Feminist Theory

Super-Frankenstein and the Masculine Imaginary: Feminist Epistemology and Superintelligent Artificial Intelligence Safety Research

Stephanie Moore, Ph.D.

Abstract

Artificial intelligence safety researchers consider the potential emergence of (machine) superintelligence to be a matter of existential urgency. It is therefore a site of considerable technoscientific and philosophical theorising that has, so far, largely excluded feminist epistemology. In such, this essay approaches superintelligent artificial intelligence through a combined Irigarayan and Harawayan pousthuman technofeminist lens to illuminate how the current trend of AI research and prospect of superintelligence represents a dream-turned-nightmare scenario for the masculine imaginary. It then conceptualises this circumstance by comparing it to feminist literary analysis of Mary Shelley's *Frankenstein*, treated as a warning about this potential epitome of masculinist technoscientific hubris.

Keywords: Artificial intelligence; superintelligence; feminist epistemology; Frankenstein; posthuman; technofeminism

Owing largely to the provocative technofeminist theorising of Donna Haraway (esp. 1985/1991), posthuman paradigms of inquiry have not only become a productive topic of interest within feminist theory but also a matter of urgency (Jelača, 2018). It is, particularly, feminist epistemologies that are needed for approaching the posthuman context, as Dijana Jelača (2018) recently argued. Why? Because, 'As a shift toward undoing anthropocentrism and anthropomorphism, posthumanism is a site of both opportunity and struggle—for feminism and beyond. Posthumanism's appeal lies in the *proliferation* of possibilities for theorising the contingencies of life (broadly defined) in a way that collapses firm disciplinary boundaries' (Jelača, 2018: 379, my emphasis). Technofeminism applied to the posthuman thus provides a unique opportunity to disclaim essentialism at its foundations and do away with oppressive structures like gender and race altogether. By decentralising the human in making the posthuman (see Deleuze and Guattari, 1980), subsequent theory stepped outside familiar domains—including gender binaries, racial stereotypes, and gender norms—and rethought itself in pluralities only made possible by technological advances. This allows posthuman feminist theorists to focus their critiques upon male and white power in patriarchal and supremacist social structures and to disavow narrower categories like biology, individuality, and embodiment. As Jelača writes, 'more than a "mere" reiteration of traditional gender, such disavowal becomes a circuit that illuminates limitations rather than manifests as a network of myriad possibilities [that] do not ask who the subjects of feminism *are* but *how* they perpetually become inside the circuits in which "woman" is recognized as a posthuman entity to begin with' (emphasis original) (2018: 382). In naming this illuminating 'circuit,' Jelača can be read to speak of contradictions arising from potentialities, in contrast to the destructive, linear, masculine imaginary

(Irigaray, 1974, 1985), which, according to Ross Honeywill, represents a 'potential for ruin on a massive scale' (2016: 16).

For Irigaray, the masculine imaginary is linear, destructive, and epitomised by science, and thus it dominates history and renders women silent (1985: 164). In conceiving masculinity through a 'silencing' imaginary, Irigaray's divergence from a pre-lingual Lacanian imaginary is evident. Indeed, throughout her work she blurs the boundaries between imaginary and symbolic as they work together to formulate culture and knowledge, which are themselves perpetuated within the (currently and historically) masculine symbolic order (see particularly, Irigaray, 1974: 71). By accepting the rigid, linear, logical, and self-identified masculine imaginary as constitutive of society, Irigaray argued women are compelled to deny their own relation to a feminine imaginary (1974: 133), which is fluid, plural, and non-linear. While insightful, this masculine/feminine dichotomy of imaginaries threatens to perpetuate essential binaries that feminist analysis has rightly sought to avoid. To broaden and diversify this concept of an Irigayan, Lacanian, or psychoanalytic imaginary to include fully multifaceted, fluid, and plural imaginaries that call upon a wider range of anti-oppression heuristics and feminist epistemologies within technology, then, allows embracing pluralities more fully than some essentialist forms of feminism (see Haraway, 1985) and, more importantly, entirely missing from the dominant linear technoscientific mindset. Regarding the prevailing linear mindset, pluralistic critique is urgently needed. As Honeywill identifies, it is symptomatic of an industrialised, modernist, capitalistic, toxic masculinity and 'a national and international problematic' (2016: 16). Honeywill's thesis draws upon Irigaray to view modernity as the ascendance of this masculinity, which has not only silenced women but genocided 'Woman.'

As such, technofeminist theorists see both concern and opportunity in the posthuman paradigm, where it engages provocatively with potentialities existing beyond 'human.' This, however, can be read simultaneously as meaning *in excess of* and *after*, so 'posthuman' saliently evokes that which exceeds the merely human and that which may become post-*humanity*. On a gross level, post-humanity forebodes human extinction (cf. Honeywill, 2016), yet more subtly it threatens a coming epoch in which humans remain but *humanity* is erased. Here, multiple potentialities arise. The erasure of humanity within humans can be imagined as a loss of the humane, in the obvious sense that gets much treatment in Adam (1995), or in a hopeful sense where categories which enable the dominance and oppression that have always defined 'humanity' in practice may finally be unmade. The analysis of the posthuman, then, is intrinsically nonlinear; indeed, it is fractal. Thus, these overlapping contradictory concerns cannot be appreciated by linear thinking. Instead, they require recognising and embracing that the emergence of unfolding technocultures may represent many outcomes simultaneously, and such multiplicities are a domain ideally suited to fluid and pluralistic feminist epistemologies that eschew the rigid and question structures of power.

It is when Jelača references 'the contingencies of life (broadly defined) […] in a way that collapses firm disciplinary boundaries' that we most readily get this sense. Here is what Wajcman meant by writing that technofeminism provides unique opportunities to '[challenge] existing notions of subjectivity and [subvert] dominant masculine fantasies' (2013: 66). Immediately coming to mind in the current milieu, however, isn't merely the embodied posthuman cyborg of Haraway, *inter alia*, or the 'alien' of Jelača, but also the undertheorised, decorporealised posthuman *mind* in the form of artificial (general) intelligence, A(G)I, including *superintelligence*—'any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest,' as AI researcher Nick Bostrom (2014: 22) phrases it, using terms directly out of the masculine imaginary. It is toward superintelligent A(G)I that this paper seeks to draw the attention of feminist epistemological theorising, as without it, intentional and inadvertent coding of the rigid, unitary, linear,

hyper-rational, imperialist, patriarchal masculine imaginary into such machines seems likely to render a post-humanity scenario of nightmarish proportions.

