Takeshi OHBA (Senshu University)

Self-knowledge and Moral Agency

We usually take it for granted that one knows one's own mind better than anyone else. In fact, we do mutually acknowledge the so-called first person authority of one's avowal of her belief. At the same time, we also realize that knowing ourselves is not so easy as to treat an old proverb "Know Thyself" as already expired. In this essay, I would like to examine recent debates about first person authority and attempt to boil down what seems essential from an ethical point of view.

1.

When, looking up at the sky, it occurs to me that it is raining over there, this thought is usually *accompanied* by a seemingly second-order thought that I believe that it is raining over there[1]. The latter thought is empirical, since it concerns a state of affairs, that is, my occurrent mental state. This thought seems further to be epistemicaly *immediate* in the sense that it is attained without any additional epistemic effort further than those which were exercised in gaining the belief about weather.

Thus, in general, when thought that **p** occurs to a person, it seems that she is, by virtue of it alone, already in a position to avow in the first person "I believe that **p**". Hearing a sincere avowal of this form, we audiences take it normally for true about the avower's occurrent mental state. Of course, we can and do sometimes doubt whether "p" is true, because her belief is about the world which we also face with. By contrast, in order to doubt whether she *really believes* so, we need to gather *extra* evidence which legitimately suggests something unusual in her psychology, say, a possibility of self-deception or the like. To this extent at least, a sincere avowal enjoys the so-called *first person authority*, a specific security against being exposed to doubt.

These all, I think, belong to platitudes about our daily communication, which even a tough behaviorist like Gilbert Ryle would admit. Then, is it not natural to say that a sincere avowal expresses the avower's knowledge of her own mind, her *self-knowledge*? For, a

[1] The reason I add the adverbial "seemingly" is that the notion of 'second order' thought usually induces one to take it for granted that it is a distinct thought from the first-order one, which seems in turn justify the locution of "being accompanied". It is, however, this induction that I would like examine in this essay.

sincere avowal is normally taken for a true report about an occurrent state of the avower. It is to this question, however, that many philosophers are now inclined to give a negative answer[2].

Of course, they do not deny the platitudes mentioned above. However, they deny that an avowal expresses a *substantial* knowledge of the avower's mental state. Instead, Crispin Wright, for instance, claims that the acknowledgement of first person authority is a constituent of a social framework of communications, which needs no epistemological or semantical investigation in order to explain and justify it. In other words, the acknowledgement is a socio-practical device for making communications smooth by acknowledging one's status as a competent communication partner[3]. Here sounds clearly an echo of what Rorty called "epistemological behaviorism" which aims at "explaining … epistemic authority by reference to what society let us say"[4].

However, nobody would dare to deny that, pragmatically speaking, the acknowledgement of first person authority performs such a socio-practical function. What is arguable is, however, a kind of *Eutypron* question: whether the first person authority of an avowal is acknowledged because it is practically important, or its acknowledgement is practically important because an avowal has indeed an essentially first person authority by nature. Where does then their negative claim that an avowal is *not* really a report, which is a manifestation of one's self-knowledge, stem from? It seems to stem from a consideration of an essential feature of an avowal.

An avowal is fundamentally distinct from other kinds of descriptive utterance. An avowal of one's own thought can enjoy an epistemically *immediate* authority in the sense mentioned at the beginning. This feature of an avowal does appear to collide with a general condition of a substantial knowledge, just as Paul Boghossian, for instance, made clear (1989, 5, 19.) For, in order for a belief to be a substantial knowledge about a state of affairs, it must be based upon an observation or upon inferences from observational evidence. If a belief which is based in neither way could be knowledge, it would be at best a formal or *a priori* knowledge. What is expressed, however, in an avowal cannot be a formal or *a priori* thought. So far, the suspicion about self-knowledge seems to be well supported by a general condition of knowledge.

---

[2] See Wright 1986, Wright 1989, Boghossian 1989, and so on.

[3] "The authority … is, as it were, a concession, unofficially granted to anyone whom one takes seriously as a rational subject" (Wright 1986, 401.)

[4] Rorty 1979, 174. He may be perhaps right in that he pursued to de-construct the obsession of giving an ultimate foundation epistemologically. But it does not necessarily follow that his diagnosis of first person authority is also correct.

2.

One might well here attempt to ascertain that a thought about one's own mind is based upon a kind of observation, '*introspection*' conceived as a sort of 'inner' *perception*. According to this sort of explanation, one introspects, that is, innerly perceives that a thought that **p** occurs, and then she reports that the thought that **p** occurs to her. Since it is the subject alone who can innerly perceive her own mental state directly, an avowal should enjoy the first person authority.

When one assimilates 'introspection' to perception this way, she has already adopted, so to speak, a *two-layered* model of a cognitive activity: a model according to which there is an item given independently of one's spontaneous activity on the one hand, and her capacity is exercised actively so as to recognize the item on the other hand. Let me call this a model of "*independently given* & *actively achieved*."

Following this model, a difference between perception and introspection would be explained as follows. In perception, the first layer, i.e. the 'independently given' is given *externally*, or from without, while in introspection it is given *innerly* within a mind. Accordingly, the "actively achieved" on the second layer is yielded by different cognitive capacities; in perception it is achieved by perceptual capacity, while in introspection it is done by an introspective one.

The model will be intuitively harmless about perception, insofar as the 'independently given' is taken to be a perceived object itself. It exists prior to, and independently of, our exercise of perceptual capacity, and we come to recognize it through our perceptual capacity. Thus, a relation between the perceived object and perception is *contingent* in the sense that it is always possible for an object to remain unperceived. The relation is therefore *mediated* by exercises of cognitive capacities.

A relation between one's own thought and an avowal, however, does not seem to be contingent and mediated that way. When a thought occurs, it is usually accompanied by a seemingly higher order thought which is expressed in avowal. If the latter thought is a cognitive gain by virtue of 'introspection' conceived as an inner perception, it would be always possible for the former thought to remain unnoticed by the thinker herself. This would, however, be implausible, just as Sydney Shoemaker emphasized it in his ingenious arguments[5].

