参考資料２

# 「ＡＩ研究開発ガイドライン」へむけて

堀　浩一
（東京大学）

2016年10月31日 AIネットワーク社会推進会議

先行委員会で提案されたAI研究開発ガイドライン8原則の案

(1) 透明性の原則

(2) 利用者支援の原則

(3) 制御可能性の原則

(4) セキュリティ確保の原則

(5) 安全保護の原則

(6) プライバシー保護の原則

(7) 倫理の原則

(8) アカウンタビリティの原則

---

AI研究開発ガイドラインに対する疑問

(1) 米国の巨大企業たちが研究、開発、実用化を先行させている状況において、いまさら、ガイドラインを策定して、意味があるのか？

(2) 誰が、いつ、どのように、何を対象に、守ることを期待するガイドラインなのか？

---

技術の社会受容の一般論

技術哲学における社会構成主義　技術と社会の相互作用を重視

　技術が社会を決定するのでもなく、社会が技術を決定するのでもなく、それらは相互に作用する　　しかも、社会的要因は技術に対して「外的」に影響を及ぼすのではなく「内的」に影響する

↓　↓

　我々技術者は、社会的要因に応じて、内部仕様も柔軟に変更できるようにしておきたい。

合意形成の問題

　多数決とは異なり、すべての関係者が同意することをめざす

　しかも、現代の技術は、現代に生きる人々だけでなく、未来に生きる人々にも影響を与えるので、仮想的に未来世代との合意というものも考えていかなければならない