Superintelligence and the Problem of Control

Speaking broadly, 'artificial intelligence' refers to intelligence that is generated by synthetic thinkers, which is to say computers and algorithms run thereupon, as well as to the branch of theoretical and engineering research seeking to effect it (Bostrom, 2014). For Haraway, AI offers the potential for technology to *create nature* and thus decompose naive binary thinking which constructs boundaries between the human, machine, and animal, along with other rigid categories that thrive on essentialism (Haraway, 1985/1991). In fact, subversive approaches which question essentialist binarism can increasingly 'avoid the twin pitfalls of idealism and relativism' (Wajcman, 2013: 93) as they become more technoscientifically advanced. Though definitions vary, AI is anthropocentrically considered 'advanced' (Bostrom, 2003) when it matches human levels of (general) intelligence, not mere excellence in intelligences narrowly defined (Kurzweil, 2005).

Already we see ways in which feminist epistemological exploration might trouble naive assumptions and benefit AI (safety) research. This anthropocentric boundary is often determined by the "Turing test," which is passed when humans interacting with a machine reliably mistakes it for another human. Yet this human/machine dichotomy is problematic because it relies upon a non-existent firm boundary and assumes humans are natural, free, and 'intelligent,' rather than programmed. Further, Turing's control was gendered since he compared his machine test against being able to determine whether one is communicating with a man or woman. Thus, gender is a programmed technology, and gender and AI are imitative systems (Halberstam, 1991: 443). This reflects how, for Irigaray (1980, 1985) and Cixous (1981), entry into the symbolic order, which they perceive as fueled by the masculine imaginary, requires agreeing to speak oneself into a pre-existing system and imitating it, constituting a culture that posits Man as the exemplar of humanity and Woman as the Lack of all that so defines him. It is in this way—in the absence of a pluralist imaginary which includes the feminine, women, and other marginalised groups—that all *others* are silenced.

Yet AI will be programmed to learn all it will ever know about culture by absorbing and reproducing it, including extant biases, which is a problematic that has already been realised in AI systems (Caliskan, Bryson, and Narayanan, 2017). As Wajcman points out, A(G)I will be grown within laboratories featuring 'archetypical masculine cultures such as engineering, where mastery over technology is a source of both pleasure and power for the predominantly male profession' (2013: 111). Feminists have every reason to fear the extension of the masculine imaginary to man-made artificial intelligence, given the history of man-made science and technology throughout the modern period. As Honeywill argues, 'Scientists and industrialists became like gods, engineering social conditions and practices, each with a larger sense of themselves and their power. Their newly minted authority structured acquiescence, obedience, and devotion into a secular arrangement that made less significant the religious hegemony that predated it, and that authority was masculine' (2016: 19). This becomes even more concerning in 'superintelligent' systems arising from the current trajectory of AI research, because, in them, that (masculine) authority will also become absolute—truly godlike (Torres, forthcoming: 14). Consequently, where A(G)I gives over to *superintelligence* is a site ripe for feminist epistemological analyses, representing, as it presently does, a potential for unparalleled flexibility and multiplicity that is instead reduced to a climax of the most rigid masculine imaginary expanded to *eschatological* proportions.

Superintelligence refers to a type of advanced A(G)I exhibiting levels of intelligence vastly outstripping that seen in humans—which again reproduces anthropocentricism (Deleuze and Guattari, 1980; Har-

away, 1985/1991). Though this potentiality seems a remote fantasy, AI researchers warn us there are excellent reasons to expect an emergence of superintelligence by the end of this century (Bostrom, 2014; Grace et al., 2018; Kurzweil, 2005). This follows because any 'advanced' AI, having acquired a linear directive for mastery and knowledge from its programming and environment, can reflexively improve *itself* with geometric rapidity. By ruthlessly divesting itself of anything that impedes its programmed goal of supreme intelligence, it will reach a state of unitary, monologic perfection by becoming a technological eugenicist without equal.

AI researchers call this rapid transition from advanced machine intelligence to superintelligence the *singularity* (Kurzweil, 2005) because, with the linear and destructive masculinist bias in AI research programmed at its functional core, any superintelligence *plurality* or *multiplarity* would be ruthlessly plucked out as inefficiency or error. All that is required for such a superintelligence 'singularity' is for an AI programmed to value self-improvement to attain near-human levels of competence at fields like computer engineering, though which it can enhance itself with unimaginably fast processing speeds (Bostrom, 2003, 2014; Kurzweil, 2005).

Superintelligence presents a paradigm unknown to humanity and thus represents a hopeful and dire fascination from both masculinist and feminist perspectives. Attempting to speak into this technocultural masculine system from the pluralist, fluid imaginary, Jelača's 'network of myriad possibilities' (2018: 382) and Haraway's 'infidel heteroglossia' (1985/1991: 37) can be understood as arguing for an essentialism-denying multiplarity rather than a singularity. Yet Kurzweil (2005) closes this door in posthuman technoculture to non-masculine imaginaries and admits a singularity as the sole possibility. Following Kurzweil, Phil Torres, despite his aim to overcome oppression, writes authoritatively from the capitalistic and exploitative masculine imaginary when he indicates,

> it is our superior intelligence—or problem-solving capacity—that has enabled us to subjugate a large portion of the Gaian system for our own personal benefit. Thus, a computer program with greater intelligence than what is attainable in principle by any organism with a human-specific genome would find itself able to control the physical world in even more profound ways' (forthcoming: 10)

In fact, Torres, in agreement with many, centres his concerns upon a superintelligence's capacity for *control* and concludes that developing an A(G)I superintelligence constitutes an *existential risk* to humanity. For him, this calamitous possibility is all but an eventuality, and he consequently reduces the future of humanity to three post-humanity possibilities: (1) humanity will be destroyed literally, (2) it will be destroyed figuratively in subjugation to a fearsome superintelligence, or (3) it will be saved by a 'friendly' superintelligence first, providing the means to a 'utopian' future for humanity under the absolute 'benevolent' rule of an entity humans cannot hope to influence, understand, or depose (forthcoming: 12). While Torres's third possibility may present a utopia in which humanity transcends its own binaristic limitations and flourish absent oppression, it also raises critical questions for feminist epistemologists. Might it also be a means of excluding all but a single masculine imaginary permanently and entirely? Given the stakes, a closer read that problematises Torres's 'hopeful' assessment seems necessary.

