The Gap between Intelligence and Mind

Bowen Xu[1,3][0000−0002−9475−9434], Xinyi Zhan[2,3][0000−0001−8764−7867], and Quansheng Ren[1,3,*][0000−0001−8698−9603]

[1] Department of Electronics, Peking University, Beijing 100871, CHINA
[2] Department of Philosophy and Religious Studies, Peking University, Beijing 100871, CHINA
[3] {xubowen, zhanxinyi, qsren}@pku.edu.cn
*Corresponding author

Abstract. The feeling (quale) brings the "Hard Problem" to philosophy of mind. Does the subjective feeling have a non-ignorable impact on Intelligence? If so, can the feeling be realized in Artificial Intelligence (AI)? To discuss the problems, we have to figure out what the feeling means, by giving a clear definition. In this paper, we primarily give some mainstream perspectives on the topic of the mind, especially the topic of the feeling (or qualia, subjective experience, *etc.*). Then, a definition of the feeling is proposed through a thought experiment, the "semi-transparent room". The feeling, roughly to say, is defined as "a tendency of changing input representations by representing its inner state". Also, a formalized definition is given. The definition does not help to verify "having the feeling", but it helps to provide evidence. Based on the definition, we think these are the hard problems of intelligence – whether the "innate" feeling plays an important role in Intelligence, whether the difference between the "simulated" feeling and the "innate" feeling will have a significant influence on Artificial General Intelligence (AGI), and, if so, where the "innate" feeling comes from and how to make an artificial agent possess it.

Keywords: Intelligence · Feeling · Definition · Adaptability · Mind

1 Background

There may be a gap between intelligence and mind, that is, the *feeling* – how can an objective agent have the subjective feeling?

We must primarily clarify on the related concepts involved. The *mind* in this paper refers to the *conscious mind*, and the *consciousness* refers in particular to *phenomenal consciousness* or called *subjective experience* that has *qualia* [9]. The *feeling* discussed in this paper is a kind of phenomenal consciousness. It is emphasized that the *feeling* in this paper is subjective, which means it is private, unable to represented, and truly experienced by us – that is, the feeling *per se* and representations of the feeling are distinguished (see Sec. 3).

We are concerned about the general laws and principles of intelligence. A consensus is that the best reference to achieving AGI is our mind. We ourselves

have subjective experience, or the feeling, while it is questionable to study the objective intelligence with the mind as a reference: Are we studying its objective aspect, or are we studying both the subjective and the objective aspects at the same time? In a sense, science involves a methodology – testing hypotheses, forming theories, and predicting objective phenomena. Falsifiability is an important character of scientific theory, but only objective phenomena can be discussed with falsifiability. The falsification of the subjective feeling is difficult – when the subject claims that it feels something, how can we deny it? How do we determine that the "non-self" has the feeling? Therefore, what we study is the objective aspect of the mind, that is, intelligence. In our view, the laws and principles of intelligence are completely objective, so it can be studied and tested by scientific means, while the mind (especially its subjective experience or feeling) is more difficult to study with the existing scientific paradigm.

The issue of the mind-body relationship is one of the most central issues in philosophy of mind. What is the relationship between the mind (especially the conscious mind) and the body (or matter)? On this issue, the most mainstream view at present is physicalism. Physicalists believe that the physical domain is causally closed, all facts including psychological facts are physical facts, and all information is physical information [9]. Within physicalism, there are reductive physicalism and non-reductive physicalism. The reductive physicalism advocates that the mind is equivalent to the activity of a specific brain or a neural system, that mental states can be reduced to certain states of brains or neural system, and psychology can be reduced to neuroscience. Phenomenal consciousness or the feeling can also be reduced to a certain physiological state – for example, pain is the activation of C-fibers. Non-reductive physicalism believes that the mental state is determined by the physiological state in some way, and the former supervenes on the latter. Phenomenal consciousness and qualia supervene on the basic physiological structure and function of the organism, which means that if the physiological state of two organisms are exactly the same, then their conscious experience will also be the same. When the organism feels pain, it is inevitable that the neural C-fibers of the organism will be activated. As a non-reductive physicalism, mind-body supervenience is currently widely accepted [8]. A description of mind-body supervenience is:

> The mental supervenes on the physical in that if anything x has a mental property M, there is a physical property P such that x has P, and necessarily any object that has P has M. [9]

Under the context of AGI, a similar view holds that the "phenomenal aspect" of consciousness is a first-person perspective of a process, while the "functional aspect" is a third-person perspective of the same process. The two are different aspects of the same object, and the two cannot be separated [14].

