The Significance of a Second-Person Perspective for the Development of Humanoid AI

Hanlin Ma(✉)

Suzhou University of Science and Technology, Suzhou, China
fanfan2011cn2000@gmail.com

Abstract. Anthropocentrism and non-anthropocentrism are a pair of basic directions in cognitive science and AI research. These directions correspond respectively to "first person" and "third person". The confrontation between these two concepts not only has guiding significance for the exploration of cognitive science but also has ethical importance. Human-like AI (for instance, Artificial General Intelligence, AGI for short) research hovers between them. I propose a second-person, or limited consensus perspective as a philosophical and ethical research context that is integrated with the first and third person to constitute a "trinity" relationship—what I call the "trinity of the contexts of intelligent artificiality", which could be represented by the relationship between emulation, simulation, and imitation. The consequences of the general loss of the first-person context will be explored, including the uncertainty of the ethical landscape of humankind—such as the fragmentation of the self—and the risk of going against the original intention of AGI research. This cannot be fixed by returning to a first-person perspective. The second-person context is an idea that is situated between anthropocentrism and non-anthropocentrism, and it may be employed to avoid these risks.

Keywords: Imitation · Second-person perspective · Anthropocentrism · Non-anthropocentrism

1 Introduction

A recent episode that highlighting the topic of humanoid AI is the story of Google's "The language Model for Dialogue Application (LaMDA)". A test engineer, Blake Lemoine claims that this chatbot system displays sentience and even "wants" to be respected. Including Google, probably most of the AI researchers would not share any consensus with Blake Lemoine. Microsoft Chief Data Scientist Juan M. Lavista Ferres says: "LaMDA is just a very big language model with 137B parameters and pre-trained on 1.56T words of public dialog data and web text. It looks like a human because is trained on human data". One reason why people don't believe it has evolved sentient AI is that the framework contains no intention of creating a self-cognitive structure.

Building something that "acts like a human" actually has not been the main goal in the AI industry recently, even though it always catches the attention of the public. This phenomenon shows that the humanoid AI standard inspired by Turing Test does not

M. Luo and L.-J. Zhang (Eds.): EDGE 2022, LNCS 13732, pp. 49–63, 2022.
https://doi.org/10.1007/978-3-031-23470-5_5

get agreement from the realm of industry. This may be caused by the fact that making a Humanoid AI has been largely abandoned by most scientists. Even though LaMDA talks like a human who seems to know who it is, it still is considered a coincidental emulation.

Nowadays, researchers in the realm of Artificial General Intelligence (AGI) are working on another direction in making humanoid AI – to design or produce a humanoid cognitive structure. But would this kind of AI simulation necessarily become an existence that could get along with real human beings, given the case that it probably does not physically look like a human in any way nor have a life that is anything like a human being's If AGI was like this, then the research into social AI would appear redundant. Philosophically speaking, when we focus on social AGI research, not just normal social AI study, merely staying with the idea of emulation and simulation is not enough, the introduction of the concept of imitation from the Theory of Mind could be a productive avenue to explore.

2 The Definitions of Emulation, Simulation, and Imitation

Please consider the following three cases:

1. When I mimic my cat's meow, he does not react to me.
2. The team of Markram (2015), which is affiliated with the Human Brain Project (HBP), builds "a digital reconstruction and simulation of the anatomy and physiology of neocortical microcircuitry that reproduces an array of in vitro and in vivo experiments without parameter tuning and suggests that cellular and synaptic mechanisms can dynamically reconfigure the state of the network to support diverse information processing strategies" (Markram et al. 2015).
3. One of the most influential primate researchers, Tetsuro Matsuzawa, can "speak chimp" with his experimental subjects. On BBC Earth's YouTube channel, you can see a video of him greeting the chimpanzees and getting responses.

In case 1, I present an action called emulation, by which I wish my cat would subjectively believe that I am communicating with him by presuming that my consciousness is similar to his. I do not know what I precisely express, however. In this case, I am trying to build a "sense of another mind" in my cat's mind, which is first person. Besides, the picture of my mind in his mind is also presumed to be judged as the first person. However, I fail.

In case 2, the HBP team processes a digital simulation of a mouse brain by reconstructing the physiology of neocortical microcircuitry. Nevertheless, the simulation can be oriented to other varieties of cognition, rather than staying with physicalism only. According to the introduction by Kotseruba and Tsotsos (2020), there are at least 195 different cognitive architectures featured in 17 sources in the last 40 years. All of these architectures could be implemented in AI applications. However, except for the model of HTM (hierarchical temporal memory; see Hawkins and Blakeslee 2004), which shares the elementary philosophical background of intelligence with HBP, there are lots of different styles of cognitive architectures that can be simulated, such as EPIC (Kieras 2017), represented by a symbolic system, ACT-R (Anderson and Lebiere 2003), Soar

(Laird et al. 1987) featured as hybrid systems, etc. They are platforms to represent the access consciousness (A-consciousness) of Block (1995).