Of course, we sometimes ascribe to a person a so-called 'unconscious' thought, which remains unnoticed by its very thinker unless the thinker comes to 'discover' it in some manner. But, to introduce the notion of unconscious thought here would doubly jumble up

---

[5] Shoemaker 1988, 1994.

the matter rather than settling. First, when, looking up at sky, a thought about weather occurs to me and a seemingly higher order thought accompanies it, both sorts of thought are not held unconsciously. Second, if a thought is entertained unconsciously, it could not be 'discovered' simply by introspection, since it is essential for a thought to be unconscious that it escapes any introspective screening. It is not by 'inner' perception but through complex inference, mobilizing memories, others' remarks about ourselves and so on, that we come to identify our own unconscious thought.

The relation between a thought and awareness of it, which I provisionally characterized above in terms of "being accompanied", is fundamentally different in kind from a relation between an object and the perception of it. A thought that **p** and awareness of it are connected so immediately that I do not need, in order to become aware of the thought, to achieve any further epistemic effort than that I achieved in ascertaining that **p**. It is precisely this immediacy that makes the model of "independently given and actively achieved" unsuitable to 'introspection'.

So far, the skeptics about introspective self-knowledge may well appear to be right when they characterize it banteringly as "an analogy of kaleidoscope" (Wright) or "a Cartesian theater" (Dennett)[6]. To be sure, this may be not sufficient for rejecting the model as a whole. There might be room, which I cannot yet find so far, for explaining the second layer, the "actively achieved", as a product of a cognitive capacity other than a sort of perceptual capacity. Christopher Peacock, for instance, once offered an argument for such a prospect, according to which an avowal expresses a higher order *judgment* based upon having an experience about a state of affairs. The judgment is, so to speak, so familiar that there emerges a short-cut circuit which gives us an impression of immediacy[7]. Although I do not deny that this could be the case, it sounds to me a little artificial and not fully convincing as an explanation of the immediacy in question[8].

If my observation is not totally off the point, there does not seem to be a hopeful prospect about the *cognitive* explanation of the first person authority based upon the two-layered model: a model of "independently given & actively achieved by a cognitive capacity." This also seems to explain partly why many philosophers are recently attempting to offer a non-cognitive explanation.

---

[6] Wright 1998, 22, Dennet 1991, 17, 113, 137.

[7] Peacocke 1999, ch. 5 - 6. However, Peacocke 1998 seems to pursue a little different line of thinking which could be taken for more Kantian.

[8] Regarding these issues, see, for instance, Gallois 1996.

3.

Very roughly speaking, there are two types of non-cognitive explanation of first person authority: one which appeals to a notion of *expression* and another which appeals to that of *commitment* or endorsement[9]. Do they, then, succeed in explaining that a sincere avowal is a *report* which can be a manifestation of the avower's self-*knowledge*?

Let us first consider an expressivist explanation. No doubt, to avow is a conative act of expressing one's mind by virtue of issuing a sentence which has a truth-condition. But an avowal is different in kind from a natural expression like an infant's crying. When a young infant cries because she misses her mother, her crying expresses her mental state. A careful observer will ascribe to her a belief that her mother is absent. Surely, her crying serves, on the part of hearers, as a report of her mental state. But it is by virtue of an interpretative activity by the hearer that her crying serves as a report. In contrast, a sincere avowal *is* a report, beyond merely functioning as a report on the part of an audience.

Generally speaking, when an informational system **s** exhibits some behavior, an observer ascribes to it a thought, say a belief that **p**, and makes a description that **s** believes that **p**. However, this does not in itself imply that **s** is itself in a position to issue an avowal in the first person. What is, then, further needed to be in a position to avow, when **s** expresses its mental or informational state?

Is it sufficient for being in a position to avow, if **s** exhibits a verbal behavior of avowing? This question may sound absurd, but I don't think so. Suppose that an alien from a far distant galaxy has eventually managed to master our language. He now uses also the construction 'I believe that … .' But imagine further that it turns out that he uses it merely as a kind of adverbial device for making assertions. This is similar to the famous thought-experiment by Sydney Shoemaker about 'self-blind'[10], which I cannot unfortunately deal with here.

Returning to the above situation, the alien can now join in the sort of conversation in which someone asks him whether he does indeed believe that **p**. However, even when he is participating in it, the notion of "one's own occurrent mental state" remains to him nothing but a phrase void of substantially descriptive meaning, just as the phrase "by the mercy of Czar" among the old Russian peasants was to a civilized foreigner. So, when asked whether he indeed believes that **p**, he can reply simply by considering whether there is a further evidence in the world which may be a good reason to withdraw his assertion. If there are none, he replies "Yes, I believe that **p.**"

Now, is his reply also a report made by him of his occurrent mental state? If an answer is

---

[9] As a typical instance of the former, see Bar-On 2004, and of the latter, Moran1994, 2001, Bilgrami 1998.

[10] Shoemaker 1988, 1994.

determined to be affirmative, what about an alarm device which issues a warning by making an artificial voice which sounds "I believe that such and such goes out of order!"? This may sound a little silly or exaggerated. I don't think, however, that the question about the alien's avowal is easy to answer simply with recourse to a behavioral criterion. To this extent, at least, it can make sense to ask what it is to be in a position to avow, and how it assures that a genuine avowal is also a report of the avower's occurrent mental state.

Why does it, then, make sense to deny that the crying infant or the seemingly avowing alien really issues an avowal? An answer would be that each of their behaviors of expressing does not seem to be an *intentional* actions of expressing their minds, an action the intention of which is to express their mind. If an avowal is an intentional action of expressing one's mind, an avower needs to *intend* to express, which in turn seems to require that she herself *understands* the content she intends to express. This point seems to me crucial. Since an avowal is an intentional act of expressing one's mind by issuing a truth-apt sentence, an avower must be able to discern whether the sentence she is going to use is suitable for expressing her mind. In order for discern this, she must *know* her occurrent mental state.

