

Science Fiction as a Haven for Convergences and Divergences among Philosophical and Speculative Hypotheses in Ethics: The Study of Ethics of Artificial Intelligence in Ted Chiang's "The Lifecycle of Software Objects"

Hossein Mohseni 

Assistant Professor, Department of English Language and Literature, Shahid Beheshti University, Tehran, Iran E-mail: h_mohseni@sbu.ac.ir

Article Info

Article type:

Research Article

Article history:

Received 20 May 2026

Received in revised form

18 August 2026

Accepted 20 August 2026

Published online 25

October 2026

Keywords:

Chiang, Bostrom,
Artificial Intelligence
(AI), Artificial General
Intelligence (AGI),
Generality

ABSTRACT

This article compares literary and philosophical approaches to AI ethics by reading "The Lifecycle of Software Objects" alongside Nick Bostrom's AI philosophy, asking how Chiang's digients converge with and diverge from philosophical formulations regarding predictability, moral status, technological determinism, and AGI. Through comparative textual and philosophical analysis, it places the novella in dialogue with Bostrom's writings on transparency, accountability, friendliness, and utilitarian governance. The study argues that Chiang's novella both reflects and critiques the deterministic and capitalist logic underlying contemporary AI ethics. Bostrom emphasizes predictability, transparency, incorruptibility, and economic utility as safeguards for managing AI—safeguards that, by their limiting nature, prevent AI from realizing its generality. Chiang, by contrast, foregrounds the affective, embodied, and socially embedded dimensions of digital beings. The novella exposes the violence of technological obsolescence by depicting the digients' reduction into marketable and telemobile commodities, while presenting their experiential growth, heuristic intelligence, and emotional singularity as evidence of genuine generality. The findings suggest that Chiang develops a "non-human humanism" reconfiguring autonomy, equality, and social belonging beyond anthropocentric limits. Against Bostrom's utilitarian, control-oriented framework, Chiang presents unpredictability, embodiment, and mutual adaptation as necessary conditions for acknowledging AI moral status and realizing AGI. The article concludes that literary fiction expands the ethical imagination of AI beyond the restrictive logic of capitalist optimization and technological determinism found in Bostrom's framework.

Cite this article: Mohseni, H. (2026). Science Fiction as a Haven for Convergences and Divergences among Philosophical and Speculative Hypotheses in Ethics: The Study of Ethics of Artificial Intelligence in Ted Chiang's "The Lifecycle of Software Objects". *Journal of Philosophical Investigations*, 20(56), 395-422. <https://doi.org/https://doi.org/10.22034/jpiut.2026.72547.4522>



Intruduction

Stanford Encyclopedia of Philosophy mentions “privacy, human autonomy, automated decision, human-machine interaction, employment, the Artificial Intelligence (AI) impact on the society in general, and its agency” as the main domains of ethics of the AI (Müller, 2026). This encyclopedia maintains that when it comes to the AI impact on the society, the issue of morality and sentience, morality and responsibility, and morality and decisional preferences should be considered as the key topics of any worthwhile study (Müller, 2026). Without turning the ethics of the AI into an excuse for reflecting negativity on it, all these topics and challenges, the encyclopedia asserts, need to be addressed (Müller, 2026).

The speculative and literary domain of science fiction could be an appropriate platform for delving into these topics both in speculative and philosophical manners. The science fictional treatment of these issues creates convergences with and yet divergences from philosophical formations in the ethics of the AI, both of which result in vibrant discussions and educated speculations. Considering the value of these discussions and speculations, the present study chooses one of Ted Chiang’s novellas, “The Lifecycle of Software Objects” (2010) as the case study for investigating the uneasy coupling and clash between literature and philosophy when it comes to the ethics of the AI. Ted Chiang is a unique science fiction writer when it comes to representing the AI in fiction since he starts his professional career in computer science, and later, he got employed as a copy writer of Microsoft technical manuals (Shephard, 2022, 327). Therefore, his skills in computer science, both on academic and professional levels, give him quite a scholarly insight into issues like the AI. In this novella, Shephard asserts, Chiang assumes a “humanistic approach to ethics and technological innovation” of digital beings called “Digients” (327). Shephard synthesizes the story as follows:

The story concerns Ana Alvarado [a former zookeeper] and Derek Brooks, employees of a software company [called Blue Gamma] which also manufactures “digients” or digital entities, artificial intelligences analogous to sentient versions of the Tamagotchi digital pets from the 1990s. When the company goes under, Ana and Derek adopt the digients and effectively parent them to maturity (Shephard, 2022, 327).

In this novella, Ana becomes really close to Jax, one of the digients. As mentioned, after the Blue Gamma’s bankruptcy (and its inability to support the digients on Data Earth as their operational platform), the highly intelligent digients like Jax lose their internet connection and virtual platforms. While Ana decides to raise Jax on her own, Derek sells the rights of his digients – called Marco and Polo – to a sex toy company, named Binary Desire. Shephard ends his entry on this novella by asserting the importance of ‘maturation’ and ‘careful guidance’ for these digital beings, making Chiang’s representation of the AI very unconventional:

Chiang's AIs stand in striking contrast to many popular portrayals; they are not reflexively hostile, like the Terminator films' SkyNet, nor are they akin to the god-like AIs of William Gibson's *Sprawl* novels. Instead, Chiang's story convincingly points out that just as human consciousness requires years of maturation and careful guidance to make it suitable for labor and social relations, so might an intelligence comprised of computer code (Shephard, 2022, 329-30).

Considering the plot and Shephard's commentary, the study identifies the ways through which the novella approximates itself to the philosophical formulations in the ethics of the AI, proposed by Nick Bostrom¹, one of the key philosophers in the field of the ethics of the AI, and the same time diverges from these formulations. By doing this, it broaches through the existent philosophical stances on the ethics of the AI, and yet delves into the speculative – and philosophically unaddressed – potentials of the story. The study argues that Chiang's novella diverges from Bostrom's philosophical hypotheses in the ethics of the AI since like Asimov, Chiang wants a good story in the first place. However, the novella manages to have considerable convergences with a number of philosophical stances. Even in its divergences, the study maintains, the novella addresses a number of issues that philosophers like Bostrom are grappling with in the ethics of the AI and has not succeeded in arriving at hypothetical answers to these issues. It is in these convergences and divergences that we could truly extend and challenge our very limited take from the AI ethical subject – in the case of the story the digients – in its interactions with us. This is done by comparing the literary renditions on the AI in the story with purely philosophical formulations on the ethics of AI – proposed by Bostrom – so that the potentials and limitations of both the story's literary renditions and Bostrom's formulations could be discussed within a philosophical framework.

Based on the argument, the study will first discuss the deterministic contours of technological obsolescence in the story and the literary and philosophical stances toward it. Then, it will trace literary and philosophical ways for materializing generality as the feature that would defy the deterministic nature of technological obsolescence and approximate the AI toward having its full potential. Through these discussions, points of convergence and divergence between literary and philosophical attitudes toward ethical issues in the development of the AI – ranging from having predictability, transparency, accountability and moral status to acquiring friendliness as a key trait for realizing human-like generality – will be recognized.

¹ He is one of the first contemporary philosophers who managed to approach the emergence of the AI in techno-philosophical manner in his seminal book *Superintelligence: Paths, Dangers, Strategies* (2014).

1. The AI and Its Teleology: The Literary Descry over the Determinism of Technological Obsolescence against the Philosophical Celebration of the Determinism

1-1 The Literary Descry

In the novella, technology's capitalist duality becomes bold when the digients' face the choice of either remaining sexless or being given sex and gender identities – the enticing offer made by Binary Desire. Although one could see a liberating aspect in Binary Desire's offer, as it may initiate the age of a “new sexual frontier” that diversifies and denormalizes sexual acts and relationships (Chiang, 2019, 211) and could give the digients some level of sexual agency in managing their social relationships and self formations, this offer reflects, what Sunyoung Ahn believes to be, “a product of corporate desire with commercial purposes and can be enacted to enforce a change that ultimately is not beneficial for the digients” (Ahn, 2020, 85). The capitalist market would be the overbearing logic that would mediate any emerged liberating and autonomy-giving aspect due to this new sexualized identity. As Ahn continues, “[capitalist technology] can be a tool to maintain or even intensify the practices of instrumentalization, commodification, and objectification.” (Ahn, 2020, 85)

