Journal of Social Complexity (1)4: 14-19, February 2004, Bandung Fe Institute.

INTERNAL CONFLICT
natural self-deception of human cognition

RENDRA SUROSO
Dept. Cognitive Science
Bandung Fe Institute
Jl Cemara 63
Bandung, West Java
+62 22 203 8620
raysuroso@pulp-fiction.com

ABSTRACT: How one can believe that he knows something is found to keep changing, even conflicting from time to time. When an occasion occurs, one may have belief differently as she usually does, as reflected in actions she chooses, and hence, at least in appearance, she seems to deceive herself. That this is not pathological but just natural, with the help of intentionality and intentional stance, will be proven in this paper.

Keywords: self-deception, intentionality, intentional stance, modularity of mind, theory of mind.

1. INTRODUCTION

Back to 1981, Bach (1981) introduced a concept called 'self-deception', a phenomenon that is highly in relevance to various cognitive traits, e.g.: Blaise Pascal who believes in his logical rejection to the existence of supreme being but also believes that it exists in order to find ultimate heaven; act of abruptly criticizing a TV show based on one's belief that the show is useless while never stop seeing and probably enjoying the entire show; prior believed knowledge that shows an evidence of the damaging effect of corruption which actually does not stop a person doing it, etc.

This conflicting internal state of human cognition, particularly human propositional attitudes ranges from beliefs, desires, intention, wishing, wanting, to hope and fear that in turn will affect human rational standard.

Bach's definition of thinking that *p* that for many philosophers is not distinct from believing that *p* is 'occurrently believing' that *p* while his definition of believing that *p* is 'dispositionally believing' that *p*. Thinking hence refers to occurrence, while believing to state, so both are the same but having different nature as he suggests that thinking that *p* does not imply believing that *p*. Bach gives an illustration provided by the case of phobia: someone who believes that traveling by air is as he thinks as safe as other means of transportations cannot help thinking that it is dangerous, even if he realizes that it is irrational. Of course, an observer cannot justify his degree of rationality, but if such degree is applied to other reasoning, the result does not indicate a self-deception.

Later on, he enter 'desire' distinct from belief as a motivation that helps the belief to be deflected, but not very clearly as he adds in footnote:

> Some cases might be more aptly (and specifically) described in terms of an emotion. For example, a self-deceiver might dread that *p* and thereby desire that not-*p*.

Finally, Bach's analysis suggests a quiet cumbersome description of being self-deceived that not-*p* of a person *S* over period of time $t_1 - t_2$ iff:

1. *S* desires not-*p*;
2. *S* believes that *p* (or has strong evidence that *p*);
3. 1) and 2) combine to motivate *S* to *avoid* (as he does) the sustained or recurrent thought that p; and
4. if 3) is satisfied, then even if the sustained or recurrent thought *of p* were to occur during $t_1 - t_2$, 1) and 2) still motivate *S* to avoid (as he would still) the sustained or recurrent thought that *p*.

In Bach's definition, thinking *of p* and thinking *that p* are different in terms that thinking of *p* can implies either thinking that *p* or thinking that not-*p*.

Bach indicates an inadequate explanation of such deception. If his motive is just to pose the phenomenon as a logical defect among many kinds of irrationality, then he succeeds doing so, but then, so what? Furthermore, for the sake of avoiding contradiction, Bach's analysis of self-deception does not imply intentionality though motivated. This is another central point under discussion of this paper.

In this short paper, differed from Bach's non-intentional viewpoint, we will examine the possibility to gain explanation of such thing in terms of its intentional stance in a way people usually deal with 'belief' and 'desire' and how they interact.

In short, human conflicting state is found to be ordinary in sense that it is the characteristic of many sorts of behavior. How it appears, how it prevails, how it eventually vanishes, and how it is found to be just natural, those are our questions.