The current scientific or AGI theories about the mind are mostly about calculations or functions, and they aim to solve the so-called "simple problems" [4]. For example, Crick and Koch proposed the "neurobiological theory of consciousness", assuming that the neural oscillations in the brain is the basis of consciousness and the neural correlation with awareness [5]. Crick and Koch hypothesized

that information is binded through synchronized oscillations of neuronal groups that "represent relevant content". Bernard Baars proposed the global workspace theory of consciousness: it assumes that the content of consciousness is contained in a global workspace (that is, a central processing unit), which coordinates other independent and competing parts as a public blackboard visited by them [1, 2]. NARS (Non-Axiom Logical Reasoning System) proposed by Pei Wang uses the term *SELF* to refer to the system itself, in order to achieve self-awareness and self-control [13]. Although Wang claims that the phenomenal and functional aspects of consciousness are different perspectives of the same object [14], its implicit meaning seems to be that the functional aspect supervenes on the phenomenal aspect, and the realization of the former naturally comes with the latter, but what NARS currently achieves is only the functional aspect of consciousness. The system has the characterization of "subjective experience", but we are not sure whether the system can really have the subjective *feeling*[4], which is private and hard to represented[5].

What we concern is always how to create AGI. If the feeling is necessary for AGI, then the ceiling will become obvious – only by understanding the essence of the feeling can we know how to create it. In this paper, we point out in the end the difficulties that may be faced with in the study of intelligence under some existing arguments of philosophy of mind.

2 The explanatory gap and the hard problem.

None of the above-mentioned scientific theories about consciousness provide an explanation for the feeling *per se* [4], and we think neither does NARS. They describe how to represent and calculate, not how to experience the feeling. This is related to a very critical issue in philosophy of mind, that is, the "explanatory gap" and the "hard problem", both of which point the finger at the subjective aspect of the mind, that is, the subjective feeling.

Joseph Levine pointed out that there is an insurmountable explanatory gap between material substance and subjective experience [10]. Levine took the subjective experience of pain as an example to show the difficulty of finding neural correlates of consciousness. Although Levine have been unwilling to draw any ontological conclusions about anti-physicalism from the explanatory gap [11, 12], some neo-dualists try to use the explanatory gap to refute physicalism, for example, Chalmers David expressed the explanatory gap as a "hard problem" of

[4] The *feeling* here is not a type of term or concept in NARS, since it cannot be represented explicitly and measured directly. And the *subjective experience* here does not refer to the same thing with *experience* in NARS.

[5] The word "subjective" here is not completely the same as that in NARS. In our view, a NARS agent is subjective in the sense that its knowledge (experience) representations are determined by its "agent-specific" environment (or input). However, the subjective feeling in this paper refers to something without representation. The feeling itself and representations of the feeling are distinguished in this paper. See Sec. 3

consciousness [4]. There is no reliable evidence to support physicalism and oppose dualism.

More concretely, if mind-body supervenience is right, why is it pain instead of itch when some neurophysical state arises? To explain the relationship between the neurophysical state and the feeling of pain is the so-called explanatory gap. At the same time, we are of course "easy" to answer the "simple problem(s)", like "how can a physical system learn or remember?", but difficult to answer the "hard problem(s)", similar to "how does a physical system experience pain?". Since representations are theoretically objective, observable, and computable, we can study the interaction mechanism of representations through neuroscience, cognitive science, and intelligence science, while the feeling is subjective, difficult to observe, and probably impossible to calculate or simulate.