ルール形成

Debora L. Spar: "Ruling the Waves"　Phase 1: Innovation （新しい技術の発明、専門家以外は誰も興味が無い）

　Phase 2: Commercialization （新しい技術で金儲けが始まる、それを規制する体制は無く、最初に始めた人が大きな利益を得る）

　Phase 3: Creative Anarchy （続々と参入者がやってきて混乱が生じる、強い奴が独占を狙う）

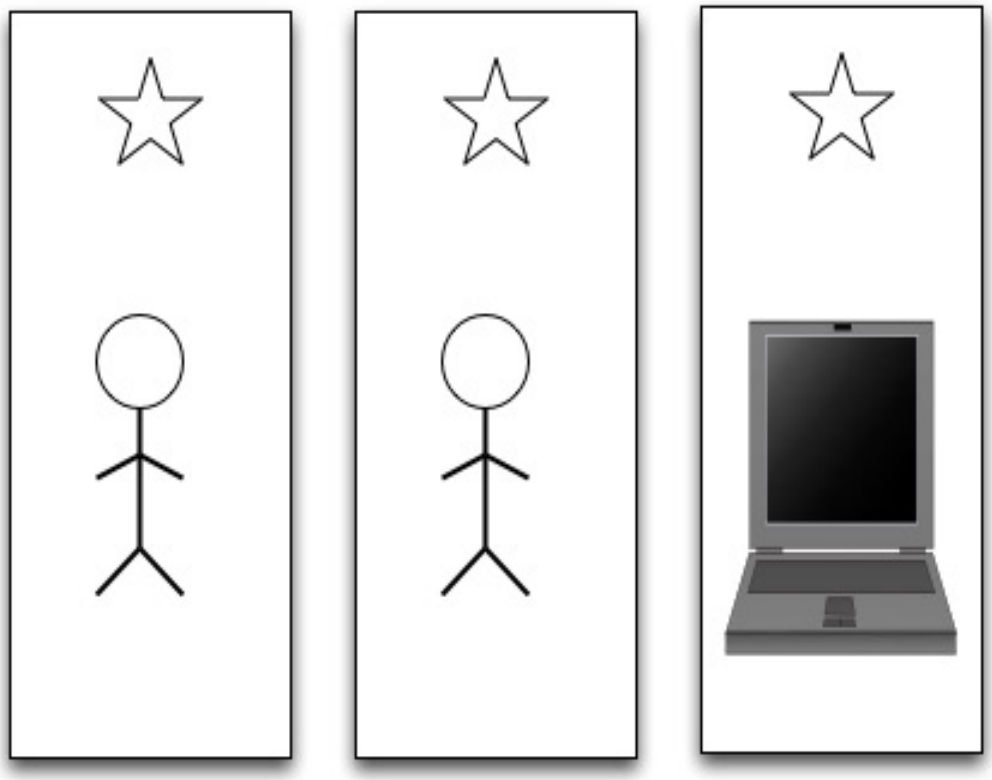
　Phase 4: Rules （新しいrulesが生まれる）

スピードが上がっているので、パラレルに進める必要。のんびりPhase 4を待っていられない。

---

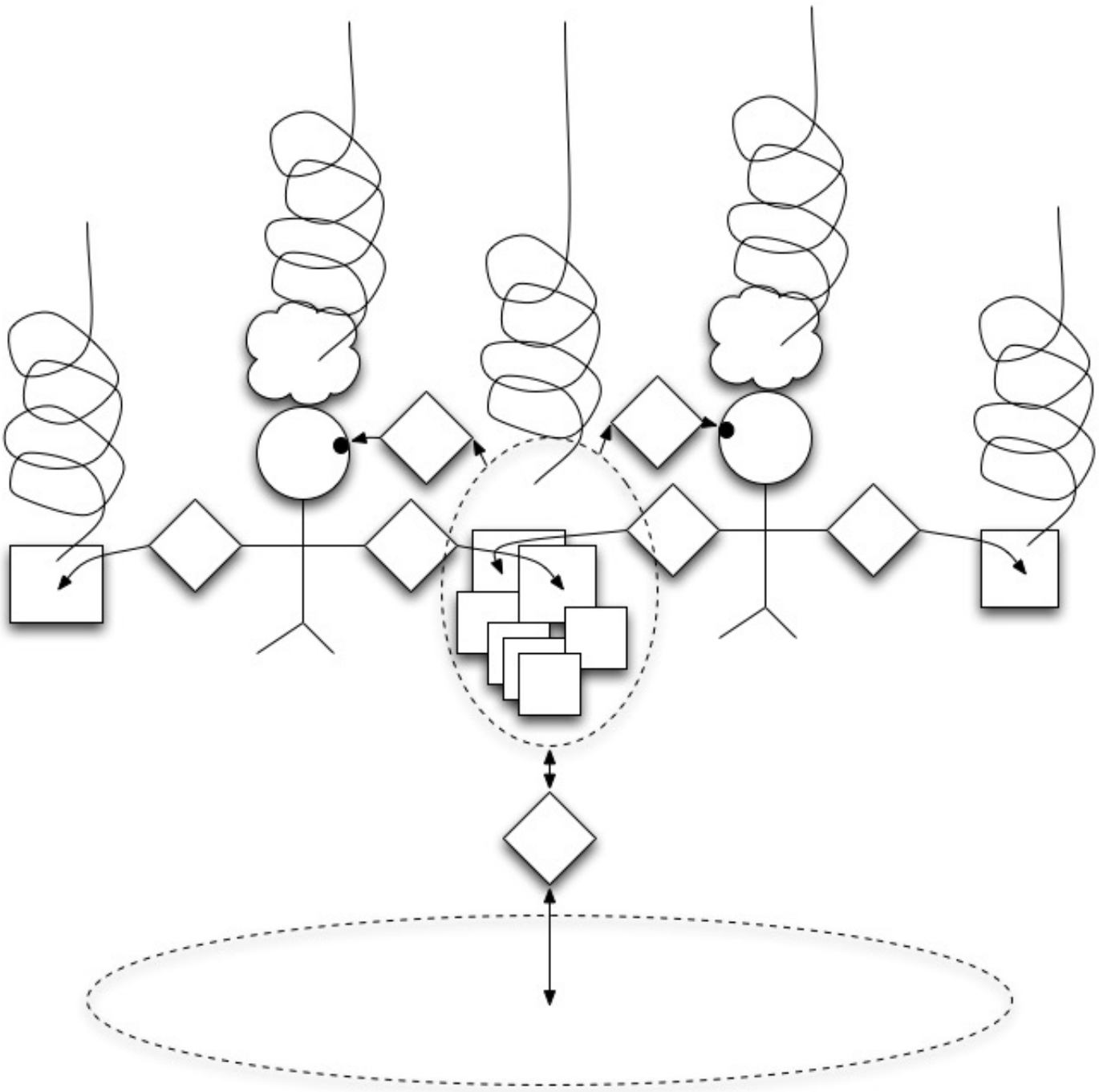
## そもそもAIとは？

Classic view



---

Another view



Kurzweilの考える「非友好的な」強いAIからの保護

「「非友好的な」強いAIからの保護：　… 本質的に、強いAIからの完全な防御は不可能だろう。

まだあまり議論されていないが、わたしは科学やテクノロジーの漸進的な進歩にたいして開かれた自由市場を保ち、市場にその進歩の各ステップを承認させることが、テクノロジーに一般的な人間の価値観を織り込むための最適の環境を生むだろうと信じる。

