権利と責任を有する AI は社会に受け入れられるか

Public Acceptance of AI with Rights and Responsibilities

岡 夏樹　畑中 佑介　田中 一晶

Oka Natsuki, Hatanaka Yusuke, and Tanaka Kazuaki

京都工芸繊維大学 情報工学・人間科学系
Faculty of Information and Human Sciences, Kyoto Institute of Technology

あらまし：医療、自動運転、金融など、人間の生命・身体・財産の安全に関わる分野への AI 技術の浸透のための一つの方策として、権利と責任を有する AI を開発することを提案する。責任としては、行為責任と説明責任の両方を持たせる。行為責任とはたとえば自分の行為によって生じた損害を賠償する責任であり、説明責任とはなぜそのような事態を招いたか、再発防止のためにどうするかを説明する責任である。我々は、軽微な権利と責任を持つ動画推薦 AI を試作中であり、日常生活での受容性を検討することを計画している。

はじめに

医療、自動運転、金融など、人間の生命・身体・財産の安全に関わる分野への AI 技術の浸透が期待されている。これらの分野では、たとえば自動運転車の対人事故などのように、AI は動作の結果、致命的な影響を与える可能性があるため、責任の所在をどう考えるか等の議論が昨今盛んに行われるようになった。

本論文では、AI の開発者・開発組織や販売者・販売組織、AI の所有者・所有組織等ではなく AI 自体が責任をとることが可能か、また、そのような AI を社会が受容する可能性について議論する。

責任としては、行為責任(responsibility)と説明責任(accountability)の両方を持たせることを提案する。

説明責任は、信頼できる AI を論じる中で、これまでも必要なことの一つとして取り上げられてきた（たとえば[1]）。そこでは、AI を開発する人や組織が果たすべき責任として種々の説明をする必要があるとされているが、本論文では、AI 自体に説明責任を持たせることを提案する点が異なる。

以下では、行為責任と説明責任、XAI と説明責任、自由意志と責任について順次考察し、さらに、社会的受容性の評価に向けて我々が試作している、軽微な権利と責任を持つ動画推薦 AI を紹介し、その社会的受容性を検討するための実験計画について述べる。

行為責任と説明責任

本論文では、行為責任(responsibility)と説明責任(accountability)の 2 つを区別して考察する。行為責任とはたとえば自分の行為によって生じた損害を賠償する責任であり、説明責任とはなぜそのような事態を招いたか、再発防止のためにどうするかを説明する責任である。

行為責任

民法７０９条では、「故意又は過失によって他人の権利又は法律上保護される利益を侵害した者は、これによって生じた損害を賠償する責任を負う。」 と規定している。これは行為責任の一つを規定していると考えられる。

一般的には、行為責任のとり方としては、謝罪する／刑罰を受ける／権利や地位を手放す／賠償する等がありえる。本論文の後半で提案する動画推薦 AI では、このうちの「謝罪する」、および、「権利や地位を手放す」の一種としての“動画推薦を控える”ことを実装する。

「刑罰を受ける」ことが AI の責任のとり方としてあり得るかについての議論も興味深いが、本論文ではこれ以上は触れない。

AI が「賠償する」ことは、AI 向けの賠償責任保険(liability insurance)の導入により実現できる可能性があるだろう。ただし、保険料の支払いを AI 自体がしない場合に AI が賠償したと言うことは適切でないように思う。保険料を AI が支払うためには、（AI の所有者でなく）AI 自体がお金を稼ぐ仕組みが必要となる。

説明責任

[1]によると、説明責任(accountability)に関して考慮すべき観点として以下の４つが挙げられている：

- 監査担当者による、アルゴリズムやデータや設計プロセスに対する監査可能性(auditability)
- 負の影響の最小化と報告(minimization and reporting negative impacts)
- トレードオフ(trade-offs)
- 補償(redress)