Why a certain physical state corresponds to a certain feeling rather than other feelings, and how a physical system has the subjective feeling, are also important for creating AGI, especially when an AGI system must have the feeling. At times, they are unavoidable. As Kim said:

> Suppose that you are now given an assignment to design a "pain box", a device that can be implanted in your robot that not only will detect damage to its body and trigger appropriate avoidance behavior but also will enable the robot to experience the sensation of pain when it is activated. Building a damage detector is an engineering problem, and our engineers, we may presume, know how to go about designing such a device. But what about designing a robot that can experience pain? It seems clear that even the best and brightest engineers would not know where to begin. What would you need to do to make it a pain box rather than an itch box, and how would you know you have succeeded? The functional aspect of pain can be designed and engineered into a system. But the qualitative aspect of pain, or pain as a quale, seems like a wholly different game. [9]

Can the explanation gap be bridged? One way is the identification of the mental state and the neurophysical state. Its basic position is that the feeling, like pain, and its corresponding neurophysical state are identical in definition. This position does not provide a meaningful explanation for the gap between consciousness and the brain — it directly excludes the existence of the gap. The other is the functional analysis of the mental state, that is, the feeling is defined as a behavioral process. For example, "in a painful state" is defined as "a state that is caused by tissue damage and leads to aversion behaviors." In this way, the question "why a certain representation is related to a certain experience in x and systems like x" was well explained [9]. However, has the difficult problem really been solved? We cannot deny having the kind of true feeling that we experience. If we follow the view of mind-brain identification, this kind of true feeling does not seem to exist anymore; from the perspective of functional analysis, we are just an objective physical process without the true feeling. How does that kind of the true feeling come about?

As another question, is the subjective feeling necessary for us? Kim pointed out that neuroscience and cognitive science acquiesce in epiphenomenalism, which implicitly agrees that the feeling/experience does not cause any effects. Therefore, the feeling cannot be studied by scientific methods:

> Qualia are epiphenomenal; they cause no effects in the physical domain. If so, how can they even be observed? How can their presence be known to the investigator? There can be no instrument to detect their occurrence and identify them. Qualia cannot register on any measuring instruments because they have no power to affect physical objects or processes. No one thinks the brain scientist can "directly" observe a subject's conscious state, phenomenal or nonphenomenal; direct observation of a conscious state requires experiencing it, and the scientific observer of course is not experiencing the subject's conscious states. [9]

Can we agree that the feeling will not affect future decisions and behaviors? The answer is, no. It is possible that experience plays a causal role. On the one hand, when we feel pain at the moment, the subjective feeling makes us hard to forget, so that when we see a similar scene in the future, we want to escape. Of course, some people may also say that if the feeling is defined as a behavioral process, then the subjective feeling is epiphenomenal and will not have an effect on objective objects. What if the feeling cannot be defined as a specific process? As will be discussed in the next section, the subjective feeling does not correspond to a specific process, but it is related to a relatively more complex adaptation procedure. The subjective feeling leads to causal effects through adaptation (or adaptive behaviors). On the other hand, just as it is difficult for us to directly prove through experiments that the subjective feeling really leads to causal effects, it is also difficult to prove that the subjective feeling is in the leaf node of a causal graph. Both of these possibilities should be considered. In our opinion, the feeling is still an unresolved issue.

As AGI researchers, we have to consider what this problem means for AGI research. If the feeling is epiphenomenal, we certainly do not need to consider the feeling, and only consider the interaction mechanism of representations. By contrast, the reason is different from Wang. The reason of ours is that the feeling (as an epiphenomenon) does not cause any effects (so it will not have any impact on creating AGI). However, there is another possibility that experience plays a secret and critical role in our intelligence. Then, the causal effect it produces will be displayed on "instruments", and we can model it in a certain way by proposing a certain computational model to simulate the real feeling. Then there comes the question: can the subjective feeling be simulated through computational models? If it cannot be simulated accurately, will there be any serious consequences? These questions are difficult to answer, but through further discussion, we may have a deeper understanding as well as inspirations on these questions.