2. INTENTIONALITY AND MODULARITY

Dated back to philosopher Franz Brentano, the notion of intentionality straightforwardly as a word that stands for the phrase 'directedness upon an object' or as Searle (1980) puts it, feature of many of our mental states about states of affairs in the world, intentionality comes to the point that if one has a belief or a desire or a fear, there must always be some content to that belief, desire or fear, existent or non-existent, for real or pure hallucination.

As an evergreen philosophical idea, intentionality has far gone through vindications and predispositions. Vindications come mainly from eliminative materialism, while predispositions especially focus on how we can interpret it and generate useful explanation from it. Under what mechanism do we interpret the behavior of intentionality of an entity?

How we intentionally treat the compounds an entity (not necessarily human since we can do the same thing to an artifact) is what Dennett (1987) calls intentional stance as opposed to physical stance or the real nature of such compounds. To put it bluntly, by pretending as if an entity were a rational agent, then intentional stance is how we interpret the behavior of such entity under consideration of its beliefs and desires. This approach is considered to be merely instrumental for some researchers, but its relevance to biological evidences are vital.

Another view to look at the problem (or potential) of intentionality is through linguistic account that begins with linguistic representation, and to explain the content of mental states in terms of the content of sentences that brings mental states.

For many reasons, we owe much to Chomskian revolution (Chomsky, 1966) on generative grammar that successfully states a ground on which all languages in the world can be put into the same mode of treatment and possible transformations that later even gives a notion of linguistic innateness and the difference between deep and surface structure. What seems to diminish in early effort to dive into intentionality through linguistic accounts is merely the deep structure as a source of semantic interpretation (Tarigan, 1984). To cope with such problem on semantics, finally a new idea or class of ideas based on a hypothetical 'language of thought' is proposed (Fodor, 1975) despite the inconvenient nuances around its mentalistic properties.

If we try to put both view into one single theme, we can either generate a domain specific view on how cognition works or particularly, how such specific domains operate and are characterized under the notion of modularity.

Modularity captures the existence of separated mental modules that operate throughout cognition, consist of computational mechanism that is innately specific, special-purpose, and informationally-encapsulated (do not have total access to all information the cognitive subject receives) (Fodor, 2000). This partition of cognition to various separated modules, at a glance, seems to wreck more domain general account of cognition, especially those under the name of connectionism that have strong confidence that neurally constructed cognition centered on cortical areas is the main line of bottom-up-built building blocks of linguistic cognition.

This debate is out of the concern of this paper since we adopt the implementative view of connectionism on cognitive architecture (Fodor, Pylyshyn, 1988). This way, despite empirical evidences both sides claim, we simply refer to the difference of level of description between neuronal architecture and linguistic operation. This way, one does not dismiss another but only through implementation of connectionist architecture the linguistic operation is expressible according to several findings and propositions, e.g.: in the study of consciousness and qualia (James, 1890; Edelman, 2002).

3. MODULES AND THEORIES OF MIND

How to put both approaches to intentionality together? On what ground should intentional stance and modularity should meet and tell us something in advance? Do not both have their own conflicting assumptions?

Those questions are not really a deep philosophical speculative conundrum after all. Both share the same or similar biological evidences, or at least, reinforcement. In his intentional stance, Dennett points out to the work of developmental neurophysiology of Alan Leslie and Simon Baron-Cohen. Intentional stance finally comes out with the idea Leslie called 'theory of mind mechanism' or 'theory of mind module' as Dennett puts it (Dennett, 1995), designed to generate second-order beliefs (i.e.: beliefs about the beliefs and other mental states of others).

Experimentally, Baron-Cohen, Leslie and Frith show a requirement of a child to have specific theories about others' minds that in general can have specific impairment such as autism, a state where the theories are considered to be deflected so autistic child fails to represent mental states of others (Baron-Cohen, 1985). This is exactly one of Dennett's intentional stance adoptions called folk psychology. Even further, Baron-Cohen in his review on various studies and experimentations on the theory of mind, suggests that his modularity theory will eventually refute non-modular or general purpose theories, in accordance with Fodorian modularity, and that there is at least a minimalist innate modularity theory involving lower-level *social perception mechanism* (Baron-Cohen, 1998).