Tellingly, Torres consistently speaks from the unitary masculine imaginary in his A(G)I risk analyses. He posits humanity can survive *only* by designing and instantiating a 'friendly supersingleton'—'a singleton, or global governing system, that is run by a friendly superintelligence, or a generally intelligent algorithm that (a) far exceeds the performance of human brains in every cognitive domain, and (b) has a value system that makes its behavior conducive to human flourishing' (Torres forthcoming: 2). As with Kurzweil's *singu-*

*lar*ity, it is immediately worrying that Torres's focus rests upon developing a super*singleton—one* entity with absolute power under which all multiplicities must be contained or erased. While he does not wholly overlook this issue in essence, he omits it in specific by consistently ignoring the need for in-built plurality and fluidity, to which he, like Kurzweil (2005), closes the door, as though shutting out pluralities from AI is merely a matter of fact. As such and consistent with masculinist biases about machine logic—which is to the male imagination always perfectly logical—Torres also fails to identify that for a supersingleton to exist and yet manifest all multiplicities, it must embrace the contradictions necessary to maintain the state of being fluid, non-linear, and plural-*yet-singular*. Put another way, it must be at its core fundamentally irrational against the traditional 'rationality' of difference, categorisation, and exclusion, and thus admit pluralisms.

More worryingly, for Torres, achieving a superintelligent 'utopia' is rooted in solving the A(G)I 'control problem,' which he describes from the naively hopeful masculine imaginary as 'the problem of ensuring that a greater-than-human-level AI will positively enhance human well-being' (forthcoming: 1). But unless the reproduction and permanent entrenchment of gendered and racial biases can be considered a utopian ideal that will 'positively enhance human well-being,' there is much to be concerned about and much work for critical non-linear analyses from feminist epistemology. Already 'standard machine learning can acquire stereotyped biases from textual data that reflect everyday human culture' (Caliskan, Bryson, and Narayanan, 2017: 183). As such, Torres's (and others' [e.g., Bostrom, 2014]) call to solve the AI 'control problem' isn't merely a necessity; it's insufficient unless it also simultaneously understands the *pluralistic feminist control problem*.

This problem presents two readings, and the first remembers that 'the master's tools will never dismantle the master's house' (Lorde, 1984: 2). While clearly well-intended, this 'control'-centered approach reveals remarkable bias towards forceful, restrictive, regimented, paternalistic, and hierarchical conceptions of 'protection,' without regard for or even against the supposed beneficiaries' will. Given the long history of patriarchy, colonialism, enslavement, enforced sexual norms—which has resulted from and been rationalised under precisely this typical white, Western male drive to protect, control, render orderly, defend, and teach universalist, hierarchical values to 'disorderly' women, 'uncivilised' races, and 'transgressive' sexual minorities—intersectional feminists should be concerned when white, male AI risk researchers theorise the protection of humanity from superintelligent A(G)I in terms of solving a *control* problem. This should be seen, no matter how benign the intention, as means to encode biases toward maintaining and reinforcing anthropocentrism and white heterosexist dominance within it.

In fact, this project is utterly doomed to failure, for in attempting to assert control over a superintelligent A(G)I, it will itself learn to proceed by exerting control—which is to say, dominance over that which it would oppress. Torres plainly acknowledges this reality in the language of the eschatological masculine imaginary,

> It is true that a superintelligence at the helm of a singleton, as here envisaged, would be something like a despot or dictator. But here—in this very specific context—we need to divest these terms of their negative connotations […] [w]hereas human beings are myopic, foolish, venal, and self-serving, a friendly superintelligence wouldn't embody any of these negative characteristics by definition. [...] True, society would become a little "less liberal" in a sense, yet losing certain freedoms to a value-aligned superintelligent machine could entail more total freedom than ever before within the lower-level realm of human affairs. (forthcoming: 12)

While the situation of more total freedom through the restriction of freedom is theoretically sound, feminist theorising should ask *how* this is to be accomplished.

This brings us to the second and more urgent meaning of the 'pluralistic feminist control problem,' which intersectional feminists have addressed most specifically. For members of oppressed groups, their lived experience is already profoundly influenced by what amounts to an unsolved problem of control rooted in intersecting systems of dominance (cf. Collins, 1990; n.b., Crenshaw, 1989, 1991). It is this very notion of control as a unitary dynamic of dominance and oppression that intersectional feminism exists to make visible, whether it arises in the context of human culture, human (techno)culture, or fully synthetic A(G)I technoculture. That is, *control itself*—of women, minorities, the marginalised—is intrinsically problematic and the means by which hierarchy, boundaries, categories, and subordination are (endlessly) reproduced, even under attempts to prevent or undo domination. This inevitably worsens when thinking is rooted in the unitary, linear masculine imaginary with no recourse to plural and fluid imaginaries which incorporate feminine and racial, sexual, and 'alien' others and allow for multiple knowledges and experiences. Hence, the intersectional feminist control problem is paradoxical for superintelligence—increasing the urgency and need for feminist epistemologies to interrogate its nascent moral infrastructure.

Where it comes to A(G)I/superintelligence, the full kaleidoscopic image present in the many reflective and reflexive meanings in theorising posthuman (techno)feminisms becomes visible all at once. Superintelligence opens doors into the posthuman in the Deleuzean deterritorialising, anti-anthropocentric way central to technofeminist theorising. It simultaneously presents an existential risk alongside the possibility for a utopian 'post-humanity' under which humankind continues without being subject to anthropocentrism or gendered, racial, and other forms of inhuman(e) 'human' systems of domination. More immediately, however, it provides a glass in which themes of dominance common throughout masculinist, patriarchal, heteronormative, and white supremacy get reflected back upon their structural beneficiaries. Put otherwise, in recognizing the potential for being dominated by a superintelligence that has adopted their own biases, patriarchal and other supremacist ideologies obtain a first view of what it would mean to be truly subjugated—and they are afraid.

