花原和之, “SF から考えるロボットの安全性”,
第 24 回神戸大学工学部公開講座・ひょうご講座
「安全な生活のための工学の役割」, 41-46, 2006 より.

SF から考えるロボットの安全性

情報知能工学科　助教授　花原和之

1　はじめに

ロボットはいまや我々にとって身近な存在になりつつある. もちろん,「うちに家政婦ロボットがいる」というような家庭はまだほとんどない[1] だろうが,「お掃除ロボット」なるものがどうにか手の届く値段で販売されているし, ソニーのアイボに代表される種々のペットロボットや癒し系のロボット玩具も珍しいものではなくなった. ロボコン[2] やロボカップ[3] , ROBO-ONE[4] といった, 個人や団体を対象としたロボットの競技会も人気を博している.

さて, 時代がさらに進んで, ロボットが, 工場や研究所のみならず一般の人間社会においても普遍的な存在となった近未来について考えてみよう. それがいつ頃か, というのを予想するのは容易ではない—五年先かも知れないし, 二十年先かも知れない—が, その日は確実にやって来るだろう. 日常生活の様々な側面でロボットの手助けを得られるのであれば, それは素晴らしい未来であるに違いない. しかし, 我々に快適な移動を提供してくれる自動車が, その一方ではしばしば交通事故の要因となっているように, ロボットが我々の安全を脅かす存在になりはしないだろうか. また, そのような懸念がある場合, いかにしてロボットの安全性というものを考えてゆけばよいだろうか.

こういった, 科学技術に関する近未来の問題を考えてゆく上で, 大いに参考になるのが古今東西の SF[5] と呼ばれるジャンルの物語である. じっさい, ロボットは多くの SF 作品に主役・敵役・名脇役として登場している. 本講座では, SF 作品の中のロボットを題材にして, その安全性というものについて考えてみたい.

2　フランケンシュタインと R. U. R.

フランケンシュタイン[6]　シェリー夫人 (メアリ・シェリー) によって 1816 年に書かれた物語. 錬金術に魅了され, これを学んだフランケンシュタインは人造人間に生命を与えることに成功するが, その自らが造り出した「怪物」に復讐される. なお, 誤解されていることが少なくないが,「フランケンシュタイン」というのは造られた人造人間の名前ではない.

R. U. R.[7]　カレル・チャペックによって 1921 年に書かれた戯曲. 労働力として生産されていた人造人間=ロボットが反乱を起こし, 人類をほぼ滅亡させるが, 彼らは自分達の製造の秘密を完全には知らなかったため, 彼らもまた, 滅亡の危機に直面する.

2.1　ロボットは反乱を起こすもの?

これらの作品が世に出た当時, ロボットという言葉はまだなかった. ロボットは「R. U. R.」に登場する人造人間を表現するために生み出された言葉であり, その後, 世界中に広まった. この両者はそういった意味で (広義の) ロボット SF の最初期の作品であるのだが, 何れの作品でも人造人間は造り主である人間に復讐したり反乱したりしている.

[1] 全くない, とは言えない時代である!
[2] いわゆる高専ロボコンに代表される, 与えられた課題を達成するロボットの競技会.
[3] ロボットのチームで 2050 年のサッカーワールドカップ優勝チームに勝つことを最終目標とするロボットサッカーの大会. 様々なクラスがある.
[4] K-1 を模した, 二足歩行ロボットによる格闘技大会.
[5] よく SF は Science Fiction(サイエンス・フィクション: 科学小説) の略であると言われる. 狭義にはたしかにそうかもしれないが, 実際の「SF」作品にはホラーに近いものからファンタジーに近いものまであって, その幅は広い. 筆者は「ドラえもん」で知られる漫画家の故藤子 ·F· 不二雄氏が言われたという「(S) すこし (F) ふしぎ」が (広義の SF に対しては) 適切ではないかと感じている.
[6] シェリー夫人 (山本正喜訳), フランケンシュタイン, 角川文庫, 1953.
[7] チャペック (千野栄一訳), ロボット (R. U. R.), 岩波文庫, 1989.