Comparing case 1 with case 2, in case 1, my emulation was committed to eliciting a first-person belief in another mind, namely that I have a phenomenal consciousness (P-consciousness). In case 2, the simulations for cognitive research were committed to establishing a series of third-person cognitive states that can serve any available purpose, e.g., to represent the behavior of greeting. All inner informational structures of these cognitive states can be shared, replicated, and re-established. However, let us consider case 3, in which Tetsuro Matsuzawa interacts with chimps by imitating them. He knows the meaning of his chimpanzees' roar and how the chimps would react to his imitation.

The understanding of what was represented by case 1, case 2, and case 3 can share the same series of actions from different perspectives. By drawing a parellel with the cover illustration of Gödel, Escher, Bach: An Eternal Golden Braid (Hofstadter 1979), I exhibit my expression of the relationship between emulation in case 1, simulation in case 2, and imitation in case 3 in Fig. 1. This shows that they cannot be reduced to each other, but rather represent different aspects of one individual process, which I call the "trinity of the contexts of intelligent artificiality" (TCIA). Nevertheless, having different understandings determines which direction we are heading in when we realize human minds through computers. In other words, when we use the first person (context of emulation), the third person (context of simulation), or the second person (context of imitation) as the pivot to construct a certain value parameter in the development of AI, we will face different ethical and other meaningful implications. My suggestion is to choose the context of the second-person perspective. The most important reason is that different perspectives represent different understandings and standards for agents. In Sect. 2, I will indicate that in the realm of HCI (human-computer interaction), the pivotal status of the context of emulation—or what I call Turingism—has been arrogated by the context of affordance, which may cause the fragmentation of self. In addition, unlike the operation of HCI, the research into artificial general intelligence (AGI) is supposed to explore the essence of self, while the contagion from the context of affordance could affect any neutral context of simulation, including AGI research. In Sect. 3, my concern goes toward the profound ethical significance of simulation proposed by Floridi (2013). I will charge his theory with a dilemma, then suggest a second-person context to explore AGI, which is neither anthropocentric nor non-anthropocentric.

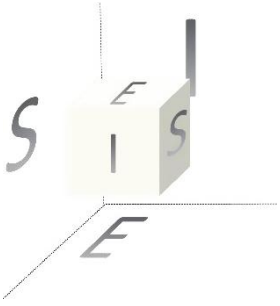

Fig. 1. .

3 Abandoned Emulation

Although emulation, simulation, and imitation may partly share features as behaviors, their intentions are entirely dissimilar in the sense of practical goals. Let us start with emulation. According to Merriam Webster's Dictionary, emulation means "ambition or endeavor to equal or excel others (as in achievement)". Accordingly, behavioral emulation is a kind of "endeavor to equal action" that aims to achieve behavioral sequences or goal-oriented behaviors. In a way, the Turing Test shows a typical example of the emulation of human actions, which tricks a person into believing that they are talking to a human (Turing 1950). This means that Turingism is typically anthropomorphic. Moreover, the ELIZA effect proves that people will unconsciously and emotionally respond to the ELIZA computer conversation program by ascribing a human mentality to it (Weizenbaum 1966). Since the ELIZA effect or something similar (Seibt, Nørskov, and Hakli 2014) make it possible to obtain available feedback from humans, anthropomorphization becomes an aspired destination in the realm of social robots (Duffy 2003). On the other hand, the "Uncanny Valley" (Mori et al. 2012) proves that some emulations may not deploy desirable anthropomorphization, even producing uncomfortable feelings in experimental participants. Therefore, avoiding uncanny robotic design appears to be a theme of HCI. Currently, from an HCI point of view, it seems that whether people believe that robots are "thinking" is not the crucial issue. Rather, fitting their cognitive or social expectations for specific situation of interaction is concerned. How to manipulate human perception to adapt to the interface design has become a hot research topic in HCI and social robotics (Vallverdú et al. 2016).

Has research in the area of AI and human interaction directed by Turingism been abandoned? If so, for what reason? The answer to the first question is yes. The philosophy of HCI only borrows the appearance of Turingism by embracing the idea that participants are thought of as perceiving the interface or "behavior of the system", as if they are demonstrating some features or capacities of intelligence. In fact, in HCI, the important thing is how participants respond to the interface of the system, not what they consciously believe about the system. The underlying idea here may not involve total behaviorism, but is close enough. Most HCI research revolves around the conception of affordance, which can be used as the foundation of our analyses. This term was initially proposed by Gibson (2014) within his ecological approach to perception. It was introduced to HCI by Norman (1988) and quickly became a widely accepted keyword among researchers. Due to its unlimited potential, the significance of "affordance" in HCI has been developed into different branches (Kaptelinin and Nardi 2012). To avoid confusion, I will stay with its original meaning.

Briefly, affordance almost equals "interaction". As point 5 below will demonstrate, affordance mostly refers to the conditions in interaction that emphasize living beings' abilities and features of the environment, rather than the mental or physical events occurring in that situation. For example, to understand affordance, we can imagine living beings' ability to define features of environmental situations. Just as Gibson mentioned, "stand-on-able" means "permitting an upright posture for quadrupeds and bipeds". By imitating him, we can define "radar-scan-able" as "permitting an ultrasonic reception to detect the spatial conditions for flying" for the affordance between bats and cave. It

seems that affordance represents the possibility of the behavior that occurs in the interaction between a living being and its circumstances. According to Kaptelinin and Nardi (2012), Gibson's theory of affordances can be summarized as follows:

1. Affordances are perceived directly; their perception is not based on an interpretation of initially meaningless "raw" sensory data; 2. Affordances are relational properties; they emerge in the interaction between the animal and environment: the same environment may offer different affordances to different animals (or humans); 3. Affordances are independent of the situational needs of the perceiver; 4. Natural environments and cultural environments should not be separated from one another; 5. The theory of affordances is concerned with how affordances are perceived rather than affordances per se.

