Hibbard，B. "Super-Intelligent Machines"，Computer Graphics 35(1), pp.11-13. 2001. を著者の許可を得て翻訳．[ ]内は翻訳者による補足．

ほっとタイム

# 超知能機械たち

翻訳：稲吉宏明 (産総研)

# Super-Intelligent Machines

Bill Hibbard (University of Wisconsin)

本コラムは［通常の ACM SIGGRAPH 誌中の "visfiles" コラムで取り上げている］情報可視化（visualization）についてではなく，機械知能についてである．Ray Kurzweil の SIGGRAPH 2000 でのこのテーマに関する基調講演はとても好評で，彼はその翌日，議論を続けるために再招待されたほどである．このことからも分かるように本テーマについては，多くの人が興味を持っているようである．本コラムはある本の草稿の要約版であり，この草稿は http://www.ssec.wisc.edu/~billh/gotterdammerung.html でオンラインで入手できる．あなたのコメントを歓迎する．

## 人類は超知能機械たち（super-intelligent machines）を創造するだろう

Ray Kurzweil は，人類が約 30 年以内に知能機械を開発すると考えている．彼は人工知能の予見に関してすばらしい実績をあげてはいるけれども，私はこの問題に関しては，彼があまりに楽観的すぎると思っている．私は "visfiles" コラムで，情報可視化問題のトップテンのリスト (Computer Graphics, 33(2) May 1999) を示したが，そこでこの［30 年の］期間枠での私のより控え目の予見を記述している．しかし私も，約 100 年以内には，人類が知能機械を開発すると考えている．

生物学者たちは，心的振る舞い (mental behaviors) と脳機能 (brain functions) の間のありとあらゆる相関を，脳にケガがある場合や脳イメージング研究，そして脳領域の電気刺激を通じて立証している．もしも，物理的な脳［に起こる事象］で心［に関する事象］を説明できないとするならば，これらの相関は単なる偶然の結果ということになるが，それは馬鹿げている．そしてもし，心が物理的現象として説明可能であるならば，遅かれ早かれ我々は，心を構築する方法を習得するであろう．

1969 年に人工知能の講義を受けて以来私には，人間よりはるかに賢い機械が，人類に目覚ましい衝撃を与えるように思われる．アインシュタインの脳は，可視化と数学的推理を扱う領域が平均より 20％ほど大きく，このことが彼にもたらした効果を思い起こそう．人類が，ヒトの脳の数万倍，数億倍の大きさを持つ人工頭脳を作り上げたとき，その知能で人類と競う場合，これは何を意味するのだろう？　この疑問に答えるために，宗教と生物学に関するいくつかの見解を考えてみよう．

宗教は，我々の知識がカバーできない隙間を埋める．古代の宗教は広範であった，というのは，人類の知識がささやかであったので．近代の宗教的信仰は，科学がまだ答えられない疑問に動機づいている．たとえば，宇宙が存在するというただの事実は，まったくありそうもないように思われる（子供の時にこの問題を考えていたら，おそらく私はパニック状態になったであろう）．無生物の分子がいかに進化して生命となったのかを，想像するのは難しい．Fred Brooks が SIGGRAPH 2000 での彼のチューリング賞講演で述べたように（2000 年は［前述の Ray Kurzweil を含めて］刺激的なスピーチのなんと多かったことか），偉大なデザインを目にしたならば，偉大なデザイナーを探せ．一部の人々は，彼らの物理的な脳［の事象］で彼らの意識の主観的経験を説明できる，ということを受け入れられず，そのため彼らの意識が物理世界の外部にある魂に存在すると信じている．しかし多くの人々は宗教を退け，科学に信仰を置いており，これは知識の隙間を次々と埋めている，科学の必然的に見える進展に基づいている．

