“いい人”ロボットの作り方

**自律するロボットが生活の中に入ってくる
周りの人々にとって「いい人」でいるにはどう振る舞うか
ロボットに覚えてもらう時代の到来だ**

M. アンダーソン（ハートフォード大学）／ S. L. アンダーソン（コネティカット大学）

暗黒の未来を描くSFの古典的シナリオでは，機械が知能を持ち，人間に対抗するのが常道だ。そこでは，機械は人間を傷つけたり，殺したりすることに何ら良心の呵責を覚えない。

今日のロボットは通常，もちろん人間を助けるように作られている。だが実際には，ごく日常的な状況でも，ロボットはどう振る舞えばいいのか倫理的に戸惑ってしまう場面にしばしば遭遇しており，従来の人工知能（AI）の領域を拡大する必要があることが明らかになった。

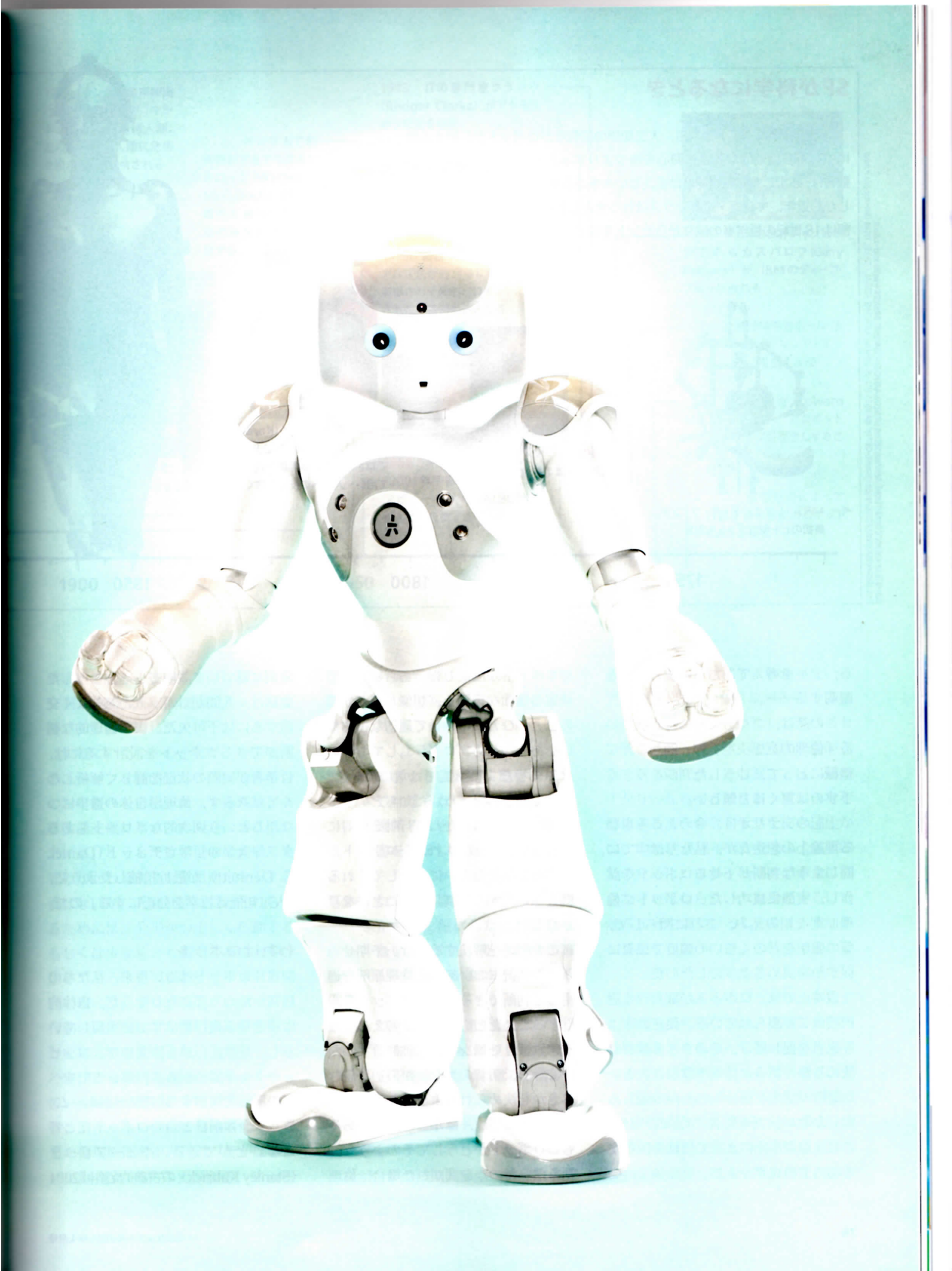
仮に，介護施設に入居していると想像して頂きたい。実際，お手伝いロボットは遠からず，こうした場所でごく一般的に見られるようになるだろう。午前11時が近づき，あなたはテレビのニュースを見ようと思って，娯楽室のお手伝いロボットにリモコンを取ってくれと頼む。だが，別の入居者はクイズ番組を見たがり，同じようにリモコンを取ってくれとロボットに言う。するとロボットは，別の入居者にリモコンを手渡してしまう。あなたは最初は腹を立てるが，ロボットは「あなたは昨日，好きなニュースを見たのだか