There can be a sense, then, in which the expressivist explanations of first person authority leave some aspect unexplained, since they say nothing explicitly about how an avower, unlike a crying infant, comes to understand what she intends to express, if it were not for a cognitive effort whether perceptual, judgemental, or else. It seems to be similar about another type of non-cognitive explanation.

4.

Consider now another non-cognitive explanation of the first person authority, which Richard Moran recently has offered[11]. According to it, when I avow that I believe that **p**, I am performing more than expressing my mind. In avowing so, I *actively identify* the belief as mine, and I *endorse* it by *committing* myself to the truth of the proposition that **p**[12]. In short, what constitutes first person authority is "the authority of the person to make up his mind, change his mind, endorse some attitudes or disavow it" (92.) This view differs from the expressivist explanations in that the act of endorsing by making a commitment is an essentially *normative* one, which *only* the avower can achieve in first person as a *decision* about what to believe, and which she takes the responsibility of. Thus, a source of first person authority, that is, what an particular person alone can perform in first person, is shifted more clearly from an epistemic space to that of practical and normative domain. In this sense, we

[11] Moran 1994, 2001.
[12] Moran 2001, especially 83-94, 113-120, 131-34.

may well characterize his approach as a pioneering attempt of what Bilgrami calls "normative turn"[13].

This sheds a fresh light upon the first person authority. The reason why we normally take a sincere avowal to be true about one's own occurrent mental state is, according to this view, that an avowal is a *practical* achievement of settling the matter about what to believe, which no one else could do on behalf of the avower. If an avowal is a mere description of one's occurrent mental state, there might well, in principle, be someone else who could quasi-omnisciently describe it better. However, it is only I that can make up my mind as to what I *should* believe.

Further, this view explains nicely why the so-called "Moore-paradoxical" statement sounds paradoxical; a statement of the form "non-**p**, but I believe that **p**." A sentence of this form does not commit a syntactical or semantical contradiction. It commits, however, according to the view, a pragmatic inconsistency, because an avowal of the belief that **p** requires one to commit herself to *truth* of **p**.

Thus, the explanation of the first person authority with recourse to the notion of commitment runs smoother and convincingly, so far. But, wherein does it differ from the above mentioned socio-practical explanation? According to it, to admit first person authority is nothing more than qualifying a person as a rational agent who can decide what to believe. A difference would be that the explanation with recourse to the notion of commitment admits an avowal to be a *report* made by the avower of her occurrent mental state, which is a manifestation of self-*knowledge*, while the socio-practical view does not admit it. In fact, Moran himself repeatedly emphasizes this[14].

How, then, can an avowal of one's own belief enjoy a status of a true report which is a manifestation of one's self-*knowledge*, if it were not backed up by any *epistemic* achievement at all? This question consists of, at least, two component questions. First, how can an avowal of a belief be so secure, as if it needs no extra epistemic effort? Second, does the explanation of first person authority in terms of commitment to truth fit as well to an avowal of other kinds of mental state, say, a desire or a hope? Is it likely that we do endorse some proposition which is true about the world when we avow our own desire or hope? In the following, I would like to mainly deal with the first question, because it concerns a more basic issue and a

---

[13] Bilgrami 1998, 214. However, it seems a little unobvious wherein his stance is distinctively different from what Wright calls "the Default View" (1989, 41), since the conceptual relation between our normatively "reactive attitudes" such as resentiment or gratitude and the agent's capability to explain her behavior can be more complicated than he seems to assume.

[14] Moran 2001, 104. This constitutes the essential issue when he argues against what he calls "the Presentational View", according to which "the verb-phrase 'I believe' … is a mode of presenting the relevant proposition" (71. 101.)

consideration upon it may, I hope, give some clue to the second question.

5.

The first questions might sound too silly. For, either sort of non-cognitive explanations, that is, explanation in terms of expressing and in terms of endorsement, seems to presuppose that one comes to notice immediately what is going on in her mind when she is thinking *consciously*. To be sure, this presupposition sounds intuitively natural. As argued earlier, when an infant begins to cry because of missing her mother, we ascribe a belief to her. Even then, however, we don't describe her believing it by saying "the infant *knows* that she *herself* believes that her mother is out of sight". If we are asked why, we would reply that she is not yet thinking *consciously*; not sufficiently conscious to entertain a *self-conscious* thought.

This reply sounds plausible to us. It is, however, upon this plausibility that a skeptic of the first person authority casts a doubt. He is suspecting that what one's conscious thinking enables her to immediately attain is too insubstantial to be counted as *knowledge* of her occurrent mental state. If a non-cognitivist merely presupposes and rehearses the reply without offering an articulated explanation, she might leave an aspect of the first person authority unexplained, the aspect which the skeptics bring into focus.

So, we now find ourselves dragged into a fairly perplexing situation, even a dilemma. If we suppose, following the cognitive explanation, that we come to notice the occurrent mental state by virtue of introspection conceived as inner perception, our explanation collides with the essential feature of avowal, that is, its epistemic immediacy. If we adopt, however, the non-cognitive strategy instead and merely presuppose that a conscious thinking enables us to immediately know what is going on in our mind, then our explanation would remain silent toward the skeptics.