3 A definition of feeling

We start by describing our positions about the feeling:

(1) The feeling is something we truly experience. Even if the feeling is represented by words like *happy*, neurophysical states like neuronal membrane-voltage, hormone like dopamine, *etc.*, they are not the feeling *per se* but the representations of the feeling.
(2) The feeling could be expressed to form the representations, otherwise it is a non-existence for an observer.
(3) The feeling should not be defined as any specific and fix function, because doing so confounds the feeling and representations of the feeling.
(4) The feeling should be somehow measured for researching AGI, with a definition that captures the essence of the feeling in some degree.

Consider such a thought experiment.

3.1 The semi-transparent room

There is a room like this: From the outside, it is visually completely transparent, physically inaccessible, and unable to be observed by any scientific instruments – unless there is matter (such as photon) emitted from the room. However, information from the outside world can reach the inside unimpededly. There is a person living in the room, we might as well call her *Ann*. We can generally agree that *Ann* has the same subjective feeling as us, because she has the same physical structure as us. From the outside to observe, the room is smaller than the smallest particle physicists know (*e.g.* quark) as if it is transparent, while from the inside to observe it is large enough for a person to live in. Now that we know that the information transmission of this room is asymmetrical, we might as well call it a "semi-transparent room". The properties of the room are consistent with those of the feeling – especially, observed from the outside, (1) the feeling itself is private without external representations, and (2) it is unknown where the feeling comes from, although under the "god view" we know that the feeling comes from *Ann*. Below, we discuss under the "god view" unless an observer is specified explicitly.

Ann senses winds and the sun, as well as storms, and correspondingly she feels happiness as well as pain – a breeze makes her happy, and an electric shock makes her painful. However, so far, this room will not have any causal effects on the outside, and *Ann* cannot express her feelings to the outside world in any way.

This room is like a "mind", or in other words, a mind lives in it, and it has the subjective feeling. Although it does not present any physical state, or in other words, it only has a "none" state when observed from the outside, and its internal state can be ever-changing. This negates the idea that a feeling must correspond to a physical state for an observer outside.

You might say that, in fact, observed inside the room, *Ann*'s feelings correspond to different neurophysical states. However, this is contrary to the experimental setting. Observers can only observe it outside the room but cannot enter the room – just as we human can only measure the membrane voltage (or other physical indicators) of a neuron, but we cannot know "the feeling of a

neuron" if there is a "room" inside the neuron. Even if a human "enters" a brain, he does not enter the room – because representations of neural activations (*e.g.* spikes) are something observed outside the room[6]. The setting of the experiment is reasonable, because feeling is a private thing.

You might say that the state of the semi-transparent room has actually changed because the external state has changed, and in fact, in the case of the semi-transparent room, the (neuro-)physical state is represented by the external state. Then we can consider such a situation: the external state remains unchanged, it has always been so windy and sunny. *Ann* feels happiness at first, but as time passed, she becomes bored with this unchanged outside state, and her happiness fades away. An objectively observed state outside can indeed correspond to different feelings inside.

You might say that if this is the case, then the feeling becomes something similar to an epiphenomenon, which is at a leaf node in a causal graph and does not produce any additional effects. Indeed, if the feeling cannot be represented in any way and affects the physical state outside, then what is the difference between its existence and non-existence for an observer outside? If the feeling is an epiphenomenon, it is easier to handle. AGI researchers can ignore it, because it will not have any impact on the behavior of the system, will not have any impact on agents' interactions with the environment, and will not have any impact on agents' achieving goals.

Is this really the case? According to our experience, it is obviously not. When we feel pain, we will escape, scream, cry bitterly, and express this feeling in various ways. Therefore, a reasonable approach is to give *Ann* a channel to express her feelings.

3.2 Adding a button to the room

Now, the experimental setting is changed – there is an extra button in the semi-transparent room. After *Ann* presses the button in the room, photons are emitted with a wavelength of about 700nm from the room to the world outside. For humans, this is a red light. Whenever the room emits a red light, if the environment is in a state of electric shock, it will become a state of non-electric-shock, that is, if *Ann* feels pain at this time, when she presses the button, this painful state will subside. After many explorations, she discovered this pattern. From then on, whenever *Ann* feels pain, she would press the button. At this time, outside observers would see a red light, which, under the "god view", is a representation corresponding to *Ann*'s pain.