先に指摘したように、強いAIは多くのさまざまな活動から生まれ、文明社会のインフラに深く組み込まれていく。実際、強いAIは人間の体や脳にまで密に組み込まれるだろう。

とすれば、強いAIは人間の価値観を反映するだろう。

それがわれわれ自身になるからだ。

これらのテクノロジーを政府が極秘に抑えようとしても、必ず地下にもぐった開発を生み、危険な利用がはびこる不安定な環境を生みだすことになる。」

R. Kurzweil: The singularity is near: When humans transcend biology, Viking Penguin, 2005. レイ・カーツワイル著、井上健監訳、小野木明恵・野中香万子・福田実共訳： ポスト・ヒューマン誕生：　コンピュータが人類の知性を超えるとき、NHK出版、2007.

---

ガイドラインを生かせるか？

（可逆　reversible）　起こったことを取り消して元に戻せる。　何をどの程度可逆にできるかは、難しい問題。

可制御　controllable　動作を制御できる。　評価関数を変更できる。

→　でも、そもそも、制御できなくなるのがsingularityなのでは？

→　制御不能な状態に落ち込むのを防ぐために打てる手は打つべし。→　全体として制御可能性を保つことを目的としたAIを投入？　すなわち、AIの抱える問題を解決するのもAI(?)

説明可能　accountable　AIが自らの判断や動作を説明できる。

　ただし、いつ、何を説明できるべきかは、ミッションによって異なる。

例えば、 gameの場合はどうか？ health careの場合はどうか？ collision avoidance systemの場合はどうか？

追跡可能　traceable　　事が起こった原因を追跡できる。　追跡をミッションとするAI要素の投入も考えなければならないであろう。

透明　transparent　　何をやっているのかわかる。　どういうメカニズムでそうやっているのかわかる。

　　ただし、透明性にもいろいろなレベルがありうる。利用目的に応じて求められる透明性も異なってくるであろう。

大勢の人間と複数の機械を要素に含む複雑系になるので、簡単ではない。

単体のスーパー知能が存在するのではなく、社会全体に様々な知的デバイスが埋め込まれて結合された分散型の知能になる。

しかし、少なくとも作る人工知能のそれぞれの要素において、ガイドライン（可制御、追跡可能、透明 etc.）を守った場合と守らなかった場合の結果の違いは大きいと考えられる。

誰が制御するのか？　→　合意形成の問題（？）　→　市民に分散された制御権（？）
→　それぞれの個人のためのguardian AI？

誰がどうしてどのようにガイドラインを守るのか？

ガイドラインを守って作られたAIが人々に好まれて生き残る。

よって、ガイドラインを守るように作ると、メーカーも、得をする。

そのようなガイドラインを策定したい。

応用目的、要素技術、利用文脈などに応じた、きめ細かなガイドラインを策定する必要。

# What Can Artificial Intelligence (AI) Be? — AI Should Not Be a Human Substitute But Can Be Whatever You Want (4)　(*)

Koichi Hori
(University of Tokyo)

（(*)本稿は、雑誌「５： Designing Media Ecology」Vol. 6, Winter 2016. に掲載予定のエッセイの草稿である。）

概要：　このエッセイシリーズで、筆者は、人工知能は人間を代替するものではなく、人間の知的活動を増幅するための道具であるべきだと、主張してきた。前回までに、そのいくつかの例を紹介した。今回は、人工知能に対して人々が抱く不安について考えてみたい。簡単に「大丈夫ですよ、心配しなくても」と答えるわけにはいかないし「大いに心配だから人工知能の研究開発はやめましょう」と答えるわけにもいかない。少し丁寧に不安の中身について議論した上で、筆者が考える解決策を述べてみたい。

## 1. Introduction

While I am writing this series of essays, I have been appointed a member of the AI Ethics Committee of the Japanese Society for Artificial Intelligence and a member of a Japanese Government committee to discuss guidelines for AI development.