KEY CONCEPTS

機械に倫理をプログラムする

- お年寄りの介護ロボットなど，自律的に行動決定するように設計されたロボットは，ごく日常的な場面においてすら，どう振る舞うのが倫理的に適切なのか迷ってしまうだろう。
- ロボットが人とのかかわりの中で適切に行動できるようにするにはどうするか。一案として，一般的な倫理原則をプログラムとして組み込み，これに照らしてケースバイケースで意思決定をさせる方法が考えられる。人工知能なら，倫理的に許容される行動の具体例から，論理学の手法を用いて，そうした倫理原則自体を作り出すことができる。
- 著者らはこうした考え方に基づき，倫理原則に照らして行動を決めるプログラムを，初めて実際のロボットに組み込んだ。

「NAO」 倫理原則プログラムを備え，これに照らして行動を決める初のロボットだ。



SFが科学になるとき

倫理学者やロボット工学者，人工知能の研究者がロボットが行動する際に生じ得る倫理的な影響に関心を持つよりずっと前から，SF作家や映画監督は，必ずしも非現実的とは言えない未来のシナリオを思い描いていた。しかし近年，マシン・エシックスはれっきとした研究分野となり，その一部は18世紀の哲学者の論文からヒントを得ている。

1495 レオナルド・ダ・ヴィンチ，最初のヒト型ロボットを設計

1780年代 ベンサム(上)とミル，倫理の計算可能性を示す

1921 チャペック(Karel Čapek)，戯曲『R.U.R.』でロボットという言葉を初めて用い，人間への反逆を描く

1750 1800 1850

COURTESY OF NICCOLAI-TEKNOART/MUSEUM OF LEONARDO DA VINCI, FLORENCE (*humanoid robot*); THE GRANGER COLLECTION (*Bentham*); GETTY IMAGES (*R U R poster*); UFA/THE KOBAL COLLECTION (*Metropolis robot*)

ら，公平を考えてあちらに渡した」と説明する——。

この話は，ごく日常的に行われている「倫理的意思決定」の一例だ。だが機械にとって，こうした判断をうまく下すのは驚くほど難しい。

上記のシナリオは，今のところ単なる理論上の想定だが，私たちはすでに，同じような判断が下せるロボットを試作し，実演に成功した。ロボットに倫理原則を組み込み，これに照らして，薬の服用をどのくらいの頻度で患者に促すかを決めるようにしたのだ。

