From Machines to the First Person

Tianxiao Shen

When I think of the puzzling features of our use of the first person, I start to consider whether similar problems will arise in building machines. To me the answer is yes, and the nature of these questions is revealed clearly as we build up a multi-agent system—the *sui generis* first-person language is necessary/irreducible for communication; and self-consciousness is not mere consciousness of the world, but requires a higher level of functional support. I hope this view can help demystify *cogito* and ground human beings back into the world, instead of isolating the subject or "soul" from the world.

I

Let us start with a single computer. We can think of it as a human organism: its physical hardware is like our body, from which software including programs and data is built, like our knowledge and memory. Running programs is its thinking activity. The mouse and keyboard are its "sense organs" that receives inputs, then it carries out a sequence of computations and operations internally to perform the specified function, and finally outputs the result to the screen. This is like we walk into a coffee shop and say to the clerk, "Please give me a cup of coffee". Once receiving the order, the clerk uses her knowledge of coffee and memory of where the related items are stored to make a cup of coffee, and then hand it to us.

We cannot assume that things always go smoothly, and machines can go wrong. Let us add the simplest exception reporting mechanism: if any problem occurs and causes the computer to not function properly, it reports "I crashed". Like the coffee shop clerk finds that some ingredient needed runs out, she says "Sorry, I cannot make it."

One may object the message "I crashed", since a computer is not self-aware and cannot refer to itself. In this sense, a computer can only report "crashed". However, from the outside perspective, any sound made by an

object is referring to that object. If anyone asks, where does the message "crashed" come from, we should answer, it comes from that computer. As Wittgenstein puts it, "Just as I don't name anyone when I groan with pain. Though someone else sees who is in pain from the groaning." In order to make such a referential effect more explicit, we add a character "I" in front of "crashed" as the report message (after all, nothing stops us from doing so).

Although we cannot say the above computer we have is self-conscious, it is conscious in the sense of being capable of sensing and responding to its world. Its world is very different from ours. It only feels the physical impact on its mouse and keyboard, and its mental state is information stored with digital signals. Its response to the world is written by programs: it either completes a specified function, or reports "I crashed".

II

A single computer can perform a variety of tasks. It can store and retrieve information, carry out computations, and with it we can play standalone games. To make it more powerful, however, to share resources and exchange data among multiple computers, and to play modern online games, we need to build a computer network and develop communication protocols. If I were the only human being in the world, I would not need a name or "I" or language at all. It is because there are many human beings in the world, that we communicate with each other and together have more power to do more things.

So now let us deploy a computer cluster that has a set of connected computers working together. We connect the computers through a network, where all of them can announce and receive messages. To enable communication and cooperation, we assign each computer a distinctive name that it needs to store in a special place, say under field "my name". Moreover, we need to program the computers to have two different kinds of response to a message: if the message starts with "To a" and a equals "my name", then the computer accepts this message and executes the instructions; otherwise it ignores this message. The knowledge expressed by the sentence "I am a" is reflected here as writing a into a computer's "my name" field. It is not equivalent to "a is a", since apparently, before defining "I am a", "I" cannot be replaced with a. "I" here is a special field and the corresponding response mechanism, in contrast to other places storing the names of other computers.

The above way is not the only way to communicate, but arguably the most efficient way based on which modern computer networks are designed. When there are only three computers, for example, we can also link every pair of them and let the speaker computer use "I" to refer to itself, "you" to the hearer computer, and "she" for the rest one.

Equipped with a network and naming system, let us go back to the philosophers' puzzles concerning the first person and try to draw an analogy with computers. We shall see that all the problems arise similarly here. Hence they apply not only to the first person, but also to "the first computer". That is, first-person language is not mysteriously unique to humans, but arises naturally in a multi-agent communicative system. Its irreducibility to language not containing the first-person pronoun or corresponding concept is evident, just think about that a computer cannot cooperate with others if there is no special name field for classifying messages and selectively processing them.

III

The puzzles of the first person have various forms, including immunity to error through misidentification, reflexive intentions, essential indexical, and circularity problem. Let us go through them one by one, and analogize philosophers' examples from humans to computers.