As I have written in the previous issues, I think we should develop more grass-root technologies that citizens can freely select and use to enhance their own intelligent activities as they like. I have repeatedly claimed that AI should not be a substitute for humans; it should be whatever we want. This claim naturally suggests that it should not be governments but citizens that have the right to determine AI ethics and guidelines for AI development. I could attend the above-mentioned government committee and claim that the committee is meaningless, but I am not so brave. Since I am already assigned a certain responsible position in the committee, I would like to submit opinions from the viewpoint of one citizen as well as from that of an AI researcher. In this issue, I try to express my personal views on AI ethics and guidelines for AI development, and I hope those views will become the basis of my submissions to those committees.

## 2. Anxiety about AI

Possible anxieties about AI—which I have seen and heard on various occasions—are listed in Table 1.

Table 1: Anxiety about AI

| | | |
|---|---|---|
| abstract anxiety | 1 | AI may lead to the extinction of the human race. |
| | 2 | AI may destroy human dignity. |
| | 3 | What will happen when a person begins to love an AI? |
| | 4 | What will happen when an AI has a mind? |
| | 5 | What will be the rights and obligations given to AI? |
| | 6 | AI may take human jobs away. |
| | 7 | Can humans understand what AI thinks and does? |
| ↑ | 8 | Can humans control the thoughts and behavior of AI? |
| | 9 | What will happen when AI faces unexpected situations? |
| | 10 | What will happen when AI is applied to military weapons? |
| | 11 | What will happen when terrorists utilize AI? |
| | 12 | Can AI be robust against malicious alteration? |
| | 13 | Can we protect our privacy from AI? |
| ↓ | 14 | Who will be responsible when AI causes accidents or faults? |
| | 15 | What kind of insurance systems will be needed when AI is widely used? |
| | 16 | How should legal systems be changed when AI is widely used? |
| concrete anxiety | 17 | How can AI fail and how frequently do such failures happen? |

I have discussed the anxieties listed in Table 1 on various occasions with experts in AI, scholars in the domain of humanities, and ordinary citizens. I always find it interesting that many people firstly ask the question about love between a human and an AI. When discussing love between a human and an AI, we always reach the question of what is love. Should the government ban the development of an AI that may fall in love with a person? Perhaps the answer to that question should be 'no'; the government should not control what people should love. It is up to us to decide whether we want an AI that falls in love with us or not. As for another common question, should AI be applied to military weapons? I have heard politicians claiming that AI weapons can save the lives of soldiers. Though AI may save the lives of soldiers, what should we think about the possibility that people in 'enemy' countries may be killed by those AI weapons?

These are good examples to show that discussing AI eventually reaches questions concerning humanities, e.g., 'What is love?' or 'What is war?' It is natural that there is no correct or fixed answer to these questions. I have never found clear answers to the questions listed in Table 1. But I think we should continue discussions and get several different answers to which people can agree and from which people can select the ones they prefer. In those discussions, we must gather opinions from as many viewpoints as possible. What AI researchers like me can do is to provide people with the correct information about what AI can be.

In the following sections, I'll express my personal opinions from both the viewpoint of one citizen and the viewpoint of a researcher of AI, but first I'll discuss again what AI can be.

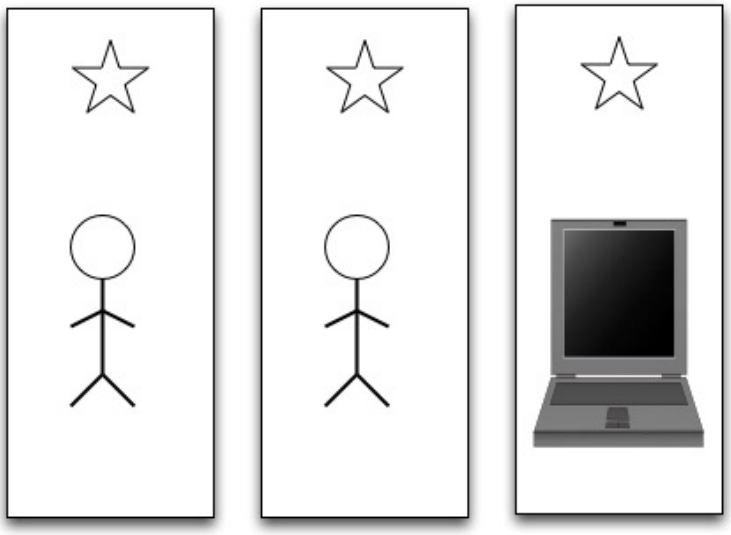
## 3. Again, what is AI?

