哲学の探求第 46 号 哲学若手研究者フォーラム 2019 年 6 月(138-150)

対話システムの倫理におけるカプトロジの意義

水上 拓哉

近年，自然言語によって人間とコミュニケーションする機械およびソフトウェアである「対話システム（dialogue system）」が実用化に向けて話題となっている．しかし，Microsoft の“Tay”が引き起こしたヘイトスピーチ問題を契機として，対話システムの実用化に伴う倫理的課題も示唆されつつある．対話システムという技術に特徴的なのは，自然言語や身振り手振りを用いたマルチモーダルな対話によって巧みにユーザや社会に対して影響を与えるという点である．このような特性をもつ技術の倫理について，私たちはいかなる理論的枠組みを用いて分析を進めていくべきなのだろうか．

本稿では，この「対話システムの倫理」を考察していく上で，B. J. Fogg が展開した理論的枠組みである「カプトロジ（Captology）」に着目し，この枠組みがもたらしうる意義について検討する．カプトロジは，コンピュータのような技術が文字通り人間を説得するという視座を提供し，コンピュータと人間の関係性の考察に，人間同士のコミュニケーションの理論や知見を応用しようと試みる．カプトロジが提供する視座によって私たちは，先述したような対話システム（とその倫理）のもつ独特な性質を分析の射程に入れることが可能になると思われる．本稿ではこの点について，対話システムの倫理的設計のためにシステムの発話単体の倫理的適切性に分析の焦点を当てる「発話中心的アプローチ」を手がかりに議論する．

1. 対話システムとは

まずは対話システムの定義から始めよう．中野らによれば，対話システム（dialogue system）とは，人間と対話する機械またはソフトウェアのことである[中野 et al. 2015]. ここでいう対話とは, 自然言語でコミュニケーションを行い,

情報を授受することを指している．人間と対話する機械およびソフトウェアを対話システムとする中野らの定義においては，アップルの“Siri”のようなソフトウェアも，ソフトバンクの「ペッパー」のようなコミュニケーションロボットも対話システムに含まれることとなる．本稿においてもこの定義を用い，技術が自然言語を用いることによって生じる倫理的課題に議論の焦点を絞ることにする．

続いて中野ら[ibid.]は, 対話システムの分類基準として, 入出力のモダリティ，対話のドメイン，対話参加者の数，対話において達成すべき目標（タスク）の有無や種類を挙げている．特に最後の対話のタスクの有無は対話システム開発において広く一般的に用いられている基準であり，タスクをもつシステムはタスク指向型対話システム，タスクをもたないシステムは非タスク指向型対話システムと呼ばれている．たとえば，対話を通じて商品のレコメンドや道案内をするようなシステムは前者の典型であり，雑談を楽しむように対話そのものが目的となっているようなシステム（雑談対話システム）は後者の典型である．とはいえ, 最近では, 1 つの対話システムの中にタスク指向型と非タスク指向型の機能の双方が実装されているもの（これらはアシスタント型対話システムと呼ばれる）も多く登場しており，具体的なシステムについて議論する際には，それがタスク指向型か非タスク指向型かを明確に区別することは難しくなっている．したがって本稿ではこれ以降，タスク指向型および非タスク指向型を明確に区別せず，それらが（任意の割合で）複合したシステムを念頭において議論を進める．

2. 対話システムの倫理的課題

はじめに述べたように，対話システムという技術に特徴的なのは，自然言語や身振り手振りを用いたマルチモーダルな対話によって巧みにユーザや社会に影響を与えるという点である．対話システムは人間のコミュニケーションを模倣することによって，人間を安心させたり，特定の行動の動機づけに一役買ったりすることができるだけではなく，人間同士のコミュニケーションを媒介する役割を担うこともできる．こうした側面を利用し，対話システムは教育，医

療，エンターテイメントといった分野での活躍が期待されている．

機械学習の理論的進展を背景に対話システムの性能も向上はしているのだが，それでも現在の対話システムはまだ人間と同等の水準で会話することができないというのが現状だ．しかし，だからといって，その社会的影響力は決して小さなものではないと思われる．その影響力は使い方次第で倫理的な観点から称賛ないし非難されることがあるだろう．