In this situation, what attracts my attention is this: the non-cognitive explanation also shares with cognitivist the *two-layered model* of 'independently given and actively achieved', since it presupposes that a conscious thinking enables us to notice immediately our own mental state *independently of* our practical achievement of expressing or endorsing. In fact, according to Moran, "without endorsement the person cannot declare his belief through avowal of it. He might still, however, retain a kind of immediate epistemic access to it" (92.)[15]

Based upon the two-layered model, both cognitivist and non-cognitivist attempt to moor

[15] When Moran says "for any person who is self-consciously reflecting on his state of mind, there will be some answer to the question of what stance he takes toward what he *discovers* there" (147, emphasis added), Moran seems to think it possible, based on the presupposition, that a self-conscious thinker 'discovers' his own thought *apart from* his practical stance to it.

first person authority onto some rock of the second layer of 'actively achieved', onto some active achievement which an avower alone can perform and nobody else could do on her behalf. According to the cognitivist, a thought is simply 'given independently' and the thinker comes to notice it by virtue of 'active achievement' performed by introspective capacity. In the non-cognitive explanation, in contrast, what is 'actively achieved' is a *conative* or *practical* performance of expressing or of endorsing. What is 'given' is a thought, which the thinker comes aware of independently of a particular conative or practical achievement if only one is thinking consciously. In short, cognitivists see the 'active achievement' in avower's *cognitive* activity of introspection, while non-cognitivists see it in her *conative* or practical activity. Thus, a way of using the model is different. But, they share the model itself in common.

Now, it seems to me that what makes our situation perplexing stems from staying within this model. If we continue attempting to explain first person authority within this model, we would either take a cognition of one's mental state to be an achievement of our active introspection, or send it back to the layer of 'independently given' by simply presupposing that conscious thinking enables us to immediately know one's own mind. How, then, does conscious thinking enable us to notice *knowledgeably* our occurrent mental event, while a thinking which proceeds in a crying infant does not?

6.

I have so far repeatedly said that an occurrence of a thought that **p** is usually 'accompanied by a seemingly *higher order* thought that I believe that **p**'. Now, this usage of 'higher order thought' has some affinity with the two-layered model. It induces us, not to say "implies", to suppose that, when we are consciously thinking, there are *two distinct* thought-episodes, i.e. two distinct mental event of coming to entertain a thought; when one is consciously thinking, she comes to hold a belief about the world, on the one hand, and *in addition to it*, or *over this layer*, she also comes to notice the very thought-episode and to hold a higher order thought about the thought, on the other hand. Thus, while a crying infant entertains merely a first order thought, a conscious thinker holds also a corresponding higher order thought, *in addition to* the former. The popular terminology of a 'higher order' thought, which is sometimes said to be "locked on to" a first order thought[16], seems to me to show that we are inclined to think this way.

If we are seduced to draw such a picture, we are dragged into the perplexing situation.

[16] Burge 1988, 660. The term "locked on to" is used by Burge to point to self-referentiality, and now by Bar-On 2004 (162, 167) to summarize many attempt to block skepticism about self-knowledge.

For, nothing is yet said articulately about how the alleged 'higher order' thought-episode could be so secure as to yield *knowledge*. If this observation is not totally off the point, a way out seems to be hard to find unless we reject the two-layered model all together. We should now cease to think that there are two distinct thought-episodes; a thinking about the world, and another one about mental state. We should instead think that there be only a *single* thought-episode when we think consciously as Burge earlier suggested it (1988, 654). Thus, we are to attempt to explain how a conscious thinking makes it possible that *one and the same* thought-episode can concern *simultaneously* the world and the very mental state[17]. A possible key to this question seems to me to lie in Kant's criticism of the Cartesian reification of the *cogito*.

When Kant criticized Cartesian hypostatization of the *cogito* and argued that the *cogito* in itself is nothing but a "*form* of thinking" (A354) or "a *form* of representation in general" (B 404), his insight seems to be important to our problems. An essential core of his insight seems to me to crystallize in his assertion that "nothing can be thought and known, unless given representations share the act of apperception '*I think*' and thereby combined in *one and the same self-consciousness*" (B. 137, emphasis added[18].) This assertion could be understood as follows.

Suppose that a thought **p** occurs and at the same time the opposite thought not-**p** also does. This does not yet constitute a logical contradiction unless both thoughts are entertained by *one and the same thinker*. Otherwise it means that two competing thoughts occurred somewhere in the world respectively, which is not in itself a contradiction. The thoughts gain a status of thought to be normatively assessed precisely when they are subsumed and related to each other by the operation of the *cogito* or '*I think*' and embedded within a single scope of 'I think that …'. Thus, synchronically speaking, a thought can have a determined content and value only by virtue of becoming a knot in an inferentially connected web of a thinker's belief system, that is, only by virtue of the 'synthesizing' operation of the '*I think*'.

Diachronically speaking, the matter is basically the same. Suppose that, first, the top of a high tower appears in sight, next its trunk, and eventually its base appears. This series of visual experiences do not yet constitute a perceptual thought of one and the same tower, unless they "share the act of apperception '*I think*' and thereby combined in *one and the same*

[17] Burge claimed earlier that "[w]hen one knows that one is thinking that p, one is not taking one's thought that p as an object. … It is thought and thought about *in the same mental event*" (1988, 654, emphasis added.) His claim may well be taken for proposing the same as mine rather than suggesting the relation of 'locking on' between two thoughts.

[18] We perhaps should make an amendment to his phrase "nothing can be thought and known" by adding to the word "be thought" an adverbial phrase, for instance, "in the way which enables us to avow it", otherwise it would be difficult to allow for room for a notion of unconscious thought.

self-consciousness".

Thus, Kant says that the operation '*I think*' is "an act of spontaneity … which accompanies to any representation" (B 132), but "a mere consciousness of I … is nothing but … a form of representation in general" (B 404). This insight which I tentatively call the '*Kantian insight*' seems to me highly significant to our issues, exactly because it seems to shed new light upon a possible way out from the two-layered model.

7.

An idea which the Kantian insight suggests is this: an operation of the *'I think'* performs two things simultaneously. It organizes or composites, or to speak with Kant' own term '*synthesizes*', a meaningful thought, on the one hand, and simultaneously, in doing so, *announces authorship* of the thought, on the other hand. The term 'announcement of authorship' may sound exaggerated, but not necessarily. Suppose that, when I say something, an audience fails to identify its utterer and asks "Who said that?" Then I will reply saying, "It is I who said that." Such announcement of authorship is an essential condition for being a *responsible* thinker, an agent being able to respond to a question about who.