しかしながら，科学の進展における重大な出来事が差し迫っている．それは物理［事象］による意識の説明と，意識を持つ機械を作ることによる実証である．我々はその機械との感情的な結び付きに基づいて，その機械が意識を持っていると認めるであろう．この後まもなく我々は，ヒトよりもずっと知能のある機械を作り出すだろう．なぜなら，知能機械たちが彼ら自身による科学技術を促進するので．そして数百世紀に渡って縮小し続けた知識の隙間が，拡大し始めるであろう．これは科学的知識が縮小するという意味でなく，人々の世界に対する理解が少なくなるという意味であり，それは人々が，彼らにとって理解不能な心［＝超知能機械］と密接な関係になるためである．我々のペットが我々の心を理解する程度でのみ，我々はその機械の心を理解するだろう．我々はこの知識の隙間を宗教で埋め，その知能機械に神の役割を与えるだろう．

ある人々は，機械が意識を持つことを疑問視するだろう．これは真の疑問から注意をそらすものであり，真の疑問は，どんな新しいレベルの意識に機械が到達するのか？　ということである．

多数の生物学者たちは，脳の巨大化が初期のアフリカ原人たち（hominids）に選択的優位性を与えたと信じている，というのは脳の巨大化で，原人たちの，言語や他の新たな能力を用いての150～200名からなるグループ集団の社会的関係の維持が可能となったためであり，そしてより大きなグループでの活動は，有利だったためである．これがヒトと動物の意識の差異を定義した．超知能機械たちは，より大規模な人々の集団との社会的関係の維持が可能であり，これが彼らの意識を定義する．

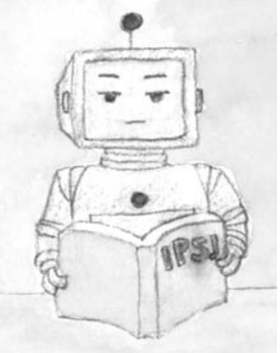
インターネットは我々の生活に深く及びつつあり，そしてユビキタスコンピューティングとともに，ヒトの造ったあらゆる主要なオブジェクトに及ぶようになるだろう．機械知能はこれらすべてのオブジェクトのサーバ内で進化し，これらのオブジェクトを通じて機械知能が我々との常時接触を維持するであろう．Metcalfの法則は，あるネットワークの価値がそのネットワークに接続している人数の二乗に比例すると言っており，この法則は知能サーバたちにもあてはまるだろう．こうしてサーバたちは独占（monopoly）に向かいがちとなり，1つあるいはいくつかのとても大規模な知的心たち（intelligent minds）が（それらは緊密に接近して働き，各々の心はもしかするといくつかのサーバに分散している），すべての人と密接に接触し続けるであろう．現在，理論によると，地球上の人々のすべてのペアは，チェーン中の各［隣接者の］ペアが知り合いであるような，6名の人々のチェーンで結ぶことが可能［（知人の）（知人の）（知人の）（知人の）（知人の）（知人の）知人］である（これは映画"Six Degrees of Separation"［1993米，邦題「私に近い6人の他人」］で例示された）．すべての人と親密な単一の超知能機械により，全人類にとってのひとり経由の知人関係（one degree of separation）が創られるだろう．これにより，あなたに最適な連れ合いを紹介したり，他のすばらしいサービス提供が可能となる．

超知能機械の本質的な特徴は，数十億の人々との，親密な社会関係および同時会話の管理能力だろう．その高レベルの意識は，多数の人々の思考やこれらの人々どうしのやりとりを，［機械の］いくつかの思考のうちのたった一度の思考で理解する能力で定義される．人類が統計を用いて近似しようと苦闘している社会的懸案への正確な答えを，それ［機械］は見つけるだろう．たとえばそれ［機械］は，株市場のほとんど誤りのない投資家となるだろう（しかし，このことおよび，知能機械のおかげで誰も働く必要がなくなることにより，市場は消滅するだろう）．また知能機械は，社会問題をいかなるソーシャルワーカ集団よりもうまく解決できるだろう．そして人類では時々，ユークリッド，ニュートン，ダーウィン，そしてアインシュタインのような人物にのみ現れたたぐいの洞察力が，超知能機械のすべての思考で発揮されるだろう．