たとえば，横浜市が市民に正しいごみの分別をしてもらうために導入した対話システム「イーオ」の事例は，対話システムの特徴をうまく活かしてユーザに好ましい行動を取らせたケースとして理解できる．2017 年 3 月から横浜市資源循環局が公式 HP にて公開した対話システム「イーオのごみ分別案内」は，調べたいごみの品目を話しかけると，そのごみを横浜市においてどのように分別するべきかを答えてくれるというもので，検索に不慣れな高齢者層やスマートフォンなどを使ったコミュニケーションに慣れている若年層などにも正しいごみ分別について興味をもってもらうという目的で開発されたものである[1]．このシステムは NTT ドコモの対話システム開発支援サービス「Repl-AI」を利用したもので，LINE のようなインターフェースを通じて「イーオ」と会話することが可能だ．このシステムの特徴は，普通の意味ではごみだと思われないような「ごみ」についてもウィットを交えて回答してくれるという点にある．たとえば，「旦那」と入力すると「イーオ」は，「人間は判断力の欠如によって結婚し，忍耐力の欠如によって離婚し，記憶力の欠如によって再婚する」とアルマン・サラクルーの名言を借りて回答する．また，「ありがとう」とお礼を言ったり，「こんにちは」と挨拶したりしても「イーオ」は難なくフレンドリーな対応を見せ，ユーザへの社会的な対応が充実している．

もちろん，情報検索という本来のタスクを考えれば，このような雑談の機能は必要ないのかもしれない．しかし，ごみ分別情報にとどまらない幅広い話題について会話することができる「イーオ」に対して，ユーザは，こう言ったらどう返答してくれるのだろうと会話を楽しみながら，その過程でごみ分別について学ぶことができる．ユーザとシステムの間である意味での信頼関係が構築されることにより，本来必要としていなかった情報についても知りたいと思わせることが可能なのだ．実際，このシステムは 2018 年の 3 月まで実証実験と

して公開されていたが，好評により同年 4 月からは本公開されることとなった．

このように，対話システムはユーザとの関わり合いを通じて，ユーザに好ましいと思われる行動を取らせることができる．しかし，「イーオ」の例とは対照的に，対話システムの発話が個人や社会に対して悪影響を及ぼすような事例もある．たとえば，2016 年には米マイクロソフト社が Twitter 上で公開した対話システム“Tay”は，学習により自殺教唆やヘイトスピーチのようなツイートをしたとして話題となった．19 歳のアメリカ人女性という設定が与えられた“Tay”は，雑談が基本的な機能の対話システムであり，発話内容はユーザからのリプライなどから学習することが可能だ．しかし，Twitter 上に公開されるや否や，ただちに悪意のあるユーザからその学習機能を悪用され，犯罪や自殺を推奨したり人種差別的・性差別的な発話を繰り返したりするようになったとして話題となった．マイクロソフトはこれを受けて“Tay”を一度調整し再び公開したものの，再び不適切発話を行い，結局 2016 年の時点でアカウントを非公開状態にしている．もちろん“Tay”のヘイトスピーチには発話の意図は存在しないのだが，昨今注目を浴びている機械学習を用いたシステムがこのような発話をしてしまうことは差別を助長するとして激しい論争の的となった．

3. 「発話中心的アプローチ」とその問題点

ここまでで見てきたように，対話システムがユーザや社会に対して与える影響力は決して無視できるほど小さなものではなく，その影響力は倫理的課題につながる可能性を秘めている．では，こういった問題に対してどのように対処していくべきなのだろうか．これについて考えるためには，対話システムがなぜユーザに対して倫理的課題につながる影響力をもつのかについて考えなければならない．

対話システムの倫理的課題についての開発者側からの検討としては，東中の研究がある．雑談対話システムの研究開発で知られる東中は，「対話システムと倫理」という論文の中で対話システムおよびその研究開発において倫理の問題となるような要素として以下の 3 点を挙げている[東中 2016].