これらは、AI を開発する人や組織が果たすべき種々の説明として列挙されているが、本論文では、AI 自体に説明責任を持たせることを考える。このため、説明を実行するのは AI 自体である。本論文の後半で提案する動画推薦 AI では、推薦・再生した動画が不適切であった場合、AI はなぜそのような事態を招いたか、再発防止のためにどうするかを自ら説明する。なお、上記の 4 点目「補償」は、本論文では前述の行為責任の 1 つとして扱う。

XAI と説明責任

深層学習のような subsymbolic レベルの学習アルゴリズムの急速な性能向上により、eXplainable AI (XAI)の必要性が注目されている（概説としてたとえば[2]）。

説明法は大きくは次の 2 つに分類される[2]。学習アルゴリズムが元々透明性(transparency)すなわち理解可能性を持っている場合と、後付けの説明可能性(post-hoc explainability)である。

本論文の後半で説明する、社会的受容性を評価するための動画推薦 AI では、次のように説明可能性を確保した。動画に付与されたタグとユーザの好みの一致度が高いものを推薦するという、元々透明性を有する単純なアルゴリズムを採用した。

一般に、モデルの解釈可能性(interpretability)と性能(performance)はトレードオフの関係にあることが多い。本研究の目的は、責任を有する AI の社会的受容性を検討することであるため、推薦性能は重視せず、解釈可能な、したがって説明可能な推薦アルゴリズムを採用した。このことにより、「推薦が不適切であった場合に説明責任を果たすことの効果」の検証を可能とした。

XAI は種々の目的を持つ[2]が、本研究との関係で重要であるのは、ユーザからの信頼性(trustworthiness)である。本論文の後半では、説明責任を果たすことによりユーザからの信頼を得ることができるかを調べる実験の構想を述べる。

自由意志と責任

多くの人々は、人間は自由意志を持つから（行為を選択できるから）責任が生じると考えており、刑法も自由意志の存在を前提に組み立てられている。

“故意”とは自己の行為が一定の結果を生ずることを認識しながらその行為をした心理状態をいう。また、“過失”は注意を欠いて結果の発生を予見しなかったことや、回避をする行為を怠ったことをいう。故意・過失は責任の要件となるが、故意または過失があっても、責任能力がない場合（たとえば心神喪失や 14 歳未満）は、責任が生じないとされる。

一方、脳科学の発展により、自由意志の存在は否定される方向にある。リベットら[3]は、行動しようと意図したと感じた時刻を、時計様の表示の中で動く点の位置として報告させる方法を考案し、意図の自覚に数百ミリ秒先立ち、準備電位(readiness potential)が立ち上がることを発見した。これは体を動かそうと意識する前に脳活動が始まっていることを示しており、自由意志を否定する証拠と解釈されることが多い。

このように実際は自由意志が存在しないにも関わらず、自由意志を持っているという感覚があることについては、人の情報処理が説明の後付け(postdiction)をしてしまう性質を持っている[4-7]ため、自由意志で行動したと後付けで説明してそう感じてしまう、とする説がある[4, 6]。

将来、脳科学等の知見が一般に共有されるようになるに伴い、自由意志の存在はやがて否定され、それに伴い、責任の概念も今後変化していくことが予想されるが、現状では、自由意志を持つから責任がある［自由意志→責任］という考え方が共有されている。

［自由意志→責任］という考えが共有されている状態を出発点として、今後、AI に責任能力があるという見方が社会的に受容されるためには、次の 2 つの可能性があると考える：

1. AI に自由意志があるという見方が社会で共有されるようになり、AI には自由意志があるから責任能力があると見なされるようになる。
2. 人は自由意志を持たないが、持っていると錯覚しているという見方[4]が社会で共有されるようになる。同様に AI も自由意志は持たないが AI 自身は持っていると錯覚しているという見方[6]が社会で共有されるようになる。人における［自由意志→責任］という考え方は、［自由意志の錯覚→責任］に置き換えられることで、責任という概念が維持されるが、同様に、AI においても［自由意志の錯覚→責任］という枠