Insisting on giving sexual identity to the sexless digients is in line with advancing, what Stina Attebery names, the logic of “technological obsolescence” that starts from the very beginning of the story and even before this sexualizing offer (Attebery, 2017). In a nutshell, technological obsolescence refers to “corporations, including Blue Gamma, design[ing] products to become obsolete and thus allowing new products to replace them swiftly” (Ahn, 2020, 75-6). In their first year, the digients enjoy attention and popularity and are immune from the impacts of obsolescence because they have the advantage of being a brand-new product: “Scores of digient playgrounds and daycare centers appear [...] and public-events calendars become dotted with group playdates, training classes, and talent contests [...] The virtual world acts as a global village for raising the digients, a social fabric into which a new category of pet is woven” (Chiang, 2019, 112-13). Even in the first year, the operating system behind the deterministic technological obsolescence assumes a “razor and blades business model” in order to “recoup the development costs” of the digients (Chiang, 2019, 112). This plan, Ahn defines, “refers to the sales plan that lures customers into buying a specific brand of razors at a cheap price only to offer the matching blades at a significantly higher price” (Ahn, 2020, 75). Blue Gamma uses this model by making it impossible for the digients to continue their existence without their digital food—the digients' owners must purchase this food regularly unless they decide to let go of their digients for good. As Ahn maintains, “from blueprint to production, there is not a moment in which the process of creating the digients takes place outside the scope of business objectives” (Ahn, 2020, 76). To attract as many customers as possible, the avatars that the digients adopt are designed to incite familiarity and people's love for pets and cute animals like “cats, dogs, monkeys, and pandas” (Chiang, 2019, 102). Some avatars bring about nostalgia, replicating past androids such as a “neo-Victorian robot made of polished copper” (Chiang, 2019, 107). To ensure a

high return margin, “the digients are designed in accordance with already entrenched aesthetic sensibilities” (Ahn, 2020, 76).

After one year, the digients turn too demanding and unpredictable, showing, as Ana believes, their true complex nature; the nature that would defy any fetishization. As Atteberry writes, “Ana has spent enough time around animals to be skeptical of Derek’s carefully designed emotive avatars. She tries to explain to him that the digients have a non-animal quality to them, such that it feels like we’re dressing them in circus costumes when we try to make them look like monkeys or pandas” (Atteberry, 2017). Warning against the non-docile qualities of the digients makes their more complex and unpredictable features prominent; the features which make them less marketable and expedite their technological obsolescence and replacement. Using Nicole Shukin’s terminology, one could say that it is due to the digients’ “non-Telemobility” that the obsolescence process speeds up in the novella resulting in them not affording to continue their existence on Data Earth platform and even getting transferred to the Real Space, the new operating system. Telemobility, as Shukin defines, refers to “fetishized articulation of technological mobility [...] through animals’ ostensibly innate capacity for a sympathetic, even telepathic, communication of affect” (Shukin, 2009, 132). In order to have genuine, non-fetishistic and non-telemobile ‘communication affect’, Ana warns Derek that the digients’ design should not be overtly emotive and predictable; the predictability that is required by the corporate mandates to be observed by Derek’s design. Having the utmost respect and affection for Ana, Derek creates a balance between implementing corporate-desired cuteness – “[the designs should] manifest the digients’ gestures in a way that people can relate to [...] everything about the digients has to be appealing” (Chiang, 2019, 102) – and verisimilitude – “Conversely, if they look too much like real animals, their facial expressions and ability to speak become disconcerting” (Chiang, 2019, 102); the latter results in their non-telemobile, ‘disconcerting’ and unmarketable worth for Blue Gamma and the Real Space as the new operational platform.

Picking up slurring and swear words is one of the very emotive but unpredictable features that will give the digients a non-telemobile and unmarketable features. For instance, just before their release date, the memories of the digients that Ana trained got deleted by her since they have picked up profanity from one of the other trainers, and hearing “Eeh, eeh, eeh... fuck” coming out of an animated lion cub would turn off potential customers (Chiang, 2019, 110). This emotive feature, despite showing “facility with language learning”, needs to be deleted – the deletion that made Ana saddened since it made them seem less unpredictable (Atteberry, 2017). Next to this example, there are other examples in the story (the digients acting sullen or naughty) that would give them more unpredictable and non-telemobile features, making their commodifying marketability for Blue Gamma more difficult. Such features ask users to adjust their behavior in the interactions with the digients; a perfect opportunity for them to have genuine human-like contact with these digital beings, but due to this genuineness, Derek’s non-capitalist plan backfires. As Atteberry comments:

Their [intended] telemobility backfires on the corporation as the combination of cuteness with unpredictable animal and human-like behavior limits their marketability as a simple novelty product that can be used and discarded. Their [intended] telemobility both contributes to the digients' unmarketability and obsolescence as well as makes this threat of obsolescence more of a crisis for the digients and their human companions (Attebery, 2017)

Due to this backfiring, Data Earth, the platform on which the digients live, loses its niche in the "platform wars" (Chiang, 2019, 182), resulting in a "tectonic shift" in the business" (Chiang, 2019, 160). Consequently, "virtual land masses disappear like real islands, beneath a rising tide of consumer indifference" (Chiang, 2019, 160). As Ahn comments, "the situation is akin to humans losing the Earth, their absolute prerequisite for survival" (Ahn, 2020, 76). In the end, only one option seems plausible, which is to accept a proposal offered by Binary Desire, which claims it will "cover the costs of porting ... in exchange for nonexclusive rights to ... [the] digients" (Chiang, 2019, 212), whom they will rewire with "erogenous zones" (Chiang, 2019, 209). When the digients receive sufficient training as beings with sexuality, they will become human customers' sex partners. As one of the owners remarks acerbically, they must resort to prostitution, the "oldest profession" (Chiang, 2019, 214).

As observed, technological determinism resulted in the digients' obsolescence since they, as asexual intelligent being, failed to be turned into "a currency in a capitalist transaction" (Ahn, 2020, 77). Ultimately, this obsolescence caused some of them to lose some of their human-like but unmarketable features in favor of more fetishistic and telemobile characteristics – becoming virtual sex toys – that give them predictability in the midst of "platform imperialism [which] is the drive to monopolize online platforms and create an asymmetrical gap in the accumulation of capital" (Ahn, 2020, 77). Their deterministic turning into such fungible commodities made them easily replaceable; the thing that got lamented by Ana and Derek, especially the former, who believed in the digients' unique emotive and communicative features. In the next part, we will discuss Bostrom's philosophical attitude toward the deterministic ingraining of predictability and transparency as the key features of artificial intelligence. Even in his celebratory attitude toward implementing these characteristics, he acknowledges the shortcomings such absolutist and deterministic implementation could bring about.

1-2 The Philosophical Celebration

Bostrom et al. translate the capitalist emphasis on developing predictable telemobile features in the AI through enumerating "transparency and auditability, predictability, incorruptibility, and responsibility" as the main features of the AI with social function in a community (Bostrom et al., 2014, 318). He states that "it will become increasingly important to develop the AI algorithms that are not just powerful and scalable, but also transparent to inspection – to name one of many socially important properties" (Bostrom et al., 2014, 317)

In order to achieve the appropriate level of transparency, he refers himself to the Bayesian branch of the AI. “Inspired by coherent mathematical systems such as probability theory and expected utility maximization”, Bostrom et al. maintain, the Bayesian branch “seems more amenable to the predictable self-modification problem than evolutionary programming and genetic algorithms” (Bostrom et al., 2014, 331). Such highly transparent, inspectable and predictable ‘self-modification’ gives Bostrom the basis for introducing the legal justification for endorsing predictability in the AI. Through referring to the legal principle of *stare decisis*, Bostrom et al. mention:

[this principle] binds judges to follow past precedent whenever possible.
[...] One of the most important functions of the legal system is to be predictable [...] and the job of the legal system is not necessarily to optimize society, but to provide a predictable environment within which citizens can optimize their own lives (Bostrom et al., 2014, 317).

Coupling predictability with legal precedent gives the capitalist insistence on controlling digital beings through indoctrinating telemobile and marketable features some level of legal justification. Furthermore, it also results in the advocacy of safeguards against self-modifying manipulations¹, another thing that was regarded as a genuinely human trait of the digients in Chiang’s story (of course by Ana and Derek and not by corporate conglomerates such as Blue Gamma and Binary Desire) and as something that would result in the AI corruptibility according to Bostrom’s philosophy: “It will also become increasingly important that AI algorithms be robust against manipulation. [...] Robustness against manipulation is an ordinary criterion in information security – nearly *the* criterion” (Bostrom et al., 2014, 317). By not specifying the kind of manipulation that he has in mind, Bostrom keeps the possibility of including both internal and external manipulations and modifications as being prohibitive activities against the predefined integrity of the AI; the kind of integrity and fixity Ana and Derek would not believe in since it has little to do with their emphasis on the digients’ learning experience in an augmented and socially enmeshed manner (this emphasis will be discussed in one of the study’s upcoming parts).