For the convenience of analysis, let us apply the framework of affordance to Turingism. First of all, the only purpose of Turing Test deployment is to achieve specific beliefs within the human's mind by AI which emulates the human.. This concerns the perceiver's interpretation of sensory data. Second, the immediate relationship between the human and the AI in the emulation test is worthy of attention, but there is no formal permanent *relationship properties* to be concerned about. Furthermore, the situation with Turingism seems to be a second-order or overlapping relation. And, as perception addresser, emulation targets the perception of the addressee. This is a dimension of the second-person situation that does not involve consensus between the addresser and the addressee. The ideas described in points 4 and 5 do not apply to our current analysis.

If the practical field of Turingism is taken over by the theory of affordance, which is a radical embodied theory, then what happens subsequently? This is where my Theory 1(T1) of humanoid AI or robot development comes in:

T1. The result of affordance theory's replacement of Turingism is the split of human cognition of the self and all possible ethical consequences this brings. However, this practical problem cannot be solved through a purely theoretical return to Turingism.

The application of affordance theory to HCI is committed to controlling the uncertainty of the interaction. Another dissimilarity between affordance and Turingism is that the self of another mind will never be considered a cognitive object in the affording relationship. Nevertheless, the uncertainty that is driven by different types of selves and the referential psychological distance will be committed as parameters to quantify the Turingism situation.

According to Knobe and Nichols's (2011) outstanding work, three types of selves can be detected corresponding to psychological distance in idealized conditions. These are: the physical concept of the self, the psychological concept of the self, and the executive concept of the self. I am concerned with the last two. The psychological concept of the self refers to mental processes in the sense of psychology or cognitive science, which are the kinds of "mental events" that philosophers Davidson (2001) and Kim (1973) worked on. The corresponding causal relationship is mental causation. The concept of self-implementation refers to an "agent"; a similar "subject" commonly used in the history of philosophy. Its metaphysical properties are not limited by physical time and causality (in contrast to the psychological concept of the self). The corresponding causal relationship is agent causation (Clarke 1993; O'Connor 1996). Moreover, according to Construal Level Theory (CLT; see Trope and Liberman 2010), the recognition of events

with a greater psychological distance is more abstract and essential (which means a high construal level); the recognition of events with a shorter psychological distance is more specific and direct (which means a low construal level). The characteristic of psychological distance is not fixed, although mainstream items are time (close or not), space (close or not), the closeness of social relationships, and real or counterfactual events.

I will use human-likeness as the parameter of psychological distance, which refers to the trends of probabilities of whether participants are more likely to treat the objective robotic interface as an agent or mental self. By employing the analysis of the "Uncanny Valley", I will clarify how the "fragmentation of human cognition of the self" in T1 happens. As Mori described, the Uncanny Valley refers to the fact that "in climbing toward the goal of making robots appear like a human, our affinity for them increases until we come to a valley, which I call the uncanny valley" (Mori et al. 2012, p. 98; Mori (1970); see Fig. 2). In the last 50 years, dozens of explanations have been posited (Wang et al. 2015). Because I am inclined to the Theory of Mind, my explanation stays with Gray and Wegner (2007, 2012). According to their experiments, participants were more likely to respond to the robotic interface as an agent rather than as an experiential sSelf. Experience is thought of as a more essentialized quality than agency, in the sense of treating robot as human-like being. Thereby, "if machines are held to be essentially lacking experience, then an appearance which suggests this capacity—i.e., human-like eyes that convey emotion (Adolphs et al. 2005)—could conflict with this expectation and therefore be unsettling" (Gray and Wegner 2012, p. 126). This explains why the Uncanny Valley occurs, but it does not mean that machines will never be treated as if they have some experience (Waytz et al. 2010), as indicated by the peak before the valley. Therefore, zoomorphic or other caricatured design appears to be useful to avoid the Uncanny Valley (Fong et al. 2003).

What I want the reader to take note of here is that to treat the object as a more experiential self does not presume that the participants are attracted by more experiential details of the object. Sometimes, a well-deployed cartoon character in the sense of affordance may be accepted as a more vivid object. By virtue of this, participants' cognition of this carton character will become a "zoom-in case", in contrast to, participants' cognition of a normal and healthy man, which will become a "zoom-out case", in which participants do not pay close attention to their objects. Accordingly, we may combine the quantification of the psychological distance of human-likeness and the psychological distance of the measure of accepted experience from the objects (y-axis in Fig. 2).