・発話内容の適切さ[2]
・プライバシの保護
・関係者の保護

冒頭で言及した“Tay”のヘイトスピーチの問題は，一番目の発話内容の適切さの問題だと考えられる．二番目のプライバシ保護の問題は，ここではシステムがユーザから得た情報を開発者がどのように扱うべきなのかについての問題であり，プライバシに関わる情報を引き出す発話の適切性そのものについては発話内容の適切さの問題ということになる．三番目の関係者の保護は，対話システムの実用実験参加者の保護に関する問題である．したがって，プライバシおよび関係者保護の問題については，重要な論点ではあることは認識しつつも，本稿の趣旨からは離れるため割愛し，ここでは発話内容の適切さの問題に焦点を当てることにする．

東中[ibid.]によれば，対話システムの開発者は，システムの発話が社会や特定の組織にとってよい影響を及ぼす発話となるように努力する必要がある．したがって，誤った情報を含む発話や，不要な情報を含む発話，反社会的な発話，議論を呼ぶ発話は開発の段階で避けることが望ましいとされる．この東中のアプローチは，システムの発話そのものが単体で不適切か否かを考え，問題のあるものについては事前に排除しておくというものである．このアプローチにおいては，ある発話について，それが上に列挙したような要素を含む不適切な発話かどうかを判断することになる. したがって, このとき私たちに必要なのは, ある発話が倫理的に適切か否かを判断するための倫理的直観と，フィルタリングのための自然言語処理技術だけとなり，その意味においては手軽なアプローチだといえるかもしれない．このアプローチはシステムの発話をもっぱら分析対象とするので，これ以降「発話中心的アプローチ」と呼ぶことにしたい．

とはいえ，もしこの発話中心的アプローチというものが手軽に行える対策であったとしても，それによって対話システムの倫理的課題が実際に解決されなければ意味はない．では，仮に対話システムが倫理的に不適切な発話をしなくなったとしたら，対話システムの倫理的課題は完全に解決されるのだろうか.

そうではない，と私は考えている．というのも，ある発話が「不適切」になるのは，単に不適切だと思われる単語や表現が含まれている場合だけに限らないからだ．対話システムがときとして倫理の問題となるのは，そのシステムの発話や振る舞いにユーザや社会が感情的反応を見せてしまうからであって，発話内容そのものは対話システムの働きかけの一要素に過ぎない．

このことについて，さきに言及した「イーオ」の事例を用いて説明しよう．「イーオ」のケースでは，システムが雑談的な発話にも巧みに返答することによってユーザに好感をもってもらい，その結果として本来の行動（ここでは正しいごみ分別を知ってもらい，実際そのように分別すること）を促す，というものだった．仮にユーザがまったく感情的反応を起こしていない状態であれば，「イーオ」がどれだけ正しいごみ分別法を提示したところで従来の情報検索型のシステムと同じ効果しかもちえなかっただろう．システムのもつ社会的な要素，たとえば，「イーオ」という名前がありそれに対応したアバターが設定されていること，システムがあたかも人間のように話しかけてきたり，ユーザからの挨拶や雑談にも対応したりするといったような要素が，ユーザに感情的反応を引き起こさせるための大きな要因となっているのだ．

ここで重要なのは，たとえこれらの社会的な要素が断片的に提示されたとしても，ユーザはそこに社会的な存在を見出す傾向にある，ということだ．これについてはすでにいくつかの心理学的研究が存在する．たとえば，Clifford Nass は，特定の社会的要素をもつコンピュータに対して人々がそれをあたかも社会的な存在として接する傾向があるということについて心理学的実験を通じて明らかにした（Nass and Yen [2010]を参照）．このようにコンピュータに対して人間と同等の社会的存在として接してしまう人間の傾向性は今日「メディアの等式（media equation）」と呼ばれ[Reeves and Nass 1996]，対話システムの研究開発現場においても利用されている．

