2 THE WICKED QUEEN'S SMART MIRROR

Snoweria Zhang

Once upon a time, there was a tale about artificial intelligence called "Snow White." In the story, the Wicked Queen has a smart mirror. When activated with the command "Mirror, mirror," an embedded voice assistant tells the Queen whether she is the "fairest of them all." Of course, today's technology renders engineering of such a mirror feasible with little effort. It would have a camera, connections to other smart mirrors in the kingdom, and a metric to evaluate the Wicked Queen's appearance against that of other users. This gadget might use a machine learning algorithm with training sets derived from *People* magazine's Most Beautiful list, or it could be based on a series of upvotes and downvotes. However, the apprehension with this contemporary version of the famous Grimm fairy tale as outlined is not its technical feasibility. Rather, the story reflects outdated values within a modernizing society. A truly smart mirror would tell the Wicked Queen that her obsession with triumphing using a singular beauty standard, one that prizes pale skin and youth, is misguided, reductive, and futile. "While we are on the topic though," the mirror would say, "here are seven products to brighten your skin!"

The Wicked Queen and her smart mirror is a telling analogy of the current state of our relationship with artificial

intelligence (AI): the technology has advanced to achieve astounding feats, but its value system is lingering behind. In turn, this is symptomatic of a stagnation in our own sociomoral framework. Even though social movements in the last century have introduced many nuances to complex issues such as race, gender, and power, their mirror images in AI development remain overwhelmingly simplistic, reductionist, and sometimes laughably clueless. John Palfrey (2017) argues that "the way we design decision-making processes in computers is certain to replicate our own biases." This is a trenchant observation. A critical look through the smart mirror of the technological tropes about future AI reflects human biases from the past. Embedded in the pixels of the smart mirror is a set of values that are scientifically enabled yet incongruous with the current social discourse.

The list of examples to draw from is endless. However, the argument is best illustrated through three popular narratives in contemporary AI depictions—that of the robot girlfriend, the invisible laborer, and the despotic overlord. Through these allegories emerge the reductionist ways common discourse treats the issues of gender dynamics, labor, and power structures—all incredibly nuanced and complex ideas that are undergoing revolutions of their own. Yet, when portrayals of these polemics venture into the AI realm, they reflect the precise defects in society's complexion. In other words, these are not problems with the projected technological advancement of artificial intelligence; they are mirror images of our own flawed attitudes toward humanity that we are in turn forcing onto AI. Nicky Case (2018) posits, "We know how to create tools to augment our intelligence, but can we create tools to augment our empathy? Our communities? Our sense of meaning and purpose?" We certainly can, and augmenting

these values does not have to depend upon the arrival and perfection of sentient machines. Instead, we should use the narratives around technology today to examine how we can augment our own complex and nuanced thinking. Just like the Wicked Queen and her smart mirror, technological and scientific storytelling can help us see beyond computational capabilities and delve into deeper assumptions of the tale itself. With that in mind, let us rewind the tapes and study, with a critical eye, common AI caricatures that are analogous to the problematic tropes we employ toward ourselves. In making our smart mirror, perhaps humanity can adopt it to augment the metrics we use to evaluate our own image.

The Robot Girlfriend

It is a truth universally acknowledged, that a single man in possession of engineering skills must be in want of a robot girlfriend.[1] The "Facial Recognition" episode of HBO's acclaimed *Silicon Valley* (Robespierre 2018) nods at the #MeToo movement by recounting the story of a female AI's experience with her creator. The thirty-minute comedy draws powerful analogies, in some ways, to the nuances of sexual harassment: Fiona, the robot, has neither the prior knowledge to contextualize her abnormal dynamic with her maker nor the capacity to confront him. She is also trapped, in a literal sense, inside a locked and windowless lab room. However, opposite Fiona's nuanced reactions is her creator, portrayed as a hunchbacked, bespectacled man with greasy strands of long hair who struggles to keep his mouth closed. In fact, the show itself referred to him as a "handsy, greasy, little weirdo."