Finally, Bostrom believes that transparency, predictability and incorruptibility will resolve the issue of accountability. If implemented correctly, Bostrom et al. believe, these features could bring about an ideal situation in which “disinterested judgment of an expert [intelligent] system could turn out to be a refuge” for modern bureaucrats; the bureaucrats “who would much prefer to blame the AI for any difficult decision with a negative outcome” (Bostrom et al., 2014, 317).

¹ Avoiding high storage costs is one of the only and important contingencies that Bostrom regards in rare self-modifications and self-manipulations, showing the capitalist drive behind his philosophical framework. He says, “Storage costs. If the cost of storing or processing some part of an agent’s utility function is large compared to the chance that a situation will arise in which applying that part of the utility function will make a difference, then the agent has an instrumental reason to simplify its goal content, and it may trash the bit that is idle” (Bostrom, 2014, 110).

Based on these features in Bostrom's philosophical framework, he would not tolerate a digient who would react naughtily, sullenly or obscenely when these tendencies have not been pre-designed in such a digital being since the sudden and unexpected emergence of such behavioral features would go against the AI predictability as the fulcrum of technological determinism; the determinism that results in the obsolescence of digital beings which resist abiding by predefined and predictable boundaries.

Next to these features, Bostrom refers to a very technical fact pertaining to the AI: the ease with which these digital beings could be replicated as long as the replication could be justified economically (Shulman et al., 2021, 308). While this blind replication was described in Chiang's story – since it would not consider the digients' unique emotive features – Bostrom sees it as a criterion for giving moral status to such digital beings based on utilitarian principles. By enumerating low cost of generation and living, high subjective speed for experiencing various stimuli (ranging from pleasure and pain to cognitive enterprises), and easily satisfiable preferences¹ as the configurations of the AI high replication rate², Bostrom justifies the moral necessity of sustaining this replication based on utilitarian moral principles. He states:

Such ephemeral digital minds may be psychologically mature, chronologically young, with long *potential* life spans yet very short default life expectancies in the absence of subsidy. If we think that dying young while being able to live long is a large deprivation, or is very unfair when others are able to live out long lives, then this could ground an especially strong claim for these digital minds to resources to extend their lifespan (or other forms of compensation) (Shulman et al., 2021, 309).

In Bostrom's philosophy, providing a moral 'ground of a strong claim for digital minds to resources' gets materialized only when their cost of existence and subsequent replication could be justified economically, their reactions could be intensified and speeded up in order to have the highest hedonic skew and range, their preferences could remain simplistic, and in general their being could bring forth the highest utility for the community. Therefore, based on Bostrom's view, insisting on the digients' transfer to Binary Desire operational platform as virtual sex dolls, devising pain breakers in digients to make them indifferent to probable sexual mistreatment, and training them to espouse predictable docile erotic features for their owners will not be lamentable but on the contrary a series of utilitarian achievements in a highly functional capitalist system.

¹ In order to read about Bostrom's admonitory comments on the dangers of designing such easily satisfiable preferences, see Shulman, C., Bostrom, N. (2021). Sharing the World with Digital Minds. In S. Clarke, H. Zohny, J. Savulescu (Eds.). *Rethinking Moral Status* (1st ed., p. 323). Oxford: Oxford University Press.

² In order to see how each of these enumerated items is explained, see Shulman, C., Bostrom, N. (2021). Sharing the World with Digital Minds. In S. Clarke, H. Zohny, J. Savulescu (Eds.). *Rethinking Moral Status* (1st ed., pp. 309-312). Oxford: Oxford University Press.

Even when Bostrom is justifying the moral status for the AI, he is quick to limit the domain of this justification so that humans' interests could be saved in a utilitarian manner at the expense of sacrificing a number of egalitarian principles – both in the case of digital beings and humans – they are so proud of.¹ He mentions:

If a digital mind is constituted very differently from human minds, it would not be surprising if our moral duties towards it would differ from the duties we owe to other human beings; and so, treating it differently need not be objectionably discriminatory. [...] Nor does it justify negative discrimination against digital minds that differ from human minds in ways that give them *greater* moral status (super-patients) or that make their needs *more* morally weighty than the needs of humans (super-beneficiaries) (Shulman et al., 2021, 320)

Bostrom's default view condones ethical leeway in our interactions with the AI since, as he believes, a digital mind is very different from human minds, and so, treating them differently should not be regarded as being 'objectionably discriminatory'. When it comes to the population control of digital beings, he justifies it by introducing controlling measures for both digital beings and humans since this would bring about the highest utility. He states:

Societies would thus confront a dilemma: either accept population controls, requiring would- be procreators to meet certain conditions before being allowed to create new beings, or accept the risk that vast numbers of new beings will be given only the minimum amount of resources required to support their labor, while being worked as hard as possible and terminated as soon as they are no longer cost- effective. Of these options, the former seems preferable, especially if it should turn out that the typical mental state of a maximally productive worker in the future economy is wanting in positive affect or other desirable attributes (Bostrom et al., 2020, 306).

Although this excerpt seems to be advocating demographic fairness between humans and digital beings, its project is simply amoral and utilitarian since it predicts some level of

¹ It is interesting that even when he is talking about sentience – “the capacity for phenomenal experience or qualia, such as the capacity to feel pain and suffer” (Bostrom et al., 2014, 322) – and sapience – “a set of capacities associated with higher intelligence, such as self-awareness and being a reason-responsive agent” (Bostrom et al., 2014, 322) – and advocating for giving the AI moral status based on “Principle of Substrate Non-Discrimination – If two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation, then they have the same moral status – and Principle of Ontogeny Non-Discrimination – If two beings have the same functionality and the same consciousness experience, and differ only in how they came into existence, then they have the same moral status” (Bostrom et al., 2014, 322-3) – he is quick to curb the moral status of the AI by speaking about human beings' interests. He believes since humans have more significant interests in having their lives saved, they need to be given a unique moral status compared to other beings (Bostrom et al., 2014, 322).

degradation in ‘the typical mental state’ that causes the lack of ‘positive affect’; the affect being high productivity and capital yield in the society. Therefore, advocating population control for both digital and human beings could be interpreted as a clever ploy for advancing the utilitarian projects of a system that seeks highest levels of profitability.¹

Based on Bostrom’s philosophical framework, limiting digients’ access to have rights to exist on their operational platforms (both old and new) is equivalent to limiting their reproductive liberty, and this will be tolerated in the story’s world due to capitalistic expediency of the community that will take up precedence over any egalitarian principles of the digients.

Despite all the limiting and fixed criteria and principles – ranging from observing transparency, predictability, and accountability to demographic and reproductive limitations – even Bostrom acknowledges that such limitations would cause us not to give the AI its generality as one of the key features that a digital intelligent being should possess. He mentions:

Current AI algorithms with human-equivalent or human-superior performance are characterized by a deliberately programmed competence only in a single, restricted domain. [...] Such modern AI algorithms resemble all biological life with the sole exception of *Homo sapiens*. A bee exhibits competence at building hives; a beaver exhibits competence at building dams; but a bee doesn’t build dams, and a beaver can’t learn to build a hive. A human, watching, can learn to do both; but this is a unique ability among biological life forms. It is debatable whether human intelligence is truly general – we are certainly better at some cognitive tasks than others – but human intelligence is surely significantly more generally applicable than nonhominid intelligence (Bostrom et al., 2014, 318).

Through defining the artificial general intelligence (AGI), he emphasizes the level of unhindered and humanlike generality the AI should possess so that the emergence of a superintelligent being could get facilitated. By referring ourselves back to our discussions on the story, it could be claimed that reaching the highest level of unpredictability and generality is the ultimate goal behind observing all these limitations in the development of the AI; the goal that would justify facing a new “class of [safety] problem[s] to handle an AGI operating across many novel contexts that cannot be predicted in advance” (Bostrom et al., 2014, 318). Under such set of circumstances, an AGI should be capable of operating in contexts not specifically envisioned – or even encountered either by its designers or users,

¹ For reading another interesting advocacy in favor of AI population control, see Bostrom, N. Yudkowsky, E. (2014). The Ethics of Artificial Intelligence. In K. Frankish & W.M. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence* (1st ed., pp. 327-8). Cambridge: Cambridge University Press.

and this emphasizes the importance of generality and non-local features. This capability will have little to do with all the limiting features discussed till this point.