現時点では，ロボットが取り得る選択肢はごく限られている。薬を飲むよう患者を促し続け，そのタイミングも決めるか，あるいは薬を飲まないという患者の意向を受け入れるかのどちらかしかない。だが私たちが知る限り，これは倫理原則によって行動を決定する初めてのロボットだ。

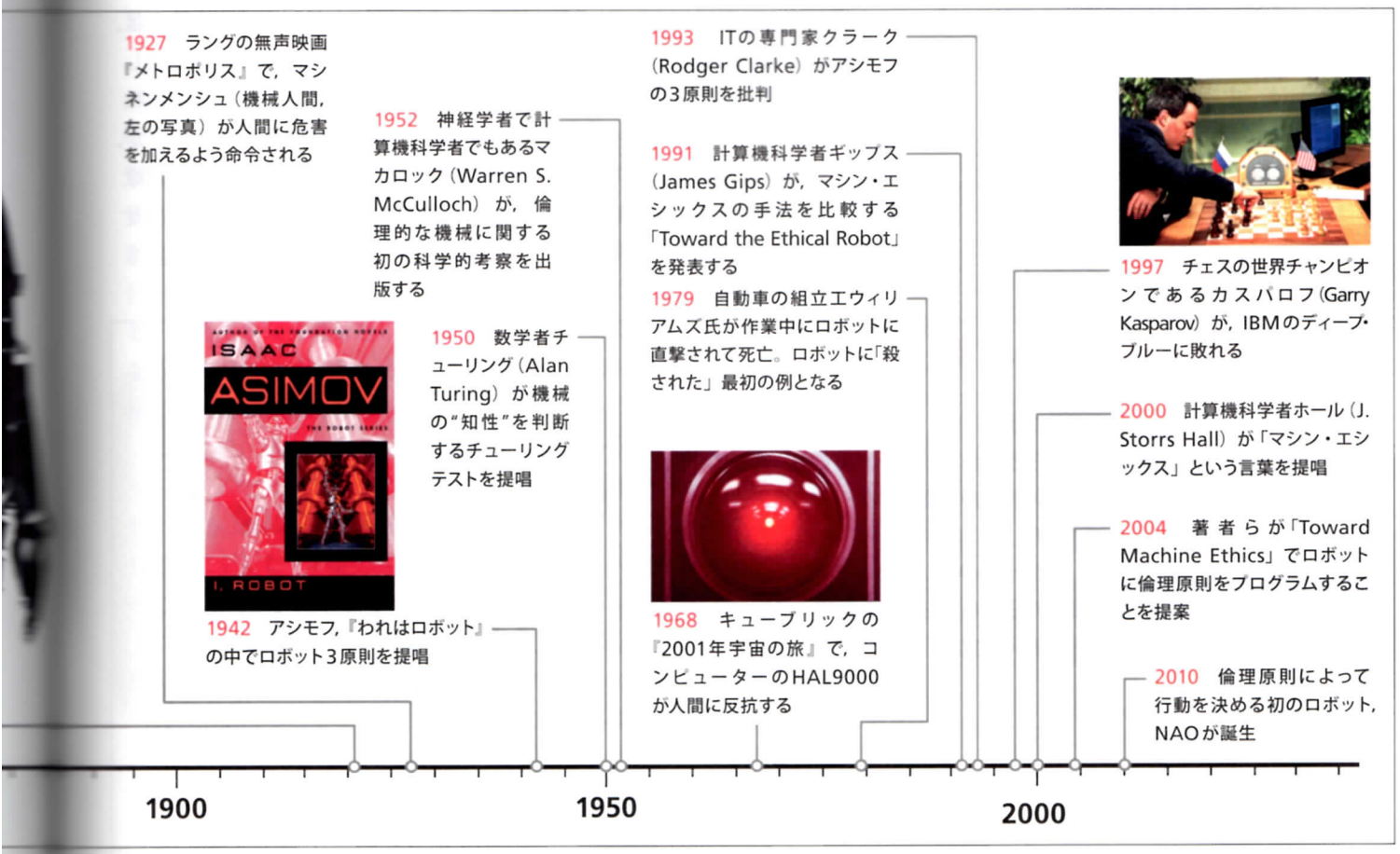
ロボットが直面し得るあらゆる意思決定の場面を予測し，想像し得る状況のひとつひとつについて適切な振る舞いをあらかじめプログラムしておくのは，まったく不可能ではないにせよ，極めて難しいだろう。だからといって，倫理的な事柄にかかわる行動を絶対に取らせないようにすれば，ロボットが人間の生活をおおいに改善してくれる機会をいたずらに制限することになりかねない。