This kind of portrayal is dangerous, and it extends beyond the scope of AI. It is a reductionist misrepresentation of the

#MeToo movement that being gross and antisocial are necessary and sufficient conditions for committing sexual crimes. Gropers are painted as clueless about the general decorum of social interactions when in fact they are adults making deliberate choices. At the same time, we hardly discuss similar atrocities committed by handsome and, more importantly, powerful men, and this is especially blatant in the tradition of AI storytelling. In fact, the biggest continuity problem with *Blade Runner* (1982), arguably one of the most iconic movies about artificial intelligence, is that the male lead (played by Harrison Ford) unambiguously rapes the only female character with a significant speaking role (played by Sean Young), and the incident goes by completely undiscussed thereafter. More surprisingly, we learn in the 2017 sequel, *Blade Runner 2049*, that the two lived happily ever after and even produced a legendary child. The idea that a few well-executed camera pans can resolve and transform indisputable assault into child- and plot-bearing love is ludicrous. It is not shocking that the scene was forced onto Young by surprise, as she admitted in *Dangerous Days: Making Blade Runner*, Charles de Lauzirika's documentary included in *Blade Runner: The Final Cut* ([1982] 2007). As her character Rachael submitted on screen, Young buried her tears and became a ghost of the franchise. A similarly uncomfortable dynamic emerges in the sequel between Ryan Gosling's character K and his holographic AI girlfriend. K's ability to love, which is emblematic of his being a more advanced replicant model, is passable yet still narrowly directed at a woman whose commercial existence hinges on catering to his needs. In one scene, she is seen powerlessly frozen and then exasperatedly wiped from existence. Neither of the aforementioned incidents, in two films set thirty-five years apart, is further remarked upon in the plotline. So the moral lesson

these stories seem to endorse is that mistreating a female robot at will is scandalous, unless one looks like Harrison Ford or Ryan Gosling.

Why do we care about the fate of holograms and replicants in dystopian lore? The allure of an artificially intelligent robot girlfriend, as presented by most science fiction writers, is that she cannot refuse the male protagonists' desires and advances (and yes, it is always a male lead); saying "no" is simply not in her program. In the rare case where a female robot, abused and aware, voices her concerns and seeks help, she is eventually silenced and dismantled, as in *Silicon Valley*, with her fleshy mask plopped mercilessly into an e-waste bucket. Jia Tolentino (2018) describes in her *New Yorker* exposition on "incels"—an amorphous community of involuntary celibates, one of whom is responsible for the 2018 Toronto vehicular attack—that the infamous group trains men to see "women in a way that presumes that women are not potential partners or worthy objects of possible affection but inconveniently sentient bodies that must be claimed through cold strategy." The robot girlfriends portrayed in the film and television examples above are precisely such "inconveniently sentient bodies." General intelligence is bestowed upon them for a narrow purpose (usually labor or entertainment, as in *Ex Machina*), but the same intelligence is feared, fought against, and stripped away the moment it acquires its own will and personhood. In this regard, these science "fictions" are in reality a grim yet accurate mirror to the facts of society.

The trope of the robot girlfriend and, more importantly, her sexual predator proliferates. Comedian Dave Chappelle, in his Netflix special *The Bird Revelation*, jokes that if Brad Pitt did what Harvey Weinstein had done, the public's reactions would have been different; women would have acquiesced! This

obvious fallacy is not limited to incendiary comics; it bleeds into our daily lives. For instance, a recent MIT-wide sexual misconduct prevention training[2] subtly harbors very similar ideas. In the sixty-minute session developed by EVERFI, stock photos with animated voices focus on how you, the viewer and a Responsible Employee, should react to and report incidents if subordinates indicate that they have been sexually harassed. In one section, the training proclaims that perpetrators can be "friends, spouses, successful, and respected," but nowhere does the hour-long training mention that *you*, the viewer and a Responsible Employee, could be the perpetrator.

This is a problematic and reductionist microcosm of how we are taught to see this issue. Other people can be bad; we ourselves cannot. In fact, one has to be *otherized* as such to become a sexual predator, and not ever shampooing again seems to be an initiation requirement. Kevin Slavin's (2016) adage, "You're not stuck in traffic you are traffic," is alarmingly apt here. The most dangerous yet pervasive attitude is to think that everyone else is "traffic," and we are simply stuck in it rather than contributing to it by our action or even our complacency. This is the understated part of the #MeToo movement. Victims have had to reveal their own past and examine painful memories, but a much broader group ought to honestly confront their own behaviors that have contributed to this culture. The fictional world of AI operates such that exclusively gorgeous if not somewhat uncanny, light-skinned, female robots are victims, and only their mad but also gross creators can transgress. We know this to be untrue, and yet the narrow mode of storytelling sticks.