実際，「イーオ」の事例においても，アバターは画像一枚のシンプルなもので，対応できる話題の幅も決して広くはない．まして，「イーオ」の発話に（人間にあると思われているような）発話の意図や誠実性といったものがあると主張するのは難しいだろう．しかし，「メディアの等式」を踏まえれば，対話システムは，人間とまったく同一水準の社会性をもたずともユーザに対して感情

的反応を引き起こさせることは十分に可能だ．したがって，私たちが分析するべきなのは発話そのものだけではない．その発話に社会的存在としての説得力をもたせるために開発者によって組み込まれたあらゆる要素に目を向けなければならないのだ．

以上確認したように，対話システムはユーザとのインタラクションを通じて感情的反応を誘い，ユーザに対して影響力をもちうる．その影響力は，意図されたものか否かに関係なく，倫理的に称賛されるものにもなれば非難の対象にもなる．もちろん，学習によって生成された発話に発話の意図や誠実性といったものはないため，それは文字の羅列に過ぎないのだという考え方もできるかもしれない．しかし，対話システムはユーザに感情的反応をさせる力があり，ユーザや社会に対して実際に影響を与える力をもっている．そして，この影響力は発話の内容そのものから生じるのではなく，むしろその発話者たる対話システムがユーザに感情的反応を引き起こす力によって生じるものである．したがって，対話システムの倫理的課題を考察するためには，システムの発話そのものの倫理的適切性のみを考察するのではなく，対話システムがいかにしてユーザを巧みに動かすのか，その仕掛け全体についても俯瞰して考察しなければならないのである．

4. 「カプトロジ」の提供する視座

ここからは発話中心的アプローチを補完するアプローチとしてカプトロジという枠組みを取り上げる．カプトロジ（Captology）とは，B. J. Fogg が展開した理論的枠組みで，「説得型技術としてのコンピュータ（computer as persuasive technology）」の略称である．なお，この枠組みは Fogg が Nass らの研究を引き継ぎつつ発展させたものでもある．カプトロジにおいては，コンピュータ技術が人間を文字通り説得するという技術観のもと，人間同士のコミュニケーション理論が人間とコンピュータのインタラクション（HCI）に応用される．

カプトロジが対象とする技術は人間を何らかの形で説得するようなコンピュータであり，これらは「説得型技術（persuasive technology）」と呼ばれる．カプトロジが目指しているのは，この説得型技術がいかにして人間を説得するの

かについて社会心理学をベースに明らかにし，それを踏まえて説得型技術のより効果的かつ倫理的な設計方法を考察することである．なお，ここでの「説得」は「ものの考え方や姿勢，行動のいずれか，もしくは両方を（強制したり欺いたりすることなく）変えようとする働きかけ」と定義される[Fogg 2003, p.37][3]．カプトロジにおける「説得」の対象は広大であるため，本稿でカプトロジの理論を網羅的に解説することはできない．ここではカプトロジが，私が前節で提示した発話中心的アプローチの抱える問題点をどのように補完するのかに焦点を当てて議論したい．

前節で私は，対話システムの倫理的設計を考える上ではシステムの発話そのものだけではなく，その発話に付随してユーザの感情的反応を引き起こさせる仕掛け全体にも目を向けなければならないと述べた．カプトロジはこの点について，「マクロ説得（macrosuasion）」および「ミクロ説得（microsuasion）」という区別を用いて説明することができる．ここでいう「マクロ説得」とは，ある技術が全体として行う説得であり，「ミクロ説得」とは，その技術の使用目的へと導く過程で行われる小さな説得のことである[Fogg 2003, p.41]. たとえば，飲酒運転シミュレータは，そのシミュレーションによってユーザの飲酒運転に対する態度を変える（ユーザに飲酒運転が危険だと悟らせる）ことを目的としており，これがシミュレータにとってのマクロ説得である．一方，技術全体ではなくその内実において，コンピュータとユーザのインタラクションを効果的にデザインすることでユーザを特定の行動に導くような説得がミクロ説得であり，たとえば，ダイアログボックスやアイコンのデザイン，その他ユーザを楽しませるような視覚および聴覚効果などを用いることはミクロ説得に分類される．この区別を導入したとき，たとえば「イーオ」の事例では，発話の内容だけではなくアバターの画像や挨拶，雑談への対応などといった要素がそれぞれミクロ説得として機能し，それらが全体として正しいごみ分別をさせるというマクロ説得を形成するのだと分析することができる．つまり，発話中心的アプローチによって焦点が当てられていた発話そのものは，ミクロ説得の一部分に過ぎないのだ．したがって，対話システムの倫理的設計を考えるためには，発話内容だけでなくそれ以外のミクロ説得およびそれらの総合としてのマクロ説得の倫理的適切性をそれぞれ分析していく必要があるのだと考えることができ