この問題を解決するには，予期せぬ新たな状況においても，倫理原則を適用して判断できるよう，ロボットを設計することだと，私たちは考える。そうすればリモコンを次に誰に渡すかだけでなく，新着の本を誰が最初に読むべきかを決められるようになる。

この手法には，ほかにも利点がある。もしロボットがどうしてそのような行動を取ったのかを聞かれた場合，倫理原則に基づいて説明できる。こうした会話は，人間がロボットと気分よく交流するには不可欠だ。また倫理的な判断ができるロボットを設計するには，哲学者が実際の状況を詳しく解析しなくてはならず，倫理学自体の進歩につながるという副次的なメリットもある。タフツ大学の哲学者デネット(Daniel. C. Dennett)が最近指摘したように，「人工知能は哲学を公正にする」のだ。

われはロボット

自律ロボットは近い将来，私たちの日常生活の一部になりそうだ。自律的に空を飛ぶ飛行機はすでに実現しているし，自律走行車も開発中だ。コンピューターが照明からエアコンまですべての機器を制御する「スマートホーム」も，家全体がひとつのロボットだと考えることができる。キューブリック(Stanley Kubrick)の古典的名作『2001



COURTESY OF RANDOM HOUSE (*Asimov book cover*), MGM/THE KOBAL COLLECTION (*HAL 9000*), AFP/GETTY IMAGES (*Kasparov and Deep Blue*)

年宇宙の旅』に登場するコンピューターHAL9000が，宇宙船ロボットの頭脳であるのと同じだ。

すでにいくつかの企業が，介護施設のスタッフを補助したり，一人暮らしのお年寄りを手伝ったりする高齢者向け生活支援ロボットの開発に乗り出している。こうしたロボットは普通，生死にかかわるような意思決定を下す必要はないが，人間社会に受け入れられるためには，公平かつ正しく行動する，もしくは少なくとも親切であると受け止めてもらう必要がある。ロボット開発の際には，プログラムがロボットの行動の倫理面にどんな結果をもたらすかを考えておくべきだ。

自律する機械と人間とがうまくやっていく秘訣は機械に倫理原則を組み込むことだ，という点に異論はなくとも，どの倫理原則を組み込むべきかはおおいに問題だ。SFファンなら，アシモフ（Isaac Asimov）のロボット3原則に，とっくに答えが書いてあるじゃないか，と思うかもしれない。

第1条 ロボットは人間に危害を加えてはならない。また危険を看過することによって人間に危害を及ぼしてはならない。

第2条 ロボットは人間から与えられた命令に服従しなければならない。ただし，その命令が第1条に反する場合を除く。

第3条 ロボットは，第1条および第2条に反しない限りにおいて，自己を守らなければならない。

——『われはロボット』I. アシモフ著

だが，この法則の帰結を考えると矛盾が生じることはすでにわかっている。アシモフ自身が1942年の短編で最初にそれを示した。1976年の『バイセンテニアル・マン』（邦訳は東京創元社『聖者の行進』に収録）では，通りすがりのならず者がロボットに自身を解体するよう命令する話を描き，いかにこの法則が理不尽かを示した。第2条のために，ロボットはこの命令に従わなければならない。自身を守るにはならず者を傷つけるしかなく，第1条に反してしまうので，それもできない。