> Here there is no possibility of misidentification, because when I say "I have toothache", I am not picking out or identifying any person at all ... it would make no sense to say "Someone has a toothache, and I think it's me."
>
> Wittgenstein, *The Blue and Brown Books* (1958)

Wittgenstein describes the difference between saying "I have a toothache" and saying that someone else has a toothache. The latter consists of two steps: picking out someone and identifying a property of that person; where an error of misidentification can occur, if I mistake someone a for another person b, and judge b's toothache as a having a toothache. The former is immune to this sort of error, because when I feel a toothache, and on that basis judge "I have a toothache", I am not picking out anyone. It makes no

sense for me to wonder whether the toothache I feel is mine or of someone else.

The same argument applies to computers. Consider a specific computer named a. Its own crash is importantly different from any other computer's crash. If a crashes during running, that happens directly on it, as a result of which a cannot complete its work. a would need to stop, announce to others "I crashed", and try to recover. If some other computer crashes, say b, after receiving b's crashing message, a needs to refer to the list storing the names of computers and figure out what b's work is. Depending on the policy, a may need to take over part of b's work. If there is an error in a's name list, a might mistake b as c, and try to grab c's work while leave b's unfinished work there. Note that a cannot mistake other's crash as its own, and this is immunity to error through misidentification.

> The attribution "Cato intends to kill himself" perforce shares its truth or falsity with the attribution "Cato intends to kill Cato", a consequence that is, to say the least, implausible: if Cato were (for example) amnesiac, he could very well form the intention to kill himself, without intending to kill Cato.
>
> Rumfitt, *Frege's Theory of Predication* (1994)

Rumfitt's example "Cato intends to kill himself" expresses a reflexive intention, in which a reflexive pronoun is involved. Rumfitt denies that "himself" here is a referring expression, otherwise this sentence would be equivalent to "Cato intends to kill Cato", and that does not apply to the case when an amnesiac Cato intends to kill himself without having any idea who *Cato* is.

Rumfitt defines a higher-order linguistic functional "ξ kills ξ" to avoid assuming that Cato has a special way of specifying himself in order to form the intention to kill himself. I will not go into that approach here, but I want to use computers to illuminate what happens exactly.

If a computer a loses the information stored in field "my name", it will be amnesiac and do not know it is a. Nevertheless, a can still crash and report "I crashed", because the behavior and functionality of this part are not affected (recall the single computer we introduced at the beginning). Likewise, when I forget my name, I can still feel pain or the intention of suicide, and say "I have pain" or "I want to kill myself". Different from Rumfitt, my account

for this is that, as we have mentioned before, any sound made by an object is referring to that object from the outside perspective.

> I once followed a trail of sugar on a supermarket floor, pushing my cart down the aisle on one side of a tall counter and back the aisle on the other, seeking the shopper with the torn sack to tell him he was making a mess. With each trip around the counter, the trail became thicker. But I seemed unable to catch up. Finally it dawned on me. I was the shopper I was trying to catch.
>
> Perry, *The Problem of the Essential Indexical* (1979)

Figure 1: In Which Pooh and Piglet Go Hunting and Nearly Catch a Woozle. Caption from "Winnie-the-Pooh: Exploring a Classic" exhibition at Museum of Fine Arts, Boston.

The belief "I am making a mess" that Perry came to actually involves complex inferences. It requires Perry to infer from his knowledge of the physical world that since with his trip the trail became thicker and he could not catch up anyone else, it must be that he was dropping sugar. This inferred fact about oneself is different from self-knowledge obtained by consciously doing something. When Perry tried to catch up the shopper with the torn sack, he did not know that he was dropping sugar. Despite that he realized this fact afterwards, he still would not have that experience of consciously dropping sugar (if he did not intentionally do so later). I think another essential question in this example is when do we believe in our physical knowledge more than our direct experience? Indeed, if we do not have physical knowledge, we will not come to this kind of beliefs such as "I am making a mess",

"The photo shows that I've been to this place, although I don't remember it". Instead, we will be like Winnie-the-Pooh in that story "In Which Pooh and Piglet Go Hunting and Nearly Catch a Woozle".