How about linguistic modules? Chomsky as the father of generative grammar suggests that linguistic capability is innate (Pinker, 1997) because with the least effort, even a mentally-impaired child can acquire a mother tongue. But Chomsky does not stand for such innateness or tacitness if based on evolutionary accounts, but insists more fundamentally on physical accounts.

Taking Chomsky's view of innateness of language based on physics rather than biology and leaving it as a puzzle or unsolved mystery is awkward, but taking Fodor's view that modularity is a result of adaptationism is helpful to construct further explanation. But still awkwardly, the faculty of language as a hypothetical module is an exception of adaptationism. This point is still debatable and intriguing (see Okasha, 2003). One way to explain this is probably by totally separating syntax and semantics completely, the former is part of Chomsky's innateness; the latter is probably Fodor's language of thought free from deep structure.

A newly developed branch of science called evolutionary psychology (Tooby, Cosmides, 1989) denounces the demise of such linguistic non-modularity and in return, proposes a new viewpoint to cognition as massively modular. Consequently, this massive modularity finally comes to specific modules such as facial recognition, and presumably many more. Central processes collapsed, and there comes many specific modules such as cheater detections on social exchange, and we do not need any kinds of pre-linguistic or quasi-linguistic language of thought to get involved. Unfortunately, massive modularity at a glance tends to disparage cognition into pieces that operate independently since module is encapsulated. Without an explicit mode of interaction among encapsulated bundles of information brought by each module, the significance of linguistic representation that brings difference between human and other species tends to fall apart.

Fodor on the other hand, insists that modules are separated from central processes, so the central processes such as thinking and reasoning and generating abstract thoughts and principles are non-modular. Dennett once call it a dysphoria of top-down architecture (Dennett, 1994) when such non-modularity of central processes is built upon another field of research studying visual imagery that for some researchers is cognitively impenetrable (Pylyshyn, 1999). (Pylyshyn's construction on visual imagery is non-modular though tacit. For an early introduction of this topic, see Pinker (1984)).

From the issue of adaptationism at least we can have two viewpoints: that non(or quasi)-adaptationist viewpoint on language faculty only implies the existence of some hypothetical quasi-linguistic language of thought that is not directly related to or does not have the same characteristics with modules, and; that adaptationist viewpoint on language faculty does not occupy the capabilities of language to represent states generated by other modules.

Finally, following Baron-Cohen's work, theory of mind as social perception mechanism is modular. We do not have to put into semantic processes, any kinds of social perception mechanism underlying intentional stance. Furthermore, this kind of theory of mind, if it is modular at all, then must follow the impenetrability constraints. Following this notion, we will have to say that intentionality is not determined by cognition or any sort of higher level process or central processes, or even simpler, it has its own 'logic'.

4. CONCLUSION

The starting point of this paper showing self-deception as ordinary characteristics of behavior thus comes to verification that conflicting state across beliefs during a period of time.

Not only we make a slight progress from Bach's earlier departure in analysing such behavior, but also, we suggest that self-deception has its own logic among other logic found in various modules. Evolutionary psychology may give an evolutionary account to this self-deceiving module or class of modules (since self-deception, according to intentional stance, can also be directed to artifacts), but without sinking it too deeply into massive modularity and eliminating the capability of abstract reasoning.

Further one might argue that the learning and gaining factual knowledge about the world lie on language of thought. This is neither helpful nor efficient unless the real thing about language of thought has been established along with its interface with modular inputs, e.g.: between visual recognition and visual imagery.

Different from Bach, we suggest that self-deception is primarily intentional. It occurs because there are many kinds of results of adopting intentional stance through various modules. One of the modules is theory of (other's) mind that according to intentional stance, can also extended to become theory of artifacts. Since each particular module is tacit and has its own logic or process, input it gives to central processes along with other modules are various, and sometimes distorted, and even conflicting. Through this explanation, we do not have to take into account the ideas of multiple selves Bach and in different fashion, Schwitzgebel (1997) decisively attack, despite the difference of 'location' of propositional attitudes that they and we take.