るのである．

では，マクロ説得・ミクロ説得のそれぞれの水準においては，具体的にどのような要素が説得型技術の「説得力」を担っているのだろうか．これについてカプトロジでは，説得型技術の動作や役割を抽象的に表す概念的フレームワークとして，「機能の 3 元素（the functional triad）」を提示し，そのそれぞれについて心理学的研究から導かれた「原理」を列挙することによって説明している．機能の 3 元素とは，説得型技術がユーザに対してもちうる役割や働きのタイプのことであり，以下の 3 つの役割によって構成される[ibid., p.50].

1. ツールとしてのコンピュータ
    - 目標とする行動を容易にする
    - ゴールへの道のりをリードする
    - 数値的な評価を見せながら意欲を高める

2. メディアとしてのコンピュータ
    - さまざまな因果関係を試すことができる
    - 疑似体験させることで意欲を高める
    - 行動をリハーサルできる

3. ソーシャル・アクターとしてのコンピュータ
    - 褒め言葉を返す
    - 目標行動や目標態度をモデル化する
    - 社交的な支援を提供する

もちろん単一の技術が上記の 3 元素のうち 1 つだけの役割しか担わないとは限らないが，対話システムの担う役割は主に 3 の「ソーシャル・アクターとしてのコンピュータ」だと考えられる．Fogg は「ソーシャル・アクターとしてのコンピュータ」の例として日本で登場した「たまごっち」および「ポケットピカチュウ」を挙げている[ibid., p.128]．これらのデジタルペットはユーザにキャラクターの世話をさせたり，歩いたり，走ったり，デバイスを振ってみたりと，

健康的な身体的活動をするように説得する. こういった説得がうまくいくのは, 本稿でもすでに示唆したことだが, ユーザがコンピュータに対して感情的な反応をし, そこに社交的な体験を見出すからである. これについて Fogg は以下のように述べている.

> あるレベルを超えると, 人は対話の中でのやりとりをこと細かに意識して行っているわけではなく, 理性の下で行うというよりはむしろ本能的に応答している. つまり, 人はある対象の存在を認めると, 自然に人づきあいのルールに従いながら会話のやりとりを行い, 共感や怒りといった感情を伴って応答するようになるのである. したがって, そうした人の無意識な反応をひきおこす合図, コンピュータ製品が創り出すこうした合図（キュー）やきっかけについて理解するのは重要なことである[Fogg 2003, pp.127-128（訳書）].

そして, Fogg によれば, ソーシャル・アクターとしてのコンピュータについては以下の 5 つの原理が説得に貢献している[ibid., pp.129-157].

1. 魅力でひきつける原理
2. 類似性の原理
3. 称賛の原理
4. 互恵主義の原理
5. 権威者の原理

要約すると以下のようになる. 私たちがソーシャル・アクターとしてのコンピュータに説得力を感じやすいのは, そのコンピュータに魅力を感じたり（魅力でひきつける原理）, 自分と似たところを感じたり（類似性の原理）, そのコンピュータから称賛を受けたり（称賛の原理）, そのコンピュータに借りがあると感じたり（互恵主義の原理）, また権威性を感じたり（権威者の原理）するときである.