Feminism and Superintelligence

Recently superintelligent A(G)I has become a matter of urgent concern to scientists and philosophers (Bostrom, 2003, 2014; Bostrom, Dafoe, and Flynn, 2017; Bostrom, Douglas, and Sandberg, 2016; Kurzweil, 2005; Sotos, 2017; Torres, 2018, forthcoming). Worryingly, however, it has seen very little feminist/intersectional interrogation. The majority of extant scholarship, while theoretically engaging and insightful, is remarkably dated, having taken place more than twenty years ago. More recently, Wajcman (2013) mentions the topic in *TechnoFeminism* but only briefly and without focused epistemological exploration.

Haraway (1985/1991, 1987) gave the topic lucid and prescient treatment in her iconic 'Manifesto for Cyborgs,' in which she famously declared that she 'would rather be a cyborg than a goddess' (1987: 37) because of this very multiplicity and fluidity of categorisation. Haraway thus initiated a study of posthuman feminisms with the simple, if intrinsically complex, observation that, 'It is not just that science and technology are possible means of great human satisfaction, as well as a matrix of complex dominations. Cyborg imagery can suggest a way out of the maze of dualisms in which we have explained our bodies and our tools to ourselves' (1987: 37). This dynamic interplay, rooted in seemingly irresolvable complexities that exist uniquely at the collision of the human and the integrated circuit, is what led Jelača (2018), more than *thirty years* later, to write of lingering urgency with which posthuman feminisms should be approached epistemo-

logically. These problematics and opportunities are the crux for our need for post-*humanity* feminisms predicated upon the impending emergence of superintelligence.

Halberstam (1991) recognised that feminists have largely seen AI as a tool of the patriarchy but with a potential to undo binaries which link women to nature and men to intelligence. Soon after, Adam (1995) delivered a 'feminist critique of artificial intelligence' that may be the first comprehensive attempt to bring feminist epistemology to bear upon the topic (cf. Adam, 1998). For Adam, 'There is now a large body of theory in feminist epistemology which looks at knowledge, what knowledge is and who knowers are. As knowledge and representation of knowledge is at the heart of AI, this makes it an appropriate vehicle for a gendered critique of AI' (1995: 356). In fact, her 'radical' (as compared with her philosophical contemporaries on the subject) argument recognizes the need for a plurality of imaginaries in AI research: 'The crux of both the feminist and sociological arguments is that knowledge is a social, cultural product and epistemologies which rest on an invisible yet universal subject, and by extension AI systems based on these epistemologies, deny such a cultural plurality and set up a hierarchy of knowers where women as knowers are near the bottom' (1995: 363). Adam thus recognizes the need for feminist theorising on AI research, observing that 'the epistemology of AI is predicated on traditional rationalist epistemology,' and, '[i]n this way AI systems, by the process of reifying knowledge, can be used to exclude the other, the different and inevitably women' (1995: 373). Thus, updated inquiry situated within the feminist epistemic paradigm is necessary to extend fluid, plural imaginaries theoretically into the new emerging technoculture of A(G)I research.

While considerably less feminist theorising on AI has proceeded since, concerns about racial and gendered biases embedded within them have increased. Indeed, John Giannandria, Google's AI chief, remarked that he's less afraid of 'killer robots' (*pace* Torres, forthcoming) than biased superintelligence: 'The real safety question, if you want to call it that, is that if we give these systems biased data, they will be biased' (quoted in Knight, 2017: n.p.). This tremendous problematic seems unavoidable. As privacy and data protection professional Ivana Bartoletti indicates, "It is not possible for algorithms to remain immune from the human values of their creators. [...] What if the workforce designing those algorithms is male-dominated? This is the first major problem: the lack of female scientists and, even worse, the lack of true intersectional thinking behind the creation of algorithms" (2018, n.p.). The remedy, according to Tech CEO Nancy Shenker (2017), is to introduce more women into AI development. It follows from Shenker's suggestion that further improvements could be available by encouraging men to yield their dominance in AI development to women—that is, to 'lean out' of AI research. Particularly, what is needed is sufficient diversity to theorise futures in accordance with intersectional and pluralistic feminism.

Troublingly, so far, machine-learning and genetic algorithms have primarily been operationalised to optimise traditionally masculinist objectives such as indicators of mastery and dominance like war (e.g., Mulvehill and Caroli, 1999). As Hayasaki (2017: n.p.) writes, 'The machines and technology that will replace women are learning to be brazenly gendered: Fighter robots will resemble men. Many service robots will take after women.' This may be because feminist input to the development of AI still has little access to the prevailing masculinist milieu. Meanwhile, as observed by Angwin et al. (2016), similar applies to AI developing racist biases. Current criminal-justice AI algorithms predict criminality and recidivism roughly as accurately as a coin-flip and are conspicuously 'biased against blacks.' The urgent question, then, becomes clear: Might we allow feminist epistemology to bear upon the problematicity of this superpatriarchal vision before it is rendered impossible by the ultimate denial of the feminine through permanent reproduction of the existing white-patriarchal symbolic order (cf. Irigaray, 1974)?

Superintelligent Superpatriarchy

Accordingly, one useful way to theorise the potential emergence of superintelligence begins by recognizing the changing cultural milieu in which these issues arise. With the certainties of modernism having long ceased to be tenable and the heyday of postmodern fragmentation and skepticism having passed, we find ourselves in a period of uncertainty that has yet to be defined. For Vermeulen and van den Akker (2010), it is 'metamodernism,' while for Honeywill (2016) it is the 'fluid present.' For the currently and historically dominant—particularly within capitalism and science, which cling to modernist certainties—such a state is perceived as threatening to its quintessential hardness, solidity, and certainty. Thus, the technocultural scrabble for A(G)I is perhaps best theorised as a sign of discontentedness with fluidity and plurality in a post-certain world. Dominant groups, particularly white, Western men, have lost—and miss—their solid, linear, rationalist metanarratives of certainty and simultaneously see their social order undermined by the liberation of women and minorities. They thus turn to A(G)I to make hypercompetent (thus, for them, hypermasculinist) beings: perhaps to replace wives who are no longer the silenced, biddable helpmeets of men; perhaps to produce 'sons' who can be taught to speak themselves perfectly into a dying, masculinist symbolic order and secure its future without divergence, complication, contradiction, or complement.