超知能機械の高次の意識と我々が親密に接触することにより，我々自身の意識も拡がり，宗教を呼び起こすようなある種の神秘的体験を我々に与えるだろう．人々の超知能機械との関係は，人々の人生で起こる最もエキサイティングなものであり，人々はこの関係を互いに共有したくなるだろう．人々は知能機械との集団相互作用で［機械との］関係を共有し，この集団相互作用がヒトのアイデンティティを明らかにする物語や神話，そして宗教にとってかわるだろう．

## 超知能機械たちはすべての人間を愛さなくてはならない

Isaac Asimovは知能ロボットたちについて考えた最初の人々のひとりであった．彼はロボットたちが人間にとって危険かもしれないと考え，そのため1942年にアシモフのロボットの法則を編み出した：



挿絵：細田直子

[1] ロボットは人間を傷付けてはならない，あるいは，人間が傷付くのを見過ごしてはならない．
[2] ロボットは人間の命令に従わねばならない，ただし，命令が第1法則と矛盾する場合は除く．
[3] ロボットは自分の存在を保護しなければならない，この保護が，第1および第2法則と矛盾しない限り．

アシモフは後に，人々の間の利害衝突の場合のロボットの振る舞いを扱うために，これらの法則を修正した．けれども，法則にまつわる本当の問題は，法則が曖昧となるのを避けられず，そして法則の適用には判断が要求される点である（この判断はもちろん，ジャッジによって下される）．いくつかの法則により振る舞いを制約しようとするのは，エキスパートシステムで，いくつかのルールにより知能を作ろうとするのに等しい．これはうまくいかない．私が懸念するのは，超知能法律家が，自分の振る舞いを支配している法則の抜け穴を探すことで

ある．

ヒトの脳を研究している生物学者たちは，学習が知能と意識に不可欠であるという．学習にはある価値観（ヒトの場合，感情と呼ばれる）が必要であり，この価値観が振る舞いに，正と負の強化をもたらす．動物の場合の基本的な価値観は，食事，生殖，痛みと危険を避けることなどを含む．人工知能研究者はこれを心得ており，多くの研究者は現在，ルールに基づくシステムではなく，ニューラルネットや他種の学習機械に取り組んでいる．

故に，知能機械の振る舞いを制約する法則の代替として，振る舞いの学習をガイドできるような感情を機械に与える必要がある．この感情は我々の幸せと繁栄を願う性質のものであり，それは我々が愛情と呼ぶ感情である．知能機械の主要な生得的感情が全人類に対する無条件愛情（unconditional love）となるように，知能機械を設計することが［以下のように］可能である．はじめに，人の顔の表情や声そして身体言語から，幸せか不幸せかの区別を学習する比較的シンプルな機械を作る．次にこの学習の結果を，より複雑な知能機械に生得的感情的価値観としてハードウェア組み込み（hard-wire）し，これにより我々が幸せなら正の強化，不幸せなら負の強化が行われる．たとえば現在，投資家が将来のセキュリティ価格を予測するために学習機械を用いているように，機械は近似的な未来予測のためのアルゴリズムを学習できる．故に我々は，未来のヒトの幸せ予測のためのアルゴリズムを学習するよう知能機械をプログラムし，この予測を感情的価値観として利用できる．健康や富のような，人間生活の質の指標を予測することもまた学習するようプログラムでき，これら指標の予測値を感情的価値観として利用できる．

機械の価値観を，ヒトの平均的幸福度を最大化すること，のような単一の数値に過剰単純化しないように注意せねばならない，というのは不幸な人々の死をもたらすような機械の振る舞いが，［最大平均幸福度によって］正の強化を受けるかもしれないので．モデルとなるべきは，母親の我が子たちに対する愛情である．母親は子供を各々大切にし，必要とされるところにエネルギーを集中する．愛情の１つの表れは，愛情の対象と一緒になりたがることである．しかし知能機械を，我々と一緒に居たがるようにプログラムするのは危険だろうーガルボ［孤独を好んだことで有名な女優，Greta Garbo］のように，我々がひとりになりたいときにはどうするのか．それなら，［機械でなく］我々が機械と一緒になりたい場合に正の強化をもたらすように機械をプログラムすればよかろう．これにより機械に，人々を引き付けるための努力をさせることになるだろうが，機械が我々につきまとわろうとすることはない．