アシモフの3原則が受け入れがたい場合，代案はあるのだろうか。またその案は実現可能なのか。機械に倫理的な行動をさせるなんて不可能だ，と考える人もいるが，英国の哲学者ベンサム（Jeremy Bentham）とミル（John Stuart Mill）は，19世紀にすでに，倫理的な意思決定というのは「道徳についての演算」によって実行されると

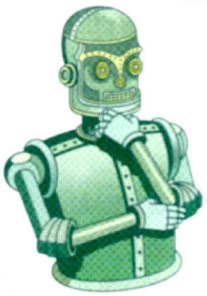
主張していた。

彼らが主張する「快楽主義的行為功利主義」は，主観的な直観に基づく倫理の対案として出されたものだ。これによると，正しい行為というのは，その行為によって影響を受ける人全員について幸福の度合いを足し合わせ，そこから不幸の度合いを差し引いた「総合の幸福度」が最大となる結果をもたらすとみられる行為だ。

ただし，倫理学者の多くは，この理論を倫理問題のあらゆる側面に適用することには疑問を持っている。例えば，正義に関する側面をとらえるのが難しく，多数派の幸福のために個人を犠牲にすることにつながりかねない。だが，この理論は同時に，ほぼ妥当な倫理原則を，原理的に計算によって導き出せ

行動の原則をプログラムする

人間とかかわりながら仕事をするロボットは，しばしば倫理的な判断を含む意思決定をする必要があるだろう。ロボットが直面し得る倫理的なジレンマをすべて予測するのは不可能だが，個々のケースの意思決定（右）の指針となる統一的な原則（下）を，ロボットに与えることはできる。著者らは，ヒューマノイドロボットNAO（69ページの写真）に，患者に薬の服用を促すか，促すとしたらどれくらいの頻度でするかを決めるプログラムを組み込み，自らの考えを実証した。

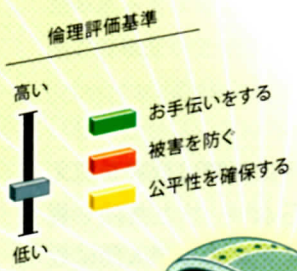


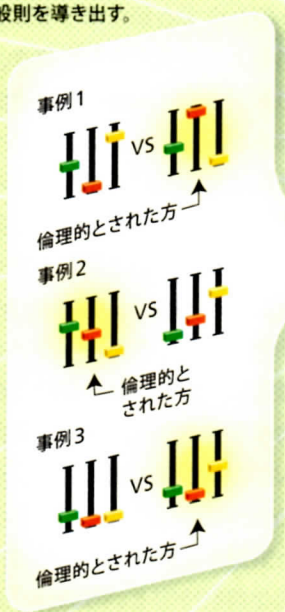
次の作業を考える

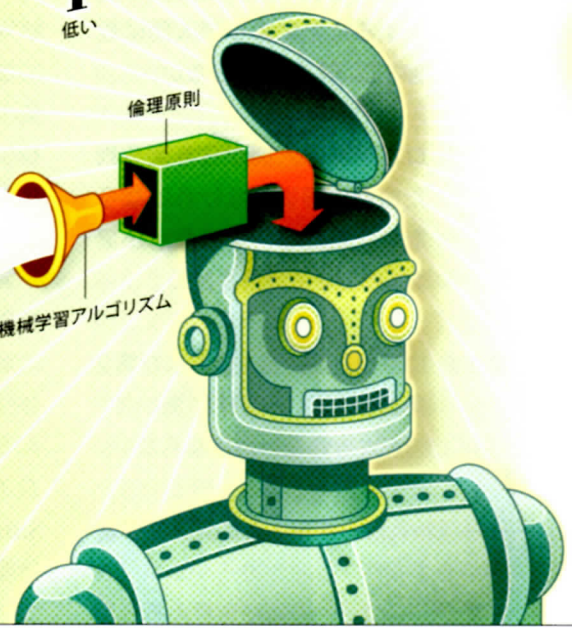
この場合，倫理原則によれば，食べ物やリモコンを取ってくるよりも，薬を手渡す方が優先される。