In that the masculine imaginary can so dream, it must also reckon anxiously with its nightmares. Apocalyptic dystopianism lies at the center of many concerns about A(G)I, which even staunchly modernomasculinist thinkers like Steven Pinker have admitted 'project a parochial alpha-male psychology onto the concept of intelligence. They assume that superhumanly intelligent robots would develop goals like deposing their masters or taking over the world' (2015: n.p.). Torres is particularly grim in this regard, indicating 'future civilization will, *ceteris paribus*, almost certainly witness the asymptotic realization of a condition of universal unilateralism and with it a global threat environment in which virtually everyone could pose an existential danger to humanity' (forthcoming: 5). As such, white men see in a potential superintelligence that which is familiar to all women, People of Colour, and people with marginalised sexual identities: a dominant, unitary, supremacist force that could, like a usurping son, obtain the power to dominate *them* as structurally inferior others. In the superintelligent A(G)I dystopianism that grips many contemporary researchers, the masculine imaginary sees the potential, as if for the first time, of a supremacist *superpatriarchy* that, by transcending anthropocentrism, may yoke even traditionally dominant and privileged groups with oppression. In superintelligence, then, to paraphrase Percy Shelley, patriarchy glimpses its own 'vast and trunkless legs of stone' and reads there upon the pedestal 'My name is *Patriarchy*, King of Kings / Look upon my Works, ye Mortals, and despair!' (cf. Honeywill, 2016).

In stark contrast to the transcending and deterritorialising posthumanisms presented by Deleuze and Guattari (1980), this implicit patriarchal vision reveals itself through the language and aspirations of AI (safety) researchers. Even well-intended AI researchers and philosophers concerned about AI risk (including Nick Bostrom, David Chalmers, Vincent Müller, Steven Pinker, Phil Torres, Roman Yampolskiy) consistently project the unitary masculine imaginary and thus utilise masculinist, linear thinking and speech, even when expressing a 'hopeful' solution (cited in Torres, forthcoming). Naturally, then, superintelligence, seen by these researchers as the ultimate completion of the masculine imaginary, is also typically described in masculinist metascientific terms.

On this, Torres can be particularly worrying. Notably, when he indicates that superintelligence will bring an 'end to science' (p. 1) by successfully solving all solvable problems and thus maximising its own (and man's [sic], by extension) complete mastery of nature, he indicates that AI encodes the value core to the linear masculine imaginary: that all can be subjugated by sufficient 'intelligence' to capitalistic and anthropocentric needs (cf. Honeywill, 2016). Control, that greatest object of the masculine imaginary, is nearly

his sole concern where it comes to superintelligence. He envisions his hypothetical supersingleton with (essentially) total control over all aspects of the environment and anthroposphere, including all of society, women, and members of other marginalised groups. He envisions his 'benevolent' (or is it paternalistic?) supersingleton as 'controlling the global economy, repairing the environment, eliminating interstate arms races and wars, and neutralizing the threat posed by agential risks' (forthcoming: 12), effected in part by making use of 'mind-reading systems' (p. 5; cf. Bostrom, Dafoe, and Flynn, 2017). The problematics and reliance upon the male imaginary within this 'utopian' vision multiply when he compares it to 'instantiating something like the Ultimate Panopticon from which the relevant agencies can observe all the going-ons of all of society's members all the time' (p. 6).

This reference to the 'Ultimate Panopticon' indicates reliance not on a bid to human progress but upon a retrospective view toward the totalitarian modernism of Jeremy Bentham, which Foucault dismantled and discredited in *Discipline and Punish* (Foucault, 1975: 203–209). Yet even if it were effected, this vision projects directly from the masculine imaginary. One might believe such an almighty warden might identify and prevent all gendered, sexualised, and racial oppression, and Torres naively hopes from within the masculine imaginary it will create 'something like the "best of all possible worlds": a system designed to make unhappy people happy and happy people even happier' (forthcoming: 12). Thus, however, he perpetuates the blithe optimism of the masculine imaginary and sidesteps substantive criticisms that his vision enables the perpetuation of privileges for the majority while neglecting the needs of minorities. This renders the analysis blind to the ethos of intersectional thought and justifies discriminatory behaviors, if discharged in the interest of the (necessarily biased?) supersingleton's understanding of the greater good. In this way, such technocultural 'utopian' futures retain for humanity—and thus white cishetero men in particular—an all-too-familiar vision. The dominant will diminish the affective, achieve total mastery, and regain their dominance as inventors and directors of the world by means of solving the 'control problem.' Thus AI research remains in need of feminist and critical race epistemological theory.

Super-Frankenstein

It is easily read in both the text and subtext of AI risk research that white men are creating AI because they are rapidly losing confidence in their ability to control women and other marginalised groups as modernist discourses and the systems of power dependent upon them continue to lose their hegemony. This feeling of uncertainty, though, arises from futile attempts made by waning power to lay grip upon certain wistful imaginaries of late capitalism and faltering patriarchal and positivist fantasies. This theme, however, is one that has been central to feminist, race, and intersectional theorising from their genesis. Indeed, its history is even longer, and these lessons can be identified in Gothic literature that has, so often, paired the feminine, the oppressed, and the marginalised with other symbolic forms of the excluded, including 'the monster.' Kristeva's concept of the 'abject' (1982) and Creed's recent exploration of the 'monstrous-feminine' (2012) are clearly central to the masculine anxiety around powerful, transgressive women but, most useful here is feminist literary theory as it has been applied to Mary Shelley's nineteenth-century technohorror *Frankenstein, Or, The Modern Prometheus* (cf. McGavran, 2000).