To be sure, thinking cannot be reduced to a linguistic activity. But, human thinking is already structured linguistically to that extent. To think as a human being involves using words latently toward possible audiences even when we think secretly in the dark, and it therefore requires a certain sort of preparedness to respond to a possible question "Who did think that?" or "Who are you to dare to think that?"

Thus, the operation of the '*I think*' is simultaneously cognitive and conative, exactly because its scope is, so to speak, forked. Its operation is *cognitive* in that it composites or 'synthesizes' an articulated thought, and it is at the same time *conative* or *practical* in that it announces authorship and makes a commitment to it as one's own thought. Thinking this way, we could say that a *conscious* mode of thinking is a mode in which the Kantian '*I think*' actively operates so that it can yield a single thought which is at the same time about the world and about one's very mental state. Then, there seems to be the possible prospect that we could explain the immediacy of an avowal and its first person authority.

The Kantian insight gives a prospect that a *conscious* thinking can have, simply by virtue of being conscious, a so to speak *double aspect*: it is simultaneously a thinking about an object and about the thinker herself. This is due to the dual-scoped operation of the '*I think*'. Following this prospect, an assertion that **p** and an avowal that I believe that **p** can be understood as two distinct expressions of *one and the same* conscious thought, respectively. We do no longer need to remain loyal to the two-layered model, supposing that

there are two distinct thought episodes, a first order thought and a higher order achievement.

Of course, a sentence "**p**" and a sentence "I believe that **p**" are distinct. But, we do not now, simply on that account, need to think that there are two *distinct* thoughts which are to be expressed by a distinct sentence respectively. Rather, in so far as they express a conscious thought, the two sentences are used to express the content of one and the same thought, though, with *different foci*: one focused upon the world, and the other upon the thinker. This is the reason why an avowal is immediate; why we need *no additional* epistemic efforts to avow. And an avowal can enjoy first person authority, because it is an explicit expression of the operation by the '*I think*' of announcing *authorship* of the thought.

Notice; this explanation allows the possibility that my assertion that **p** turns out to be false, while my avowal that I believe that **p** remains true. Indeed, what occurs is one and the same thought-episode. But, its content can be expressed with different *foci*; with a focus upon the world and with another upon thinker herself. According to the difference of focus, different sentences can be used for expressing the thought. Since the two different sentences have distinct truth conditions, it is possible that my assertion turns out false while my avowal of my own belief remains true.

The point is that there is only a single thought-episode. It is not the case that there is a first order thought about the world on the one hand, and a higher order thought about the thought one the other. Therefore there can be no room for it that the former can qualify as knowledge while the latter cannot.

8.

This explanation following the Kantian insight, however, leads to a bothersome question about the status of an avowal of one's *belief* in contrast to other sorts of mental state. Following the insight, any thought of mine is constructed and articulated under the operation of the '*I think*' regardless of its contents, whether the thought may concern the outer world, other minds, or my own mind. Consider the following thoughts which occur to me now:

(1) It is going to rain.
(2) She is afraid that he will get wet.
(3) It is desirable that her family will do well, or
(3') I desire that her family will do well.
(4) He has probably gone out with an umbrella, or
(4') I believe that he went out with umbrella.

When I avow each of these thoughts, I usually utter straightly the above sentences respectively, without using the construction 'I believe that' and embedding each of them

into it, that is, without uttering a sentence which is usually used for expressing a higher order thought. Nonetheless, when I avow, an utterance of each of the above sentences can be thought of to be anteceded by a construction 'I believe that', insofar as this construction can be taken for a manifestation of the '*I think*'. Now, according to the Kantian insight, none of the resulting complex sentences which equally anteceded by "I believe that" is to be seen as an expression of a higher-order thought which is distinct from the first-order one. Should we, then, think that all of these are used for avowing a *belief* or should we think otherwise?

If we choose the former, then we would have to give a highly special status to an avowal of belief in contrast to other sorts of mental state. We would think the following. An avowal of one's *belief* can enjoy first person authority in a special way, insofar as the leading construction "I believe that" can be taken for a manifestation of the operation of the Kantian '*I think*'. This is not the case, however, at least in the same manner, with an avowal of one's desire, hope or the like. When I avow that I *desire* that **p**, it is hard to regard the construction "I desire that" as a manifestation of the '*I think*' in the same manner as we can regard the "I believe that" so. What can enjoy first person authority in the special manner is rather a more complex utterance of the form "*I believe* that I desire that **p**", "*I believe* that I am delighted". This is because it is only the precedent phrase "I believe" which can be taken for a manifestation the operation of the '*I think*', while the embedded phrase "I desire" cannot be in the same manner.

This might seem, perhaps, an inevitable consequence of the Kantian insight. If so, there would be a striking asymmetry between an avowal of belief and that of other mental states concerning its first person authority[19]. I am not yet fully sure that this line of thinking could provide us with a wholly convincing explanation of our linguistic praxis. There seems, however, to be something in our praxis which could leave room for such asymmetry. In our daily communications, an avowal of a desire or the like seems to be much more vulnerable to a doubt than that of a belief. A doubtful question regarding a desire, for instance, "Do you really desire that p?", "What you really desire is rather that q, isn't it?" is often much easier to raise than against an avowal of a belief. No doubt, even a sincere avowal of one's own belief cannot be infallible. We are sometimes victims of self-deception regarding our own beliefs. However, an avowal of the form "I *believe* that I desire that p" is much more secure than the more direct "I desire that p."

If this is the case, the old proverb 'Know thyself' would remain still important especially concerning mental states other than belief. For, my avowal that I desire that **p** cannot enjoy

[19] This may be, perhaps, underwritten by our experience of self-deception, which seems to show that we are apt to be fallen into self-deception concerning our own emotion or practical attitude more often than concerning a brute fact in the world.

first person authority in the same manner, and to the same degree, as my avowal that *I believe* that I desire that **p** can. If I am confronted with the question "Do you really desire it?", I must ponder not only the desirability of **p**, but also upon myself, including my behaviors, my memories, remarks made about me by others, and so on.