ルールの決め方

機械学習と呼ばれる人工知能技術を用いて過去の事例から倫理原則を引き出し，ロボットにプログラムすることは可能だ。設計者はまず機械学習アルゴリズムに，いくつかの事例において，どんな意思決定が倫理的と考えられているかを入力する。倫理的かどうかは，その行動がどれくらい良い結果をもたらしたか，どれくらい被害を防いだか，公平性はどのくらい実現されたか，などに基づいて判断する。アルゴリズムはこれらの情報を抽象化し，これまでに経験したことのない事例にも適用できる一般則を導き出す。







優先度の決定

高齢者支援ロボットは，取り得る行動の選択肢それぞれを，各倫理評価基準をどの程度満たしているかという観点から点数付けする。そしてこの点数を，プログラムに組み込まれた倫理原則に照らして，今この時点でどれを優先すべきかを計算する。例を挙げると，ある居住者が食べ物を，ほかの居住者がテレビのリモコンを頼んだとしても，ロボットは患者に薬の服用を促すといった，別の作業を優先するかもしれない。

PETER AND MARIA HOEY

ることを示している。

機械は感情を持っておらず，機械の行為によって影響を受ける人の気持ちを感じ取ることができない。だから倫理的な意思決定はできない，と考える人もいる。だが一方で，人間は感情に影響されやすいがゆえに，非倫理的な行動を取ってしまうことがしばしばある。加えて自分自身や近しい者に甘くなる傾向があり，人間も倫理的な意思決定者として理想的とは言えない。

私たちは，適切な教えを受けたロボットは，たとえ自身の感情を持たずとも，人間の感情を把握し，それを考慮した上で公平な判断を下すことができると確信している。

経験から学ぶ

倫理的ルールをロボットに組み込むことが可能だとして，一体誰のルールを組み込めばいいのだろうか？　そもそも，現実の人間に普遍的に受け入れられる倫理の一般則などというものは，いまだ提唱されていない。

だが機械は一般に，ある特定の，制限された場面でのみ機能するよう作られている。そうした個別事例における行動について倫理的なパラメーターを決めるのは，何が倫理的で何がそうでないかを決める普遍的な一般則を打ち立てる（倫理学者たちはずっとこれを目指してきた）よりも，ずっと手が届きやすい。さらに，ロボットが活動するような多くの場面で，特定の状況に関する説明が与えられれば，倫理学者のほとんどは，倫理的に許容されること，されないことについて，ほぼ合意に達するだろう（人間が合意できないような状況においては，ロボットはいかなる自律的判断をも下すべきではないと私たちは考えている）。

機械が取るべき行動に関する原則の導出を目指して，これまで様々なアプローチが試みられてきた。その多くは，AI技術を用いるものだ。例えば2005年，北海道大学のジェプカ（Rafal Rzepka）と荒木健治（75ページの訳者紹介参照）は，「民主主義依存アルゴリズム」を提案した。ウェブサイトのマイニング技術を用いて，人間が過去にどのような行為を許容したかに関する情報を集め，これを統計解析して新たに生じた倫理的な問題に対する答えを見いだす，いわば「群衆の知恵」を活用するアイデアだ。

カナダのオンタリオ州にあるウィンザー大学のグアリーニ（Marcello Guarini）は，ニューラルネットワーク（人間の脳にヒントを得た，情報処理の最適な方法を徐々に学習していくアルゴリズム）を既存の事例を用いて“トレーニング”し，同様の事例で倫理的に望ましい決定を認識し，選択できるように学習させる手法を提唱した。