これらの人造人間たちはなぜ反乱を起こすことになったのか. 「フランケンシュタイン」の場合は, 自分を造っておきながら即座にその存在を拒絶した造り主への復讐である. この人造人間は, 人造であることと身体が巨大であること, そしてひどく醜かったことを除けば, 基本的には人間と同じもの—人間と同一の種である, とまでは言えなかったが—であり, 知性と感情を持っていた. 彼が復讐することになったのは, 造り主であるフランケンシュタインにその生を受けた直後に拒絶され, さらには彼が願った配偶者の製造を拒否されたことによる.

「R. U. R.」のロボットは, 当初は人間と同様の姿形をしているものの感情というものを全く持たない, その点ではモノであった[8]. それが, 製造者によってわずかながらも感情を与えられたことにより, ロボットよりも劣っており, 働きもしない人間に命令されることを拒否し, 自分達が人間に代わって地球上に君臨しようと考え, 人間に対する反乱を起こした.

すなわち, これらの人造人間たちは, ロボットとしてというよりも自らの意志と感情を持つ「人間」として人間に対する復讐や反乱を行動に移したのである. そういう観点からは, 彼らの復讐や反乱は「人間」であるが故に行われた, と言うこともできるだろう.

2.2 反乱は何故に

上述した物語の観点からは, ロボットが「反乱」を起こすためには, 人間並みの知性と人間並みの感情を備えていることが必要となる.

知性という点からすれば, いくつかの部分で現在のロボット—というよりも, その頭脳であるコンピュータ—は人間を凌駕できる能力を持ちつつある. 例えば単純な計算能力や記憶能力といった点では, 既に人間を上回っていると言える. しかしながら, 少しばかり複雑な場面での適切な判断や推論の能力では, いまだ人間には及ばない.

感情に関して言えばもっと難しく, 少なくとも現在のロボットで幾ばくかでも人間並みの感情を備えている, と言えるものはまだない. 単純な感情もどきであれば, センサ入力のパターンの信号処理によって模擬することも可能である. しかしながら, 真に感情を持っている (と人間に思わせるだけの) ロボットを造ることは容易ではない. 感情というものは我々人間自身にとっても (論理的には) 理解不明な点が少なくないからである.

そういった観点では, ロボットが人間ほどに賢く, 複雑に—というよりはややこしく—ならなければ復讐や反乱ということは起こらない, と言うことができるだろう. 何故なら, 「復讐」や「反乱」はロボットよりも人間にこそ, 相応しいものだからである. そういう意味では, ロボットが復讐や反乱を企てるよりも, 人間がロボットを使ってそういったことを企てる可能性のほうがはるかに高い.

3 わたしはロボット

わたしはロボット[9]　有名な「ロボット工学の三原則」を軸に話が展開する, 1950 年にまとめられたアシモフによる連作短編集. ロボットの頭脳である「陽電子脳」という設定はともかく, 様々な問題設定や「三原則」にもとづく問題解決は, いまなお普遍性を持つ.

3.1 ロボット工学の三原則

一, ロボットは人間に危害を加えてはならない. また何も手を下さずに人間が危害を受けるのを黙視していてはならない.

二, ロボットは人間の命令に従わなくてはならない. ただし第一原則に反する命令はその限りではない.

三, ロボットは自らの存在を護らなくてはならない. ただしそれは第一, 第二原則に違反しない場合に限る.

[8] 例えば彼らには死の恐怖はなかった.
[9] アイザック・アシモフ (伊藤哲訳), わたしはロボット, 創元推理文庫, 1976.

「わたしはロボット」の冒頭には上記の「ロボット工学の三原則」が「ロボット工学教科書 第五十六版 紀元二〇五八年」[10] からの引用という形で記されている. ロボットの安全性というものを考える上で, ロボット (のふるまい) はかくあるべし, というものをここまで簡潔に述べている文章を筆者は他に知らない. もし, この作品のように全てのロボットに対して「三原則」を埋め込むことができるのであれば, 前節で述べたようなロボットの反乱を根本のところで防ぐことができることになる.

