should we direct ourselves to allow this technology to play a productive, societal role without consuming our humanity in the name of automated profit, or nullifying knowledge in the chasm of self-referentiality? Up until now, our almost religious belief in the possibility of a neutral position has caused us to oscillate between windows and mirrors; between self-sameness and annihilation in an attempt to bridge the gap between probability and reality. Today, this oscillation holds the trajectories of Artificial Intelligence hostage to a particular cultural frame, and the power structures built upon it. Against it, the question is: how we can imagine AI trajectories and ends in the absence of neutrality, *with* the wobble?

material entanglements

If neutrality is a modern myth once dreamt up by white men to wrest dominion from their God, then how can we live without it? We've been stuck between windows and mirrors for a long time; trapped between Nietzsche's 'eternal return of the same' and our transcendental erasure in technology. We've been attempting much like Sisyphus to roll the rock of representation up over the threshold of meaning[24]. In the 1980s, Donna Haraway made sense of this situation, pointing out to us pattern-seeking animals that:

"*Technology is not neutral. We're part of what we make and it is part of us. We live in a world of connections - and it matters which ones get made and unmade.*"

— Donna Haraway:
Simians, Cyborgs, and Women: The Reinvention of Nature (1985)

Haraway points to a state of entanglement with technology as a part of being not-neutral. How can the particular kind of 'connection' she describes help us to create alternate images and ends for AI? The assumption of neutrality is of particular importance to Artificial Intelligence and machine learning techniques, which are by their probabilistic nature dependent on clean models to make accurate predictions in relation to complex systems. A clean, or neutral state entails a lack of interference to the computational system defaults, which create the optimal condition for such systems to execute their function; i.e. provide 'effective computation'[25]. This state however, assumes the neutrality, or absence of variability of the operator and of connected systems in the execution of the system in question. Our state of entanglement with such processes seems to suggest the scandalous conclusion that the proper functioning of such technologies is impossible. Where would the evidence

---

[24] Albert Camus raises the question of what makes life worth living in a world without religious meaning by way of suicide in "The Myth of Sisyphus". For him, the absence of God confronts us with an absurd situation. As with Nietzsche, the paradox arises out of our separation from the world in terms of meaning: "in this particular case and on the plane of intelligence, I can therefore say that the Absurd is not in man, nor in the world, but in their presence together."

[25] Friedrich Kittler deals with this issue in 'Real Time Analysis, Time Axis Manipulation' (1990). As its introduction outlines: "Passing through symbolic, analogue, and digital stages, media technologies appear to be moving ever closer to natural objects by moving ever farther away from their human subjects. But like any good tale of intimacy, it is also one of deception. The ability of digital media to store, process, and communicate levels of the real inaccessible to human perception comes at the cost of humans no longer being able to determine whether that which is allegedly processed by media is not in fact produced by them."

be for such scandalous malfunctioning, and does this evidence indeed point us towards entanglement? Actually, evidence abounds. In practice, the 'bugs' that prevent the smooth operation of these systems in relation to AI systems are much discussed in terms of bias[26]. Bias is usually understood not as a failing of the system, but as the pesky traces of human shortcomings, undue influences such as racism and sexism that should be suppressed to achieve full functionality.

Bias

Bias might have a commonsensical ring to it, yet its meaning is subject to extensive categorisation. The sheer multitude of extant categories suggest that suppressing bias might not such a dependable method to achieve a neutral vantage point and a transparent view of the world[27]. From general categories such as cultural bias, cognitive bias and statistical bias spawn ever-more complex and particular features of human reality. Ironically, these include many types of bias that are very applicable to the hyperbolic interpretations of AI: *apophenia*, or pattern bias; attribution bias; framing; self-serving bias; status quo bias; observer-expectancy effect; agenda-setting; sensationalism; funding bias; etc. What do all these categories of bias have in common? "Biased means one-sided, lacking a neutral viewpoint, or not having an open mind. Bias can come in many forms and is related to prejudice and intuition."[28] Extensive categorisation and wide applicability show that bias is not so much an outlier, as it is a gigantic, jettisoned 'remainder' of stuff that doesn't fit in a neutral frame. Spanning both the natural and the cultural, bias seems to refer to things that we might consider exemplary aspects of human life and history. Not so much a worrisome outlier, or a bug on our shiny new windows, but a critical part of our reality that has been - by no little effort - excluded from our modern frame in the name of neutrality and objectification. Rather than something to suppres, bias is of fundamental importance to an integrated understanding of ourselves in all our tangled connections

---

[26] Among scores of examples: https://www.theguardian.com/commentisfree/2019/jun/10/the-guardian-view-on-digital-injustice-when-computers-make-things-worse

[27] The contingent, constructed nature of classification and consequences were laid out by Michel Foucault in 'The Order of Things' (1965), with his famous introduction on our efforts "to tame the wild profusion of existing things" by referencing Borges 'Celestial Emporium of Benevolent Knowledge': "in which it was written that 'animals are divided into (a) belonging to the Emperor, (b) embalmed, (c) tame, (d) sucking pigs, (e) sirens, (f) fabulous, (g) stray dogs, (h) included in the present classification, (i) frenzied, (j) innumerable, (k) drawn with a very fine camelhair brush, (l) *et cetera*, (m) having just broken the water pitchere, (n) that look from a long way off like flies'." What would a database according to these categories, and a subsequent model trained on it, look like?