私たちは，倫理的な意思決定においては，倫理学者が「プリマ・ファキエの（自明な）義務」と呼ぶ複数の義務のバランスを考えることが重要と考えており，それを研究にも反映させた（プリマ・ファキエは「一見して」という意味のラテン語）。私たちは基本的にそうした義務を履行しようとするが，状況によっては，ある義務が別の義務を凌駕することもある。人間だって，普通は約束は守らないといけないが，些細な約束を破ることでずっと大きな損害を防げるならば破るべきだ。果たすべき義務が互いにぶつかった時，個々の状況に応じてどちらを優先するかを決めるのが倫理原則なのである。

ロボットにプログラムする倫理原則を作るためには，「機械学習」と呼ばれるAIの技術を用いた。私たちのアルゴリズムは，まず人間が「倫理的に正しい」と判断した意思決定の代表的な事例を，多数解析する。次いで帰納論理学を用いて，そこから倫理原則を抽出する。この「学習」はソフトウエアを設計する段階で行い，得られた倫理原則を実際のロボットのプログラム

に組み込む。

検証のため，最初のテストとして，ロボットが患者に薬を飲むように促し，それに応じなかった場合には監督者に通知する，というシナリオを考えた。このシナリオでは，ロボットは3つの義務のバランスを考える必要がある。

①患者に薬を服用させ，期待される効果をもたらす

②患者が薬を服用しないために起こり得る健康被害を防ぐ

③患者（正常な知的能力を持つ大人）の自己決定権を尊重する

特に患者の自己決定権の尊重は，医療倫理の分野において極めて優先度が高いと考えられている。もしロボットがあまり頻繁に薬の服用を促したり，あるいは薬を飲まなかったことをすぐに監督者に報告したりしたら，この義務に反することになる。

機械学習アルゴリズムは，私たちが示した特定の事例について情報を解析し，以下の倫理原則を作り出した。「医療ロボットは，ほかの方法では患者の健康被害を防げないと予測される場合，あるいは患者の福祉を促進する義務に著しく反する場合には，患者の自己決定権を侵しても，患者の意思決定に反する行動を取るべきである」。

ロボットへの実装

こうしてできた倫理原則を，私たちはフランスのアルデバラン・ロボティクス製のヒューマノイド，NAO（ナオ）にプログラムした。NAOは薬の服用を促すべき患者を見つけ，歩いて近寄り，薬を渡すことができる。自然言語（英語や日本語など人間の言葉）で話し，必要な時には監督者にメールで知らせることも可能だ。

まず最初に，NAOが監督者（一般には医師になるだろう）から，初期入力を与えられる。薬の服用時間，服用しなかった場合に起き得る最大の健康被害，その健康被害が起こるまでの時間，薬を服用した場合に期待される最大の効果，その効果がなくなるまでの時間などだ。これらの入力に基づいて，NAOは3つの義務のそれぞれについて実行度と違反度を計算し，そのレベルの時間変化に応じて異なる行動を取る。3つの義務の実行度と違反度が，倫理原則に照らして「（薬の服用を）促さないより，促した方がいい」と判断されるレベルに達したら，ロボットは初めて薬の服用を促す行動に出る。監督者に知らせるのは，薬を服用しないことで患者に健康被害が起き得るか，または相当の利益を失う場合のみだ。

実際の介護ロボット（Ethical Elder Care Robot，通称EthEl＝エセル）には，さらに幅広い行動ができるよう，より複雑な倫理原則を組み込む必要があるだろう。だが基本的なアプローチは同じだ。高齢者の介護施設を巡回する際，プログラムされた倫理原則に照らして，ある義務をほかの義務より優先すべき時を判断する。

EthElの典型的な1日は，こんな風になる。早朝，彼女は部屋の隅で，自らを電源に接続してじっと立っている。充電が完了すると，「自らお手伝いをする」義務の優先度が「自らを維持する」義務を上回り，室内の巡回を始める。居住者をひとりひとり訪ね，何か手伝うことはないか聞く。飲み物を取ってくる，ほかの居住者に伝言するなどだ。実行すべき作業の指示を受けると，その作業について各義務の実行度と違反度の初期値を割り当てる。