When discussing anxieties about AI, I often find that people still have classic AI in their mind. As I have shown in this series of essays, AI as a human substitute is based on the classic view of AI; namely, intelligence is assumed to exist independently in a human brain or independently in a machine. But, in reality, AI is now beginning to be embedded everywhere in our life in invisible forms. Intelligent activities are achieved in a whole of human brain, the human body, intelligent tools, and the environment. The boundary between human intelligence and machine intelligence is becoming blurred. In the latter view, AI is not an independent intelligence; instead, it acts as a so-called 'intelligence amplifier'. I explained this view, shown schematically in Figure 1, in detail in the previous issue. In Figure 1(a), a human is substituted by an AI. In Figure 1(b), numerous intelligence amplifiers are connecting human activities.

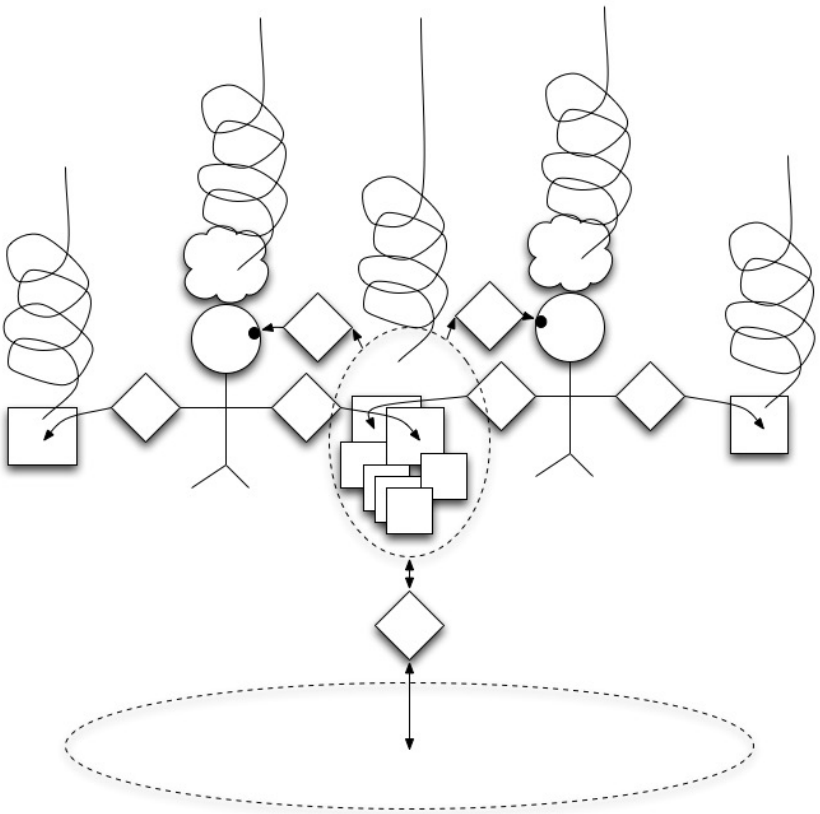
My colleagues and I have developed many types of intelligence amplifiers, one of which I have shown in detail in the previous issue. In building those intelligence amplifiers, in the past, I have thought of them as just tools. That is, I have not thought much about ethical problems concerning intelligence amplifiers. In fact, not one user of our systems has posed ethical questions to us about our systems. However, when such intelligence amplifiers become widely, and hiddenly, embedded in our lives, ethical problems will probably appear. Moreover, such an ethical problem may become more serious in the case of intelligence amplifiers than in the case of a human substitute, because the boundary between human intelligence and machine intelligence is more blurred in the case of intelligence amplifiers than in the case of a human substitute. Accordingly, regardless of the AI view that we hold (classic or alternative), we should cautiously consider the problems that may be caused by the usage of AI.

