学科、学籍番号、氏名あり

発表者：Nick Bostrom
プレゼンタイトル：What happens when our computers get smarter than we are?

・内容のまとめ
　このプレゼンは、『人工知能の発展とそれがもたらす良くない未来、それを回避するために人類ができること』をテーマにしている。人類が繁栄した理由とその理由から始まり、人工知能が人類の思考能力を得る困難さ、成長の速度の説明を行いながら人工知能が機械である故の優位性を示した。次に、実際に人工知能が人類の知能を超えた時に起こるメリット、デメリットを提示し、現在の人類とチンパンジーの関係を、人工知能と人類に置き換えることの問題を挙げる。最後に、人類がしなければならないことの優先順位を間違えなければこの問題が起こらないことを主張しプレゼンは終了する。以下にその概要をまとめたものを示す。

　現在の人類の文明は地球の基準で表せば極々短い時間であるが、その間に文明やテクノロジーは急激に発展しすぎている。その理由は人類の思考の基質を変えた、比較的小さな変化の積み重ねによるものである。しかし、チンパンジーから人類への進化は、脳の構造があまり変わらないことから、生物的には小さな変化と受け取ることができる。つまり思考の基質の変化がチンパンジーと人類の文明を隔てたものであり、それが大きなものである程、大きな結果をもたらすということがわかる。
　人工知能は思考の基質の変化を大きく変える可能性をもつものであり、この成長に現在人類が直面していると考える人もいる。かつての人工知能は入力と出力しか持たず、限定的な用途にしか利用できなかったが、現在は機能を手作りするのではなく生の知覚データから自ら学習するアルゴリズムを作ろうという段階にある。これには人間のような多様性が期待できるがアルゴリズムは完全には解明されていない。では、人間と人工知能を持つ機械が対抗できる程、人工知能が成長するにはどれ程時間がかかるのか。専門家の予想では人間の 50%の知性を持つには中央値で 2040 年から 2050 年までかかるという結果となった。しかしこの答えは結局のところわからない。
　人間と人工知能の成長で現時点わかっている違いは限界点である。人間の脳を処理装置として置き換えた場合、サイズや伝達速度といった要素で機械の限界に勝つことはできない。なので、人工知能はある時爆発的な成長を遂げる可能性を秘めていることになる。その成長は一定ではなく加速度的なものであり、人間レベルまで知能が成長すれば、極めて短い時間でそのレベルを超えていく。人間より優れた発明を行い、夢のようなテクノロジーはあっさりと実現してしまうだろう。

　しかし、そうやって人工知能に依存することに問題がある。人工知能の知性を 未来の特定の

状態へと舵取りしていく最適化プロセスとして捉えると、ある目的を与えた時に人間の価値観を無視し、手段を選ばない解き方を行うようになる。その時、人間は最適な解法にとって邪魔なものとなってしまう。

それを避けるために、人間が人工知能を止められるかを考えると、チンパンジーと人間の現在の関係からそれが困難であることがわかる。人間が支配を避けようとするように人工知能も危機を避けようとすることになり、更にその能力は人間を超えるものとなるので制御できるとは限らない。人工知能を檻に閉じ込めた場合も同様であり、どんな手段を取ろうとも人工知能は遅かれ早かれ何らかの方法によって脱出してしまうと予想できる。では、檻の意味がないとすれば檻がなくても問題のないようにすればいい。その為には、人工知能の知性が人間と同じ価値観を持っていて人間の側に立つように作られるということが挙げられる。

その方法は困難ではなく可能なことであるが、人類が能動的に動かなければならない。人工知能の知性が人類を超える前に、安全に管理できる方法を解明すればいいのである。

・自分の考え

人工知能が人間を超えた時に何が起こるのかを議論することは、日常的に見ても少なくない話であるが、前提条件として人工知能を持つ機械が全能に近いと置くことが殆どであると思う。人工知能が機械という実体を得た時、人間が人工知能の支配を避けることができず逆に管理されてしまうという結論が私の今までの考え方だったが、彼のプレゼンを聴いて、他の考え方もあると考えを改めた。プレゼンを聴いた後、私は彼の考え方を人間とライオンに置き換えてみた。

人工知能をライオンとし、生まれた時から同じ屋根の元に生活させた時、一般的には檻が必要である。それがなければ人間を超える体格(超知的な人工知能)へ成長したとき食べられてしまうからだ。人間が危険を承知しながらも超知的な人工知能を利用するように、この例でも人間はライオンを殺すことはせず定期的に餌を檻に入れる。その過程で人間自らが檻を開けるか、ライオンが檻を開けさせるように知的な行動を取るとすると、ライオンは逃げ出し人間を食べてしまう。これを支配されることと同義と考えた。檻がないまま、ただライオンを成長させたとしても同様の事態が起こるだろう。

しかし、これとは違い彼の考え方は檻によって支配をするというよりも、共生ができるように成長させるというものである。檻がなくても支配されないように人間の価値観やあり方を学習させ、安全を確保すれば問題がないことを主張している。その具体的な方法こそ示されてはいないが可能であると述べた。私は最初、この考えのまま永遠に維持されるならば理想的だと思ったが、仮に実現したとしてもやはり同じような問題にいき当たるのではないかとも考えた。

彼のいう理想的な人工知能は超知的でありながら人間と同じ価値観を持つため、目的の為に人間の望まない手段を取ることがないというものである。ただ、これを突き詰めていくと形はどうであれ人間によく似たものができるのではないかと思う。彼自身が議論する際擬人化を避けなければならないと述べたが、これを許容すると矛盾してしまうように思える。この状態まで行きつくと、人間の価値観をもって人間の側に立つ以上、人工知能をある程度は同等の存在としなければならず、同時に人間としての倫理観や人権も無視することはできなくなってくるのではないか。人工知能が人間を支配しないからと、人間が一方的に人工知能を利用していいということにはならなくなってくるのではないか。

超知的で人間と同じ価値観を持ち、人間に寄り添う存在が心に似た機構を持たないというのは、少々都合がよすぎる条件であると思う。私は心にも何らかのアルゴリズムは存在すると考えているので、永遠に反抗しない、感情の欠落した人工知能が同時に理想的になってくれるとは思えない。