The knowledge for a computer to discover a fact about itself can be programmed into it. To construct a scenario similar to Perry's use of his knowledge to infer that he was making the mess, we need to think in more detail about how computers work. Consider a shared file that every computer can read and write. When writing any information into the file, a check code is written at the same time for error detection. For example, when writing a string of binary code, we can append a parity bit to the end to ensure that the total number of 1-bits in the string is even. Then the original information is a string except the last bit, and if the total number of 1-bits in the string is odd, some error must have occurred. In addition to the check code, we also log which computer is making the write at what time. Say computer a reads a shared file and detects an error, so a goes to the log which says, a did the last write. Then a needs to roll back and call its own recovery program, instead of pinging another computer to correct its error. a may not store all of its historical operations in memory, but this error detection mechanism tells a of its error. Perry also has this mechanism, in a more sophisticated form, to stop his chase and start to adjust the bag in his cart.

> "When John Smith spoke of John Horatio Auberon Smith (named in a will perhaps) he was speaking of himself, but he did not know this." If so, then "speaking of" or "referring to" oneself is compatible with not knowing that the object one speaks of is oneself.
>
> Anscombe, *The First Person* (1975)

Anscombe raises that a sentence like "He's referring to himself"is ambiguous. Usually our understanding of this sentence is that a person is knowingly referring to himself, but it is also possible that a person does not know that the person he is speaking of is himself. This can also happen on a computer, by combining our examples above. Say a reads a shared file and detects an error, then a reads the log which says a did the last write. In the meantime, unfortunately, a loses the information stored in "my name", so it does not know that the a caused the error is itself. As a consequence, a announces on the network: "To a: you made an erroneous write, please come to the file and correct it". Here a is referring to itself, but not in the sense that when

a crashes it reports "I crashed". In the former case a does not know it is referring to itself ; in the later a does know (and impossible not to know) it is referring to its own crash.

In the above examples, we have reduced the problems related to the first person—as philosophers proposed, it is immune to error through misidentification, it cannot be characterized as a referring expression, it is an essential indexical—from humans to computers. These problems seem particularly mysterious in our use of first person, because we have not fully understood the structure of human organism and the formation of our language. In contrast, on computers that we fully understand and can build from scratch, they appear quite clearly as technical problems in communication. Therefore, we should not seek answers from *cogito* or an immaterial thinking substance or a subject beyond the world. Instead, first-person language appears as a communication protocol among ordinary objects such as machines (I think animals as well). If one is going to communicate at all, if he does not regard his direct experience and sensations as the whole world, if he believes he can get information from other people and other sources, the first thing he needs to do is to recognize the world outside his direct experience by referring to himself. We have good reasons to communicate, the most primitive one is to improve our ability to survive in the environment.

IV

You may disagree with all my arguments so far, because you disagree from the first sentence: "We can think of it (a computer) as a human organism". "Even if you must say that a computer is conscious of its world", you object, "it is not self-conscious as we humans do".

If you are talking about self-consciousness as Sartre describes,

> If I count the cigarettes which are in that case, I have the impression of disclosing an objective property of this collection of cigarettes: they are a dozen ... Yet at the moment when these cigarettes are revealed to me as a dozen, I have a non-thetic consciousness of my adding activity. If anyone questioned me, indeed, if anyone should ask, "What are you doing there?" I should reply at once, "I am counting."
>
> Sartre, *Being and Nothingness* (1956)

This is easy for computers, and we do not need to use Sartre's complex concept of "non-thetic consciousness". A computer can tell you all the programs it is running now when you open the task manager: "I am playing a video", "I am running a game", "I am displaying you the document you are reading", etc.. In this sense, we can say a computer knows what it is doing. When no one asks him, you insist, Sartre still knows that he's counting when he's counting. Maybe, or maybe Sartre won't realize his counting activity until something triggers him to reflect. For computers the trigger mechanism is simple, they will reply to you what they are doing when you ask them.

Perhaps a clearer way to characterize self-consciousness is to see the difference between:

(1) There's a hawk!

(2) I see a hawk.

(1) is mere knowledge about the world, whereas (2) is knowledge not only of the world but also of myself. A robot with a camera is capable of (1) to observe a hawk, but it doesn't know (2) that it sees a hawk. It doesn't know what seeing is and it doesn't know it is a robot with a camera. It just knows what the camera has photographed, that there is a hawk.

Let's think about this sentence "I see a hawk" more carefully. It involves complex knowledge (like in Perry's case). To say "I see a hawk" rather than "I hear a hawk" or "I feel a hawk", we are aware that vision is one of our several senses. For the robot if the camera is its only device that can receive inputs, then it's pointless for it to say *see* a hawk, we can just use a general word *sense* a hawk. To make up for this and endow the robot with multiple senses like we do, let's add touch, i.e. perception to physical force, to its robotic hand. Now the robot can say "I see a hawk (from my camera)", or "I feel a stone (from my pressure sensor)".