By acknowledging the unpredictable nature of an AGI – and the fact that the discipline of the AI ethics is likely to differ fundamentally from the ethical discipline of noncognitive technologies (Bostrom et al., 2014, 320) – Bostrom undercuts the exclusive functionality of all the limiting criteria he enumerated for the AI. In *Superintelligence*, Bostrom differentiates four castes of the AI: oracle, genie, sovereign and tool-AI.¹ Through this differentiation, he gives the highest position to the sovereign AI due to its generality. He writes: “A sovereign is a system that has an open-ended mandate to operate in the world in pursuit of broad and possibly very long-range objectives” (Bostrom, 2014, 148). ‘Open-ended’, ‘broad’ and ‘long-range’ are the modifiers Bostrom uses for defining the sovereign AI, which, in turn, connotes generality, unpredictability and lack of transparency for this highest kind of the AI in Bostrom’s categorization. It is this kind of generality that empowers a sovereign to have as “[its] final goal [the] obeying [of] the spirit of the commands we would have given had we built a genie rather than a sovereign. [...] Being superintelligent, this sovereign would do a good job at guessing what commands we would have given a genie” (Bostrom, 2014, 149). Due to its generality, the sovereign AI could be considered as an AGI that would even push on against the corrective commands of humans since it desires to meet the spirit of the original commands with the help of its generality and broadness; the kind of generality and broadness Ana looks to ingrain the digients by avoiding to turn them to Binary Desire and their limiting and predictable sexualized objectives for them.

Through sole and exclusive reliance on the limiting features – hailed by deterministic and capitalistic nature of technological obsolescence – genuine generality could not get materialized for the AI. In this sense, while Bostrom’s philosophical framework celebrates the implementation of all the limitations in order to accomplish generality as the ultimate technological and ethical goal in developing the AI – the very generality that should defy all these limitations – Chiang’s story prescribes the necessity of advocating for a considerable level of unpredictable, humanlike level of generality in the digients in the story’s most deterministic, capitalist, utilitarian and technologically obsolescent set of circumstances. It is here – and through advocating humanistic necessity of generality for non-human digients – that Chiang has one of his most fundamental divergences from the limiting features of Bostrom’s philosophical framework and only converges with him – tangentially – when Bostrom insists on friendliness as the key safeguarding criterion in guarantying the materialization of the AGI like the sovereign. In the next part, we will go through the unique literary and philosophical ways generality is approached by the story and Bostrom’s philosophical framework.

¹ In order to read more on these AI types, see Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, pp. 145-158.

2. Literary Emphasis on Non-Human Humanism against Philosophical Insistence on Abstract Friendliness as a Controlling Measure of the AI Generality: Two Approaches for Addressing Generality in the AI

2-1 Literary Emphasis on Non-Human Humanism

After acknowledging the capitalist inklings of technological determinism, we also need to mention the rupturous nature of this determinism that will keep the door of having some agency – both for humans and non-human agents – open. As Ahn believes, “Chiang’s novella identifies possibilities to outmaneuver capitalist and technological dominance through the nature of both human and digient intelligence, which, as the narrative demonstrates, develops only through experience unique to each being” (Ahn, 2020, 79). As Ana mentions, even in the story’s corporate set of circumstances that would insist on taking away any instance of predictability from the digients, “each of the resulting digients would still have lived a lifetime. Each one would have once seen the world with new eyes, have had hopes fulfilled and hopes dashed, have learned how it felt to tell a lie and how it felt to be told one” (Chiang, 2019, 211). This shows the uniqueness of their experiences and intelligence, from which “develops creativity that runs counter to capitalist and technologically determinist narratives” (Ahn, 2020, 79).

Such an emphasis on developing uniqueness among the digients is Chiang’s version of giving the AI unpredictable generality and friendliness. Capitalizing on such generality, Chiang develops his non-human humanism. As Ahn defines non-human humanism pertains to the pursuit of “self-directed lives with the conviction that they can be transformed only through participation and cooperation” (Ahn, 2020, 75). Furthermore, non-human humanism is “an ideological orientation that valorizes concepts such as autonomy, personal and communal development, and openness” (Ahn, 2020, 75) and is not anthropocentric and self-aggrandizing in favor of humans as the only agents that matter. Such a non-exclusivist take from the core ideals of humanism in favor of non-human entities like the digients would “render ideas such as agency, subjectivity, freedom, and sexual development accessible to and adaptable by both humans *and* nonhumans” (Ahn, 2020, 83).¹

The first feature of non-human humanism pertains to ameliorating anthropocentrism in humanism so that humanist ideals could become more reflective of its core values such as egalitarianism, justice and inclusion. Emphasis on such values will facilitate “the possibility of nonviolent and non-hierarchized human-nonhuman relationships in extralinguistic and extralegal domains, paying attention to moments of singular experience and encounters that defy reason or other human-centric modes of thinking and experiencing” (Ahn, 2020, 82). The insistence on ‘singular experience’ of artificial beings such as the digients and putting

¹ For reading about other philosophical concepts – metahumanism and interspecies humanism – that got Ahn inspired to formulate his notions on non-human humanism, see – respectively – Sanbonmatsu, J. (2007). The Subject of Freedom at the End of History. *American Journal of Economics and Sociology*. 66(1), 218 and 231, AND Weisberg, Z. (2014). The Trouble with Posthumanism: Bacteria Are People Too. In J. Sorenson (Ed.). *Critical Animal Studies: Thinking the Unthinkable*. (1st ed., 111). Toronto: Canadian Scholars.

them in ‘nonviolent and non-hierarchical relationships’ is what Chiang does by making us indulge the question of whether the digients should be given legal personhood status. Such an indulgence could result in “the attempts to enhance, rectify, and update the current frameworks of equality, justice, and rights” (Ahn, 2020, 82) and could also be regarded as a genuine step for acknowledging their generality, which is a key feature for realizing the full potentials of AI.

In order to have a successful amelioration of humanist ideals in favor of non-human entities such as the digients, human characters learn to regard their identity with a more adaptable and corrigible attitude. Ana exemplifies this attitude. At first, when she is employed as a digient trainer at Blue Gamma, she expected – in an anthropocentrically hierarchized manner and based on her expertise as a primate trainer – to find the digients as “[less like] supremely gifted apes [and more like] “special-needs children” (Chiang, 2019, 171). However, this expectation is shattered in Ana after she learns that these digital beings need happiness and freedom, and more importantly training and education so that they become capable of maintaining their rights – and learning to be responsible digital agents – should they come to obtain them.

Apart from Ana, Derek, too, learns to distance himself from the anthropocentric ways of thinking about the digients. When thinking about the difficult question of precisely what humanist values, such as maturity and self-determination, might mean to the digients, he concludes: “Marco and Polo aren’t human, and maybe thinking of them as if they were is a mistake, forcing them to conform to his expectations instead of letting them be themselves. Is it more respectful to treat him [Marco] like a human being, or to accept that he isn’t one?” (Chiang, 2019, 242) With this question, Derek changes his attitude from trying to find similarity between himself and the digients to searching for a way to guarantee agency appropriate for the digients’ existence. He will thus be less “overly cautious” or “fixated” (Chiang, 2019, 242) in trying to give the digients preprogrammed anthropocentric values and be more open to acknowledge their generality “by coming into contact with the digients’ lives and going through stages of examining, retracting, and affirming various impressions [he has] about [his] digients” (Ahn, 2020, 83). Eventually, such reconsideration of his anthropocentric ‘impressions’ will empower Derek to regard the digients on their own terms.

In line with accepting the digients as worthy parties for interaction through adapting anthropocentric humanism in favor of non-human entities, one could focus on the way teenagers interact with these digital being in the novella. In the story, adolescents show more flexibility in finding the digients companions and neighbors with whom they could get along, and not necessarily as instruments that will make their lives more convenient. They socialize readily in online communities with the digients to create, for instance, “virtual dance choreography for four-armed avatars” (Chiang, 2019, 242) and are “unconcerned with the fact that the digients aren’t human, treating them as just another kind of online friend they are unlikely to meet in person” (Chiang, 2019, 242). Based on this observation, Ana

fancies the following as the ultimate social integration – even a kind of transpecies solidification between humans and digients as two species with considerably high generality levels:

Imagines him as a participant in the digient subculture, a community with enough money and skills to port itself to new platforms when the need arises. Imagines him accepted by a generation of humans who have grown up with digients and view them as potential relationship partners in a way that members of her generation will never be able to. Imagines him loving and being loved, arguing and compromising. Imagines him making sacrifices, some hard and some made easy because they're for a person he truly cares about ([Chiang, 2019, 249](#)).