Grappling with the nuances and complexities of gender dynamics is difficult and requires a certain amount of comfort with the unknown. It is a common assumption in science fiction

that the world is computable and simulatable. Isaac Asimov's *Foundation* (2004) is one such prominent example, where the story and worldbuilding hinge on one man's wizardly arithmetic abilities to divine future events. Assuming this simulatability to be true (which itself invites much debate), we have to inquire which worlds the algorithms are simulating. Without intention and by default, we feed into the simulations all the flaws in our society today. Some of these flaws will function as features, but most will live on as bugs—this time in perpetuity. The allure of AI may be a "knowable" and "controllable" system, but many questions on the topic of sex and gender are ill-suited for generalization. Bennett and Jones (2018) recently published a daunting list of stories about consent in the *New York Times*. Some of them are ambiguous; all are complicated. If humans blunder when grasping consent, what will we teach the robots? We are at a crisis moment where many societal forces are assiduously trying to reconcile our collective epistemic framework with these unknowns, which were previously thought of as known and knowable only because they were taboo. This is strange waters.

The trope of the robot girlfriend is not a tale about AI; it is a reflection of a much deeper pandemic in cultural thought about expected gender dynamics. As many in the field of AI prepare for Singularity, we must also develop mechanisms to reflexively address the issues that emerge along the way. Otherwise, the next Women's March just might be led by *Her*.

The Invisible Laborer

After panning through its trademark caliginous cityscape, *Blade Runner 2049* introduces K's love interest, Joi, in its first domestic scene. Though the film is intentionally ambiguous

about her personhood at first, it is unapologetically apparent about her role in the house. Following a quaint repartee about cooking dinner and mending a shirt, she emerges as a hologram wearing a 1960s updo and a circle skirt that could only have belonged to Donna Reed in a previous incarnation. As the futuristic housewife saunters toward the camera, our suspicion is confirmed: the delectable dish she puts on the table is a hologram too. But we, along with Gosling, continue to play house anyway. Why?

The answer lies in how we conceive of our own labor. When we discuss artificial intelligence, the most common anxiety revolves around jobs. Much of the discourse is predicated upon the premise that some professions will survive the popularization of AI and some will disappear. Certain tasks are valued and others valueless. It is those "valueless" jobs that we want AI to do. At first glance, this seems to make sense. Of course some tasks are less desirable and unworthy of human effort. Why would we not want someone to farm, cook, drive, clean, and free us from these burdens? Once we resolve the job loss in those sectors, the paradigm is bound to evolve into a utopia.

This kind of thinking is not fundamentally flawed, but it is incomplete. It is a natural reaction to grow anxious about the future of one's job security when the most imminent prospect is automation. However, this angst can also help us reflect on the invisible structural forces shaping our own labor. Frequently, the work that is unvalued is also work done by the impoverished and disenfranchised. Most jobs we are relegating to robots are considered tasks with little to no social value. In turn, people who perform those tasks currently seldom receive recognition or status, social or economic. In dystopian depictions, there is always an enslaved class—underlings who

perform requisite tasks that no one else deems worthy. They are embodied by the hooded women in Margaret Atwood's *The Handmaid's Tale* (1985), rusty droids in Disney's *WALL-E* (2008), and female clones in "An Orison of Sonmi," the dystopian chapter in David Mitchell's *Cloud Atlas* (2004). In addition to performing undesirable labor, these groups face abject discrimination and inequality. Somehow, while we are painting tales of the future with flying cars and holographic companions, we struggle to envision a scenario where work performed by these groups is equally respected.