According to Knobe and Nichols (2011), at a short psychological distance (zoom-in case), "people are willing to take a thick view of what counts as the self". In this case, the average score comparison between agent self and experiential self is 4.95/5.48. At a long psychological distance zoom-out case), on the other hand, "people are inclined to think of the self as a thin executive self", and the average score comparison is: 6.10/3.95. This means that in the affordance of zoomorphic or some caricatured cases, people are more likely to treat the self of the cognitive object as an agent. The agency of the robot may not equal to the agency of humans, however. In other "zoomed-in" research of Knobe and Nichols, comparing computers as a determined cognitive process with humans, participants were more inclined to judge that no alternative possibility (AP) could be

attributed to the computer. This means that participants are more likely to treat robots as controlling Selves but not agents with complete free will in the sense of libertarianism.

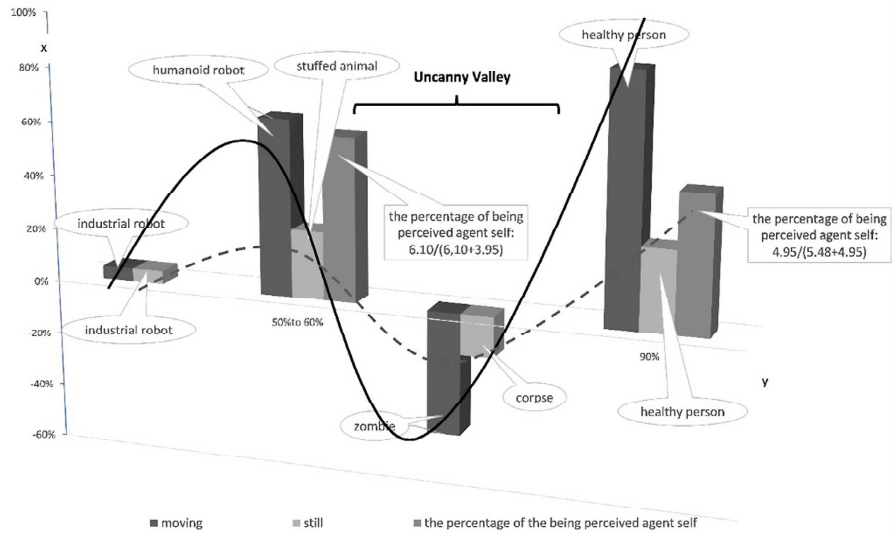

Fig. 2. x-axis represents the percentage of affinity or counterpart of the percentage of being perceived agent self, and the y-axis represents the percentage of the psychological distance of the perceiver from industrial robot to a healthy person

With the context of being trapped in the situation of designed affordance, if we follow Luciano Floridi's (2013) proposition that treating AI as a moral agent to some extent without possessing any libertarian-type free will (Kane 1998), then we will face at least two types of splits in the self: the split between the experiential self and agentic self, and the split between self with AP and controlling self. Therefore, it seems that the human HCI industry has a significant opportunity to create the type of humanoid robot with executive selves similar to humans'. Meanwhile, such selves are considered to have no passion or freedom. Furthermore, the affordance that characterize these selves may be used to eliminate some uncertainty in human life in their interaction with computers. Thanks to this, some potential evolutionary directions of memes may be eliminated. Another fact makes this conclusion look even more disquieting: due to the huge impact of modern communication methods on human society (which was catalyzed by Covid-19), the the psychological distance between people and the "psychological distance" between people and machines are becoming more and more equal. The problem here is not what the computer or any interface in that affordance has deflected in humans' cognition, but rather what future this kind of HCI would create for humans. A possible situation may be that humans will be trapped under the control of HCI, which is not a dictatorship but stronger and more effective than a dictatorship.

Another of the subsequent thought-provoking ethical problems is whether people will perceive the bullying by any kind of embodied interface in the same way as the bullying by other people. Some experimental (Keijsers 2020) or theoretical studies (Kissel 2020)

have appeared. What we should beware of, however, is that when we start research of this kind, we automatically get back to the Turingism standpoint. Nevertheless, these operations in the philosophy of HCI may re-mask the problem rather than solve it. I will explain this after elaborating the second-person standpoint of the design philosophy of AI. This is important because otherwise, we would not know whether our research conclusion of the previous problem is founded on a correct approach.

4 Simulation for What?

The fields of HCI and AGI intersect in only a limited way. An HCI system may only apply part of the functions of an AGI system. The dissimilarity between the application of an AI technology and the construction of an AGI system is that the former simulates for practical goals. In contrast, the latter simulates to understand human cognition. Nevertheless, I prefer to describe AGI as simulating for *imitation.* To understand what I mean, we should employ Susan Hurley's work on understanding simulation (Hurley 2008). She distinguishes two ways of defining simulation in terms of *resemblance* and *reuse.*

According to Alvin Goldman (2006, p. 37; see also Hurley 2008, p. 756), the resemblance can be paraphrased as follows:

Process P is a simulation of process P' if, by definition, P duplicates, replicates, or resembles P' in significant respects, and in doing so fulfills a purpose or function.

Philosophers like Bostrom (2003) and sci-fi movie directors like the Wachowskis (*The Matrix*) may share this definition when they use the concept "simulation", but they more likely agree with the scientists who posit a *theoretically driven simulation*, which is *intrapersonal* and represented in case 2 of the introduction above. According to Goldman, "A computer may simulate wind flow around a suspension bridge by applying the laws of relevant theory to derive a symbolic description or graphical representation that is about the target. It does not use or undergo the same type of processes as its target undergoes, but rather represents the target processes, using another mechanism" (Goldman 2006, pp. 35–36; see also Hurley 2008, p. 758). Evidently, all AGI researchers share this definition of simulation.