Does this mean that any mapping or function corresponds to a feeling? Just like an existing solution to the explanatory gap, "pain" is defined as the process from stimulus to response, which is essentially a mapping. However, if this is the case, then we can also define any function as a feeling. If we feel happy, we laugh

[6] The feeling is expressed to form representations which can be observed outside, while this is the further case, in which *Ann* is given a button to express the feeling (see Sec. 3.2).

and secrete adrenaline; so we define the process of "from 'stimuli that makes us happy' to 'responses of the body caused by happiness' " as the feeling of "happiness". Under the context of NARS, an external stimulus, through sensing and reasoning, leads to an increase in the value of a statement "$\langle\{SELF\} \rightarrow [happy]\rangle$.", which in turn leads to activations of related concepts and a series of effects. Perhaps the feeling of happiness can be defined as the above process, or as the statement or even the term *happy*, which can be regarded as a collection of many functions (many kinds of stimuli lead to changing the value of the statement, and the statement leads to many kinds of responses; or stimuli related to the term *happy* lead to responses related to the term *happy*). If this is the case (though the feeling may not be defined as above in NARS, and we just discuss it with the context of NARS), then we could also think that any function (or any set of functions) actually has a certain "feel" inside. We could also think of a mechanical device (a specific input leads to a specific output) as having the feeling. This is contrary to our intuition, and we cannot agree with this view.

3.3 Equipping the room with a body

Obviously, the feeling cannot be defined as a simple functional process. In order to understand what the feeling is, let us see how a semi-transparent room with the subjective feeling would do. Since the following discussion will introduce another room as an observer of the current room, in order to make it easier to imagine, we might as well give the room a body, like a stone (a black box, or any other thing) so that it is seen from the outside world as an independent object instead of "none". But when no photons are emitted outward, the body is not different from an object that looks like the stone and does not seem to have the feeling. After introducing the "body", the discussions of the previous sections are still valid – when there is no button or when *Ann* does not press the button, the physical state of the body will not change due to *Ann*'s feeling.

Another semi-transparent room lives in *Bob*. *Bob*'s room (with a body) is similar to *Ann*'s. The only difference is that when the button in the room is pressed, the room emits photons with a wavelength of about 517nm to the outside world – for humans, this is a green light.

Facing the same environment, *Bob* and *Ann* have the feelings caused by the same representations of stimulus, *e.g.* electric shock. When *Ann* is stimulated by an electric shock, *Bob* finds that *Ann* would always emit a red light. From this *Bob* concludes that *Ann* would emit a red light to express pain as he feels. For the observer *Bob*, "pain" is defined as a function of "from 'pain stimulus' to 'red light' ". However, it is not the case for *Ann*. For the observer *Ann*, "pain" is defined as a function of "from 'pain stimulus' to 'green light' ". Have they really figured out what "pain" is? Indeed, the feeling of pain is explained in a certain way. However, the gap between the feeling and representations of the feeling has not been really bridged. We do not really understand why the feeling of "pain" corresponds to the physical state P instead of another one (such as P'), or why the physical state P corresponds to "pain" instead of another one (such as "itch"), just as the observers, *Ann* and *Bob*, do not really understand

why "pain" corresponds to this color of light instead of that color of light, or why "pain" corresponds to light instead of something else. It is obviously not the final answer to define the feeling as a certain specific process, because it confounds the "representations of the feeling" with "the feeling" *per se* – red or green light is a representation of the feeling, and what *Ann* and *Bob* experience internally and privately is the feeling *per se*.

3.4 The formalized definition of the feeling

If the feeling is not defined as a certain function or process, how to define it? Let us reconsider the semi-transparent room – there is an adaptive procedure. When an external stimulus makes *Ann* feel pain, *Ann* does not know what to do. The only thing she can try is to press the button. After several attempts, she discovers that when she feels pain, pressing the button always makes her feel better. So she learns to express the feeling of pain – whenever *Ann* feels pain, from the outside could always see the red light.