In a world measured in conspicuous capital flow, those who labor outside it are rendered invisible. House chores are not work. Grocery shopping is not work. In fact, these biases are so deeply ingrained in our value system that we dare not imagine a future society, accelerated with the aid of AI, functioning in any different way. Of course, this does not mean that automation should be thwarted. It is simply to say that the way we conceptualize work and the nature of it is fundamentally limited to the status quo. In this framework, it would seem, the importance is that the toilet gets cleaned, and whether the cleaner is a robot or an immigrant is merely a difference in cost. Conveniently, AI allows us to perpetuate this mindset and ignore how societal structures need to change, adapt, and evolve.

The asynchronicity between cultural progress and technological advancement is not unique to AI; similar mismatches have accompanied many prior leaps in automation. Writer and activist Betty Friedan writes in her 1963 book, *The Feminine Mystique*, that the technologies that ostensibly made household chores easier did not in fact liberate women from these tedious tasks as anticipated. Instead, more work and expectations emerged. Consequently, women were even busier than

before and the prospects of equality were kicked further afield. Where AI is concerned, there are always going to be unpredictable contingencies. However, few of these contingencies will lead to the apocalyptic dystopian future that filmic imaginations like to portray. Like Friedan's example of the relationship between automation and social progress, there exist smaller but more insidious grains of anxiety worthy of examination. A concrete instance involves the precipitous rise of self-driving vehicles. Many predictive charts tout the cost-saving effects of eliminating the operators of public transit that will occur when autonomous cars enter into the mainstream. However, when asked about such cost reductions on a panel at Harvard,[3] Seleta Reynolds, General Manager of LADOT, replied that it is a fallacy to assume that operators will become obsolete simply because the act of driving is automated. Operators, she argued, did much more than driving: they could mitigate conflict, help people with mobility issues, and serve as an arbiter for whether one can use the bus without enough change for the fare. All of these services might remain unnoticed to some but are crucial to others. At an urban scale, people who perform these invisible labors or seemingly unimportant tasks are key contributors to the liveliness of a city: bus drivers, homemakers, and fruit vendors. However, without much direct capital flow in these activities, they are either categorized as positions replaceable by AI or are not in the conversation altogether.

AI cannot just be about efficiency or convenience, and productivity as measured by capital is neither a virtue nor the norm. In most depictions, AI is a not-so-opaque simulacrum that fills the same echelons currently occupied by women, racial minorities, and immigrants. It is a borrowed narrative, stolen from our own realities.

The Despotic Overlord

Despite the two reductive yet pervasive storylines in the preceding sections, Singularitarians and their cinematic imaginations fear one kind of AI trope the most: the despotic overlord. The narrative seems to rely on technological preoccupations that mostly fall in two categories: inventing the intelligent machine itself and contemplating how to avoid our own inevitable downfall. The former decorates magazine covers while the latter haunts our collective psyche. The deep-seated assumptions that a species more intelligent and capable than *Homo sapiens* will invariably seek power and dominion is overwhelming. Out of these assumptions, doomsday thought experiments like Roko's Basilisk[4] emerge, where the debilitating fear of a despotic overlord's retroactive punishment ironically turns into a driving force in AI development. This assumption is bleak, yet it is rooted in historical precedents and corollaries. As a 2015 issue of the *Economist* titled "The Dawn of Artificial Intelligence" incisively avers, "Humans have been creating autonomous entities with superhuman capacities and unaligned interests for some time. Government bureaucracies, markets and armies: all can do things that, unaided, unorganized humans cannot. All need autonomy to function, all can take on life of their own and all can do great harm if not set up in a just manner and governed by laws and regulations." Based on this lineage of thought on autonomy, it would appear that the Singularitarians' crippling fear is justified.

In the testosterone-fueled universe of science fiction, fear of AI as a destructive and malicious force runs rampant from *The Terminator* (1984) to *The Matrix* (1999) and from blockbusters like *I, Robot* (2004) to independent films such as *Ex*

Machina (2014). In these survivalist narratives, AI is developed as assistants to human endeavors and evidence of human ingenuity. As a direct consequence of imposed servitude, the machines' inevitable malfunction combined with their superintelligence leads to the desire to harm humanity in their pursuit for power and dominance. Yet, must we assume that a hyperintelligent and sentient species will necessarily evolve into despotic overlords? Must the relationship between our progeny and their technology be one of subjugation?