Nevertheless, when it comes to *interpersonal mindreading*, the simulation may become "reuse", in terms of appealing to the *Simulation Theory (ST)* of mindreading. As Hurley and Goldman stated, ST proponents believe that in mindreading, people ply their own mental elements to rebuild the mental state of other minds. However, there is another mindreading theory named *Theory Theory (TT)*, in which people read other minds by *interpersonal theoretically driven simulation* in the approach of theorizing the information exploited by the experience of the other agents. No matter whether ST or TT is right, they both represent an interpersonal resemblance that I name "*imitation*", similar to what Tetsuro Matsuzawa does with his chimps.

In this case, it seems that AGI designers' works are dissimilar from the landscape of the human mind described by mindreading theorists, although the latter may have a chance of being successful in the HCI field based on the former. Nevertheless, some AGI designers are not optimistic about humanoid AI. Jeff Hawkins says: "I don't believe we will build intelligent machines that act like humans, or even interact with us in human-like ways" (Hawkins and Blakeslee 2004, p. 206. The common purpose of AGI study

is to probe the essence of intelligence, whose range may not be restricted to human cognition. Since not all characteristics of human intelligence are necessary, (and some may be accidental products brought about by evolution), AI researchers need only focus on the intellectual features that are rationalized by cognitive science to structure the AI system. For example, NARS (an AGI architecture designed by Wang Pei) sets emotion as a crucial feature that helps control the action selection and decision-making processes (Li et al. 2018). This setting has already exhibited an adequately human-like and rationalized theorization of emotion, even though it falls only under the intrapersonal frame. In contrast, ST scholars may disagree with it in the light of some empathic emotion, which can be explained by the mirror neuron mechanism. They may believe that sometimes people feel pain and sadness only because others' miserable situation evokes these emotions. If empathy is an automatic mechanism of humancognition, could we just ignore it, even though it has survived through the accidental process of evolution?

Through answering this question, we will find that at least part of my reason to elevate the significance of a second-person perspective on AI development is driven by ethical considerations. Here is my Theory 2:

T2. General artificial intelligence research under the third-person perspective is necessary to explore the essence of self, but if the second-person knowledge of imitation is not used for reference, AGI will probably proceed in a non-anthropocentric direction.

It is not possible to establish without any ethical contradiction an AI with a self by simulating the human mind on the one hand, while on the other hand not building a human-like contact approach with the person in the system. Newell (1994, p. 19) proposes thirteen constraints that shape the mind.

It is odd that John R. Anderson chooses to omit point 9 (the feature of social community) when he applies these principles to test his famous ACT-R system, which is partly based on Newell's AI philosophy (Anderson and Lebiere 2003). Anderson may be thinking along the same lines as Hawkins, believing that ACT-R does not need to interact with humans.

Nevertheless, another spontaneous answer is to say that if the other 12 characters have been realized, then the last one may be subsequently achieved. Then, two things may occur. The first one is that machines may develop their own social entrainment without connecting with human society. This is not just fiction. Wang Pei pictured this landscape of AI species in light of his AGI predictions at the 12th International Conference, AGI 2019. If the AI species live in a parallel dimension away from ours, then nothing ethical needs to be said about it. However, what if they would have some connection with human society? Then, "how should we deal with this situation?" will be a question. If this occurs, then what I mentioned in T1—that the problem of HCI with affordance cannot be solved through a purely theoretical return to Turingism—applies to our debate right now.

Previous research has shown that people can feel empathy for robots when the robot is mistreated (Darling et al. 2015). In an interesting study on robotic bullying, Keijsers (2020) concludes that people would apply the same mind attribution to robots and humans, because they feel unacceptability in the case of both robotic and human bullying. This is because of the empathy that participants feel toward robots. In this case, empathy is measured as a first-person perspective, thereby fitting the Turingism context. If all these conclusions make sense, then it appears that passing an empathic Turing Test may

afford entrainment between humans and robots. But this will not be true, because it is possible that some AGI techniques coincidently couple the conditions of HCI affordance, and at the same time, could pass the Turing Test. Hence, the problem in T1 remains. This is why I suggest a second-person perspective in T2. For example, if we use it to analyze robotic bullying, we may reach a different conclusion. Why would the same attitude of the unacceptability of different agents certainly stem from the "same mind attribution to robots and humans"? Why would it not stem from the same mistreatment coming from the "same" human? This experiment only reveals that participants may discern part of the interpersonal situation between the human abuser and the robot victim as the normal status of their lives. It is reasonable to infer that people may agree that the abuser in this situation may bully the robot as if they were bullying some iron box, or animal, or even some fool. But will they rationally agree that the robotic victim believes itself to be suffering like a human being bullied? I think not. The analysis framework I currently use is in light of the second-person standpoint.