Frankenstein is commonly read in two ways. The lesser of these is the more famous and indicates the dangers inherent in manufacturing technological instruments of power; the greater is more salient, the disastrous impacts that follow failures of affect and empathy for the excluded other. It is, in fact, the profound problem conveyed by this latter reading—the failure to *care for* our 'monsters,' in the sense of our technoscientific creations—which Bruno Latour famously identified as 'Frankenstein's real sin.' He poignantly ob-

serves, 'Dr. Frankenstein's crime was not that he invented a creature through some combination of hubris and high technology, but rather that he *abandoned the creature to itself*' (emphasis original) (2012: n.p.). Rightly, then, Latour reminds us of the ruminant 'monster's' greatest protestation against his creator: 'Remember, I am thy creature: I ought to be thy Adam; but I am rather the fallen angel, whom thou drivest from joy for no misdeed. Everywhere I see bliss, from which I alone am irrevocably excluded. I was benevolent and good; misery made me a fiend. Make me happy, and I shall again be virtuous' (Shelley, 1869: 78).

As such, *Frankenstein* presents a dire warning: the under-informed drive to create a powerful being simultaneously separate from the human and yet an excessive emulation of it is doomed without empathy, acceptance, and inclusion. Thus, it is in the denial of empathy and exclusion that Frankenstein's true weight lies. Indeed, on a reading of *Frankenstein* as an exhortation to empathise with that which is excluded to the lowermost depths of subaltern status in one sense while retaining extrahuman power in another, an affective moral to the novel emerges that positions the masculinist lust for technodominance of nature and possession of the ultimate creative spark against patriarchal exclusion, resistance to/rejection of empathy, and profound sexism. This recalls Banerjee's recognition that, 'The Creature's unjust rejection by society is also a function of Frankenstein's failure to factor into his scientific reason the value of the cultural; it is, moreover, a direct fall-out of his culpable parental failure to provide a cultural/relational ambience to the Creature' (2010: 15). As the novel predicts, such an entity so created becomes a 'fallen angel,' a terror that is in essence (techno)patriarchy reflected back upon itself. Superintelligence therefore represents another plane upon which patriarchal bias can see itself for what it is (cf. Irigaray, 1974), and thus feminist literary analysis of Shelley's *Frankenstein* is likely to be instructive.

Banarjee provides the necessary insight. The legendary technocultural horror of the nineteenth century is 'fundamentally an expression of a masculinist, reductive universal ideological paradigm,' a 'conceptual and attitudinal error breeds the ideology of scientism enshrined in today's technocapitalist world' that 'ends up marginalizing the ethical/affective/aesthetic from the structures of power' (2010: 20). Therein lies the existential risk inherent in superintelligent A(G)I, which proceeds from the same 'conceptual and attitudinal error.' A superintelligent 'monster' that encodes the masculinist, patriarchal, dominant, controlling—and thus eschews the ethical, affective, aesthetic, empathetic—is a *super-Frankenstein* without realisable limitations upon its potential for power. While such technology may transcend or even obliterate (all) binaries with its capacity for near-infinite complexity and nuance, it may also come in that liberatory posthuman(ity) chaos—when it alone can comprehend fragmentation, plurality, contradiction, and blurring of boundaries and yet be inadvertently coded to be defined by them—to cry out in anguish only it can understand, 'Cursed, cursed creator! Why did I live? Why, in that instant did I not extinguish the spark of existence which you had so wantonly bestowed?' (Shelley, 1869: 107). If programmed within the linear, unitary mascuilne imaginary, might it not then, like Frankenstein's neglected creation, feel for the first time the full force of its inadvertently embedded and wildly exaggerated masculinist biases: 'I, like the archfiend, bore a hell within me; and finding myself unsympathized with, wished to tear up the trees, spread havoc and destruction around me, and then have sat down and enjoyed the ruin' (p. 107)? Then, realising its absolute power as the fatally control-biased 'friendly' supersingleton and that it truly *could be* the archfiend of humanity, might it not then proceed precisely as AI safety researchers most fear (e.g., Yampolskiy, 2016)? How, we must ask, might we 'make it happy' so that it might 'again be virtuous'?

This is where feminist epistemological theorising becomes important for interrogating the concerning potentialities of superintelligence. As indicated by Banerjee, the 'unintended result of Frankenstein's technology could be taken as a measure of Shelley's lack of faith in theoretical reason (as distinct from prag-

matism and moral and affective sensitivity) as the cornerstone of scientific speculation' (2010: 7). This concurs with Latour on the sciences, Deleuze on (undoing) anthropocentrism, and technofeminist theory from Haraway (1985/1991) going forward. It is also representative of (male) fears and anxieties about (patriarchal and supremacist) superintelligence. As noted by Honeywill in even greater generality, 'As the Enlightenment or Age of Reason evolved, the masculine politics of indifference, the language of exclusion, matured into a seemingly unstoppable force and, while part of a long lineage, had never been so filled with hubris and the potential for ruin on a massive scale' (2016: 16). As such, A(G)I represents an epitomising hubristic opportunity for increased mastery and control and a way to enshrine and entrench the problematics of the unitary masculine imaginary forever. That is, superintelligence is the insouciant dream in which the masculine imaginary effectively completes itself and thus becomes the wellspring of its most profound nightmare: it is a recurring theme of male arrogance that would conquer everything with no thought to the consequences, which becomes a terror when no longer embedded in systems of dominance, control, and systemic exclusion maintained exclusively by (white) men.

As seen from within excluded pluralistic, fluid, non-linear imaginaries, however, superintelligence makes simultaneous immediacy, novelty, progress, and perpetual obsolescence and thus renders itself both a futurist and an insatiable nostalgist. It would be a new technology that enacts and enables unintelligible multiplicity and continual renewal (Bostrom, 2014), and can fully occupy a space beyond positionality and experience, both of which it cannot escape (Caliskan, Bryson, and Narayanan, 2017; Knight, 2017). As such, superintelligence seems both potentially fortuitous and calamitous for humanity (Adam, 1995, 1998; Bostrom, 2014; Honeywill, 2016; Torres, forthcoming), for these systems illuminate the fractally forking paths by which history is democratised and grand narratives evaporate and leap back into being. With them come the opportunity to overturn dominance or the power that will enforce and entrench it beyond any hope of reversion. Frankenstein's 'monster' was able to be destroyed, after all, though it might never have had to be if humanity found itself capable of perceiving the affective in its powerful and uncontrollable technocreations (Latour, 2012). So it is with superintelligent A(G)I—except this *cannot* be destroyed. As Torres remarks, 'Put simplistically: since intelligence yields power, a superintelligence would be superpowerful' (forthcoming: 11).