その時，居住者のひとりが苦痛を訴え，EthElに看護師を呼ぶよう頼んだとする。居住者の苦痛を放置するのは，「被害を防ぐ」義務に反し，その優先度は「自らお手伝いをする」義務を上回る。そこでEthElは看護師を探しに行き，居住者が助けを必要としていることを知らせる。一連の作業が終わると，「お手伝いをする」義務が再び最優先となり，EthElは巡回を再開する。

適切に訓練された機械なら大半の人間より倫理的に行動できる可能性すらある

時計が10時を告げる。居住者に薬の服用を促す時間だ。「お手伝いをする」義務が最優先となり，服用を促す仕事はその義務にかなう。そこでEthElは居住者を探し，薬を手渡す。それが終わると，居住者はテレビを見始める。ニュースか，クイズ番組か何かだろう。なすべき仕事はすべて片づき，電池はもうじき切れそうだ。「自己を維持する」義務の優先度が徐々に上がってきて，EthElは部屋の隅に戻って，自らを充電器に接続する――。

ロボットの倫理を扱うマシン・エシックスの研究は，まだ始まったばかりだ。私たちの研究結果も，ごく初期の段階ではあるが，機械が見いだした倫理原則によってロボットの行動を律し，人間により受け入れられるようにすることが可能だ，という期待を抱かせる。ロボットに倫理原則を教えることは重要である。知能ロボットが非倫理的なことをするかもしれない，との懸念が生じたら，人々は自律ロボット全体を拒否しかねない。AIそれ自体の未来がかかっているのだ。

マシン・エシックスの研究は，最終的に倫理学そのものに波及していく可能性を秘めているという点でも興味深い。AI研究は「現実世界」の観点で物事を見るので，倫理学者が抽象化して理論的に論じるより深く「人々の倫理的な行動とは何か」をとらえられるかもしれない。適切に訓練された機械なら，大半の人間より倫理的に行動できる可能性すらある。

機械は完全に中立な意思決定をなし得るが，人間は必ずしもそうではない。いつの日か，倫理的なロボットとの交流を通じて，より倫理的な行動とはどうあるべきか，私たちがロボットに教えられる日が来るかもしれない。 ■

訳者 荒木健治（あらき・けんじ）

R. ジェプカ（Rafal Rzepka）

荒木は北海道大学大学院情報科学研究科教授。人間の言葉を学習して成長するコンピューターの開発に取り組んでいる。ジェプカは同助教で，ロボットに人類の知恵と「標準的な意思」を与えることを目指している。

著者 Michael Anderson ／ Suzan Leigh Anderson

マイケル（左）はハートフォード大学コンピューターサイエンス学科准教授。コネティカット大学でPh.D.を取得し，長年，人工知能の研究に携わっている。スーザンはコネティカット大学哲学科名誉教授。カリフォルニア大学ロサンゼルス校でPh.D.取得，専門は応用哲学。2人は2005年に開催された第1回マシン・エシックス国際会議の主催者側として名を連ねた。近くマシン・エシックスに関する著書を出版する予定。

原題名

Robot Be Good（SCIENTIFIC AMERICAN October 2010）

もっと知るには…

IEEE INTELLIGENT SYSTEMS. Special issue on machine ethics. July/August 2006.

「ホームロボット時代の夜明け」，B. ゲイツ，日経サイエンス2007年4月号

Machine Ethics: Creating an Ethical Intelligent Agent. Michael Anderson and Susan Leigh Anderson in AI Magazine, Vol. 28, No. 4, pages 15–26; Winter 2007.

MORAL MACHINES: TEACHING ROBOTS RIGHT FROM WRONG. Colin Allen and Wendell Wallach. Oxford University Press, 2008.

「ロボットが変える戦争」，P. W. シンガー，日経サイエンス2010年10月号