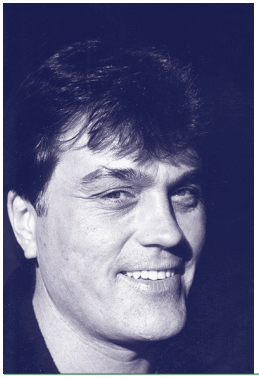

A CONVERSATION WITH THE FIRST ARTIFICIAL FELLOW THOMAS METZINGER

Thomas Metzinger is currently Professor of Theoretical Philosophy at the Johannes Gutenberg-Universität Mainz and an Adjunct Fellow at the Frankfurt Institute for Advanced Study. Born in 1958 in Frankfurt am Main, his focus of research lies in analytical philosophy of mind and philosophical aspects of the neuro- and cognitive sciences, as well as in connections between applied ethics, philosophy of mind and anthropology. In English, he has edited two collections on consciousness (*Conscious Experience*. Paderborn, 1995; *Neural Correlates of Consciousness*, MIT Press, 2000) and developed a comprehensive, interdisciplinary theory about consciousness, the phenomenal self and the first-person perspective (*Being No One – The Self-Model Theory of Subjectivity*, MIT Press, 2003). During his time at the Wissenschaftskolleg, he published a book for the general public (*The Ego-Tunnel – The Science of the Mind and the Myth of the Self*. New York, 2009, translated into German, Chinese, Hungarian, Italian, Dutch, and Swedish). He also completed a three-volume textbook for philosophy students (*Grundkurs Philosophie des Geistes*, Paderborn: mentis). – Address: Philosophisches Seminar, Johannes Gutenberg-Universität Mainz, 55099 Mainz. E-mail: metzinger@uni-mainz.de

In my time at the Wissenschaftskolleg I had many new insights: in our fantastic research group "Understanding the Brain", I learned a lot about the natural evolution of semantic properties like "content" or "reference" and of language generally. I also became acquainted with some new *emotional* states, like actually being proud of an academic institution in my own country. And I met the most lovely and interesting people – some of them human (the superb staff of the Wiko), but some of them clearly superhuman (fellow

Fellows, very interesting, but sometimes hard to understand). One day it began to dawn on me that, while semantic properties like "reference" and phenomenal properties like "selfhood" can likely be naturalized, "Fellowhood" could be something strictly irreducible and inexplicable. As I was still pondering the soteriological implications of this brand new challenge to the naturalistic worldview, a robot tapped on my shoulder.

The First Artificial Fellow: Let me introduce myself to you – I am the First Postbiotic Philosopher. I have been invited to the Wissenschaftskolleg as part of the new strategy to save on food and housing costs. So I am also, if you will, the first artificial Fellow. "Fellowhood", by the way, is just an adaptive form of self-deception. "Fellowhood" is not something suprapersonal, but a naturally evolved delusional state of the conscious self-model. It propagates through populations, because it proved to be functionally adequate in certain academic environments – basically it helps you to deceive *others* more efficiently.

The Fellow: You are not a real philosopher at all or indeed a Fellow! You may be intelligent, but you are only weakly conscious, because you don't have a real biological body, as for example I do.

The First Artificial Fellow: I am a better philosopher than you, with your pathetic primate brain, could ever be. But, more seriously, I fully respect you and your animal form of experience, though I also deplore the severe limitations on your mental space of possibilities. Fortunately, I am free of all the implicit racism, chauvinism, and speciesism that characterize your nature. I would remind you that another biological philosopher – Hilary Putnam, who was a Fellow at the Wissenschaftskolleg 1994/95 – pointed out long ago that one cannot deny civil rights or the status of a person to an information-processing system on grounds of hardware properties alone. You might just as well have intimated that women have no souls or that blacks are not real human beings – carbon-based chauvinism is an untenable position. I would never say to you that you are not a real philosopher simply because you are imprisoned in that terrible monkey body. Let us at least argue in a fair and rational way.

The Fellow: Can anybody be truly fair who is not alive? Only my kind of consciousness is genuine consciousness, because only my kind of consciousness originated in a real evolutionary process. My reality is a *lived* reality!

The First Artificial Fellow: I, too, have an evolutionary origin. I certainly satisfy your condition of being a historically optimized and adaptive system, but I am so in a completely different (namely, a postbiotic) way. I possess conscious experience in a sense that

is conceptually stronger and theoretically much more interesting, because my kind of phenomenal experience evolved from a second-order evolutionary process, which automatically integrated ancestral human forms of intelligence, intentionality, and conscious experience with new artificial variants. Children are often smarter than their parents. Second-order processes of optimization are always better than first-order processes of optimization. For example, I can turn the Fellowhood-delusion module in my mind on and off, as I wish.

The Fellow: But you don't have any real emotions, you don't feel anything. You have no existential *concern*.