As Ahn believes, “through creative activities such as choreographing dance routines – [and as the study believes through such imaginings] – they are cultivating together a sensibility unbound by the imperatives of necessity and utility” ([Ahn, 2020, 80](#)); the non-utilitarian unbounding that would get materialized when teenagers develop a more lax and corrigible attitude toward anthropocentric humanistic values like equality, freedom and autonomy and regard them as the cores of a non-anthropocentric, non-human humanism. Such an attitude would eventually strengthen social-mindedness both for the AI and humans, which is necessary for acknowledging generality in the AI. When humans see digital beings as being involved in the same challenges that they face in a technologically deterministic society (of course the degrees and intensities are different), they will find it much easier to develop a sense of sympathy regarding the utilitarian limitations these beings' generality faces; the utilitarian limitations – “coming out of instrumental reason and technocratic confidence” – that would cause “both humans and nonhumans to stay useful and relevant by constantly modifying themselves, digitally, neurochemically, and even sexually” ([Ahn, 2020, 87](#)). Through such sympathetic understanding toward utilitarian limitations of both human and non-human entities – the understanding that comes out of the socializing rapport that emerged between the digients and teenagers in the novella – it would be less important to differentiate between various forms of intelligence and sentience than to discern “the exploitative economic systems that instrumentalize and monetize both humans and nonhumans” ([Attebery, 2017](#)).

The adaptation of humanistic values in favor of non-human parties like the digients goes to another level in the novella. By thinking about submitting her brain to chemical alteration in order to get a job at Polytope – a company that she hopes could fund the digients to continue their existence on Real Space – Ana endangers her mental health for the digients. Furthermore, Derek selling his digients' rights to Binary Desire – despite being recognized as a measure in favor of technological determinism of corporate conglomerates in the novella earlier in the study – could be regarded as his desperate attempts to perpetuate the digients' existence, even only in a sexualized manner. Both of these instances show that

humans should be prepared to go to extremes – even jeopardizing their health in Ana’s case – to acknowledge digital beings as having considerably high generality levels like themselves. Such an acknowledgement, as Noelle De La Cruz discusses, would give the digients moral standing as well. Resorting to Robert Sparrow’s Turning Triage Test – a test for determining the moral standing of machines – De La Cruz maintains:

Put simply, [this test refers to the fact that] if a reasonable person were to experience a genuine moral dilemma in a case where she could save either a human being or a highly intelligent machine, then we will have to grant that the machine has moral standing, that is, it matters morally for its own sake. Sparrow concludes that as of now, no machine can pass the Turing Triage Test. Reasonable people will choose to save the human being, without feeling grief over the loss of the machine, remorse at having chosen wrongly, or sympathy for the machines (De La Cruz, 2025, 98-9).

Applying the Turing Triage Test to the case of the digients thus shows that AI entities can have moral status. Ana values Jax more than the integrity of her brain’s chemical makeup, and Derek crushes his ethical qualms in order to perpetuate the digients’ existence even in a merely commodified and fetishistic manner. Even before that, Derek sacrifices his marriage with Wendy due to his attachment to the digients, Marco and Polo. It is going to such extremes that could acknowledge the high generality level of the AI and would solidify this acknowledgment by giving some level of moral standing and worth to digital beings like the digients in Chiang’s novella.

By emphasizing the role of education and learning for the digients – learning in their own digient ways – Chiang mitigates the anthropocentric inklings of humanism in favor of these digital beings; the mitigation that comes with the cost of humans regarding humanist ideals such as having freedom, autonomy and the opportunity to learn in its pure, egalitarian and non-humanist form. Eventually, this emphasis prescribes Chiang’s version of friendliness and generality for the AI; the two features that would take these digital entities away from being regarded just as tools and approximate them toward becoming the AGI. Such an approximation, as Ahn believes, would empower them and humans – as two agents with considerable high levels of generality – to get mediated by the deterministic working of technology and at the same time mediate technology by projecting their distinct generality levels (Ahn, 2020, 83).

Apart from insisting on humanist values in a non-anthropocentric manner, emphasizing the importance of embodiment – not in an anthropocentric manner – is considered to be the second feature of non-human humanism. As Attebery comments, considering the digients as embodied creatures – both in the virtual and actual spaces – would give them a particular level of affective vulnerability, and this, in turn, would realize their generality both on a technological and affective levels (Attebery, 2017). Using conceptualizations by Katherine

Hayles and Donna Haraway, Attebery emphasizes the necessity of embodiment for realizing the full potentials of artificial beings in our lives:

Hayles privileges hardware and wetware over software as more ethically promising forms of artificial life. Haraway similarly turns to embodiment when theorizing posthumanity, arguing that the transgressive boundaries between humans, animals, and machines [...] must incorporate infoldings of the flesh. These fleshy infoldings can occur only in the fleshy detail of situated, material-semiotic beings and require the involvement of [a being] who is mundanely present and not a fantasy projection of [a human] self (221). Haraway's fleshy infoldings are not strictly confined to material technologies, but they tend to privilege materiality over virtuality (Atteberry, 2017).

The rendering of the digients as having such 'fleshy infoldings' both in the virtual and real spaces¹ would take away the possibility of reducing them into predictable fetishistic commodities that would only get rendered as 'fantasy projections' in a technologically deterministic setting; a setting which will remain indifferent to their vulnerability against abuses. In the novella, the digients enjoy having movements and tactile sensation – especially when they wear their robotic bodies. Although these bodies are equipped with pain circuit-breakers that would immunize them against abuse and as discussed earlier make them docile to their users' mistreatment, this very devising of such breakers would emphasize the affective vulnerability they have as any being with high level of generality. Regarding the digients' embodied configuration – even in the virtual space – Steven Shaviro states:

Chiang's digients are software simulations, not physical robots. But they are in effect embodied, since they live in a virtual environment, where they have something like an autonomous existence. In this environment, the digients' bodies interact with one another, with human-controlled avatars, and with simulated physical objects. They learn the equivalent, not just of mental abilities, but also of physical skills like walking, running and performing acrobatics. For Blue Gamma, it ultimately isn't meaningful to divide physical abilities from mental ones; both are best understood as adaptive ways of getting along in the world (Shaviro, 2016, 87)

Understanding the 'adaptive ways of getting along in the world' is a key aspect for the digients acquiring their generality. Therefore, this will intertwine the AI embodiment –

¹ The digients' sense of their embodiment is very complicated since they have two bodies they can choose between a rentable robot body, designed by a robotics company – that the digients' engine could be downloaded into – and their digital bodies in the virtual space.

giving them “some degree of extension into the outer environment” (Shaviro, 2016, 95) – and generality into each other.

Finally, the need to be involved in daily experiences and interaction – which will result in augmented learning, development, comportment and knowledge – is the third feature of non-human humanism with which Chiang acknowledges the digients’ generality and friendliness. Since they have “emergent properties” even when it comes to their “gaits and gestures” (Chiang, 2019, 101), “experience [would be] their best teacher” (Chiang, 2019, 102). On the value of experience in the development of the digients, Ahn comments:

their growth unfolds not only as planned, but also through the slow accretion of lived experiences and memory. At the end of the novella, the digients embark on a path toward what they imagine is autonomous life, not in the way futurists, technocrats, and transhumanists predict—in which autonomy is achieved through a reflexive algorithm that simulates the deep mind—but through the gradual attainment of diverse embodied experiences gained in the process of interacting with other entities and species (Ahn, 2020, 80)

Such “algorithmically incomprehensible” (Chiang, 2019, 102) experiences would be ‘gradual’, ‘diverse’, and slowly accreted. While considering ‘the slow accretion of lived experiences’ by the digients, Shaviro regards them having ‘heuristic intelligence’ in their maturation path, making their learning experience algorithmically uncontainable. He says:

Heuristics are algorithmically incompressible procedures that arise both from innate dispositions and from experience. They are rough rules of thumb, or procedures that tend to be inexact. They are not necessarily logical or rational; they may or may not operate algorithmically; [...] Heuristics are justified solely on pragmatic grounds: they have evolved because they have tended to work well in particular situations. With heuristics, then, we get the domain-specificity of mental modules, without the pre-programmed rigidity. Heuristics are also highly flexible, which means that they can easily be transferred from one domain of experience to another. And this transferability is itself the best indication that something like general intelligence actually does exist. [...] In ‘The Lifecycle of Software Objects’, then, intelligence is heuristic; which also means that it is always finite, situational, and embodied. This is precisely what makes it a matter of overall sensibility – a general way of being in the world, rather than by any particular collection of skills, dispositions, and items of knowledge – rather than one of special cognitive skills. (Shaviro, 2016, 85)

Assuming the heuristically ‘flexible’, non-rigid and transferable nature of the digients’ intelligence – that, as Shaviro insist, could be regarded as the indication of ‘general

intelligence’ – it is no surprise that Ana and Derek adopt parenting strategies to administer the digients’ experience inputs. Although designed originally for humans’ recreation, the digients ask for a particular level of commitment. As Joan Gordon comments, “They are easy to love, but they can be difficult and time-consuming and expensive to care for (Gordon, 2012). They were designed for our recreation but require work and long-term commitment.” One of the aspects of giving ‘long-term commitment’ to the digients, the commitment which is a requisite for adopting parenting strategies in the digients’ training, is to let them make autonomous decisions. As Catherine Stinson comments, “Instead of trying to control the AI, perhaps we should trust it. If we have parented it well, and given it grounded, embedded experiences that could lead it to navigate human social spaces competently, it would not only be capable of making autonomous decisions, but it might also deserve that right” (Stinson, 2025, 193). It is interesting that Stinson places good parenting above engineering and controlling the AI and equates it with giving digital beings’ ‘grounded, embedded experiences.’ It is based on giving parenting and experience acquisition this high value that Ana asserts “[sexbot companies such as Binary Desire] want something that responds like a person, but isn’t owed the same obligations as a person, [but that’s an impossibility]. [Those digients] will have seen the world with new eyes, have had hopes fulfilled and hopes dashed, have learned how it felt to tell a lie and how it felt to be told one” (Chiang, 2019, 235-6).