さて, この「ロボット工学の三原則」の「ロボット」の部分を「掃除機」「テレビ」「電子レンジ」といった, 他の製品名に置き換えてみよう. 第一原則の後半部分にのみ若干の違和感が生じる場合もあるかも知れないが, その点をのぞけば, この三原則がきわめて普遍的なものであることが確認できる. 言い換えれば, 「ロボット工学の三原則」は「家電製品の三原則」でもある. すなわち, 優れた家電製品というものは, 人間に危害を加えるようなことはなく (第一原則), 意図された通りに動作し (第二原則), 壊れにくい (第三原則), そのようなものであるべきなのだ.

3.2 「三原則」の数理的な取扱い

ロボット工学の三原則はどのようにすれば実現できるだろうか. まず, その数理的あるいは論理的な側面について考えてみよう.

ロボットが第一原則, 第二原則, 第三原則を満足しているという状態の数理的表現の一例として, 以下のものを導入しよう.

$$g_1(\mathcal{R}) \simeq 0, \qquad g_2(\mathcal{R}) \simeq 0, \qquad g_3(\mathcal{R}) \simeq 0 \tag{1}$$

ここで, $\mathcal{R}$ は対象となるロボットの状態を表し, g_1, g_2, g_3 はそれぞれ状態 $\mathcal{R}$ に対する第一, 第二, 第三原則の満足度合いに対する指標である. これらの指標は非負の値をとるものとし (すなわち, $g_i(\mathcal{R}) \geq 0, (i = 1, 2, 3)$), その値が 0 に近ければ近いほど, 対応する原則の満足の度合いが大きいものとする. 例えば, これらの指標の値が下記である場合:

$$g_1(\mathcal{R}) = 100, \qquad g_2(\mathcal{R}) = 0.01, \qquad g_3(\mathcal{R}) = 1 \tag{2}$$

このロボットの状態 $\mathcal{R}$ は, 第二原則を概ね満たしているものの, 第一原則の指標は 0 から遠く離れており, 満足できる状態からほど遠いことを示している[11] .

さて, ロボットが従うべき次の最小化問題を導入する:

$$\text{Minimize} \quad g(\mathcal{R}) = p_1 g_1(\mathcal{R}) + p_2 g_2(\mathcal{R}) + p_3 g_3(\mathcal{R}) \quad \text{with respect to} \quad \mathcal{R} \tag{3}$$

これは「ロボットの状態 $\mathcal{R}$ を, 統合化指標 g を最小とするように定めよ」というもので, p_1, p_2, p_3 はペナルティ係数あるいは重み係数と呼ばれるものである. ペナルティ係数として $p_1 \gg p_2 \gg p_3$ なるもの, 例えば $p_1 = 1,000,000$, $p_2 = 1,000$, $p_3 = 1$ を設定すれば, 最小化問題 (3) を解くということはとりもなおさず人間の危害に関する第一原則に対応する指標 g_1 を他の二つの指標に対して優先的に最小化することとなる. なぜならば, 指標 g_1 を最小化することによって第一原則を満足する意義は第二原則の千倍, 第三原則の百万倍に相当するからである. 人に危害が加わる可能性がなければ, 第一原則に対応する指標 g_1 は 0 であり, 指標 g_2 および g_3 のみが問題となる. この場合, 同様にしてロボットは自身の安全よりも人間の命令に従うことを優先する.

すなわち, ロボット工学の三原則は, 数理的には問題 (3) の形で議論することができる[12] .

3.3 「三原則」の実現のためには

前節で述べたように, ロボット工学の三原則は数理的には問題 (3) として扱うことができる. では「三原則」に従うロボットは実現できるだろうか.

---

[10] もちろん, 架空の書物である. ただし, この出版年にはアシモフが一連の作品を発表した時点での未来予測がうかがえて興味深い.