[28] https://en.wikipedia.org/wiki/Bias

with the world.

It would seem to me then that 'the outside' of the modern frame is not a ghostly consciousness or an unbridgeable gap, but simply our own, non-modern 'remainder'. Outside the modern frame we find many of the very things that make us human, yet need to be labelled as 'bias' and exorcised from technologies to maintain their functionality and *our* objectivity. Jean Baudrillard illustrates this mechanism of exclusion at the heart of modern technology in 'The Perfect Crime' by the example of photography: "every photographed object is merely the trace left by the disappearance of everything else."[29] This principle of mechanised capture/exclusion is applicable to any kind of technologically rendered image, including our binary datasets and the AI models they condition. Given the endless amounts of 'snapshots' processed in parallel by machine learning, no wonder that bias is currently perceived as a core risk in these spectacular systems of automation. This risk is more likely a central point of failure, foreclosing the possibility of 'effective computation', which perhaps led Jeff Dean, head of AI at Google to recently state: "It's a bit hard to say are we're going to come up with a perfect version of unbiased algorithms."[30]

Inclusion

So what use is this excluded 'everything else' to us, looking for alternate images and ends for AI? What deeper entanglements is bias showing us? As Haraway states, "We're inside what we make and it is inside of us." In stead of relations defined in terms of representation or probability, she points to very physical connections between us and the things we make. Bias is the wobble that points us towards the myriad attachments that define us in our fleshy and tangible relations to the world, before the 'disappearance of everything else'. Taking bias as an inalienable part of our material and cultural reality, we shift from a perspective that is *exclusive* by default in an attempt to be neutral, to a perspective that is *inclusive* by default. Through bias, we're able to acknowledge that we're always-already part of the situation, to acknowledge the connections our positions and perspectives engender. This means that bias is not a flaw in systems of measurement and automation, a 'remainder' to be exorcised, but a primary means of understanding, or rather *accounting for* our place and histories in the

[29] Jean Baudrillard, *Objects in this Mirror*, in 'The Perfect Crime' (1995) pp 87

[30] quoted in: https://www.bbc.com/news/business-46999443

world, starting with our place in these artificially intelligent technological systems we talk so much about. Following bias, the dividing lines between domains such as technology and society, between nature and culture start to diffract and collapse into each other, leaving us collectively stranded *in* the situation we once thought to discover, segment and control.

As Haraway asserts: "We're living in a world of connections, and it matters which get made and unmade," pointing to the relational nature of reality. If the possibility of a modern perspective is premised on the concept of neutrality that operates by means of exclusion, then bias shows how the connections spill all over the edges of the modern frame, connecting outside and in. This opens up the possibility to employ bias in AI as a tool to account for our position 'in a world of connections'. It allows AI to be employed as an inclusionary device and an ethical technology. Its 'mirror-like' qualities let us see how we cannot subtract ourselves from the world, and are therefore always-already part of the situation. Entangled with ourselves, the window-like qualities of AI mean we can harnesses our computational prowess to help explore *how* we're entangled with each other and our environments.

Once we understand bias-as-culture, the erstwhile glitch turns into a surface, or opening that allows for a reversal of agencies: rather than automating all the wrong kinds of bias by suppressing them, we open up a space to acknowledge all of our bias and design *with* the right kinds of bias. Therefore, my basic proposition with this paper that the meta-position implied by neutrality is untenable, and this moves us from 'neutral-by-default', to 'biased-by-default'. This shift entails that questions of ethics become formative in our work with technology and AI: how we incorporate ourselves, our material and cultural entanglements into our systems becomes a more pertinent design question then how to optimise our erasure.

If accounting for our place in the world involves a potentially infinite number of critters, phenomena and datapoints, then how to incorporate all these perspectives in our black boxes? In response to this question, Ed Finn proposes the notion of 'Culture Machines'[31], with which he structures this relationship between messy and material, biased-by-default cultures and algorithmic systems of computation. Finn proposes to

[31] This concept is proposed by Ed Finn in his 'What Algorithms Want' (2017), and will serve to further inform research into the relation between complex, emergent (natural/cultural) phenomena and computational systems.

understand material, implemented algorithmic systems as culture machines, operating the gap between culture and computation:

"The mythos of computation reaches its limits where it begins to interact with material reality... Computation in real-world environments is messy and contingent, requiring constant modification and supervision.....I call this problem 'implementation': the ways in which desire for effective computability gets translated into working systems of actual computers, humans and social structures... By learning to interpret the container, the inputs and outputs, the seams of implementation, we can begin to develop a way of reading algorithms as culture machines that operate the gap between code and culture."[32]