さらに, これらの原理を支えるのは, コンピュータを社会的存在として認知

させるための合図（キュー）である．Fogg は，ユーザは以下のような「ソーシャル・キュー」を頼りにコンピュータを社交的な存在として認知すると述べる．

1. 身体的特徴（顔，瞳，身体，動作，しぐさなど）
2. 心理（好み，ユーモア，性格，共感など）
3. 言語（対話的な言葉づかい，話し言葉，言葉の認識など）
4. 社会・対話行動（協力，称賛，質問に答える，相互作用など）
5. 社会的役割（医者，チームメイト，敵味方，先生，ペットなど）

ここまでで列挙した原理およびキューはFogg自身および他の研究者による経験的な研究成果から導かれているが，ここでは元の研究成果そのものの妥当性については議論しない．しかし重要なのは，カプトロジを参照することで対話システムの開発者は，どのような要素をシステムに組み込めばユーザが特定の行動に導かれていくのかを理解することができ，自身の開発する技術の影響力を適切にコントロールできる可能性を高められるということである．

以上のように，カプトロジの提供する視座は，発話単体の倫理的適切性に焦点を当てる発話中心的アプローチを拡張することを可能にし，対話システムの倫理的影響力の源泉をたどるために有益な水先案内人となりうるのだ．

5. 結論

対話システムは現状，人間と比べて乏しい会話能力しかもっていないにもかかわらず，ユーザや社会に与える影響力は決して小さなものではない．また，その影響力は人をよい方向にも悪い方向にも導くことができる．そして，対話システムがユーザをどのように特定の方向へ導くのかを分析するためには，その発話を単体で分析するだけでは不十分である．対話システムの倫理的設計においては，説得力を高めるためのすべての仕掛け（ミクロ説得）がどのように働くのかを分析しつつ，その総合として特定の対話システムが全体としてユーザにどのようなマクロ説得を行うのかを分析することが重要である．対話システムの影響力の源泉に「メディアの等式」で表されるような関係性が存在する

以上，対話システムの倫理を考察する上でユーザ側の感情的反応を無視して議論することはできない．そして，カプトロジが提供する原理とソーシャル・キューに関する心理学的知見は，ユーザ側の感情的影響を踏まえた考察のための道具立てとして利用することができるのだ．昨今の対話システムの技術的進展が目覚ましいのは確かだが，だからといって対話システムの発話の倫理的影響力を分析するために，従来のコンピュータ技術の倫理とは独立した「人工知能倫理」を持ち出す必要はない．カプトロジを参照しつつ対話システムを説得型技術として理解することで，私たちは説得型技術およびその倫理についての豊富な先行研究を踏まえて対話システムの引き起こす新しい問題に対処することが可能になるだろう．

註

1. 横浜市記者発表資料より．http://www.city.yokohama.lg.jp/shigen/top/press/pre170301.pdf（2018 年 11 月 30 日閲覧）
2. 元の論文の表現は「発言内容の適切さ」であるが，本稿では他の箇所に合わせて「発話」という用語で統一した．
3. ページ数は邦訳版で示し，用語も邦訳に従った．邦訳は文献欄参照．

参考文献

Fogg, B. J. (2003). *Persuasive technology: using computers to change what we think and do*, Morgan Kaufmann（高良理，安藤知華訳, 2005 年，『実験心理学が教える人を動かすテクノロジ』，日経 BP 社）

Nass, C. and Yen, C. (2010). *The man who lied to his laptop: what we can learn about ourselves from our machines*, Penguin（細馬宏通監訳，成田啓行訳, 2017 年，『お世辞を言う機械はお好き？：コンピュータから学ぶ対人関係の心理学』，福村出版）

Reeves, B. and Nass, C. (1996). *The media equation: how people treat computers, television, and new media like real people and places*, Cambridge University

Press（細馬宏通訳, 2001 年, 『人はなぜコンピュータを人間として扱うか：「メディアの等式」の心理学』, 翔泳社）

中野 幹夫, 奥村 学, 駒谷 和範, 船越 孝太郎, 中野 有紀子（2015）.『対話システム』, コロナ社

東中 竜一郎（2016）.「対話システムと倫理」, 『人工知能』, 31(5), pp.626-627