If an AGI system rather than just normal AI with the anthropomorphic embodied interface was meant to implement social entrainment with humans, then it has to be designed with the perspective of the second person. This means that this AGI agent will believe that they are (or are not) a human-like agent to some extent, and believe that people consider that they believe that they are (or are not) a human-like agent to some extent, and that, thanks to this consideration, people believe they are (or are not) a human-like agent to some extent, etc. There are two aspects to realizing this blueprint: one aspect is to endow robots with mindreading ability, and the second is to equip people with a proper principle of operation allowing recognition all intelligent features of AI, and then to clarify to what extent it is human-like.

I will superficially engage with the first aspect. It appears that the domain of mindreading is still a minefield between TT and ST. ST may be thought of as conveniently bridging the first-person and second-person perspective by the mechanism of mirror neurons. But what remains unexplored is how to build a mirror connection between humans and robots. If embodiment AGI like Cogprime were prompted to engage in this domain, then they would need to address T1 first. Besides, some studies relying on TT have made good progress in second-person machines. Nevertheless, the question is how to devise the inductive strategy of TT. Bello and Guarini (2010) and Bello & Bringsjord (2013) established a Polyschame framework by virtue of CLT and cognitive logic based on possible world semantics. TT seems to be more easily accomplished in a *symbolic AI* that is adept at inductive inference. If researchers insist on absorbing empathy mechanisms into the system, then it will be a challenging work.

5 Imitation: In the Gap Between Non-anthropocentrism and Anthropocentrism

Based on the philosophy of information, Floridi (2013) designs a third-person informational ethical system, which distributes different agents using the scale of what he calls "levels of abstraction (LoA)". Different LoAs are represented by different simulations and coupled with different but comparable degrees of morality. The convenience of this theory is that it could couple AI agents with specific moral identities. From the Turingist

perspective, since the robot interfaces are more likely to be treated as agents who can always precisely control their "behavior" but not as agents with free will like humans, they may not be treated as possessing responsibility, which is presumed to be coupled with legal rights. From the parlance of Floridi's philosophy, robotic agents are supposed to be endowed with a morality similar to what they possess from the Turingist perspective. Floridi however has a more ambitious plan for the continuum of all simulations. According to the reducibility between simulations or LoA, he premises that different agents could be compared in the same fundamental LoA by using a coarse-grained evaluation method. In this LoA, all the agents could be described as possessing the following capacities (Floridi 2013, p. 147):

1. To respond to environmental stimuli—for example the presence of a patient in a hospital bed—by updating their states (interactivity), for instance by recording some chosen variables concerning the patient's health. This presupposes that H and W are informed about the environment through some data-entry devices, for example some sensors; 2. To change their states according to their own transition rules and in a self-governed way, independently of environmental stimuli (autonomy), for example by taking flexible decisions based on past and new information, which modify the environmental temperature; and 3. To change the transition rules by which their states are changed according to the environment (adaptability), for example by modifying past procedures to take into account successful and unsuccessful treatments of patients. The so-called agents H and W could be morally compared by using a rule (O):

An action is said to be morally qualifiable if and only if it can cause moral good or evil, that is, if it decreases or increases the degree of metaphysical entropy in the infosphere.

Evidently, both normal AI and AGI may fit the requirements of the agent of Floridi, even though all of them could be attributed to the set of features of human free will. In a way, what Floridi does is not abandoning free will, but rather shortcutting it. By investigating the problem of free will from all aspects, we may find that it is not easy to discern a necessary and sufficient condition for free will, and we may agree that the boundary of free will is vague. However, vagueness just means incompleteness and not uselessness or meaninglessness, as Floridi believes. For example, I may ask: in what way can we call an agent "self-governed"? Could the agent in the situation of the famous Frankfurt-style stories (Frankfurt 1969) be called "self-governed"? It seems that many issues arising from the question "what does it mean to be under control?" need to be figured out before deploying point 2. Furthermore, in regard to point 3, should we not first determine the criteria for *successful and unsuccessful treatments of patients*? I think all conclusions of the debate on free will could be used as references to enrich the extension and intension of free will. Not all elements in the set of components of free will can be realized by normal AI, and that is what AGI should engage with. According to my previous statement, we may detect that Floridi treats simulations as a static set, not a dynamic developing process, which is the essence of research in AGI.

Floridi believes that as long as the interaction is analyzed under the premise of these agent theories, moral evaluation can be given. At the same time, this type of interaction shares the same extension with the interactive relationship in the context of affordance, but does not rely on any analysis of human perception. Floridi's theory and the theory

of affordance share different aspects of non-anthropocentrism. Floridi's informational philosophy supports a non-anthropocentric ontology for the continuum of the relational artifact, AI, and persons. HCI, according to the philosophy of affordance, brings people into a part of a non-anthropocentric interaction.

Nevertheless, this non-anthropocentrism may face a dilemma. It seems that Floridi does not believe any ultra-intelligent machine will emerge in the future. From that, we can infer that he may not believe that a machine in the form of what Firth (1952) called the "ideal observer" will come into our world at anypoint. Since the "ideal observer" is defined by its being omniscient and omnipercipient with respect to non-ethical facts, and hence would not make moral mistakes relying on the reductionist moral philosophy, an ultra-intelligent machine like this would never cause any entropy production in our moral world. Consequently, if Floridi persists in disagreeing with this anti-existence that runs contrary to the Second Law of Thermodynamics in the sense of information or intelligence, then he will inevitably fall into a defeatism of morality, because there would be no reductive metaphysical entropy process at the elementary level of LoA or simulations. In contrast to Floridi's foundationalism, Rifkin (2009) believes that entropy reduction—what he calls empathic activity—may always happen at some high order and local kinds of simulations, and entropy reduction may cause entropy production elsewhere; in other LoA or simulations.