But to be practical, this is the case usually occurs in formalization of cognitive capabilities interfered by propositional attitudes, commonly found formalization of (traditional) AI researches. Following the work of Newell (1990) on knowledge level analysis, in this field of research, belief is considered to exist in informational level, expressible in syntax, performs an input to particular semantic operation, and the most important is that it requires various informative actions that may yield conflicting result. A technique to expel this kind of conflicting result is by implanting a classification process known as credibility (van Linder, van der Hoek, Meyer, 2000). One way to perform this sort of process is through default beliefs. Even further, in AI tradition, known belief is considered to be the knowledge itself.

Default belief is default reasoning Bach also ever mentioned (Bach, 1984), that is an act of making implicit assumptions (that do not actually come to mind) and following rules like taken-for-granted or not-worth-considering. The strange point of Bach's point is that default reasoning does not really come to mind, or in other words, it is a form of Dennett's intentional stance or a form of another module applied the output of other prior modules, integrating various information and has possibility to give an irrational performance. Hence, it 'filters' conflicting information according to AI viewpoints, but does not necessarily give more rational performance. This is another source of self-deception during a period of time albeit it hypothesizes another kind of module namely default reasoning. Whether modular or not, covering this issue in a more general sense, Verschure and Althaus (2003) proposes, though not very clearly, to redefine Newellian knowledge level in which reasoning is definable usually by Bayesian inferences, by 'equating' it with intentional stance.

But the point of this paper is only to show that self-deception is not only possible and ordinary. And in addition, it comes from intentionality of general consciousness or intentional stance towards intelligent or rational behavior.

Human during her lifetime must deal with a lot of internal conflicts, especially in her modular output. The output is so distorted sometimes, so whenever it becomes an input of language of thought affected by rational reasoning that involves deduction or inference, it is her duty to adjust the input with some kind of factual state (What is factual and what is not are part of another philosophical debate. See Bicchieri (1987) for a short review to this debate. Whether conventionalist or justificationist, many researchers tend to assume a formal procedure to model or to test a person or an agent's ability to access factual state.) Sometimes she fails, sometimes not, either because each module has specific process, or because of failure to adopt rational thinking in central processes.

In turn, the failure or success is not a matter of moral or epistemological question. It is just natural, perhaps biologically natural.

Cognitive science is where philosophy goes when it dies.
Jerry Fodor (1994)

5. FURTHER WORKS

The early work presented here is very limitedly abstracted from the standpoint that belief and other propositional attitudes really do exist, and such existence comes from evolution. Consequently, the close relation between mental and biological properties of cognition comes again into question to prevent us from misplacing the description to each level. Hence, in the future, following this paradigm will at least to certain point, eliminate the conflicting arguments from eliminative-materialistic standpoint and the more mentalistic ones, no matter what functional approach they adopt.

It seems that the works in visual cognition represent too much – though not necessarily astray – emphasis on biology, while works on various approaches in folk theories represent too loose – though not necessarily rhetoric – description. On the other hand, the works on 'belief box' remains hypothetical, or in Dennett's words, it was widely adopted just for the sake of convenience. And we are really aware that we have taken 'belief fixation' a little bit closer to modularity, and we do not attempt to preserve modular equipotentiality. Not that because this approach was set hypothetically, but as Hilary Putnam says, too many speculations owing 'pay-off' for psychology or philosophy can be deemed extremely problematical, not to say remote from practice.

In short, our further works will be departed from the framework depicted by the existence of those three levels – with a wide possibility to extend it down to cognitive neuroscience or up to metaphoric reasoning –, each with their own relevant properties and functional traits.

6. ACKNOWLEDGEMENT

The never-ending support from Prof Y. Surya, the lively inspiration from Y. Haryadi, the strong scientific skepticism of all colleagues in BFI, and the abundant access to relevant publications provided by D. Rustandi, are the primal reason I have to express my greatest thankfulness.