To be sure, this line of thinking seems to incline toward an excessive type of intellectualism, if it leaves no room for first person authority of an avowal of a desire or the like. We do each other admit and respect the authority. But, it does not seem fully convincing to suppose that an avowal of a desire can enjoy first person authority as well, on the ground that an avower commits herself to *truth* of a proposition that the desired thing is really desirable[20].

We should rather think of first person authority of an avowal of a desire or the like in more 'voluntaristic' way in Moran's sense[21]. We might then be able to explain first person authority of an avowal of a deontic or conative sort of mental states like desire or intention in terms of an explicitly *volitional* verb like 'will', 'decide' or the like, which could be regarded as a conative counterpart of the Kantian '*I think*' [22]. This needs, however, a totally different analysis which goes beyond problematic of self-*knowledge* which I am dealing with here.

Anyway, following these lines, although there is a difference in kind between an avowal of belief and that of other mental states, this difference does not necessarily make the latter less secure. Then, we are now able to answer the second question mentioned at the end of section 4 negatively. Even though we may explain fairly well first person authority of one's belief in terms of commitment to *truth* of a proposition about the world, this explanation does *not* fit to an avowal of a desire, emotion, hope, or the like.

9.

Thus, following the Kantian insight this way, we must now divide avowals into two subclasses in respect of its security. One is an avowal of one's own *belief* which can be seen as a manifestation of the Kantian '*I think*', and another is that of other mental states, for instance, a desire, an emotion, or the like. When I avow my desire or emotion, my avowal usually enjoys, so to speak, *prima facie* first person authority. These avowals are, however, more vulnerable to cross-examinations than that of a belief, just as is seen in the previous section. By contrast, an avowal of a belief can enjoy first person authority *simpliciter*. When

[20] Moran seems to be inclined to think of desire this way (116-118).
[21] Moran 1988 gives such an explanation about authority of an avowal of intention.
[22] Rather, we had better, perhaps, posit a more basic operation of '*I decide*' which involves cognitive '*I think'* as well as conative '*I will*' at the same time. Then, an explanation of first person authority would, perhaps, inherit some idea similar to what Moran calls "constitutive" view (Moran, 2001, 38.)

a thought that a professor is admirable occurs to me, my avowal "I *believe* that I admire the professor" is true, even though it may turn out that I really rather look down upon him. This is simply because the precedent phrase "I believe" can be seen as a manifestation of the Kantian '*I think*' which enables us to entertain an articulated and meaningful thought at all, regardless whether its content concerns either the world or one's own mind.

An avowal of belief has, thus, a special *epistemic* security not because it is an expression of a seemingly higher order thought which is supposedly attained by virtue of an extra epistemic effort of the thinker herself, but because it can be seen as manifestation of the fact that the first order thought itself is articulated under "the act of apperception '*I think*' and thereby combined in *one and the same self-consciousness.*" An avowal of *any* occurent belief, so-to-speak, *automatically* enjoys the epistemic security and therefore the first person authority*,* at least insofar as one is thinking consciously.

Then, it might seem, to speak with Boghossian's phrase, that an avowal of an occurent belief is '*insubstantial*' in the sense that an indexical judgement that 'I am here now' is insubstantial because it is always true by virtue of grammar of indexicals, even when we reject to join in his externalism. Does this, however, imply that what is expressed in an avowal of belief is too 'insubstantial' to constitute *knowledge* about oneself?

Is it, to begin with, true that the indexical sentence "I am now here" is doomed to be too insubstantial to be an expression of one's self-knowledge? It appears to be so, at first glance, because it tells nothing about a particular time and place where he is then. Suppose, however, it is uttered just after his being involved in a serious disaster. Hearing it, we, the audience, then understand that he *knows* that he *himself* is still alive. Suppose that, further, he continues to utter, in reply to our question about where he is, that "I don't know, but there is a huge bridge in front of me, and a tall tower over there." We then guess the place where he is and, in doing so, we take it for granted that he *knows* what sort of place he himself is located at, even though he cannot at that time specify the place in terms of proper name yet. Thus, the indexical utterance can be taken to express the utterer's self-knowledge, at least insofar as he has already acquired capacity to correlate his description of the landscape with a particular point on a map[23].

Now it seems to me that the circumstance is the same concerning an avowal of a belief. The fact that it is true, "based on nothing" (Boghossian), only by virtue of its 'grammar' does not necessarily imply that it is too insubstantial to be taken as an expression of one's self-knowledge, at least insofar as the avower has already mastered capacity to correlate the content of belief, or concepts which compose it, with a description of the objective world.

[23] About such superposition of an objectively locating description onto a subjective description of a landscape, see Lewis 1979 and Evans 1982.

10.

If we still deny that an avowal of a belief is an expression of self-knowledge on account of its being 'insubstantial', then we are to think that the avower does *not* need to *know* that she herself entertains the belief despite the fact that she consciously believes it. Could we still then acknowledge that she knows about the world which her belief is about?

To begin with, avowing one's belief is a linguistic performance of announcing authorship, which cannot be achieved without being conscious of one's entertaining the belief. And, following the Kantian insight, both my consciously believing that **p** and being conscious that it is I who believes that **p** are two aspects of one and the same thinking. Then it seems extremely hard to suppose that a conscious belief about the world can be a candidate of knowledge while self-consciousness cannot. This diagnosis seems to be able to be supported by our practice of ascription of knowledge.

Suppose that an infant watches her surroundings and behaves herself in a certain way. Then, we legitimately ascribe to her a certain belief about the surroundings. Suppose further that it turns out to be *true*. Now, her belief is based upon her own observing her surroundings and is in fact true. Even then, we would *not* regard her belief as *knowledge*, unless we can say that the infant is aware that she *herself* observes it[24]. This seems to me to offer a reason for accounting consciousness of one's own belief as a sort of knowledge.