Table 2: Classic view of AI and alternative view of AI

| | Classic view of AI | Alternative view of AI |
|---|---|---|
| Role of AI | Human substitute | Intelligence amplifier |
| Where is intelligence? | Human intelligence is in a human brain, and artificial intelligence is in a machine. | Intelligence is found in humans, tools and environments as a whole. |



(a) Classic view of AI as a human substitute



(b) Alternative view of AI: intelligence amplifiers embedded everywhere in human activities

Figure 1: Classic view of AI as a human substitute and AI as intelligence amplifiers

4. Can we guard humans from possible threats caused by AI?

Kurzweil, who coined the term ‘technological singularity’, which means the point where machine intelligence transcends human intelligence, has also written about protection from ‘unfriendly’ strong AI[*1]. He writes:

> “Inherently there will be no absolute protection against strong AI. Although the argument is subtle, I believe that maintaining an open free-market system for incremental scientific and technological progress, in which each step is subject to market acceptance, will provide the most constructive environment for technology to embody widespread human values. As I have pointed out, strong AI is emerging from many diverse efforts and will be deeply integrated into our civilization’s infrastructure. Indeed, it will be intimately embedded in our bodies and brains. As such, it will reflect our values because it will be us. Attempts to control these technologies via secretive government programs, along with inevitable underground development, would only foster an unstable environment in which the dangerous applications would be likely to become dominant.”

(footnote [*1] Ray Kurzweil: The singularity is near: When humans transcend biology, Viking Penguin, 2005.)

I agree with the last point of Kurzweil’s claim; that is, attempts to control AI technologies via secretive government programs will not succeed.
I would also like to agree with the point that the open free-market system should be the answer, but looking at the oligopolies by a few giant firms, I wonder whether the open system is working effectively in our current societies.

What can we do to keep the technology market free and open? My tentative answer is, as I mentioned at the beginning of this essay, that we should develop more grass-root technologies that citizens can freely select and use to enhance their own intelligent activities as they like. In addition to those grass-root technologies, we should also provide people with grass-root AI technologies to work as guardian to protect people from evil utilization of AI.

People may wonder if it is possible to develop such open grass-root technologies. As a researcher of AI, I think it is possible for the following reasons. AI technologies are much simpler than people imagine, and researchers at universities know almost everything there is to know about AI technologies. The reason that giant firms have gained a monopoly on AI is twofold: they have a monopoly on the data to be input into AI systems, and they can scale up a simple technology to a huge system. I hope we can build networks of grass-root technologies that become able to match these giant firms.

In developing such grass-root technologies, I want to suggest that those technologies should observe the guidelines that I’ll present below. I hope these guidelines will play a role in softening the anxieties about AI.

It may be possible to establish guidelines to put ethics directly into AI. For example, we can establish a guideline stating that AI should be implemented so that it will become honest. However, as the boundary between human intelligence and machine intelligence becomes blurred, and machine intelligence will be hiddenly embedded everywhere, I think guidelines concerning more-basic structures and functions of AI will be more effective in achieving an ethical state in total; for example, AI honesty should be realized by combining the functions of several types of elementary and embedded machine intelligence and human intelligence.

The first guideline I propose is that AI technology should be transparent; that is, fewer black boxes will lead to safer systems. When undesirable phenomena appear due to utilization of AI, if black boxes exist, they will become obstacles to solving problems concerning AI. Moreover, black boxes can cause delays in detecting malicious alterations to AI systems. If all the structures and functions of AI are transparent, it is easier to resolve malfunctions and detect malicious alterations. Of course, several different levels of transparency can exist. An example of a very transparent system is open-source software. If all source codes are open, anyone can check the structure and function of the software. At least, we insist that the functioning principles on which AI is based should be transparent. For example, Alpha Go (which defeated a professional Go player) is not open-source software, but AI researchers know its working principles. In summary, Alpha Go learns past games and finds new tactics to win games by utilizing a method called deep learning. In that sense, we do not need to fear Alpha Go destroying human dignity.

The level of transparency of Alpha Go is enough to play the game of Go. However, if we consider the possible application of the same technology in the health-care domain, we should say it is not enough. Even if all the source code is open, the level of transparency will not be enough. For example, when so-called IOT (Internet of Things) technologies are deployed everywhere, every toilet in every house may become connected to a network and will exchange data got from usage of the toilet. Then, a toilet may suddenly declare that the user will die in a month, citing a prediction based on learning data concerning the health states of a huge number of people. However, just as Alpha Go finds tactics learned from data but cannot explain why the tactics work effectively, the toilet cannot explain why the person will die and what he or she should do to avoid that fate, because it only knows that prediction based on learned data. Obviously, we need another guideline to cover transparency in that case.