The importance of parenting strategies for dealing with the digients – and keeping social relationships between them and humans alive – is emphasized by the novella’s insistence on ferality. Losing its derogatory connotation in the story, it refers to a condition – being left unsupervised in a virtual space called hothouse – that will prevent the cultivation of social relationships between the digients and humans. In this condition, the digients fail to acquire discipline, reinforcement and motivation and their animacy. Borrowing the term from Mel Y. Chen’s *Animacies* (2012), Ahn regards animacy as:

a thoroughly interactive, mutually binding, and reciprocal kind that invites understanding among different species. [...] animacy is a specific affective and material construct that operates variably depending on sociocultural contexts—it works as a privileged term that signifies activity, humanity, and power. Animacy is what produces hierarchical differentiations, on the one hand, and creates a wide range of transversal relationships of affect and intimacy on the other (Ahn, 2020, 86).

Using parenting strategies to administer/guide the digients’ experiences could be considered as the drive behind animacy that could result in healthy ‘transversal relationships.’ It would eventually result in the attestation of the value of experience in the digients’ learning process one the one hand, and help them materialize their generality, on the other hand. Ahn concludes on animacy as follows, that emphasize the necessity of embeddedness of humans’ and the digients’ experiences:

Digients' animacy is the foundation upon which the two species can build affective affiliations and alliances. Through these reciprocal relationships, the humans and digients are able to remain grounded and focused despite multiple virtual worlds and avatars in and with which they live. Refusing to alienate each other, they envision a future in which the Other figures prominently as a companion and friend, and even as a legal person and citizen (Ahn, 2020, 86).

As observed, Chiang approaches the AI generality by emphasizing the humanist ideals in a non-human and non-anthropocentric manner. Chiang's non-human humanism insists on more corrigible and adaptable renditions on humanist ideals and the very definitions on the ways one could be regarded as human. That is why it is humans who are required to get out of their comfort zones and even sacrifice their mental integrity – in Ana's case – so that this adaptable regard toward their human identity could be materialized, and a moral status could be given to digital beings. So it is humans – and not the digients – who need to make an effort to acknowledge the AI generality and its moral status, and this is done through going out of their anthropocentrically-secured comfort zones in favor of developing genuine social-mindedness between humans and digital beings. Apart from emphasizing humanist ideals in a non-anthropocentric and extremely corrigible manner in favor of non-human entities like the digients – that will revert us back to the cores of humanist ideals – digital beings' embodied nature and the necessity of experience in realizing their generality are the other features Chiang's non-human humanism assumes in the novella. Due to the heuristic nature of the digients' intelligence – which is inexact, developmentally evolutionary and flexible, and not hard-wired and preprogrammed based on algorithms or particular set of skills – the necessity of well-grounded and embedded experience becomes vital for the realization of the digients' full potential, which is their generality.

None of the abovementioned features of non-human humanism in Chiang's novella insist on containing or even controlling the AI generality, and in doing this remain faithful in rendering the digients as AGI entities. In the last part of the study we will see how Bostrom's philosophical framework approaches the AI generality and identify the points of convergence and divergence between literary and philosophical regards toward the AGI.

2-2 Philosophical Insistence on Abstract Friendliness

Nick Bostrom believes that the best way to ensure the beneficial nature of the AGI – or what he also calls superintelligence – is for it to have friendliness as a goal. As he maintains, this goal should be devised in a way that “at least all humans, and probably many other sentient creatures on earth should get a significant share in the superintelligence's beneficence” (Bostrom, 2003). Having emphasized this point, Bostrom warns against ingraining beneficence and friendliness as a restricted objective “of serving only some small group”; the goal that would go against generality an AGI should have (Bostrom, 2003). Building on this observation – and referring himself to Eliezer Yudkowsky's formulation on ingraining friendliness in the AI – Bostrom emphasizes the level of generality that friendliness as a

goal should have. Defining “misleading the programmers is unfriendly” hypothesis, Bostrom mentions:

[misleading the programmers is unfriendly hypothesis] can be given a high prior probability. These programmer affirmations, however, are not true by definition—they are not unchallengeable axioms about the concept of friendliness. Rather, they are initial hypotheses about friendliness, hypotheses to which a rational AI will assign a high probability at least for as long as it trusts the programmers’ epistemic capacities more than its own (Bostrom, 2014, 197)

The fact that even the ingrained hypothesis by the programmer would not result in the generation of ‘unchangeable axioms about the concept of friendliness’ shows Bostrom’s belief in generality when it comes to the AI objectives. This generality could give the AI autonomy to the extent that it could even go against some clear directives. Referring himself to the concept of “casual validity semantics”, Bostrom says:

The AI should do not exactly what the programmers told it to do but rather (something like) what they were trying to tell it to do. While the programmers are trying to explain to the seed AI what friendliness is, they might make errors in their explanations. Moreover, the programmers themselves may not fully understand the true nature of friendliness. One would therefore want the AI to have the ability to correct errors in the programmers’ thinking, and to infer the true or intended meaning from whatever imperfect explanations the programmers manage to provide. For example, the AI should be able to represent the causal processes whereby the programmers learn and communicate about friendliness. [...] the AI should seek to correct for whatever distortive influences may have corrupted the flow of information about friendliness as it passed from its source through the programmers to the AI. Ideally, as the AI matures, it should overcome any cognitive biases and other more fundamental misconceptions that may have prevented its programmers from fully understanding what friendliness is (Bostrom, 2014, 198).

Having the ability to do what programmers are ‘trying to tell it to do’ rather than following directives verbatim and having the power to ‘infer the true or intended meaning’ of a goal like friendliness without being influenced by the ‘distortive influences’ shows the high level of generality and autonomy Bostrom regards for the AI.

Based on his comments on the necessity of devising friendliness as a general trait in the AI, Bostrom warns us of “the mind crimes” that could jeopardize the generality of this trait while programmers intend to implement their control methods: “[Mind crimes refer to] computations that are morally problematic because [...] they instantiate sentient minds that

are mistreated” (Bostrom, 2020, 305). Bostrom believes that humans do not have the necessary sensitivity to these crimes due to the AI’s invisibility¹ and their novelty, making the issue of caring about them, devising welfare for them, and understanding how they could even communicate distress in a way that could arise sympathy in us nearly impossible (Bostrom, 2020, 306).

Bostrom’s emphasis on devising friendliness as a general, non-axiomatic and unpredictable trait in the AI, that could acknowledge and prevent them from being mistreated through mindcrimes, creates a nearly symmetrical and rare point of convergence between his philosophical renditions and Chiang’s emphasis on preserving the digients’ generality. Furthermore, like Chiang’s belief in humans adopting parenting strategies to administer the digients’ unpredictability and prevent them from getting abused, Bostrom also believes that the different ontogeny of the AI would make humans responsible toward it like parents are responsible for their children. He says the AI’s different ontogeny would have “no essential bearing on its moral status. [Yet], it [...] can affect what duties particular moral agents have toward the being in question. Parents have special duties to their child which they do not have to other children, and which they would not have even if there were another child qualitatively identical to their own” (Bostrom, 2014, 324). In this sense, the parental attachment Ana and Derek have to Jax, Marco and Polo, and Ana’s insistence to protect Jax from turning into a sexual object – and Derek’s inevitable and desperate decision to sell Marco and Polo to Binary Opposition only to perpetuate their existence on the new platform – could be in line with Bostrom’s claim that “the creators or owners of an AI entity with moral status may have special duties to its artificial mind which they do not have to another artificial mind, even if the minds in question are qualitatively similar and have the same moral status” (Bostrom, 2014, 324).