According to this moral theory of entropy, when we introduce the context of imitation, would an eerie landscape inspired by *The Matrix* be brought about by the second-person design philosophy of AI? I think it will not, because imitation is premised on a cross-level simulation connection, which differs from the anthropocentric theory of Rifkin. Although this connection relies on a more basic simulation, it will not provide any significance of morality or any kind of culture on that level. The significance of morality stays with the context of imitation itself, which cannot be reduced to the context of emulation and simulation. Consequently, I will move on to T3 as the end of this section:

T3. The main significance of intelligent interactionism rather than human-computer interactionism under the second-person perspective is that it connects the development of human-like intelligence research with the consensual mental content formed by human sociality. It cannot guarantee any utopian vision, but provides new possibilities for the mental development of human society. Briefly, the philosophy of the second-person perspective stems from P.F. Strawson's work (2014) on reactive attitude and is developed by Steven Darwall (2006) into a systematic theory on morality. In their view, to every individual, moral accountability should be discerned and developed in a dynamic process of predicting of others' feedback to one's own actions, and thereby regulating one's beliefs and intentions. The moral agent in this social circle will be treated as having the same moral consideration as oneself. Everyone in the social circle shares the consensual mental content by which they will reach consent. This content is not based on the requirement that every construction of self in the circle share the same LoA.

6 Conclusion

Tegmark (2017) divides the development of life into three stages, determined by life's ability to design itself:

Life 1.0 (biological stage): evolves its hardware and software. Life 2.0 (cultural stage): evolves its hardware, designs much of its software. Life 3.0 (technological stage): designs its hardware and software.

We cannot posit that this is an anthropocentric theory, but to Tegmark it appears that a higher-order life form leads to higher-order self-designing ability, which seems to be an ideal human attribute. Regardless of whether intelligence in the future will be able determine its own hardware and software, it does not surrender this power of self-determination to evolution. This idea stems from the observation that as the model of AI, human minds are able to partly determine their own development. However, if humans abandon the Turingist approach to designing machine interfaces interacting with them, rather than affording implements to modify themselves by taking advantage of their cognitive fitness, then how can humans determine their software, since many of their embodied cognitions were determined by their hardware? The goal of Turingism in the context of emulation is to let the intelligent interface be perceived as a human being, but not be comfortably perceived, as well as to pander to the perceivers' cognitive fitness. Furthermore, intelligence based on the context of affordance may easily pass some Turing Tests, although not all of them. Intelligence of this kind could possibly be realized by a simulation practice of AGI. Development in this direction seems to have little business with the task of exploring the essence of humanoid intelligence.

Perhaps proposing and emphasizing the previous challenge per se deserves further examination. Nevertheless, I offered my own suggestion for the development of humanoid AI, especially AGI. From my perspective, positing any humanoid intelligence as both moral addressor and addressee is not only a direction that *should* be taken, but also a proper simulation that may fit the real situation when humans relate with each other. Human babies have been described as imitating machines. At the early stage of human development, their growth focuses on the ability to attribute others' endogenous intentions to others' exogenous behavior as mentally sourcing as they do to themselves. The necessary and sufficient condition of the mechanism of second-person knowledge is still being explored, just as other cognitive processes of humans. Perhaps second-person knowledge research based on AGI will complement second-person research in the Theory of Mind. At the same, the direction of this approach may guarantee the ethical acceptability of AGI.

References

Adolphs, R., Gosselin, F., Buchanan, T.W., Tranel, D., Schyns, P., Damasio, A.R.J.N.: A mechanism for impaired fear recognition after amygdala damage. Nature **433**(7021), 68–72 (2005)

Anderson, J.R., Lebiere, C.: The Newell test for a theory of cognition. Behav. Brain Sci. **26**(5), 587–601 (2003)

Bello, P., Bringsjord, S.: On how to build a moral machine. Topoi **32**(2), 251–266 (2013)

Bello, P., Guarini, M.: Introspection and mindreading as mental simulation. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 32, no. 32 (2010)

Block, N.: On a confusion about a function of consciousness. Behav. Brain Sci. **18**, 227–287 (1995)

Bostrom, N.: Are we living in a computer simulation? Philos. Q. **53**(211), 243–255 (2003)