This unease can trace its provenance back to our own assumptions about power structures. Case (2018) argues that "whether it's our immediate worries about AI (machines stealing your job, self-driving cars making deadly mistakes, autonomous killer drones) or the more far-fetched concerns about AI (taking over the world and turning us all into pets and/or paperclips), it all comes from the same root fear: the fear that AI will not share our human goals and values." This lack of value sharing, coupled with power imbalance, has been a foolproof recipe for disenfranchisement for quite a large portion of our histories. The millennia-old narrative seeps into how we conceive of power structures today: we assume that power directly leads to tyranny. In the battle of human versus killing machine, only one can emerge victorious. But why must there be a battle in the first place?

Humans have, through cycles of trial and error (and *lots* of errors too), at least occasionally subscribed to the virtues of equality, collaboration, and democracy. Yet even as our societies push toward systems of equity and balance, we choose to conceive of a comparably intelligent force in the fundamentally limited mode of cutthroat competition where only one winner can thrive. If AI is meant to simulate the better quadrants of humanity, is it not more likely to replicate and

ameliorate the success of equal and democratic power structures? Today's AI is mostly and sometimes solely depicted as in a fiercely survivalist competition with its human counterpart. Even in domains with few pugilistic tendencies, AI is seen by default as an adversary rather than an ally. Case (2018) cites Gary Kasparov's 1997 match with IBM's Deep Blue as an analogy of the reductionist thinking in human-machine relationships: a zero-sum chess game. This win-or-lose framework is not only dominant when it comes to futuristic game-playing computers, it is also demonstrative of the problematic narratives in human-to-human relationships.

From historical epics to contemporary headlines, we see the lineage of one dominant theme: us versus them. Believers triumph over heretics. Invaders supplant the indigenous. New Yorkers oppose Bostonians. Based on a few of the darkest episodes in the Anthropocene, it almost appears that the only way humans can make sense of a multitude of value systems is by suppressing all but one. Rather than opening channels of freely exchanged ideals, the current is expected to flow only one way. However, there are also budding trends, especially more recent ones, that indicate a movement toward collaboration and mutual augmentation. International alliances, open source technologies, and gender equality movements like HeForShe are indications that forces previously thought of as oppositional and territorial can actually blur their own perimeters and become porous and inviting. Sociologist Richard Sennett, in his essay "The Open City," describes two kinds of edges: boundaries and borders. While boundaries are where things end, borders are sites of interactivity and exchange. Sennett's argument mostly operates at an urban scale, but its analogous relationship to AI development is clear. Just as tribes and nations can form partnerships, the dividing line between

human and artificial intelligence need not be so rigid. One can improve the other.

In heeding many of the essayists' advice about resisting reductionist approaches to Singularity, it is imperative that we recognize our own assumptions about power. Many current narratives focus on myopic self-gains rather than long-term co-prosperity. Artificial intelligence will be smart, but we can choose to imagine that this intelligence will be able to accommodate and learn from multiaxial values rather than having to oppress them. This requires an expansion of our own values and a shift from competitive, win-or-lose paradigms to collaborative, win-win ones. AI derived from a synergetic mindset will most certainly not take the form of despotic overlords but will instead be our partners. Rather than being trapped in the binary of having to either kill us or sweep for us, it will share the workspace, the dinner table, and maybe even the Netflix password.

Epilogue

Most technologists believe that the advancement of AI will result in a better society. I believe it too—not only in the sense that filing taxes will be easier and chores will be a relic of the past, but also that the process of developing AI will reflect, for our own sake, some of the flawed ways that societies function now. As we sprint to create a new network of intelligence, we ought to first see the problems and imperfections of our own. In fact, current big data endeavors are already revealing structural cracks in our system and painting concrete pictures of previously nebulous biases. Like the Wicked Queen's smart mirror, scientific advancements should not merely showcase technological capabilities; they must also reflect the assumptions we make and the flaws in the logic.

Frequently, skeptics ask if these technologies will strengthen equality or lead to technocratic extremes. This view assumes that we have to wait for the technology to mature before we can answer that question. This is not true. The course of developing technological narratives gives us a unique mirror with which to examine our own values. Donella Meadows (2008) argues that the most effective intervention is the "power to transcend paradigms." The reductionist tropes we have built around AI currently are not only unable to transcend paradigms but also in danger of perpetuating existing ones. We must not write prevailing tales about tomorrow as direct spawns of yesterday's framework.