7. REFERENCES

1. Bach, K. (1981). 'An Analysis of Self-Deception', *Philosophy and Phenomenological Research, 43(3)*.

2. Bach, K. (1984). ‘Default Reasoning: Jumping to Conclusions and Knowing When to Think Twice’, *Pacific Philosophical Quarterly, 65*.
3. Baron-Cohen, S., A. Leslie and U. Frith. (1985). ‘Does The Autistic Child Have A “Theory Of Mind”?’, *Cognition, 21*.
4. Baron-Cohen, S. (1998). 'Does the Study of Autism Justify Minimalist Innate Modularity?', *Learning and Individual Differences, 10(3)*.
5. Bicchieri, C. (1987). ‘Rationality and Predictability in Economics’, *The British Journal for the Philosophy of Science, 38*.
6. Chomsky, N. (1966). 'The Current Scene in Linguistics: Present Directions' in *College English*, *27(8)*.
7. Dennett, D. (1987). *The Intentional Stance*. Cambridge, Mass.: MIT Press/A Bradford Book.
8. Dennett, D. (1994). ‘Cognitive Science as Reverse Engineering: Several Meanings of "Top-Down" and "Bottom-Up"’ in D. Prawitz, B. Skyrms and D. Westerstahl (eds.), *Proceedings Of The 9th International Congress Of Logic, Methodology And Philosophy Of Science*. Amsterdam: North-Holland.
9. Dennett, D. (1995). *Darwin’s Dangerous Idea: Evolution and the Meanings of Life*. London: Penguin.
10. Edelman, G., and G. Tononi (2000). *A Universe of Consciousness*. London: Penguin.
11. Fodor, J. (1975). *The Language of Thought.* Hassox, Sussex: Harvester.
12. Fodor, J., and Z. Pylyshyn. (1988). ‘Connectionism and cognitive architecture: A critical analysis’, *Cognition, 28*.
13. Fodor, J. (2000). *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology.* Cambridge, MA: MIT Press.
14. James, W. (1890). *Principles of Psychology*. [http://psychclassics.yorku.ca/James/Principles/index.htm]
15. Newell, A. (1990). *Unified theories of cognition*. Cambridge MA: Harvard University Press.
16. Okasha, S. (2003). ‘Fodor on Cognition, Modularity and Adaptationism’, *Philosophy of Science, 70.*
17. Pinker, S. (1984). ‘Visual cognition: An introduction’, *Cognition, 18*.
18. Pinker, S. (1997). ‘Language Acquisition’ in L. R. Gleitman, M. Liberman, and D. N. Osherson (eds.), *An Invitation to Cognitive Science, 2nd ed., vol. 1: Language.* Cambridge, MA: MIT Press.
19. Pylyshyn, Z. (1999). 'Is vision continuous with cognition? The case for cognitive impenetrability of visual perception', *Behavioral and Brain Sciences, 22(3).*
20. Searle, J.R., *The Rediscovery of the Mind.* Cambridge, MA: MIT Press.
21. Schwitzgebel, E. (1997). ‘Words About Young Minds: The Concepts of Theory, Representation, and Belief in Philosophy and Developmental Psychology’. *Ph.D. thesis, University of California at Berkeley*.
22. Tarigan, H. (1984). *Psikolinguistik*. Bandung, West Java: Angkasa.
23. Tooby, J., and L. Cosmides. (1989). 'Evolutionary Psychology and the Generation of Culture, Part I Theoretical Considerations', *Ethology and Sociobiology 10.*
24. van Linder, B., W. van der Hoek, and J.-J. Ch. Meyer, ‘Seeing is Believing (and so are hearing and jumping)’, *Journal of Logic, Language and Information*, *6*.
25. Verschure, P., and P. Althaus. (2003). 'A real-world rational agent: unifying old and new AI', *Cognitive Science 27*.