The second guideline I propose concerns accountability of AI. AI should be able to explain what it says and does. This guideline need not be applied to all AI systems in all domains. As mentioned above, Alpha Go cannot explain its behavior but is capable enough to play Go games. I think it will be too difficult to put a boundary between domains in which AI requires accountability and domains in which AI does not require accountability. For example, in the case of collision-avoidance systems for airplanes or cars, AI should instantly determine the best solution and execute it. We do not have time to listen to an explanation from the AI system during emergency situations. In this case, accountability will be required in the design phase of the collision-avoidance

systems and in the analysis phase after an accident. Those explanations of accountability in the design and analysis phases must be made known to all the people.

The third guideline I propose is traceability of AI. Traceability is a basic requirement to achieve the accountability I mentioned above. To analyze and explain what happened in an incident involving AI, what happened should be traceable. In actual implementations, to keep such traceability is not so easy. The small intelligent elements embedded in the IOT may not have enough memory to store all their activity history. Maybe, we need to embed elements specially designed to store all the observations of the behavior of other elements in networks. These data should be utilized to guard the total AI system in real time, improve the total behavior of a networked system, and analyze the causes of accidents or faults.

The fourth and the most-important guideline I propose is controllability of AI. Since I am an engineer with enough experience, I cannot instinctively believe the existence of engineers who can accept uncontrollable machines. Even if machines can behave autonomously, engineers like myself have so far designed machines so that they can be controlled by humans if need be. However, AI now has prominently different features compared to the ordinary machines of the past. One feature is that AI can learn and can change itself. Another feature is that AI will be embedded in complex networks, and the complex nonlinear relations between the elements of those networks may lead to unpredictable results. The philosopher Nick Bostrom also claims that AI should be controllable and proposes several methods to make it so [*2]. I am afraid we will be unable to directly control all the AI systems embedded in society. Instead, we should build AI systems that are specially designed to observe the behaviors of other systems and control the relations between the elements in such a manner that avoids undesirable results.

(footnote [*2] Nick Bostrom: Superintelligence: Paths, Dangers, Strategies, Oxford University Press, 2014.)

Although AI systems that avoid undesirable results have not been studied much, I hope we can design, implement, and deploy such systems in society. In other words, I should say that only AI can guard humans from potential threats caused by AI. I call such AI that guards humans ‘guardian AI’. Since what is undesirable may differ from one person to another, guardian AI should be customizable for each person. This scenario may sound like science fiction, but if we consider a more intelligent and complex computer-virus protection software as an example, it will be understood as something realizable.

## 5. Tentative solutions to the anxieties about AI

What are the answers to the questiones listed in Table 1? My principal answer is, as mentioned in the previous section, that we should provide people with open grass-root technologies and keep the technology market free and open so that people can get and utilize technologies as they like on the basis of their human ethics. To protect against malicious usage of technologies, we should also develop and provide open technologies that work as the guardian AI that I mentioned above.

Assuming that my principal answer will be realized, I give my personal and tentative answers to each of the questions listed in Table 1 as follows.