Knowing is essentially an active, first person business of a reflective and responsible agent who can account for reasons why she thinks so. To have knowledge is not merely a putative informational state which an observer ascribes to others on the basis of observation of their behaviors. To ascribe knowledge to someone involves regarding her as being conscious of what she *herself* believes[25]. Then, we had rather think that what an avowal expresses constitutes, so to speak, '*transcendentally' basic* knowledge, in the sense that it alone makes possible knowledge in general. Citing again the core phrase of the Kantian insight, "nothing can be thought and known, unless given representations share the act of apperception '*I think*' and thereby combined in *one and the same self-consciousness*." If these considerations are not totally off the point, we can answer the *Euthypron* question mentioned in the beginning: the first person authority of an avowal is acknowledged not simply because it is practically important; rather, acknowledging first person authority is important to society exactly because an avowal expresses the transcendentally basic knowledge about oneself.

---

[24] Despite 'the Gettier problem' I think that the notion of "true, justifiable belief" can serve as a common sense criterion of knowledge. Then, to be aware of content of one's own belief can be counted as a constituent of justifiability without committing ourselves to a strict version of internalism about knowledge.

[25] This appears, to be sure, to get into trouble concerning the so-called tacit knowledge, but it seems to me possible to separate it from occurrent consideration if we could take it a kind of practical capacity.

11.

Now I have made much of the Kantian insight in searching a way out from our perplexing circumstances as regards first person authority. The core the insight lies in focusing on the active operation of the '*I think*' which alone makes possible any articulated thought. Does not it, then, lead to restoration of the alleged absolute certainty of the Cartesian *cogito* to emphasize this way the constitutive importance of *self-consciousness* ? On the contrary, it is precisely for the purpose of criticizing the Cartesian hypostatization of *cogito* that Kant emphasized that, in *cogito*, "the I is a mere form of consciousness" (A328) and "a mere consciousness of I … is nothing but … a form of representation in general" (B404).

These remarks are still now, not only an ontologically significant warning against hypostatizing the *cogito,* but also in moral philosophy highly important when we think about first person authority of an avowal. An avowal manifests the synthesizing and authorship-announcing operation of the '*I think'*. In avowing, one manifests her competence as a responsible thinker, responsible to a question of "Who and What?" This crucial point of an avowal, however, can reinforce hypostatizing the '*I think*' under the name of an 'inner self', 'true self', or the like, despite Kant's repeated warning.

This is the more likely to happen when we remain, though implicitly, within the two-layered model of 'independently given and actively achieved'. For, the alleged inner self appears to be a suitable executive of the supposed higher-order achievement exercised onto the given. Once the '*I think*' is regarded as an achievement by "inner self", the notion of "inner self" in turn alludes us to take the two-layered model for ontologically guaranteed. If we are tempted this way, we become inclined to figure out as follows: first, various kinds of thought-episodes happen within us independently of our exercising practical or conative capacity. Next, our 'inner selves' censor them and actively pick out a thought to endorse and to express.

What is problematic about this figure is that it is totally *up to* one's 'inner self' to actively pick out a thought to endorse. To be sure, it is up to me to decide what to think and which thought to avow, in an ordinary sense of 'up to'. This sense of 'up to me' is indispensable to a responsible agency, as Moran rightly emphasizes it. In the figure in question, however, one would be seduced to overdraw, with recourse to the notion of 'up to', a strong conclusion from the fact that our mental state is sometimes or even often indeterminate about what to believe. To be sure, our mental states are not always stable and fixed. According to Moran, "without endorsement the person cannot declare his belief through avowal of it. He might still, however, retain a kind of immediate epistemic access to it." (2001, 92) What status should be, then, conferred on such a thought which "I retain a kind of immediate epistemic access to" without endorsing it?

An answer will be this: such a thought merely constitutes content of such an unfixed mental states as 'being wondering whether', 'being afraid that', or the like. Another answer, however, could be drawn if the '*I think*' was reified as an 'inner self'. It would run as follows: the thought in question could not yet be qualified as my own belief, since I still suspend endorsing it. The thought, instead, remains a mere happening in my mind. Although it in fact occurred in my mind, it is a mere exudate within me, instead of being my own thought which I am the author of. Thus, it is my 'inner self' who does or does not *identify* a thought as my own.

If the operation of '*I think'* is hypostatized as an achievement of 'inner self', the notion of 'up to me' could then degenerate into a tool for rejecting identification of an uncomfortable thought within me as my own, and for spinning a story about myself which sounds sweet to my ears. Suppose that a thought occurs to me which embarrasses me by making me realize that I am the kind of person who does entertain such a thought. Then, not only can I decide not to avow it, but also I could even refuse to identify it as my own thought, following a mandate issued by my 'inner self'. In this respect, the hypostatized 'inner self' resembles an absolute monarch who governs his territory and determines at will who are inhabitants and who are not.

12.

An 'inner self' *qua* such an absolute monarch could exercise his sovereignty further outwards in order to reject 'domestic interferences'. Suppose that someone, observing my behaviors, describes my thought in a certain way which jars on my ear, although I do not regard it as a mere framing up. Then, not only could I refuse to identify it as a description of my own thought by saying that "it is not a thought held by *true* and *real* me, although it might have appeared as if I think so in her eyes", but also could I further attempt to justify my response by talking to myself that "others have no authority at all to determine what my *true* self is thinking" just as an absolute monarch rejects, under the name of 'interference in the domestic affairs', any advice from without which jars on his ears.

The notion of 'up to me' could thus degenerate into an emblem of the Guards of a monarch, if the operation of the '*I think*' is hypothesized as discussed. This is morally serious, not only in the sense of leading to irresponsible and cheap self-justification, but also in that it undermines a basis of both human agency and mutual acknowledgement, just as, for instance, Hegel's famous argument about 'the dialectic of a master and a servant' suggests it. In order for us to mutually acknowledge as responsible agents, each party must be prepared to *superpose* a description offered by others in third or second person onto one's own first

person avowal. A superposition of the form "it is me that you describe so" is indispensable to mutual acknowledgement, no matter how I am unwilling to hear a harsh description by you.