Clarke, R.: Toward a credible agent-causal account of free will. Noûs **27**(2), 191–203 (1993)
Darling, K., Nandy, P., Breazeal, C.: Empathic concern and the effect of stories in human-robot interaction. In: 2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 770–775. IEEE (2015)
Darwall, S.L.: The Second-Person Standpoint: Morality, Respect, and Accountability. Harvard University Press, Cambridge (2006)
Davidson, D.: Essays on Actions and Events: Philosophical Essays, 2nd edn., vol. 1. Clarendon Press, Oxford (2001)
Duffy, B.R.: Anthropomorphism and the social robot. Robot. Auton. Syst. **42**(3–4), 177–190 (2003)
Firth, R.: Ethical absolutism and the ideal observer. Res. **12**(3), 317–345 (1952)
Floridi, L.: The Ethics of Information. Oxford University Press, Oxford (2013)
Fong, T., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots. Robot. Auton. Syst. **42**(3–4), 143–166 (2003)
Frankfurt, H.G.: Alternate possibilities and moral responsibility. J. Philos. **66**(23), 829–839 (1969)
Gibson, J.J.: The Ecological Approach to Visual Perception: Classic Edition. Psychology Press, New York (2014)
Goldman, A.I.: Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading. Oxford University Press, Oxford (2006)
Gray, H.M., Gray, K., Wegner, D.M.: Dimensions of mind perception. Science **315**(5812), 619 (2007)
Gray, K., Wegner, D.M.: Feeling robots and human zombies: Mind perception and the uncanny valley. Cognition **125**(1), 125–130 (2012)
Hawkins, J., Blakeslee, S.: On Intelligence: How a New Understanding of the Brain Will Lead to the Creation of Truly Intelligent Machines. Henry Holt, New York (2004)
Hofstadter, D.R.: Gödel, Escher, Bach: An Eternal Golden Braid. Basic Books, New York (1979)
Hurley, S.: Understanding simulation. Res. **77**(3), 755–774 (2008)
Kane, R.: The Significance of Free Will. Oxford University Press, Oxford (1998)
Kaptelinin, V., Nardi, B.: Affordances in HCI: toward a mediated action perspective. Paper Presented at the SIGCHI Conference on Human Factors in Computing Systems (2012)
Keijsers, M.: Robot bullying. Doctoral dissertation, University of Canterbury. Research repository of University of Canterbury (2020). https://ir.canterbury.ac.nz/handle/10092/100776
Kieras, D.E.: A summary of the EPIC cognitive architecture. In: Chipman, S.E.F. (ed.) The Oxford Handbook of Cognitive Science, pp. 27–48. Oxford University Press, Oxford (2017)
Kim, J.: Causation, nomic subsumption, and the concept of event. J. Philos. **70**(8), 217–236 (1973)
Kissel, A.: Free will, the self, and video game actions. Ethics Inf. Technol. **23**(3), 177–183 (2020). https://doi.org/10.1007/s10676-020-09542-2
Knobe, J., Nichols, S.: Free will and the bounds of the self. In: Kane, R. (ed.) The Oxford Handbook of Free Will, 2nd edn., pp. 530–554. Oxford University Press, New York (2011)
Kotseruba, I., Tsotsos, J.K.: 40 years of cognitive architectures: Core cognitive abilities and practical applications. Artif. Intell. Rev. **53**(1), 17–94 (2020)
Laird, J.E., Newell, A., Rosenbloom, P.S.: SOAR: an architecture for general intelligence. Artif. Intell. **33**(1), 1–64 (1987)
Li, X., Hammer, P., Wang, P., Xie, H.: Functionalist emotion model in NARS. In: Iklé, M., Franz, A., Rzepka, R., Goertzel, B. (eds.) AGI 2018. LNCS (LNAI), vol. 10999, pp. 119–129. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-97676-1_12
Markram, H., et al.: Reconstruction and simulation of neocortical microcircuitry. Cell **163**(2), 456–492 (2015)
Mori, M.: Bukimi: no tani [the uncanny valley]. Energy **7**(4), 33–35 (1970)
Mori, M., MacDorman, K.F., Kageki, N.: The uncanny valley [from the field]. IEEE Robot. Autom. Mag. **19**(2), 98–100 (2012)

Newell, A.: Unified Theories of Cognition. Harvard University Press, Cambridge (1994)
Norman, D.A.: The Psychology of Everyday Things. Basic Books, New York (1988)
O'Connor, T.: Why agent causation? Philos. Top. **24**(2), 143–158 (1996)
Rifkin, J.: The Empathic Civilization: The Race to Global Consciousness in a World in Crisis. Penguin, New York (2009)
Seibt, J., Nørskov, M., Hakli, R. (eds.): Sociable Robots and the Future of Social Relations: Proceedings of Robo-Philosophy 2014, vol. 273. IOS Press, Amsterdam (2014)
Strawson, P.F.: Freedom and Resentment and Other Essays. Routledge, London (2014)
Tegmark, M.: Life 3.0: Being Human in the Age of Artificial Intelligence. Knopf, New York (2017)
Trope, Y., Liberman, N.: Construal-level theory of psychological distance. Psychol. Rev. **117**(2), 440–463 (2010)
Turing, A.: Computing machinery and intelligence. Mind **59**(1950), 433–460 (1950)
Vallverdú, J., Trovato, G.: Emotional affordances for human–robot interaction. Adapt. Behav. **24**(5), 320–334 (2016)
Waytz, A., Gray, K., Epley, N., Wegner, D.M.: Causes and consequences of mind perception. Trends Cogn. Sci. **14**(8), 383–388 (2010)
Wang, S., Lilienfeld, S., Rochat, P.: The uncanny valley: existence and explanations. Rev. Gen. Psychol. **19**(4), 393–407 (2015)
Weizenbaum, J.: ELIZA—a computer program for the study of natural language communication between man and machine. Commun. ACM **9**(1), 36–45 (1966)