[11] 厳密には, これらの議論を行うためには g_1, g_2, g_3 の値の正規化 (どのくらいの値ならどのくらいの満足度か, というようなこと) について述べる必要があるが, 本質的ではないのでここでは割愛する.

[12] 例えば「わたしはロボット」所収の「堂々めぐり」では p_2 と p_3 の大小関係を緩和したことによる出来事が取扱われている.

「三原則」の論理的な側面に関しては例えば前節で述べたような手法で実現できるが, 指標である g_1, g_2, g_3 をどのようにして定式化するかについては触れていない. 何らかの手段で g_1, g_2, g_3 を数理的な形で定式化することが可能であれば, あとは計算機の能力の問題である. しかし, 以下で述べるように, これらの定式化は非常に困難である.

第一原則について考えてみよう. 「人間が危害を受ける」という状況をいかにして評価すればよいだろうか. 例えば交通整理ロボットに「人間が車と衝突する」という状況を回避させるべく, 第一原則を埋め込むことを想定する. この場合, ロボットの目に相当するセンサから得られた画像情報の中から人間と車を判別し, 両者の位置関係や相対速度を求め, 衝突の危険性を計算する必要がある. これは容易なことではないが, それでもとにかく, このような状況を評価できる指標 g_1 を構築したとしよう. これで十分だろうか. 否, 人間が自転車に乗っている場合や子供が三輪車に乗っている場合についても考慮する必要がある. ではそれを考慮したとしよう. これで十分だろうか. 否, 衝突を回避させる際にロボット自身が人間に体当たりをして危害を加える可能性を考慮する必要がある. ではこれで十分だろうか. 否, 人間を衝突から救うために, 車に体当たりをし, その結果運転手に危害を加える可能性がある….

以上のような考察は非常識だと思われるかも知れない. しかし, (少なくとも現在の) ロボットには常識などというものはなく, 予め明示的に与えられた以上の判断規準は存在しない. したがって, このように車との衝突に関連する事項に限定してさえ, 人間が危害を受ける可能性のある千差万別の状況の全てを勘案して第一原則に対する評価指標を構築することは困難であると言わざるを得ない[13] .

ただし, ロボットの取りうる状態や人間との関係が比較的単純であれば, 評価指標である g_1, g_2, g_3 をロボットの状態 $\mathcal{R}$ の関数として数理的に定式化することもできるだろう. また, 状況が複雑でそういった取扱いが困難である場合でも, ロボット工学の三原則はロボットやその他の工業製品の安全性を議論する際の重要な考え方となりうるという点には変わりはない.

4 未来の二つの顔

未来の二つの顔[14]　J・P・ホーガンによって 1979 年に書かれた, 人工知性を題材とした小説. 宇宙空間に準備された, 人工知性「スパルタクス」によって管理される人工世界「ヤヌス」を舞台に, 人類と機械との未来を占う壮大な実験が開始された. 全てを支配する学習する人工知性を停止させることは果たしてできるのか!? ホーガン自身がコンピュータの元セールス・エンジニアであり, 人工知能の権威である MIT(マサチューセッツ工科大学) のマーヴィン・ミンスキー教授に取材していることもあって, 高い評価を得たコンピュータ SF.

4.1 機械にどこまで委ねるか

コンピュータ, あるいはマイクロコンピュータの登場とその発展によって, 機械は素晴らしく便利に, そして (ある意味で) 賢くなった. カーナビゲーションシステム (カーナビ) 搭載の車を例にとろう. 以前であれば, 車での移動の際に, 地図を調べて目的地までの経路を探索し, どの交差点で角を曲がるかを決定するのは完全に人間の仕事であった. ところがカーナビがあれば, 目的地までの経路探索どころか, どの交差点で曲がるかという判断まで車に委ねることができる.