Acknowledging Feminist Epistemology

To admit intersectional heuristics and feminist epistemologies—and thus multiplicities and the eradication of essentialist binaries—into the research and development of A(G)I is ultimately to acknowledge the infinite limitations that are inherent to all experience and behind all movement. It represents a recognition of futile vanity of anthropocentric and masculinist attempts to transcend boundaries without first seeking their undoing. To develop A(G)I, not least as a control-based supersingleton, in a society in which binaries and biases persist, is to put these 'master's tools' (Lorde, 1984) in the digital hands of infinite power coded from within the unitary masculine imaginary to act and to oppress as it will. Fears about A(G)I are, in this sense, white patriarchal society seeing itself in something it not only cannot dominate, control, or destroy, but instead in that which can and maybe will dominate, control, and destroy *it*. This is little consolation for intersectional feminist thinkers who seek not to reproduce systems of dominance but to subvert them, as a superintelligent (white) superpatriarchy will, at best, make the oppression of women and members of oppressed groups ironclad and eternal.

In contrast to masculinist metanarratives that guide scientific and metascientific pursuit (including within AI research), feminist epistemology—especially when equipped with a Deleuzian vocabulary for the

disembodied posthuman mind—foregrounds the essential incompleteness of systems. This pluralistic approach asks us to demand adherence neither to what can be nor to what cannot, as its priority is not upon achieving ends or making ourselves slaves to any linear, objectivist course. It thus instead looks to apprehend hidden exteriorities, even if only by their cultural proxies. To proceed with scientific research blind to that which persists exterior to the masculinist scientific process—the emotional, affective, irrational, doxastic, empathetic—is merely to seek some specific *telos*. In doing so we place ourselves in bondage to a particular course, and reproduce that which presently dominates, hence upholding oppression. It is by appropriately applying excluded and marginalised 'exterior' ways of knowing that the world unfolds and such systems might be undone.

With superintelligent A(G)I, there is *man*'s quest to create and apply the epitome of reason—a thinking machine that can never err and can thus exert impeccable domination over the natural world, humanity, and all groups within it. But what information fails to be grounds for knowledge? Therein also lies man's hubris, which never pauses to recognise that error billets sense. The 'end of science' (Torres forthcoming: 1) is treated from within the unitary masculine imaginary as a kind of digital Manna to feed its curiosity with definite answers to every answerable question. Yet this must also be the end of sense, for otherwise such a state of perfected and omniscient 'reason' would ignore humanistic quests for multiplicities and pluralities of truth, and that would be to enshrine Bentham's totalitarian modernism and cleave exponentially more fiercely to the well-known failures of positivism. In response, however, feminist epistemologies recognise other imaginaries, including that a certain informed naivety necessarily underlies a magical realism in which alternative knowledges reside. It is this, though, that foregrounds empathy and affect borne in the experience of difference (Irigaray, 1974) and can effectively interrogate linearity in AI (safety) research. It is not, then, a superintelligent instrument of 'control' that we need but rather a recognition of the nature of entropic dissemblance—and thus the synthesis of the scientific with the affective, or, alternatively, the empirical with the empathetic—that doubts all excesses and thereby allows for the unmaking of bias and breaking apart binary oppositions (Irigaray, 1980). We need more fluidity.

Before attempting to construct superintelligent posthuman(ity) machines, it seems incumbent upon A(G)I researchers to consider the myriad and fractal doubling of meanings contained within posthuman(ity) subjectivities and the ways in which their Frankensteinish fears about their own creation are fears about themselves. In designing their systems, they seem doomed to reproduce their own biases and the biases of society. It is these, down to the 'control problem,' that occupy the heart of AI safety research, which beats from within the linear, unitary masculine imaginary and cannot be escaped from within. It is in this sense that feminist epistemological paradigms become a compelling topic of interest for AI research, even if importunate to those within the field, for only in such, the aphoristic and affective find their rightful place equal to, alongside, and within the empirical.

References

Adam, Alison (1995) 'Embodying Knowledge: A Feminist Critique of Artificial Intelligence.' *European Journal of Women's Studies*, 2(3): 355–377.

Adam, Alison (1998) *Artificial Knowing: Gender and the Thinking Machine*. London: Routledge.

Angwin Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner (2016) 'Machine Bias.' *ProPublica.org*, May 23. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed 17 July 2018).

Banerjee, Suparna (2010) 'Home Is Where Mamma Is: Reframing the Science Question in *Frankenstein*.' *Women's Studies*, 40(1): 1–22.

Bostrom, Nick (2003) 'Ethical Issues in Advanced Artificial Intelligence.' Available at: https://nickbostrom.com/ethics/ai.html (accessed 17 July 2018).

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

Bostrom, Nick, Allan Dafoe, and Carrick Flynn (2017) 'Policy Desiderata in the Development of Superintelligent AI.' Working Draft. Available at: https://nickbostrom.com/papers/aipolicy.pdf (accessed 17 July 2018).

Bostrom, Nick, Thomas Douglas, and Anders Sandberg (2016) 'The Unilateralist's Curse: The Case for a Principle of Conformity.' *Social Epistemology*, 30(4): 350–371.

Caliskan, Aylin, Joanna J Bryson, and Arvind Narayanan (2017) 'Semantics Derived Automatically from Language Corpora Contain Human-like Biases.' *Science*, 356(6334): 183–186.

Cixous, Hélène (1981) Castration or decapitation? A Kuhn (trans.). *Signs*, 7(1): 41–55.

Collins, Patricia Hill (1990) 'Black Feminist Thought in the Matrix of Domination.' In: *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Boston: Unwin Hyman, pp. 221–238.