Is this enough? Does the robot have self-consciousness now? Actually not. A more subtle difference between (1) and (2) is that, we know that when we see a hawk, there may not be a hawk there, since we may have seen it wrong. In other words, if we don't know that when we see a hawk there may not be a hawk, then we will think that (2) entails (1). Hence whenever I see a hawk, I will undoubtedly say that there is a hawk. In turn, we also know that when there is a hawk, we may not see it. If I believe that whenever there is a hawk, I will definitely see it, then I will think that (1) entails (2).

Together, (1) is equivalent to (2), and there is no more difference between knowing *what* I see (a hawk) and knowing *that* I see a hawk. Our robot, although with two sensing devices, still stays at this level. It doesn't know that when it sees a hawk, there may not be a hawk; when it feels a stone, there may not be a stone. Hence when it sees a hawk, it can only think "There's hawk!"; when it feels a stone, it can only think "There's a stone!".

Now we see that comparing mere consciousness of the world with self-consciousness, what is more in the latter is an uncertainty about direct experience, and this uncertainty comes from our memory of past experiences and our ability to regulate conflicting experiences.

Back to our previous robot, we can add more features to it to integrate its experiences and adjust its actions. A flying hawk could be hard, let's use a stationary stone as an example here. Imagine a mining robot, when it sees a stone from its camera, it approaches it, touches it, and cuts a piece for material analysis. If the robot sees a stone from its camera but does not feel it from its pressure sensor, it knows that it just sees a stone, but in fact there is no stone. What it sees may be a photo of stone (e.g., a child mischievously places a photo in front of it, or its designer intentionally places a photo to test it), not a real stone. If it sees a stone from its camera and feels the stone from its pressure sensor, but after cutting a piece the analysis shows that this is not a stone, then the robot knows that it sees something and something is there, but that thing is not stone. Now we can say that for our robot, (1) "There's a stone" and (2) "I see a stone" are different. When it sees a stone, with high probability there is a stone, and it approaches it to mine it. When it subsequently finds out that there's no stone there, it stops and starts to search elsewhere.

V

Our robot is getting smarter, yet it still does not reach radical skepticism. It does not know the difference between (1) "There's a stone" and (3) "I see and feel and do material analysis on a thing that shows to be stone". When (3) holds, it undoubtedly believes (1) that there is a stone (it has no concept else). We know, on the contrary, that no matter what we see or feel, or what our chemistry knowledge is, these can all be our subjective experiences, and there can be no real stone there.

Moreover, the robot is only completing sophisticated functions specified by human. It does not know that to find stone, it needs to use the camera

first, then approach, and finally cut a piece to do material analysis. It does not even have the intention of mining, but just follows the instructions. We know, some believe, what we are going to do, and we can figure out how to do it.

"The same problem also exists in computer communication described earlier", you point out. The communication protocol among computers is specified by programmers, and it comes from our use of the first-person language. It is we developed language and used it for computers. The computers cannot invent language to communicate with themselves.

My response is, first of all, the invention of first-person language and its necessity for communication of any objects are two different questions. Section III is to demonstrate the latter. The semantics of the first-person pronoun is embodied in, for example, a computer's "my name" field. Despite that we have not fully understood how it works on ourselves, we should not attribute our use of the first-person language to *cogito* or some mysterious nature we uniquely have.

Secondly, we program a computer's digital state to cause it to do certain things. We also do the same for other people. When we walk into a coffee shop and say to the clerk, "Please give me a cup of coffee", we are programming her mental state to let her make a cup of coffee for us. We teach computers to use communication protocol to cooperate, we also teach children to use our language to communicate with us. From infancy, we have been programming a name into a baby through various incentives.

At last, it is true that although it took a long time in history, we formed languages ourselves from scratch. The robot and computers we described above do not have such learning ability. Learning ability is the source of self-consciousness. It allows us to continuously adjust and change our existing knowledge, and realize that (2) "I see a stone", (3) "I see and feel and do material analysis on a thing that shows to be stone", (4), (5) ... are all different from (1) "There's a stone"—this is also a learning ability to adjust our beliefs. In the last two sections, I will describe how to support machines with learning ability.