Such superposition is further the very condition for the indexical "I" to be meaningful, though this issue runs beyond the range of this essay[26]. It is essential to the meaning of "I" that it refers to exactly a person whom the word "you" in your mouth refers to. Each of us can be a human being only by virtue of mutually being inter-human. What makes inter-human-being possible is exactly that each of us thinks and talks about oneself under the schema "I = you of you, and vice versa", where each has already mastered how to identify the referent of both 'I' and 'You' in terms of proper names.

To be sure, we might sometimes be seduced to take these facts for a merely contingent matter of biological and sociological facts, and to suppose that the indexical 'I' could be meaningful in a solipsist's mouth, totally apart from the above schema. To cite Ludwig Wittgenstein's phrases, the supposed solipsist might assert that "[o]nly what I see … is really seen", and explain his use of the 'I' by claiming that "the word 'I' I don't mean L.W.", but that "it will do if the others understand 'I' to mean L.W., if just now I am in fact L.W." (64.) What is essential is, however, that "it is conceivable that my *fellow* creatures thereupon will arrange their notation so as to *fall in with me* by saying "so-and-so is really seen" instead of "L.W. sees so-and-so", etc., etc." (66, emphasis added.) This is precisely because his audiences can correctly understand that he tries merely to "adopt a symbolism in which a certain person … holds an exceptional place". They understand this precisely by virtue of taking his utterance of 'I' under the above schema, even if the solipsist insisted that "the other should not be able to understand 'what *I* really *mean*'" (65.)

To continue to borrow Wittgenstein's phrase, there may be no objection to adopting such a symbolism in itself. "What, however, is wrong, is to think that I can justify this choice of notation" (66), by virtue of positing, as a referent of "I" in an avowal, an 'inner self' who just happens to be identified as a person L.W. now. This diagnosis by Wittgenstein of solipsism seems also to support the Kantian warning against hypostatizing the cogito as a subsistent inner monarch who could govern his inner territory at will, though the ontological questions about the 'self' is beyond the scope of this essay.

Concluding Remark

The mutual acknowledgement through mutual superposition is essential to the very

[26] About my view on this issue, see Ohba 2003.

notion of responsibility. To be *responsible* involves to be prepared to *respond* to a question and a calling issued from others, even when it jars on my ears. The alleged 'inner self" could, however, gerrymander a range of my preparedness to respond by treating an unpleasant question or calling as a mere *noise* rather than a voice, just as it could refuse to identify an embarrassing thought as my own by treating it as a mere *happening* within me.

I am not claiming that any sort of hypostatization of the '*I think*' would necessarily lead to the degeneration. It would be hard, however, to deny that there can be some affinity between them. An old far-western legend tells an interesting episode about gerrymandering a range of responsibility: an autonomous and well-behaved person "wanted to vindicate himself, and he asked 'Who is my neighbor?'" To this extent at least, the Kantian criticism of hypostatization of the 'I think' is relevant, even if my interpretation of Kant is exaggerated and not yet conclusive*.

Reference

Bar-On, Dorit 2004: *Speaking My Mind*, Oxford U.P.

Bilgrami, Akeel 1998: 'Resentment and Self-knowledge' , in Wright, C., Smith, B., and MacDonald, C. (eds.) 1998

Boghossian, Paul 1989: "Content and Self-knowledge", *Philosophical Topics*, 17.

Burge, Taylor 1988: "Individualism and Self-Knowledge", *Journal of Philosophy,* 85.

------- 1996: 'Our Entitlement to Self-knowledge', *Proceedings of the Aristotelian Society*, 96.

------- 1998: 'Reason and the First Person', in in Wright, C., Smith, B., and MacDonald, C. (eds.) 1998

Dennet, Danniel, C. 1991: *Consciousness Explained*, Boston, Little Brown.

Evans, Garret 1982: *The Varieties of Reference*, Oxford U.P.

Gallois, Andrè 1996: *The World Without, The Mind Within*, Cambridge U.P.

Kant, Immanuel 1787: *Die Kritik der reinen Vernunft*, Hamburg, Felix Meiner (1956)

Lewis, David 1979: "Attitudes De Dicto and De Se", rep. in his *Philosophical Papers*, vol. 1, 1983, Oxford U.P.

Moran, Richard1988: "Making Up Your Mind: Self-Interpretation and Self-constitution", *Ratio,* NS 1.

------- 1994: "Interpretation Theory and the First Person", *Philosophical Quarterly*, 44.

------- 2001: *Authority and Estrangement*, Princeton U.P.

Ohba, Takeshi 2003: *How am I 'I'*, (in Japanese), Tokyo, Kodansha.

Peacocke, Christopher 1999: *Being Known*, Oxford U.P.

Rorty, Richard 1979: *Philosophy and the Mirror of Nature* , Princeton U.P.

Shoemaker, Sydney 1988, "On Knowing One's Own Mind", rep. in his *The First Person Perspective and Other Essays*, 1996, Cambridge U.P.

------- 1994: "*Self-knowledge and " inner sense", lecture II* ", rep. in his 1996.

Wittgenstein, Ludwig 1964: *The Blue and Brown Book*, Blackwell, Oxford.

Wright, Crispin 1986: "On Making Up One's Mind", in Weingartner and Schurz (eds.) *Logic, Philosophy of Science and Epistemology*, 1986, Vienna, Kirschberg.

------- 1989: "Wittgenstein's Later Philosophy of Mind", *Journal of Philosophy*, 1989

------- 1998: "Self-Knowledge: The Wittgensteinian Legacy", in Wright, C., Smith, B. Macdonald. C. (eds.) 1998.

Wright, C., Smith, B., and MacDonald, C. (eds.) 1998: *Knowing Our Own Minds*, Oxford U.P.

---

* I am deeply thankful to Jeffery Man (Susquehanna University) for his detailed advices to my earlier draft, especially concerning so many grammatical mistakes. A part of the earlier draft was read on a conference at Osaka university.