This shift of perspective - from perfect windows allowing us to look 'on' the world, to being 'in' the world together with these machines, points our attention to 'the seams of implementation', and call for capacities to 'interpret the container'. We've lost access to a meta-position, and necessarily need to describe our position in relation to, and therefore from within any phenomenon under consideration. Donna Haraway, thirty years after her Cyborg Manifesto, again helps us to make sense of our dis/utopian situation, advising us to 'stay with the trouble': "In urgent times, many of us are tempted to address trouble in terms of making an imagined future safe, of stopping something from happening that looms in the future, of clearing away the present and the past in order to make futures for coming generations. Staying with the trouble does not require such relationship to times called the future. In fact, staying with the trouble requires learning to be truly present, not as a vanishing pivot between awful or edenic pasts and apocalyptic or salvific futures, but as mortal critters entwined in myriad unfinished configurations of places, times, matters, meanings."[33]

This 'clearing away the present and the past' indeed summarises the impossibility of 'neutralising' the bias that occurs in Artificial Intelligence, and the inadequacy of current, hyperbolic discourse promising grand futures for AI. By including that which was until recently 'outside of the frame' of modernity, by paying attention to connections at 'the seams of implementation', we are able to start developing inclusive,

---

[32] Ed Finn: 'What Algorithms Want' (2017) pp 47

[33] Haraway, Donna: 'Staying With the Trouble, Making Kin in the Chthulucene' (2016) pp 1

ethical AI systems as technologies of accountability and connection. What is staying with the trouble other than tracing the seams, grappling with our biases and proclivities, with our 'unfinished configurations of places, times, matters, meanings'?

As the hyperboles of present-day AI showed us, the ways in which we develop and use technology are at their core questions of ethics; of which connections are made and unmade. Indeed, as Mark Zuckerberg has exhaustively proven, not all connectivity is good. As Gregory Bateson put it in the Ecology of Mind: "There is an ecology of bad ideas, just as there is an ecology of weeds."[34] Rather than reaffirming the smooth and objectively accessible reality that came so naturally to our modern sensibilities, AI as 'culture machines' provides new pathways into the cultural and material entanglements at heart of our technologies. With it, bias becomes a constitutive part of our science and technology; acknowledging that bias is, in many senses of the word, culture. What were once understood to be dysfunctional growing pains of AI, turn out to be a core feature of algorithmic systems. This feature allows us to embrace our critical role as creators and operators, acknowledging the ethics of our actions in what gets inside the frame and what is left 'out of the cut'. Rather than shying away from a ghost in the machine, perhaps AI can allow us to chart new and complex pictures of our endless entanglements, and enable us to better account for our position and choices in this world of connections.

[34] Félix Guattari starts his Three Ecologies with this quote from Gregory Bateson in the Ecology of Mind (1982)

Bibliography

Agamben, Giorgio: 'What is Real?' (2018)

Agamben, Giorgio: 'The Adventure' (2018)

Aranda, Julieta; Wood, Brian Kuan; Vidokle, Anton: 'The Internet Does Not Exist' (2016)

Barad, Karin: 'Meeting the Universe Halfway' (2007)

Baudrillard, Jean - 'The Perfect Crime' (1995)

Broussard, Meredith: 'Artificial Unintelligence' (2018)

Camus, Albert: 'The Myth of Sisyphus' (1967)

Debord, Guy - 'Society of the Spectacle' (1967)

Finn, Ed: 'What Algorithms Want' (2017)

Foucault, Michel: 'The Order of Things' (1966)

Guattari, Félix: 'The Three Ecologies' (1989)

Haraway, Donna: 'Simians, Cyborgs, and Women: The Reinvention of Nature' (1985)

Haraway, Donna: 'Staying With the Trouble, Making Kin in the Chthulucene' (2016)

Heidegger, Martin - 'The Question Concerning Technology' (1954)

Kittler, Friedrich: 'Real Time Analysis, Time Axis Manipulation' (1990)

Latour, Bruno: 'Facing Gaia' (2018)

Manuel, George: 'Fourth World: An Indian Reality' (1974)

Nietzsche, Friedrich: 'The Will to Power' (1887)

Images

Cover image: gynandromorphic Morpho butterflies, by Nipam Patel (http://www.patellab.net/)

Image 1: via google search 'AI' - https://code.fb.com/ml-applications/facebook-to-open-source-ai-hardware-design/

Image 2: via https://en.wikipedia.org/wiki/Domain_of_a_function, by Damien Karras

Image 3: left, film still from Terminator - right, film still from Ex Machina

Image 4: left, via https://www.theverge.com/2019/5/6/18530917/game-of-thrones-got-season-8-hbo-final-last-of-the-starks-starbucks-coffee-cup-blooper - right, Adam Willaerts (1577-1669), Dutch School. Fishing Scene. Private Collection.