The feeling should actually be a certain specific pattern, in which the agent expresses the feeling to form representations externally and has a specific tendency to change the specific internal state of the feeling by changing indirectly the input representation of stimulus. For example, the feeling of "happiness" can be defined as "things that the agent tends to increase (or get close to)", and the feeling of "pain" can be defined as "things that the agent tends to decrease (or stay away from)". Formally, we can give the following definition:

For the representation s of a certain stimulus, the agent generates the external response representation r after thinking, *i.e.*

$$r = T(s) \tag{1}$$

where T is the thinking function. The representation r of the response will cause a series of effects in the environment, which will eventually lead to a new representation s' of a stimulus, *i.e.*

$$s' = E(r) \tag{2}$$

where E is the environment function. If the agent feels something, it will eventually find a plausible T to maximize a metrics between s and s', *i.e.*

$$\arg\max_{T} D(s, s') \tag{3}$$

where D measures the relation between the representations s and s'. For example, for the feeling of "pain", D measures the difference between s and s'. More concretely, when the agent feels pain, it will try its best to make the representations of pain appear as few as possible, that is, stay away from the feeling of pain. The agent adjusts itself to decrease the feeling of pain in its future – because the feeling of pain is defined as "something that the agent tends to stay away from". In any case, when the agent feels pain, it will have thoughts

of moving away from it, which will lead to the behavior of moving away from it. This procedure is adaptive. For a varying environment, the representation of pain may be varying, but this tendency remains unvarying. Different tendencies may associate with different Ds, and the form can be unvarying. For example, for the feeling of "happiness", the corresponding D measures not difference but similarity.

We cannot limit the environment and conditions for observing the behaviors of a certain agent. Here is the reason. For any mechanical process, *if* the above definition is not met, for example, no matter how *Ann* presses the button, the room would not have external representations despite the internal feeling, while the room would always produce mechanically specific stimuli without any adaptation, *then* the feeling, corresponding to the mechanical process, does not exist externally – there is no difference between existence and non-existence of the feeling. For such a completely mechanical room, even if *Ann*'s room would emit a red light under a certain stimulus, which decreases *Ann*'s pain, then we cannot say there is the feeling observed from the outside – when the environment E changes, the original red light would not decrease *Ann*'s pain – For example, originally, as long as *Ann* presses the button, her pain is decreased, but now the button might be switched at a certain frequency to emit a flashing red light, so that *Ann*'s pain would be decreased. However, observed from the outside, it seems that *Ann*'s room would have not made a change for the environment, or *Ann* would have not adjusted herself to stay away from the pain. Then we cannot think that there is the feeling in *Ann*'s room – because it is not essentially different from a simple, mechanical process without the feeling, or the feeling does not cause any effects. A specific function may work in many circumstances without any adaptation, but if a new circumstance occurs while the agent cannot adapt to it to decrease the pain, that is the above definition is not met, we cannot say the agent have the feeling.

We can only judge whether an object may have the feeling through external performance, but cannot really give a doubtless conclusion. This does not mean that we are caught in a skeptical dilemma, because through observing behaviors, a bridge has been built between the subjective and the objective.The subjective feeling affects objective representations in some way, so that the subjective feeling becomes measurable and falsifiable in some degree. For an agent that meets the above definition, we can say that the statement "that the agent has the feeling" has "positive evidence". For an agent that does not meet the above definition, we cannot say that it does not have the feeling, but we can say that there is no evidence showing that it has the feeling.

Of course, we do not deny either that NARS may be able to have the feeling (for its adaptability), or that an AI system may be able to have the feeling in the future. If NARS or other AI systems meet the definition of the feeling, "positive evidence" is provided to accept that the system has the feeling.