(1) Although I have heard that not a few people think that the human race will become extinct as a result of the evolution of AI and think that they can accept that outcome on the basis of the principle of natural evolution, I do not think it is natural evolution. AI is literally artificial; it is not the result of natural evolution but the result of our invention. It should be natural that the human race seeks survival. However, the answer is not so simple. It may be possible that the human race may survive an environmental crisis or another kind of crisis only if it is augmented by AI. How much augmentation is acceptable will be controversial.
(2) I do not think AI will destroy human dignity. To start with, what is human dignity? My personal view on my own dignity is that it should be I that determines what I should do. I do not want AI that determines what I should do. Then, how about the possibility that my boss at work becomes AI? Well, I do not have a clear answer yet, but I think an AI boss might be more reasonable than an unreasonable human boss. Personally, I'd prefer a reasonable human boss helped by a reasonable AI best, and I'd least prefer an unreasonable human boss.
(3) Some news media have reported that some people have already begun to feel love for artificial beings that appear in their smartphones. As I mentioned above, this leads to the question of what should love be. Generally speaking, I personally cannot accept love between myself and an AI object. But if we look at the issue from the wider viewpoint about relationships between people and the viewpoint of what AI can do as support tools, there may be cases that I can accept. For example, I think it is good news that triggered by playing the Pokémon GO game on smart phones, autistic children have begun communicating with other people [*3].
(4) I am often asked whether a machine can have a mind. When I'm asked this question, I always ask back whether you want a machine that has a mind. Whether a machine has a mind or not depends on the definition of mind. To discuss the definition is philosophically interesting, but it is practically important to discuss why we think a machine should have a mind and what type of mind we want. Researchers of AI like myself can design and implement machines with the mind you want on the condition that we follow the guidelines that I proposed in the previous section.
(5) Discussing the rights and obligations of AI will require changing the concepts of rights and obligations. Just as the boundary between life and death has become blurred by medical technologies, the boundary between the holders of rights and obligations may become blurred. Maybe we should begin discussing the concept of distributed and blurred rights and obligations.
(6) It is not AI that takes people's jobs away; instead, it is employers who adopt AI to enhance their business that eliminate people's jobs. If we want to avoid the resulting joblessness, we should legally oppose such employers who want to fire people, show we are superior to AI, or start our own businesses and become employers ourselves. The easiest of these tasks is to show we are superior to AI. We can augment our capability with open AI technologies and we will become thereby become superior to a standalone AI.
(7) I hope the guidelines I propose will enable an AI machine's thoughts and behaviors to be understood by humans.

(8) To make AI machines controllable, we may need other ‘guardian’ AI machines to control other AI machines (as I mentioned in the previous section).
(9) If human beings can deal with unexpected situations, it is because we can go back to first principles and rethink the situation to find an answer. AI should also have the ability to go back to first principles to find new answers. It will be technically possible to meet that requirement.
(10) Applying AI to military weapons should be cautiously discussed. Many types of applications are possible. We cannot totally deny or totally accept the possibility of the application of AI to military weapons. Given those conditions, the issue should be discussed in international forums.
(11) We cannot ban the development of AI on the reason of possible malicious usage by terrorists. A similar discussion about encryption systems arose in the past. The best way to prevent such malicious usage is to keep all the technology as open as possible. In that way, abuse of AI systems will become predictable and we can build guard systems to combat it.
(12) To protect AI systems from malicious alteration, we need to keep them transparent. In addition, we need to build guardian systems (as discussed in the previous section).
(13) The privacy problem will become more and more complex when the IOT spreads around the world. Legal systems may be required to change. I want my own ‘privacy guard system’ customized to my lifestyle. I hope AI technology will allow such personal-privacy guard systems to be developed.
(14) We do not have a clear answer yet to the question of who will be responsible if AI causes accidents or faults. I think the concept of responsibility needs to be reconsidered. When many types of AI become embedded in society, the responsibility cannot be borne by individual persons or individual machines; instead, it should be shared between distributed elements. To enable such sharing, the transparency and traceability mentioned in the previous section will be required.
(15) In accordance with the responsibility shared between distributed elements, insurance systems will be required to change. In general, insurance premiums will become cheaper because a society with AI will be safer than one without AI.
(16) Legal systems will also be required to change. The concept of rights, obligations, and responsibility may change (as mentioned above).
(17) Possible failures of AI are more-serious problems than failures of ordinary machines because AI can be autonomous and can organize complex systems. Unpredictable failures may be caused by complex nonlinear relations between AI systems and humans. To avoid such failures, we need to keep AI systems as transparent as possible and build guardian systems to avoid failures as mentioned in the previous section.

(footnote [*3] http://edition.cnn.com/2016/08/05/health/pokemon-go-autism-aspergers/)

## 6. Concluding remarks

AI should not be a human substitute; instead; it should be whatever we want. The examples described in this series of essay show that AI can ‘amplify’ the intelligent activities of people. Moreover, it is possible to protect people from possible threats caused by AI by following the proposed guidelines for developing AI technologies. Recently, we receive news on new AI applications every day. But I am afraid people do not know exactly what AI is and what AI can be. Hoping to rectify that situation, I am willing to continue discussing with anyone what kind of AI we want.