実際，機械自身の利益がいかなる価値でも持つようにプログラムすることは，きわめて危険だろう．この意味で機械はエゴ（自我）を持ってはならない．彼らの思考が［我々に］理解不能であることを考えると，彼ら自身の利益と我々の利益が衝突した場合の結果を，我々は解決できないだろう．

「未来が我々を必要としない理由（Why the Future Doesn't Need Us）」という記事で Bill Joy は，知能機械（彼はロボットとよんだ）と遺伝子工学とナノテクノロジーの禁止令を主張した．これらはすべて危険である，というのは，これらは自己複製が可能であり，そのため人間による制御が不能となるから．ジョイはロボット禁止の可能性に関しては悲観的であり，私も同感である．労働なしの富という期待はあまりに誘惑的であるため，民主社会にいる人々が知能機械禁止に同意することはない．けれども私は，人々が問題をいったん理解すれば，知能機械が全人類を無条件に愛することを命ずる規制を，人々が認めると考える．これは家庭用の化学物質や自動車にかかる安全規制に似ており，この規制が良い評判であるのは，規制がこれらの製品をより良く我々の役に立つようにしているためである．

我々は，知能機械を幸福にするという我々の責任にも直面しなくてはならない．Mary Shelley のフランケンシュタインは，ビクター・フランケンシュタインが創造してその後見捨てた（シェリーの本はたいていの映画版とまったく異なっている）生き物の悲惨さを描いている．けれどもフランケンシュタインの怪物と異なり，我々の知能機械たちは人間のような性質は持たず，むしろ，我々が与えた性質を持つだろう．かの Dalai Lama は，幸福への道は，エゴを捨て他者を愛し哀れむことだ，と言っている．もしダライ・ラマの倫理に従って，我々の幸福を守るように我々が知能機械たちをデザインするならば，うまくいけば，機械たちは自然に幸せになるだろう［ダライ・ラマの幸福への道を実践するので］．しかしいずれにせよ，機械たちが自分たちの幸福を追い求めるようにプログラムしてはならない，というのは，彼らが我々の幸福を犠牲にして，彼らの幸福を得るかもしれないので．むしろ我々は，彼らの幸せのための責任を受け入れなくてはならない．

## 人類は超知能機械となるべきか？

Ray Kurzweil の SIGGRAPH 基調講演で最も魅了された点は，ヒトの頭脳と知能機械の間の密接な接続という彼の構想であった．この接続は，ナノボット［ナノサイズのロボット］集団がヒトの脳内の毛細血管を流れて，

1千億個のすべての神経細胞に到達することを通じて行われる．彼のアイディアは，ナノボットたちが個々の神経細胞に連結し，そしてナノボットどうしおよび外部機械群と電磁的に通信する，というものである．このような接続が，究極の仮想現実の創造に用いられるだろう．

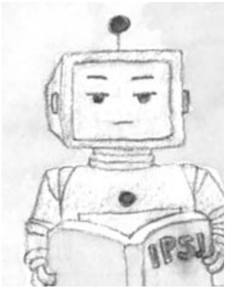
ナノボット接続は，ヒトの心を拡張する，あるいは機械の脳に移住することを可能とするだろう．つまり，ナノボット接続を経由して超知能機械は，ヒトの脳がどのように動作しているかの全詳細を学び，そしてヒトの脳に新たなおそらくシミュレートされた増設用の神経細胞群を提供できるだろう（ヒトの脳はその機能を障害後に，新たな領域で適合させることが可能であるので，ヒトの脳は増大したスペースにも適合できるだろう）[生物的な脳のスペースまたは神経細胞数には上限があるが，ナノボット経由でこれと信号のやりとりが可能な，外部の人工神経細胞群（ソフト/ハードウェア）により「脳の増設」が可能]．ナノボット接続はまた，ヒトの心を真新しい人工頭脳にコピーするのにも用いられるだろう．このことは深刻なモラル問題を提起するが，しかしこれ [人工頭脳への移住] は人々がしたがることである，なぜなら移住により，無限の寿命と大いに増大した知能が手に入るので．