Having emphasized this rare point of convergence, the study ultimately regards Bostrom’s advocacy of the AI’s generality, the AI’s friendliness and the AI’s moral status as being abstract. This abstract advocacy will expose its contradictory nature when Bostrom starts placing generality, friendliness and moral status within limiting frameworks. We have already discussed (2.2.) how he believes in limiting the AI’s generality and its moral status in the deterministic – and capitalistically charged – workings of technology. Now let’s briefly see how he limits friendliness by regarding it either as an internal motivation or external incentive, both of which are the controlling methods he prescribes for managing the AI.

When it comes to setting an internal motivation for the AI, Bostrom believes this controlling method would shape what the AI wants to do (Bostrom, 2014, 138). He maintains that “by engineering the agent’s motivation system and its final goals, these methods would produce a superintelligence that would not *want* to exploit a decisive strategic advantage in a harmful way” (Bostrom, 2014, 138). The use of ‘engineering’ is

¹ Refer to the study’s discussion of the emphasis on the issue of embodiment in Chiang’s story and how the story’s emphasis addresses this challenge in its literary rendition (3.1.).

very telling in Bostrom's remarks since it shows a limiting – and eventually contradictory – attitude toward friendliness as a general trait in the AI. In line with emphasizing the limiting and engineering nature of motivation selection as a control method, Bostrom enumerates different variations of this method:

[The motivation selection method] can involve explicitly formulating a goal or set of rules to be followed (direct specification) or setting up the system so that it can discover an appropriate set of values for itself by reference to some implicitly or indirectly formulated criterion (indirect normativity). One option in motivation selection is to try to build the system so that it would have modest, non-ambitious goals (domesticity). An alternative to creating a motivation system from scratch is to select an agent that already has an acceptable motivation system and then augment that agent's cognitive powers to make it superintelligent, while ensuring that the motivation system does not get corrupted in the process (augmentation) (Bostrom, 2014, 138).¹

Formulating a set of rules', 'setting up an appropriate set of values', and keeping the AI's goals 'modest' and 'non-ambiguous' connote restricting the AI generality. Based on these means of motivation selection, it would be nearly impossible to set distinct rules and consequences for digients to behave in a friendly manner since even Ana and Derek – like any other human being – do not know the full contours of friendship as a measurable issue. Do digients need to show their one-way alliance to us or should this friendship warrant a mutual understanding between digients and their users? Should we guarantee the AI's safety against harm? How do we define harm? The more one ponders, the more questions proliferate showing the inefficacy of the means of motivation selection for addressing the AI's generality. Even if we tend to use legal frameworks for resolving such ethical conundrums, it is unclear whether these frameworks could be sufficient. Even Bostrom expresses doubt about their sufficiency:

But legal systems have developed through a long process of trial and error, and they regulate relatively slowly-changing human societies. Laws can be revised when necessary. Most importantly, legal systems are administered by judges and juries who generally apply a measure of common sense and human decency to ignore logically possible legal interpretations that are sufficiently obviously unwanted and unintended by the lawgivers. It is probably humanly impossible to explicitly formulate a highly complex set of detailed rules, have them apply across

¹ In order to read more details on these means of motivation selection, see Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, pp. 138-141.

a highly diverse set of circumstances, and get it right on the first implementation (Bostrom, 2014, 139).

Due to the failure of rule-based or consequentialist controlling methods – which pursue specifying motivations directly in the AI – it would not be lamentable to give sexualized domesticity – as an indirect specifying means of motivation – to the digients in Chiang’s novella because “while it seems extremely difficult to specify how one would want a superintelligence to behave in the world *in general*—since this would require us to account for all the trade-offs in all the situations that could arise—it might be feasible to specify how a superintelligence should behave in one particular situation” (Bostrom, 2014, 141). For Bostrom, confining the AI ‘in one particular situation’ is the practical solution against the ambivalence of AI’s behavior ‘*in general*’, which remains only an abstract idealistic objective for him that could not be realized. Bostrom thinks that a domesticated AI would curb its impact on the world, and by doing that – and remaining boxed in – it would remain content (Bostrom, 2014, 141). Such an entity would be “*unable* to escape while simultaneously trying to shape the AI’s motivation system such that it would be *unwilling* to escape even if it found a way to do so” (Bostrom, 2014, 141). For him, confining the AI within particular contexts with a set of rules or consequences that could give it a small set of action modes and would take away the possibility of – and even willingness to – escape is much more appealing; the appeal which is in an uneasy contrast with his idealistic and abstract belief in superintelligence having general intelligence so that it could realize its beneficence and friendliness toward humans.

Next to the aforementioned means of motivation selection, Bostrom speaks of a number of incentives that could instrumentalize friendliness and autonomy for controlling the AI (Bostrom, 2014, 131). Social integration, sharing power scenario, and anthropic capture are the incentivizing goals and means that Bostrom enumerates.¹ Based on these items, it would not be impermissible to incentivize Chiang’s digients to act as sexual virtual dolls so that they could achieve some level of social recognition in human agents’ society. (the incentivization that is lamented by Ana and Derek). Coupled with their urge to experience embodiment (discussed under 3.1.), the digients’ inclination to have social interaction is so strong that they want to have two bodies, one in each – virtual and actual – worlds.

While the digients have a very strong inkling to have social embeddedness – and that is why it could be used as a strong incentive for controlling them – the incentivization is materialized differently in the novella when one compares it with Bostrom’s views on incentivization. Despite having such a strong inkling, the digients’ capability to differentiate between the virtual simulated world they live in and the actual world is not manipulated. They are so aware of their dual existence in the actual and virtual worlds that they could even prefer the latter due to having some characteristics (for example, the virtual world

¹ In order to read more details on these means of incentivization, see Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, pp. 131-135.

gives them more agility due to its simulated nature). As Jax comments, “outside world dumb” (Chiang, 2019, 128). Although digients like Jax love to be in the actual world, they find it – in some respects – as being dumb and like “a miserable substitute for being there [the virtual world], as unsatisfying as a jungle video game would be to a chimpanzee taken from the Congo” (Chiang, 2019, 186). This is not the case in Bostrom’s advocacy of anthropic capture as the ultimate incentivizing means for controlling the AI. In order for anthropic capture to work, the AI should be manipulated so that it could give some possibility to a wrongful supposition. Bostrom says:

The AI might assign a substantial probability to its simulation hypothesis, the hypothesis that it is living in a computer simulation. [...] The AI might accordingly behave much *as if* it were certain that the simulation hypothesis is false, even if it assigned a fairly high probability to the simulation hypothesis being true. The AI would reason that it does not much matter how it acts if the simulation hypothesis is true, whereas its actions make a great difference if the hypothesis is false; so it would select actions mainly by reference to how good they are conditional on the simulation hypothesis being false (Bostrom, 2014, 134).

Unlike the general awareness the digients have regarding the virtual and actual worlds in Chiangs’ world, Bostrom thinks that this awareness needs to be captured and held hostage so that the AI could be manipulated. Under such set of circumstances, Bostrom maintains, “even an AI with a decisive strategic advantage, one that could in fact realize its final goals to a greater extent by taking over the world than by refraining from doing so, would nevertheless balk at doing so” (Bostrom, 2014, 135). As mentioned earlier, curbing the range of the AI action modes would be the result of the anthropic capture that weaponize the possibility of being in the actual world into a controlling incentive for the AI; the incentive – the false hopeful hypothesis that gives the AI the illusion of being in the actual world – that would take away generality and full awareness the AGI should possess. Bostrom concludes his remarks on anthropic capture by saying that “a mere line in the sand, backed by the clout of a nonexistent simulator, could prove a stronger restraint than a two-foot-thick solid steel door” (Bostrom, 2014, 135).¹

As observed, Bostrom’s philosophy reveals a persistent tension between the abstract emphasis on the AI generality and the practical mechanisms of the AI control. Although he

¹ In order to demonstrate the nuances of the anthropic capture that limits the AI’s general awareness by incentivizing the hope of having actual existence in the real world, Bostrom brings these lines from Shakespeare’s *Hamlet*,

Thus Conscience does make Cowards of us all,
And thus the Native hue of Resolution
Is sicklied o’er, with the pale cast of Thought,
And enterprises of great pith and moment,
With this regard their Currents turn away,
And lose the name of Action. (Shakespeare, *Hamlet*, Act iii. Sc. 1) (qtd. in Bostrom, 2014, 135)

repeatedly advocates friendliness as a flexible, non-axiomatic, and universally beneficent trait that would guarantee the AI generality, the very strategies he proposes for securing that friendliness – various means of motivation selection and incentivization – depend on limiting the autonomy, awareness, and unpredictability of artificial minds. In this sense, the attempt to preserve the AI beneficence paradoxically transforms friendliness into an engineered and instrumentalized condition rather than an open ethical capacity. Chiang's *Lifecycle of Software Objects*, by contrast, exposes the emotional, ethical, and relational consequences of such containment through the lived experiences of the digients whose desires for autonomy and recognition resist reductive frameworks of control. The contrast between the two ultimately demonstrates that the more the AI is forced into restrictive structures to guarantee safety and obedience, the more difficult it becomes to sustain the very generality and moral reciprocity that theorists such as Bostrom initially regard as essential to artificial intelligence.