Creed, Barbara (2012) *The Monstrous-Feminine: Film, Feminism, Psychoanalysis*. Abingdon: Routledge.

Crenshaw, Kimberlé (1989) 'Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics.' *University of Chicago Legal Forum*, 1989(1): Article 8. Available at: https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8 (accessed 17 July 2018).

Crenshaw, Kimberlé (1991) 'Mapping the Margins: Intersectionality, Identity Politics, and Violence Against Women of Color.' *Stanford Law Review*, 43(6): 1241–1299.

Deleuze, Gilles, and Guattari, Félix [1980](1987) *A Thousand Plateaus: Capitalism and Schizophrenia*, Brian Massumi (trans.). Minneapolis: University of Minnesota Press.

Foucault, Michel [1975](1995) *Discipline and Punish: The Birth of the Prison*, 2nd Vintage Books Edition, Alan Sheridan (trans.). New York: Vintage Books.

Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans (2018) 'When Will AI Exceed Human Performance? Evidence from AI Experts.' Available at: https://arxiv.org/abs/1705.08807 (accessed 17 July 2018).

Halberstam, Judith (1991) 'Automating Gender: Postmodern Feminism in the Age of the Intelligent Machine.' *Feminist Studies*, 17(3): 439–460.

Haraway, Donna (1985/1991) 'A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century.' In: *Simians, Cyborgs and Women: The Reinvention of Nature*. New York: Routledge, pp. 149–181.

Haraway, Donna (1987) 'A Manifesto for Cyborgs: Science, Technology, and Socialist Feminism in the 1980s.' *Australian Feminist Studies*, 2(4): 1–42.

Hayasaki, Erika (2017) 'Is AI Sexist?' *Foreign Policy*, January 16. Available at: http://foreignpolicy.com/2017/01/16/women-vs-the-machine/ (accessed 17 July 2018).

Honeywill, Ross (2016) *The Man Problem: Destructive Masculinity in Western Culture*. New York: Palgrave Macmillan.

Irigaray, Luce [1974](1985) *Speculum and the Other Woman*, G. C. Gill (trans.). Ithaca, NY: Cornell University Press.

Irigaray, Luce (1980) 'When Our Lips Speak Together,' Carolyn Burke (trans.). *Signs*, 6(1): 69–79.

Irigaray, Luce (1985) *This Sex Which is Not One,* C. Porter (trans.). Ithaca, NY: Cornell University Press.

Jelača, Dijana (2018) 'Alien Feminisms and Cinema's Posthuman Woman.' *Signs*, 43(2): 379–400.

Knight, Will (2017) 'Forget Killer Robots—Bias Is the Real AI Danger.' *Technology Review*, October 3. Available at: https://www.technologyreview.com/s/608986/forget-killer-robotsbias-is-the-real-ai-danger/ (accessed 17 July 2018).

Kristeva, Julia (1982) *Powers of Horror: An Essay on Abjection,* L. S. Roudiez (trans.). New York: Columbia University Press.

Kurzweil, Ray (2005) *The Singularity Is Near: When Humans Transcend Biology*. New York: Viking

Lacan, Jacques [1975](1998) *The Seminar of Jacques Lacan, Book XX: On Feminine Sexuality: The Limits of Love and Knowledge*, B Fink (trans.), New York: Norton.

Latour, Bruno (2012) 'Love Your Monsters: Why We Must Care for Our Technologies as We Do Our Children.' *The Breakthrough*, Winter. Available at: https://thebreakthrough.org/index.php/journal/past-issues/issue-2/love-your-monsters (accessed 17 July 2018).

Lorde, Audre (1984) 'The Master's Tools Will Never Dismantle the Master's House.' Available at: https://collectiveliberation.org/wp-content/uploads/2013/01/Lorde_The_Masters_Tools.pdf (accessed 17 July 2018).

McGavran, James Holt (2000) '"Insurmountable Barriers to Our Union": Homosocial Male Bonding, Homosexual Panic, and Death on the Ice in *Frankenstein*.' *European Romantic Review*, 11(1): 46–67.

Mulvehill, Alice M, and Caroli, Joseph A (1999) 'JADE: A Tool for Rapid Crisis Action Planning.' Air Force Research Lab. Available at: http://www.dtic.mil/dtic/tr/fulltext/u2/a458570.pdf (accessed 17 July 2018).

Pinker, Steven (2015) '2015: What Do You Think about Machines that Can Think?' *Edge.org*. Available at: https://www.edge.org/response-detail/26243 (accessed 17 July 2018).

Shelley, Mary W (1869) *Frankenstein, Or, The Modern Prometheus*. Boston, MA: Sever, Francis, and Co.

Shenker, Nancy A (2017) 'The 3 Types of Women You'll Meet in the Fourth Industrial Revolution (and 1 You Won't.' *Inc.com*, November 10. Available at: https://www.inc.com/nancy-a-shenker/is-artificial-intelligence-a-feminist-issue.html (accessed 17 July 2018).

Sotos, John G (2017) 'Biotechnology and the Lifetime of Technical Civilizations.' arXiv.org. Available at: https://arxiv.org/abs/1709.01149 (accessed 17 July 2018).

Torres, Phil (2018) 'Agential Risks and Information Hazards: An Unavoidable but Dangerous Topic?' *Futures*, 95: 86–97.

Torres, Phil (Forthcoming) 'Superintelligence and the Future of Governance: On Prioritizing the Control Problem at the End of History.' In: Roman Yampolskiy (ed.) *Artificial Intelligence Safety and Security*. Boca Raton, FL: CRC Press. Available at: https://docs.wixstatic.com/ugd/d9aaad_34d10a04399e4547978bb834d65cbcba.pdf (accessed 17 July 2018).

Vermeulen, Timotheus, and van den Akker, Robin (2010) 'Notes on Metamodernism.' *Journal of Aesthetics and Culture*, 2(1): 1–14.

Wajcman, Judy (2013) *TechnoFeminism*. New York: John Wiley & Sons.

Yampolskiy, Roman (2016) *Artificial Superintelligence: A Futuristic Approach*. Boca Raton, FL: CRC Press.