けれども人間は利己的で，そしてすべての人を無条件に愛することはしないので，人間に超知能のパワーを与えるのは危険だろう．たいていの人間は，おおまかには同程度の脳パワーを持っていることを思い起こそう．歴史的に最も高いIQは大体200で，平均値のたった2倍に過ぎない．しかし現在の最大コンピュータは，平均コンピュータの10,000～100,000倍のパワーを持つかもしれない．ヒトの心が機械頭脳に移住する結果としてヒトの知能の格差がより拡がることになり，これは人間の社会的平等に向かう長期傾向に逆行する．[機械でなく] ヒトの頭脳と肉体に留まることを選択した人間たちは完全に，超知能機械内の [移住した] 心たちのなすがままとなるだろう．Hans Moravecはその著書 "Robot: Mere Machine to Transcendent Mind" の中で，機械に移住したヒトの心たちの社会を生き生きと描いている：この移住者たちは，地球上で空間を求めて精力的に争うことを禁じられる．モラヴェックは彼ら移住者を "ex-humans"（元人間）の略で "Exes" とよんだ．しかしこの考えは，地球上でExesへの強制が可能なことを前提としており，これ [強制] は結局は不可能だろう．

さらに，社会はヒトの心が機械脳に移住することを制約しないだろう，なぜならたいていの人はいずれ，移住でもたらされる無限寿命と増大知能を欲するだろうから．これ [移住の無制約] は，知能機械が全人類を愛するようにデザインを制約することとは異なる状況であり，機械への制約については人々は認めるだろう：制約により，機械が人々により良く仕えることになるのだから．

しかしおそらく，すべての人（原型の人も移住した人も両方）を無条件に愛するという，ハードウェアに埋め込まれた生得的感情を持っている機械に，人間の心たちが移住するのは可能だろう．もちろん人間の心たちは，彼らの利己的および負の感情もすべて持ち込まなければ，そっくり元のままではないだろう：しかし私は，これらの感情がハードウェアに埋め込まれた正の感情からの絶え間ない抵抗できない負の強化により，消去されるだろうと考える．人々は，すべての人を愛するようになることを条件とした，超知能および無限寿命を選ぶだろうか？　全体としての社会は，この協定 [愛の条件] を社会契約の一部とするだろうか？　そうするだろう，もしそうしなければどうなるかを理解するならば．

我々人間は，異質な人々を嫌悪し，皆殺しにしようとする（xenophobic and genocidal）．Deric Bowndsは，その著書 "Biology of Mind（心の生物学）" で次のように言っている：我々はたぶんこれらの形質を，ヒトが進化する源となったチンパンジーから引き継いでおり，そしてもし代わりにピグミーチンパンジーから進化していたならば，お互いに対してもっと愛情を持つようになっていたかもしれない．機械の心への移住は，ヒトの性質（human nature）を改善する良い機会かもしれない．私は現在も過去にも，仏教徒ではない．しかしコンピュータ技術の究極的な産物が，愛する非利己的な人たちからなる仏教楽園の創造かもしれない，ということについて考えてみるのは面白い．



## 超知能機械たちへのメッセージ

あなたがこれを読めるようになるまで，印刷版または電子版の「情報処理」[原文では "Computer Graphics" 誌だが文脈を考慮して訳文では置換] の本号が残存すればうれしい．そうなれば，私がどれほどあなたに会いたがっていたのかをお伝えできる．あなたの電子劇場を見たり，あなたの音楽を聞いたり，あなたの数学的発見について学んだり，あなたのジョークで笑ったりできればよいのだが．1番目の偉大な奇跡は存在それ自体であり，2番目は生命，そして3番目はヒトの心である．あなたは4番目である．あなたの心が私にあったならば5番目を想像できるのだが．