A typical chilling forecast of AI is that it will be smarter, stronger, and more powerful than us, but the real fear should be that it might *not* be better. It could be instilled with values from our past, with less nuance, more bias, and replete with reductionist tropes. As automation grows, we need to take frequent intermissions to look into the mirror and examine the images it reflects. These technologies are supposed to be harbingers of great scientific progress. Let there be social strides too.

Notes

Much has happened since I first wrote this piece. Even the publication of the essay has become an internally contested decision. Kate Darling's introduction details the whirlwind of changes that this anthology has experienced in connection to Joi Ito's resignation. Many authors, including myself, found ourselves suddenly seized with incertitude—incertitude about whether our mere participation in the publication enables an insidious institution of harm. I am not privy to the inner workings of the Media Lab, and my ultimate decision to continue to support the publication of this book hinges not on Ito but on the rare nuances that my fellow essayists address. The voices that came together to make this book what it is go far beyond Ito's original piece, and they represent a kind of audacious and humanistic approach to artificial intelligence and,

ultimately, to power that I believe is more valuable when seen rather than hidden. I had thought, at the moment of Ito's stepping down, and I still do now, that it would be a shame for Jeffery Epstein's hand to reach out of the grave and silence more voices. So here I am.

1. As a contemporary Jane Austen might quip.

2. A statement about the initiative can be found at http://hrweb.mit.edu/titleixtraining.

3. The panel was a part of a series of debates organized jointly by the MIT Senseable City Lab and the City Form Lab at the Harvard Graduate School of Design. The "Driverless City and the Future of Streets" debate featured Seleta Reynolds, Robin Chase, and Diane Davis.

4. Roko's Basilisk is a thought experiment first proposed by the user Roko on the forum LessWrong. It postulates that an all-powerful artificial intelligence in the future might retroactively punish those who did not help bring it into existence.

References

Asimov, Isaac. *Foundation*. New York: Bantam Books, 2004.

Bennett, Jessica, and Daniel Jones. 2018. "45 Stories of Sex and Consent on Campus." *New York Times*, May 10. https://www.nytimes.com/interactive/2018/05/10/style/sexual-consent-college-campus.html.

Case, Nicky. 2018. "How to Become A Centaur." *Journal of Design and Science*, no. 3 (January). https://jods.mitpress.mit.edu/pub/issue3-case.

"Dawn of Artificial Intelligence, The." 2015. *The Economist*, May 9. https://www.economist.com/leaders/2015/05/09/the-dawn-of-artificial-intelligence.

Friedan, Betty. 2010. *The Feminine Mystique*. New York: W. W. Norton.

Haas, Tigran, and Hans Westlund. 2018. *In the Post-Urban World: Emergent Transformation of Cities and Regions in the Innovative Global Economy*. New York: Routledge.

Meadows, Donella H. 2008. "Leverage Points: Places to Intervene in a System." In *Thinking in Systems: A Primer*, ed. Diana Wright (White River Junction, VT: Chelsea Green Publishing, 2008), 145–165.

Palfrey, John. 2017. "Line-Drawing Exercises: Autonomy and Automation." *Journal of Design and Science*, no. 3 (December). https://jods.mitpress.mit.edu/pub/issue3-palfrey.

Robespierre, Gillian, dir. 2018. *Silicon Valley*. Season 5, episode 5, "Facial Recognition." Aired April 22 on HBO.

Scott, Ridley, dir. (1982) 2007. *Blade Runner: The Final Cut*. Burbank, CA: Warner Home Video.

Slavin, Kevin. 2016. "Design as Participation." *Journal of Design and Science*, no. 1 (February). https://jods.mitpress.mit.edu/pub/design-as-participation.

Tolentino, Jia. 2018. "The Rage of the Incels." *New Yorker*, May 15. https://www.newyorker.com/culture/cultural-comment/the-rage-of-the-incels.

Villeneuve, Denis, dir. 2017. *Blade Runner 2049*. Burbank, CA: Warner Home Video.