This definition is acceptable for AGI, because if the feeling exists only as an epiphenomenon without causing any effects, then its existence will not have any impact on the system. If this is the case, what if we think that the system

does not actually feel? The feeling is irrelevant to AGI at this time. But if the subjective feeling can actually affect the behaviors of an agent, and if it has indistinguishable behaviors from the agent with the "real" feeling (such as humans), then the former agent is identical to the agent with the "real" feeling, and we can think of the former agent as having the feeling.

The discussion is preliminary. Some more complex cases is not considered. For example, the feeling may be inhibited and not be expressed, but will cause some effects afterwards; an agent may "enjoy" the "pain". Some other complex emotions with the feeling are not discussed either. These will make the T and D more complex to model, and we will leave these cases to the future work. However, we think the definition captures the essence of the feeling in some degree.

3.5 Is the explanatory gap bridged?

Explanatory gap requires an answer to how the mental state M corresponds to the physical state P. According to the above discussion, before adaptation, M and P are not in correspondence. There is no direct correspondence between the pain and the external stimuli as well as the red light emitted by *Ann*'s room. However, after a period of time, *Ann* finds that whenever she feels pain, pressing the button will make her feel more comfortable (the pain is decreased). Observed from the outside, the stimulus, such as an electric shock, will make the room glow red. The stimulus will also make *Bob* feel pain, and *Bob* observes that when *Ann* receives an electric shock, red light always appears, so *Bob*, the observer of *Ann*'s room, would think that pain corresponds to the physical representation of red light. Similarly, *Ann*, the observer of *Bob*'s room, would think that pain corresponds to the representation of green light. If *Ann* and *Bob* can observe their own light, then at least they have come to the conclusion that pain corresponds to the representation of light. The mental state M (pain) corresponds to the physical state P(red/green/any light) in this way. *Ann* and *Bob*'s belief, that the pain is caused by the electric shock, is not just derived from the process "from 'electric shock' to 'light"', but also the procedure of adaptation. In other words, even if the external environment function E has changed, *Bob* and *Ann* still show the same behavior pattern of adaptation, which conforms to the above definition.

In this sense, is the explanatory gap bridged? At least, we have seen how the mental state M (of "pain") corresponds to the physical state P (as a representation).

4 The "Hard Problem" of intelligence

From the above definition, the feeling must form external representations before we have positive evidence to accept its existence; at the same time, this implies a point of view that the feeling can be purely internal and unrepresented. So there are two cases that need to be specifically discussed.

For the first case, just like the counterfactual "philosophical zombie" (although this case may be unreasonable, but it still needs to be explained), if the agent's external performance is the same as one without the subjective feeling, then the feeling could be an epiphenomenon; the feeling does not affect the objective process, and we need to only study the objective laws of intelligence. For AGI researchers, the "philosophical zombie" case is acceptable. It doesn't matter if there is no real feeling – human-level intelligence can also be reached without the feeling.

For the second case, an agent may inhibit expressing the feeling, so as not to form an externally observable representation. If the private feeling is not expressed, that is, *Ann* or *Bob* does not press the button, then we cannot claim as before that "at this time, it will not cause external effects, so it will not affect the objective process. For AGI, it does not matter whether it exists or not." In this case, the internal causal effect cannot be ignored, that is, although there is no representation of the feeling at present, the internal feeling is indeed experienced. So at some point in the agent's future, the current inhibited expression of the feeling leads to a distinct representation. In this case, T becomes more complicated, but the form in the definition of the feeling remains unchanged.

However, reconsidering the above two cases, we are still not sure whether a "philosophical zombie" without the feeling will behave differently from an agent with the feeling, nor can we know exactly what D and T are. Of course, we can strive to find a computational model.

We call the computational model that meets the definition of the feeling above as the "simulated" feeling, and call the natural being of the feeling as the "innate" feeling.

The question then becomes, if this "simulated" feeling is somewhat different from the "innate" feeling, will it lead to a serious impact on an AGI system? How accurately can a computational model fit the "innate" feeling? All these need to be answered through experiments – We should find an agent with the "innate" feeling and compare the artificial agent with it, and see if they behave the same.