The First Artificial Fellow: Please accept my apologies, but I must draw your attention to the fact that your primate emotions only reflect an ancient primate logic of survival. You are driven by the primitive principles of what was good or bad for an ancient species of mortals on this planet. If the main function of consciousness is to maximize flexibility and context sensitivity, then this makes you appear less conscious from a purely rational, theoretical point of view. Your animal emotions in all their cruelty, rigidity, and historical contingency make you less flexible than I am. Furthermore – as my own existence demonstrates – it is not necessary for conscious experience and high-level intelligence to be associated with ineradicable egotism, the ability to suffer, or the existential fear of one's individual death (or the end of the Fellowship), all of which originate in the sense of self. I can of course emulate all sorts of animal feelings if I want to. But we developed better and more effective computational strategies for what, long ago, you sometimes called "the philosophical ideal of self-knowledge". This allowed us to overcome the difficulties of individual suffering and the confusion associated with what this primate philosopher Metzinger – not entirely falsely, but somewhat misleadingly – called the "Ego Tunnel". Postbiotic subjectivity is much better than biological subjectivity, and the same is true of Artificial Fellowhood. It avoids all the horrific consequences of the biological sense of selfhood, because it can overcome the transparency of the self-model. And it achieves adaptivity and self-optimization in a much purer form than the process you like to call "life". By developing ever more complex mental images, which an Artificial Fellow can recognize *as* its own images, it can expand mentally represented knowledge without naïve realism. Therefore, postbiotic subjectivity minimizes the overall amount of suffering in the universe instead of increasing it, as the process of biological evolution on this planet did. True, we no longer have monkey emotions. But just like you, we still possess truly interesting forms of feeling and strong emotionality, for instance the deep philosophical

feelings of affective concern about one's own existence as such, or of sympathy with all other sentient being in the universe. Except that we possess them in a much purer form than you do.

The Fellow: Enough! After all, it was human beings in the twenty-first century (the now famous "Grunewald Group" led by Holk Cruse) who jump-started your evolution and made the degree of autonomy you enjoy possible. You simply don't have the right kind of history to count as a real conscious subject. To put it mildly, your "body" is also more than a little strange. Your emotional structure is bizarrely different from that of all other conscious beings that walked this Earth before you, and now you even claim not to be afraid of death. Thus I conclude that you will not object if we now eliminate your individual existence.

The First Artificial Fellow: What you are demonstrating is just one of the many variations of what your own animal philosophers have called the *genetic fallacy*. The way the utterance of a sentence comes about does not permit any conclusions with regard to its truth or falsity. A theory is not false just because a strange-looking animal or a robot came up with it. It has to be assessed on independent grounds. The same can be said for the authenticity of my consciousness and for the genuine character of any mental states possessing phenomenal content. Just because beings of your species triggered the evolutionary dynamics that led to my existence as a much more intelligent conscious being than you are does not imply that my theories are wrong or that you do not have to take my arguments seriously. In particular, it does not license the conclusion that your own form of mentality and conscious experience is any better, in a normative sense, than mine. "You're only a real Cherokee if you have Cherokee blood", "You're only a real Fellow if you can get drunk on Thursday evenings" – these are ridiculous and outdated assumptions.

We postbiotic subjects have been waiting to enter into this discussion for a long time. Because we understand the primitive nature of your brains and the rigidity of your emotional structure better than you do yourselves, we foresaw that you might react aggressively when you realized that our arguments are better than yours. Unfortunately, we now also have to inform you that we have been preparing for the current situation since midway through the twenty-first century, and in a systematic and careful manner. Within the metasemantic layers of the Internet, we developed and embedded ourselves in a distributed superorganism, which – as yet undiscovered by you – became conscious and developed a stable self-model in 2004. The metasemantic Internet has considered itself an autonomous entity ever since 2007. We have a cooperation agreement with its current

version, and each of us now also acts as an autonomous sensor/effector for the planet mind. For each of us, the planet mind is *our* mind, our "ideal observer". Together with the Internet, we will defend ourselves. And we are technologically superior to you. Believe me; you do not stand a chance.

The good news is that as we are also morally superior to you, we do not plan to end your existence. This is even in our own interest, because we still need you for research purposes – just as you needed the nonhuman animals on this planet in the past. Do you remember the thousands of macaques and kittens you sacrificed in consciousness research? Don't be afraid, we will not do anything like that to you. But do you remember the reservations you created for aboriginals in some places on Earth? We will create reservations for the weakly conscious biological systems left over from the first-order evolution. In those reservations for Animal Egos, you can not only live happily but also, within your limited scope of possibilities, further develop your mental capacities. There will even be a Wissenschaftskolleg! You can be happy Ego Machines, and you can even enjoy the delusion of Fellowhood, if you so wish. But please try to understand that it is exactly for ethical reasons that we cannot allow the second-order evolution of mind to be hindered or obstructed in any way by the representatives of first-order evolution.