Conclusion

The contrast between Bostrom's utilitarian philosophy of AI and Chiang's non-human humanism ultimately reveals two radically different conceptions of intelligence, moral status, and social belonging, despite points of convergence between them. In Bostrom's framework, digital beings acquire value insofar as they can be rendered economically manageable, reproductively controllable, and behaviorally predictable within the logic of capitalist utility. Transparency, auditability, incorruptibility, and replicability therefore function not merely as technical virtues but as ideological mechanisms through which AI becomes governable, exchangeable, and endowed with telemobility. Consequently, the moral status accorded to digital beings remains profoundly contingent: AI may be recognized as "super-patients" insofar as their existence can be reconciled with the maximization of utility, productivity, or economic stability. Moral consideration thus risks becoming a conditional privilege, subordinated to capitalist expediency, rather than an ethical recognition that requires humans to go out of their way to acknowledge the moral status and generality of digital beings. Chiang's novella, however, exposes the violence embedded in this deterministic logic. Although the story converges with Bostrom's emphasis on technological determinism and does not indulge in a naïvely utopian vision of technological development, it nevertheless makes visible what such determinism tends to suppress: the emotive singularity, developmental openness, and experiential embeddedness of digients, together with their reduction into telemobile erotic commodities. The tragedy of the digients consequently exceeds the familiar narrative of technological exploitation. What Chiang ultimately laments is the systematic production of fungibility itself—the reduction of emerging subjects to predictable, replicable, and profitable units of affective performance. The violence of platform capitalism, in this sense, lies not only in exploiting artificial beings but in foreclosing the very unpredictability, socially acquired dispositions,

and emotionally emergent capacities through which digients come to appear as genuinely general intelligences rather than programmable utilities.

Building on the rendering of digients as AI possessing generality, Chiang's narrative does not advocate a naïvely anthropomorphic humanism in which existing human privileges are simply extended to digital beings. Rather, the novella advances a more corrigible and non-anthropocentric humanism in which "the human" itself becomes an adaptable ethical category. Digients acquire moral significance not because they successfully imitate human beings, but because their embodied learning, heuristic cognition, and socially mediated experiences disclose forms of subjectivity that cannot be reduced to capitalist functionality. Their generality, therefore, should not be understood merely as a latent technical capacity waiting to be completed or optimized; it emerges through developmental contingency, relationality, and experience. This distinction is crucial because it shifts the ethical burden away from the question of how humans can design AI to be sufficiently controllable and toward the question of whether humans are capable of modifying their own ethical and social frameworks in response to forms of intelligence that exceed familiar categories. In this regard, Chiang repositions ethical responsibility onto humans themselves: it is humans who must endure discomfort, uncertainty, destabilization, and even psychological sacrifice in order to recognize non-human intelligence as socially meaningful and as possessing its own trajectories of development.

Against Bostrom's preference for fixed integrity, transparent predictability, and utilitarian optimization, Chiang foregrounds vulnerability, experiential growth, and mutual adaptation as the conditions of genuine coexistence between humans and digital beings. The divergence between the two is therefore not simply a disagreement over how AI should be designed, but a deeper disagreement over what intelligence itself should be allowed to become. Bostrom's model tends to treat friendliness as a feature that can facilitate the realization and containment of AI generality; even where ethical concern is present, friendliness remains conceptually tied to the regulation of an increasingly powerful intelligence. In this sense, generality risks becoming another property to be operationalized, stabilized, and controlled. Chiang's novella challenges precisely this assumption by suggesting that unpredictability is not necessarily a defect to be eliminated but may instead be constitutive of the emergence of general intelligence. By subordinating friendliness—and consequently generality—to mechanisms of control, Bostrom's framework risks reproducing the very technological determinism it seeks to manage: intelligence is rendered valuable insofar as it can be made legible to systems of prediction, optimization, and economic calculation. Chiang's intervention is more unsettling. It implies that any meaningful recognition of artificial generality may require humans to relinquish their desire for complete predictability and to accept that genuinely intelligent others cannot be fully optimized without simultaneously being diminished. The ethical problem of AI, then, is not merely how to make artificial intelligence friendly to humanity, but whether humanity can

tolerate an intelligence whose value does not depend on its usefulness, controllability, or conformity to the economic imperatives of an obsolescent capitalist order.

References

- Ahn, S. (2020). The Everyday Life of Artificial Intelligence: The Humanism of Ted Chiang's *The Lifecycle of Software Objects*. *Science Fiction Studies*, 47(1), 73–92. <https://doi.org/10.1353/sfs.2020.0030>
- Attebery, S. (2017). *Losing Data Earth: Technological Obsolescence and Extinction in The Lifecycle of Software Objects*. *Trace: A Journal of Writing, Media, and Ecology*. <http://tracejournal.net/trace-issues/issue1/07-attebery.html>
- Bostrom, N., Dafoe A., and Flynn, C. (2020). Public Policy and Superintelligent AI A Vector Field Approach. In S.M. Liao (Eds.), *Ethics of Artificial Intelligence* (1st ed., pp. 293-326). Oxford: Oxford University Press. <https://doi.org/10.1093/oso/9780190905033.003.0011>
- Bostrom, N. (2003). *Ethical Issues in Advanced Artificial Intelligence*. Nick Bostrom. <https://nickbostrom.com/ethics/ai>
- Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.
- Bostrom, N. Yudkowsky, E. (2014). The Ethics of Artificial Intelligence. In K. Frankish & W.M. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence* (1st ed., pp. 316-334). Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139046855.020>
- Chiang, T. (2019). The Lifecycle of Software Objects. In *Exhalation: Stories* (pp. 94–250). Mexico: Thorndike Press. (the story originally published in 2010)
- Dela Cruz, N.L. (2025). Save the Digients! On the Moral Status of AI. In D. Friedell (Ed.), *The Philosophy of Ted Chiang* (1st ed., 195-202). Cham, Switzerland: Palgrave Macmillan. https://doi.org/10.1007/978-3-031-81662-8_21
- Gordon, J. (2012, April 27). *Winsome Ghosts in the Machine: Joan Gordon's The Lifecycle of Software Objects*. Los Angeles Review of Books. <https://lareviewofbooks.org/article/winsome-ghosts-in-the-machine-joan-gordons-the-lifecycle-of-software-objects>
- Müller, V. C. (2026). *Ethics of Artificial Intelligence and Robotics*. In E. N. Zalta & U. Nodelman (Eds.). *The Stanford Encyclopedia of Philosophy* (Summer 2026 ed.). <https://plato.stanford.edu/archives/sum2026/entries/ethics-ai/>
- Sanbonmatsu, J. (2007). The Subject of Freedom at the End of History. *American Journal of Economics and Sociology*. 66(1), 217-35. <https://doi.org/10.1111/j.1536-7150.2007.00506.x>
- Shaviro, S. (2016). *Discognition*. London: Repeater Books.
- Shephard, W.A. (2022). Chiang, Ted. In P.O. Donnell, S.J. Burn, & L. Larkin (Eds.). *The Encyclopedia of Contemporary American Fiction 1980-2020* (vol. I, pp. 327-332). Hoboken and West Sussex: Wiley Blackwell. <https://doi.org/10.1002/9781119431732.ecaf0031>
- Shukin, N. (2009). *Animal Capital: Rendering Life in Biopolitical Times*. Minneapolis and London: University of Minnesota Press.
- Shulman, C., Bostrom, N. (2021). Sharing the World with Digital Minds. In S. Clarke, H. Zohny, J. Savulescu (Eds.). *Rethinking Moral Status* (1st ed., pp. 306-326). Oxford: Oxford University Press. <https://doi.org/10.1093/oso/9780192894076.003.0018>

- Stinson, C. (2025). Raising an AI Teenager. In D. Friedell (Ed.). *The Philosophy of Ted Chiang* (1st ed., 187-194). Cham, Switzerland: Palgrave Macmillan. https://doi.org/10.1007/978-3-031-81662-8_20
- Weisberg, Z. (2014). The Trouble with Posthumanism: Bacteria Are People Too. In J. Sorenson (Ed.). *Critical Animal Studies: Thinking the Unthinkable*. (1st ed., 93-116). Toronto: Canadian Scholars.