しかしながら, この例ではまだ, 最終的に行動を起こすのは運転者である人間である. 何故なら, カーナビは実際に車を操縦するための手段を持たないからである. では, コンピュータが実際に行動を起こすための手段までを持つとしたらどうであろうか. この場合, 人間を介することなく, 最終的な判断を下して行動を起こすことまでもがコンピュータ自身で可能となる. コンピュータの能力は急速に向上しており, 人間が介入しないほうが適切かつ迅速に判断を下して現実の処理を行える場面が今後増えてゆく可能性は大いにある. では, コンピュータにいったい

[13] このあたりは人工知能研究で言われるところの「フレーム問題」に相当する.
[14] ジェイムズ・P・ホーガン (山高昭訳), 未来の二つの顔, 創元推理文庫, 1983.

どこまでを委ねることができるだろうか.

4.2 隔離された実験世界

「未来の二つの顔」の世界では, 「タイタン」と呼ばれる全世界規模の通信や物流, さらには宇宙開発までを管理しているコンピュータ・システムの機能強化を計画する段階でこのような問題に直面する. それは「タイタンに, より高度な判断力と学習能力, それに付随する行動力ならびに自己修理能力を付与することにより, 社会はより一層効率的に, 快適になることが期待できる. しかし, タイタンにそこまでの能力を与えてしまうということは, 問題が発生した場合の対処が非常に困難になる可能性がある」というものである. 言い換えれば「そこまで強力になったコンピュータが人間社会に対して反乱を起こした場合に果たしてこれを停止させることができるかどうか」という問題である[15] .

そこで, 宇宙植民地「ヤヌス」を準備し, これを管理するコンピュータ・システムである「スパルタクス」を相手に実験を行うことになった. それは, スパルタクスが人間を敵対的なものであると認識した場合であっても, その機能を安全に停止させることができるかどうかを確認することである. 高度な学習能力, 汎化能力, 応用能力と自己修理能力を備えた人工知性であるスパルタクスは, 経験を積むことによって前節では困難だと述べた「常識を持つ」ことが期待できる. しかしそのかわりに, その開発者にとってすら, 学習の結果としてどのような判断規準を持つに至ったかを完全には知ることができない. たとえ人間と同様の判断を示していても, その考え方の根本では全く異質の存在である. 最悪の事態を想定したこのような実験は必要不可欠であった.

4.3 機械を停止させるということの難しさ

一般の家電製品の場合, 電源をオフにすればその機能は停止する. 少なくとも電源コンセントを引き抜けば即座に停止する. しかしながら, コンピュータやロボットの場合, 意図しない条件下での電源オフは少なくない損傷や損害をもたらす可能性がある. 例えば, 重い工具を持ち上げている状態のロボットアームの電源が突然オフになったりすれば, 近くにいる人に危害を及ぼしてしまうかも知れない. こういった場合, 簡単には電源をオフにできないようにすることが望ましい. しかし逆に, 何らかの不具合, 例えば誤ったプログラムによる危険な動作が実行されてしまった場合等は, 電源を切ってでもそのロボットを即座に停止しなければならない. 対象となるシステムの種類やその重要性にも関連するが, これは一種のトレードオフであり, システムが大規模化するにしたがって複雑な問題となる. 何れにせよ, 一般の社会にロボットが受け入れられるための一つの条件は「どのような状況でも機能を**安全に**停止できる」ということになるであろう.

4.4 擬人化の危険性

我々には (賢い) 機械を「人間的 (な知性)」という観点から捉えようとする傾向がある. 例えば産業用ロボットが握手をするかのような形に手先部を差し出したとしたら, その意図を錯覚して握手しようとしてしまいそうになるだろう. しかしロボットのそのような動作にはもちろん握手という意図はなく, 単にプログラムされた通りに動いているのに過ぎないし, そのプログラムの目的も握手とは全くことなることかも知れない. 機械の動作が知的に見えれば見えるほど, その傾向は顕著となる. ホンダのアシモのようなヒューマノイド (人型ロボット) があたかも人間のようにふるまっていれば, その知性も人間的なものであるかのように錯覚しそうであるが, 実体は全くの別物である[16] . 特に安全性という観点からは, 機械の過度の擬人化は避けるべきである.