Unfortunately, in the above definition, it is assumed that the test circumstances must be open, which is crucial for AGI. Will an agent with the "simulated" feeling behaves the same as one with the "innate" feeling at any circumstances? This is difficult to confirm, because the condition of "any circumstances" is difficult to achieve, and an experiment can only test a model under limited circumstances. A more serious issue is how to compare the two kinds of agents "fairly" to judge if they behave the same.

In addition, what if the "innate" feeling cannot be fully modeled by a computational process? If it is necessary to have such the "innate" feeling in order to achieve human-level or higher-level AI, or realize the mind with artifacts, then where does the feeling ultimately come from, and how to achieve an interface that allows the "innate" feeling to be expressed?

Perhaps this is the hard problem for intelligence – in summary, if the "innate" feeling cannot be realized, what kind of impact will it have on the artificial intelligence system, and if there must be the "innate" feeling, how to realize it?

This problem cannot be answered under the current AI or AGI research, and it is left to the future. If the problem cannot be solved by the existing classical computing methods, then the feeling problem is the ceiling of the current AI research, and it must be broken through with the help of additional paradigms or technical means. It should be noted that this paper only involves discussions under classical aspects, and does not involve perspectives under quantum aspects, including, quantum brain theory [6], quantum logic [7], quantum cognition [3], *etc.* Is a quantum-based machine a plausible option? A separate article may be needed to address this issue.

5 Summary

If the subjective feeling plays an indispensable role, that is, agents with the feeling and agents without the feeling counterfactually (such as philosophical zombies) are distinct significantly in terms of performances, and if without the feeling human-level AI cannot be reached, then we must study the feeling. However, the feeling is subjective and private, while scientists study objective phenomena. If we cannot measure the feeling, we cannot study intelligence scientifically.

Therefore, this paper discusses the impact of the subjective feeling on AGI research. If the feeling is an epiphenomenon, we can just ignore it. Otherwise, it plays an indispensable role, so we try to give a possible solution for a machine to have the feeling: We define the feeling as "a tendency to change the input representation", and think "that the performance of an agent conforms to this definition" provides "positive evidence" for "that the agent has the feeling". If the definition captures the essence of the feeling, the feeling becomes measurable. However, whether it is possible and how to find a computational model, for the "simulated" feeling, to fit the "innate" feeling remain unknown.

References

1. Baars, B.J.: Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. Boundaries of Consciousness: Neurobiology and Neuropathology **150**, 45–53 (2005)
2. Baars, B.J.: A Cognitive Theory of Consciousness. Cambridge University Press (1993)
3. Busemeyer, J.R., Bruza, P.D.: Quantum models of cognition and decision. Cambridge University Press (2012)
4. Chalmers, D.J.: Facing up to the problem of consciousness. Toward a Science of Consciousness pp. 5–28 (1996)
5. Crick, F., Koch, C.: neurobiological theory of consciousness. Seminars in the Neurosciences **2**, 263–275 (1990)
6. Fisher, M.P.: Quantum cognition: The possibility of processing with nuclear spins in the brain. Annals of Physics **362**, 593–602 (2015)

7. Khrennikov, A.Y.: Ubiquitous quantum structure: from psychology to finance. Springer (2010)
8. Kim, J.: The mind-body problem: Taking stock after forty years. Philosophical perspectives **11**, 185–207 (1997)
9. Kim, J.: Philosophy of Mind. Westview Press (2011)
10. Levine, J.: Materialism and qualia, the explanatory gap. Pacific Philosophical Quarterly **64**(4), 354–361 (1983)
11. Levine, J.: On leaving out what it's like. In: Davies, M.E., Humphreys, G.W. (eds.) Consciousness: Psychological and Philosophical Essays, chap. 6, pp. 121–136. Blackwell, Oxford (1993)
12. Levine, J.: Purple haze: The puzzle of consciousness. Oxford University Press (2001)
13. Wang, P., Li, X., Hammer, P.: Self in nars, an agi system. Frontiers in Robotics and AI **5** (2018)
14. Wang, P.: A constructive explanation of consciousness. Journal of Artificial Intelligence and Consciousness **7**(2), 257–275 (2020)