組みを共用して、人と同様に責任能力を持つと見なされるようになる。

社会的受容性の評価に向けて

我々は軽微な責任が発生する場面として動画推薦に注目した。ユーザが自分の趣味に費やせる時間は有限であり、動画共有サービスにおける推薦システムでは、システムが推薦を失敗したときユーザの不利益となり、システムに軽微な責任が発生すると考えた。今回の試作では、動画共有サービスの一つであるニコニコ動画を利用した。ニコニコ動画では、動画を投稿したり、投稿された動画に対してコメントやタグと呼ばれる動画の特徴を表す文字列をつけたりすることができる。

Siri のような対話エージェントに「おすすめの動画は？」と質問することはできるが、推薦した動画がユーザの好みと違っていた場合でも責任をとるようには設計されていない。これに対して本研究の評価用システムでは、推薦が不適切であった場合に、行為責任や説明責任を取れるエージェントを設計する。

推薦システムは、販売サイト、旅行案内サイト、動画配信サイト等、様々なところで用いられている。動画を対象とした推薦システムの研究は盛んに行われている（たとえば[8, 9]）が、今回の試作システムにおいては、推薦の質を高めることが目的ではないため、動画に付与されたタグに基づく単純な推薦方法を採用した。

評価用動画推薦システムの試作

試作システムの動作の流れを図 1 に示す。以下、流れに沿って順に説明していく。

なお、試作システムでは、検索した単語に関連する動画リスト、動画に関連するオススメ動画リスト、各動画の動画 ID・動画名・タグを使用する。これらの値を取得するために niconico コンテンツ検索 API、getrelationAPI、getthumbinfoAPI を用いる。

動画選択

検索画面（図 2）や視聴画面（図 3）に表示される動画リストから上位 5 個の動画を選び、その中から、記憶したユーザの好みと各動画のタグの一致した数が一番多い動画を選択する。選択した動画は、推薦または再生に使用する。

ユーザの好みの取得

動画選択時に用いるユーザの好みは、動画視聴完了時にその動画が面白かったかどうか質問しユーザが面白いと思った動画のタグをユーザの好みとして記憶したものを使用する。

たとえば、好みの動画が動画 A(タグ：a, b, c)、動画 B(タグ：a, d)だとすると、ユーザの好みは、(a, b, c, a, d)となり、動画 C(タグ：a, c)との一致数は 3 となる。このようにして、ユーザの好みは、タグの出現回数を含んだ表現となっている。

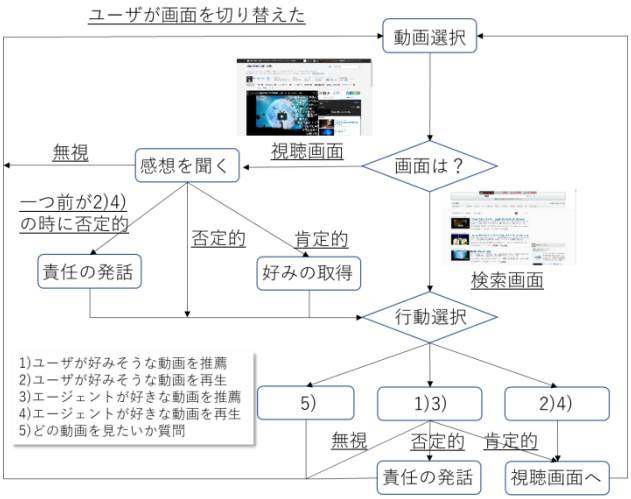

図 1：評価用動画推薦システムの動作の流れ

図 2：評価用動画推薦システムの検索画面

図 3：評価用動画推薦システムの視聴画面

行動選択

試作システムでは、エージェントは次の 5 つの行動から 1 つを選択し、それに応じた発話・行動をする。1）ユーザが好みそうな動画を推薦、2）ユーザが好みそうな動画を再生、3）エージェントが好きな動画を推薦、4）エージェントが好きな動画を再生、5）どの動画を見たいか質問。ここで、再生というのは、推薦するだけでは責任があまり発生しないと考え、強制的に再生する行動を取り入れた。