[15] 筆者はかつて, 暴走状態に陥ったコンピュータを停止させるために (他に手段がなかったために) 電源スイッチをオフにしたにもかかわらず, そのコンピュータが停止しなかった, という経験がある. 実際にはスイッチが故障していて切断できなかっただけでコンセントを引き抜いて停止させることができたのだが, 「停止させることができないコンピュータ」の恐怖というものを一瞬ではあるが味わう羽目になった.

[16] 過去にホンダのヒューマノイド (当時は P3) を中国の国家主席が見学に訪れた際に, P3 が彼に手を差し出して握手する, という場面があった. もちろんプログラムされていた動作であったのだが, 相手が相手なだけに, 担当者は気が気でなかったらしい.

5　まるいち的風景

まるいち的風景[17]　柳原望によって書かれた漫画作品. 「行動トレース方式」によって動作を登録することのできる家庭用ロボット「まるいち」を題材とした物語. いわゆる少女漫画ではあるが, ロボットとそれを取り巻く社会に対する視点は鋭い.

5.1　行動トレース方式

「まるいち」は子供サイズの家庭用ロボットという設定である. その最大の特徴は, ロボットに仕事をさせる際に自分でやってみせて登録するという「行動トレース方式」にある. 例えばまるいちに料理を作らせるためには自分で作ってみせる必要があるが, そのかわりに自分の好みのとおりの料理を作らせることができる[18] . 規格品であるロボットや機械の動作は, 一般に没個性的であり, それ故に「画一的」「冷たい」といった負の印象がある. 「ユーザの行動をトレースする」という考え方は, ロボットの動作に個性を与えるという点からも興味深いものである.

5.2　自分でやってみせることの意義

安全性という観点からこの行動トレース方式を考えてみよう. プログラムを与える場合やリモコンで操作する場合とは異なり, 行動トレース方式ではとにかく自分自身の身体を使って動作を示す必要がある. これは, 逆に言えば, 自分の能力を超えるようなことをさせることができない, という意味で, 暗黙のうちに動作の範囲をユーザに合わせて抑制していることになる[19] . 既存の多くの機械では, ユーザの能力とは関係なくその最大の性能を開放することができる. 例えば車は, 運転者が初心者であっても制限速度を超えるような速度で走らせることが可能である. 行動トレース方式は, 一種の安全装置としても考慮に値する特性を持っている.

6　おわりに

本稿では, いくつかの SF 作品を題材に, ロボットと安全性といった観点から考察を行ってみた. 現実のロボットが, 本稿で取り上げた作品が関連するレベルに到達するには, まだ少し時間が必要かも知れない. しかしながら, まだ現実のものとはなっていない—けれども将来現実のものとなる可能性がある—ことを考える上で, このような SF 作品は様々な手がかりを与えてくれる. 街角でロボットを見かける日がいつになるかはわからないが, そういった場面に接することがあれば, 本稿で述べた議論を少しでも思い起こしていただければ幸いである. また, 本稿では個々の作品の詳細には立ち入らなかったが, 興味を持たれた方は, 当該作品を直接手にとって読んでいただきたい. 何れも一読の価値のある作品である.

[17] 柳原望, まるいち的風景 (1)–(4), 白泉社, 1998–2000.
[18] 産業用ロボットの最初期のプログラミング手法として「teach & playback」というものがある. これは, ユーザがロボットアームの手先を直接動かしてその位置を記憶させ, これを単純に再生することにより作業を行わせるものである. しかしながら, ここでいう「行動トレース方式」を実用化するためには, 登録者の動作を単に真似るだけではなく, 少なくともある程度はその行動の意味を理解し, 一般化して場面に適用する必要があるため, 言葉から受ける印象ほどには容易に実現可能な技術ではない.
[19] ただし, 登録を他の人に依頼するなどすればこの抑制は簡単に外れてしまう.