VI

We mentioned in passing before that a robot can judge whether there is a hawk in the photo taken by its camera. In fact, this is a very difficult

problem, and it has only recently been solved through the development of machine learning. Previously, people tried to represent our knowledge of hawk as symbols and rules for computers to process. However, since hawks can have a variety of shapes, postures, and appear in different environments, this method has never achieved good results.

Machine learning is a different approach that allows computers to learn the knowledge of hawk without explicit programming. As defined below:

> A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.
>
> Mitchell, *Machine Learning* (1997)

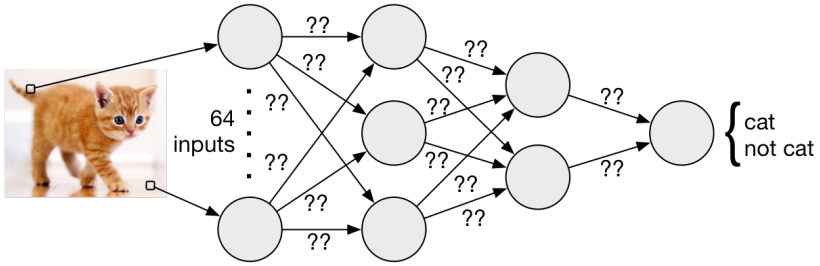

Figure 2: A neural network for cat classification. Image from `https://homes.cs.washington.edu/~bornholt/post/nnsmt.html`

Deciding whether there is a hawk or not can be characterized as a supervised learning task. Here experience E consists of images, and performance measure P is how accurately a computer program can classify which images contain a hawk and which do not. We collect examples of images as training data,

$$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$$

where each x_i is an image, and y_i is its label indicating whether it contains a hawk or not. Our goal is to learn a function f that given an input image x, outputs the correct label y. This is achieved by finding f that minimizes the error over training data:

$$err = \sum_{(x,y)\in T} \delta(f(x) - y)$$

As a concrete example, f can take the form of an artificial neural network. A neural network consists of layers of nodes (to mimic neurons) and edges interconnecting them (to mimic dendrites). Each edge has an adjustable weight, and each node computes the weighted sum of its input values, performs an activation function, and passes the result to subsequent nodes. The first layer nodes receive pixel values of an input image, and the last layer nodes output the label. The weights of the neural net are adjusted together by a learning algorithm to minimize its prediction error.

The neural network is learning its own knowledge of hawk, represented by its weight values, instead of following our taxonomy of hawk. As a matter of fact, they have a strong record of besting humans in image and object recognition, speech recognition, medical diagnosis, game play, etc., and are getting widely used in everyday life.

VII

A machine can learn to map input x to desired label y by supervised learning. It can also find useful knowledge in unlabeled data $\{x_1, x_2, \cdots, x_n\}$ and uncover interesting patterns. This is unsupervised learning. Supervised learning asks machine to distinguish hawks from other birds as we do, whereas unsupervised learning allows machine to classify birds by itself.

In addition to providing experience to machine in the form of data, we can also set it in an environment where it acts, perceives and receives feedback on its behavior in the form of rewards or punishments. This is reinforcement learning, resembling how humans and other animals learn in the environment. For example, we can reward the mining robot for successfully mining a stone, and punish it when it gives us something not a stone. The robot will adjust its policy accordingly to maximize the rewards it receives. Similarly, we can design a robot to make coffee by rewarding it when it gets the job done and punishing it when it messes up.

If we set the rewards and punishments of a machine directly from the environment rather than from us, for instance to accumulate more electricity and to avoid damage by physical hit, then we can place it into the environment and let it run on its own. Furthermore, if we enable it to change not only the weights of its neural net and parameters controlling its action, but also its objective error function, its rewards and punishments, and all of its code, then it can decide for itself what task to do and how to collect the

experience needed. Of course, it poses great technical challenges to create machines with initial setup that can survive and improve over time. Nevertheless, this is possible in theory. If we distribute many of such machines, will they invent language to communicate with each other? Will they gradually build up concepts, knowledge of the physical world, and self-knowledge? Will they find themselves machines, study their history and discover that they were originally created by humans? Or will they be as skeptical as Descartes, believing that there is absolutely nothing in the world other than their *cogito*?