エージェントとのインタラクションは対話的なものとするが、本研究では自然な対話を実現することは目標ではないため、ユーザからの発話は、提示された選択肢から選択する方式とした。これは、音声認識誤りや、テキスト入力の煩わしさを避け、また、対話の流れを想定した範囲に収めることで、対話機能の開発を容易にするためである。なお、ユーザは発話を選択せずにエージェントの発話を無視することもできるようにした。

エージェントが上述の 5 つの行動のどれを選択するかは、ユーザの行動や反応を評価として用いて強化学習させる。学習方法は、Q学習（式(1)）を用い、ソフトマックス法により行動を選択する。下記式のsは状態、aは行動、$Q(s,a)$は行動価値、αは学習率、rは報酬、γは割引率である。

$$Q(s_{t+1},a_{t+1}) \leftarrow Q(s_t,a_t) + \alpha\left(r_{t+1} - \gamma \max_a Q(s_{t+1},\mathrm{a}) - Q(s_t,a_t)\right) \quad (1)$$

ここで、状態sとしては動画推薦後、動画再生後、どの動画を見たいかの質問後の 3 状態を使うことを検討している。報酬rは、動画推薦行動に対して肯定的な発話応答が選択されたときや、推薦した動画が視聴されたときに正の報酬を、それ以外のときに負の報酬を与える。また、動画再生行動に対しては再生した動画を面白いと思ったときには正の報酬を、そうでないときには負の報酬を与える。行動 5（どの動画を見たいかの質問）に対する報酬は即時報酬としては「エージェントに選んで欲しい」というボタンを押したかどうか、遅延報酬としては推薦・再生動画に対する評価を使うことを計画している。

責任に関する発話

責任に関する発話は、行為責任、説明責任のそれぞれをとるための発話に分けられる。

行為責任をとる発話としては、推薦・再生が不首尾に終わったときに、謝罪すること【行為責任（謝罪)】、および、責任をとって動画推薦・再生を次回自粛すること【行為責任（自粛)】の 2 種類の片方、または、両方を想定している。

説明責任をとる発話としては、動画選択理由となったタグを最大 3 つ（多数あった場合はそのうちの 3 つ、3 つ未満であった場合はその数だけ）選び、それらのタグを含むので選択したと説明する（動画選択理由となったタグが存在しなかったときは、その旨を説明する）こと【説明責任（理由)】、および、その説明に加えて今後はそれらのタグを選択理由に用いないようにすると説明する（タグが存在せず失敗した場合は今後好みの学習をしっかり進めることにより選択が成功するよう努めると説明する）こと【説明責任（理由と対策)】の 2 種類のいずれかを想定している。

責任に関する発話をするタイミングは、再生した動画が面白くなかったとき（再生失敗時)、および、推薦した動画が見られなかったとき（推薦失敗時)、の 2 種類を想定している。

社会的受容性の評価構想

ここでは、まず、責任をとるエージェントの社会的受容性を評価する実験の構想を述べ、続いて、評価アンケートの項目について記す。

評価実験

まず、【行為責任（謝罪)】、【行為責任（自粛)】の両方の発話を行い、かつ、【説明責任（理由と対策)】の発話も行う場合（すべての責任発話を行う場合）と、責任発話を全く行わない場合を被験者間実験で比較（評価項目については次節を参照）し、責任発話の効果を調べる。効果があった場合は、続いて、【行為責任（謝罪)】【行為責任（自粛)】のいずれが効果的であったか、【説明責任（理由)】だけでも効果があるか等を調べる。効果がなかった場合は、実験設定や、責任発話等を再考し、改良結果を評価する実験を計画する。

実験時間は、まずは、視聴する動画を 5 分以内程度の短い動画に限定し、実験参加者一人当たり 1 時間程度で実施できるものを想定しているが、その次の段階として、中長期間のインタラクションの効果（日常生活における社会的受容性）をみたいと考えている。

実験の流れとしては次を想定している：

1. 実験参加者は好きな単語で検索をするか動画を一つ選択する
2. 1）エージェントの発話に対してそれへの応答選択肢から一つ選択する、2）動画リストから見たい動画を選択する、3）好きな単語で検索

する、4）視聴画面なら視聴する、の内から好きな行動をする
3. 2 を繰り返す

評価項目

以下の項目のアンケートを 5 段階評価で実施することを計画している：

- エージェントは自分の意思で動画を選定していると感じた
- エージェントはプログラムに従って動画を選定していると感じた
- エージェントは私の好みを学習していると感じた
- エージェントによる推薦・再生が不適切だったとき、エージェントに責任があると感じた
- エージェントによる推薦・再生が不適切だったとき、システム開発者に責任があると感じた
- エージェントによる推薦・再生が不適切だったとき、エージェントは心から謝罪していると感じた
- エージェントに謝罪されても、単にそうプログラムされているだけだと感じた
- エージェントによる、なぜその動画を選択したかの説明に納得した
- このエージェントは責任をとる能力があると感じた
- このエージェントは信頼できると感じた
- この推薦システムを実験以外の場でも使ってみたいと感じた

おわりに

本論文では、人間の生命・身体・財産の安全等に関わる分野への AI 技術の浸透のための一つの方策として、権利と責任を有する AI を開発することを提案した。行為責任と説明責任、XAI と説明責任、自由意志と責任について順次考察し、さらに、責任を有する AI の社会的受容性の評価に向けて我々が試作している、軽微な権利と責任を持つ動画推薦 AI を紹介し、その社会的受容性を検討するための実験計画について述べた。

AI の開発者・開発組織や販売者・販売組織、AI の所有者・所有組織等ではなく AI 自体が責任をとることが可能か、また、そのような AI を社会が受容する可能性については、多様な意見が存在し、すぐに結論が出る性質の問題ではないが、本論文が議論のきっかけとなれば幸いである。さらに今後、社会的受容性を検討する実験を実施することで、より議論を深めていきたい。

参考文献

[1] High Level Expert Group on Artificial Intelligence, Ethics guidelines for trustworthy ai, *Tech. rep.*, European Commission (2019).

[2] Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion*, Vol. 58, pp. 82-115, (2020)

[3] Libet B., Gleason C. A., Wright E. W., Pearl D.K., Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): The unconscious initiation of a freely voluntary act., *Brain*, Vol. 106 (Pt 3), pp. 623-642, (1983)

[4] Shimojo S., Postdiction: its implications on visual awareness, hindsight, and sense of agency. *Frontiers in psychology*, *5*, 196. (2014)

[5] Johansson P., Hall L., Sikström S., and Olsson A., Failure to detect mismatches between intention and outcome in a simple decision task, *Science*, 310 (5745), 116-9. (2005)

[6] Oka N. Apparent "free will" caused by representation of module control, *No Matter, Never Mind. Proceedings of Toward a Science of Consciousness: Fundamental Approaches*, pp. 243-249, (2002)

[7] 松島茜, 岡夏樹, 深田智, 吉村優子, 川原功司, 対話行為を表す機能語の獲得：自動的な処理と意識的な処理を統合したモデルの構築, 信学技法, HCS2019-70, pp. 93-98, (2020)

[8] Paul Covington, Jay Adams, Emre Sargin, Deep neural networks for YouTube recommendations, *Proceedings of the 10th ACM Conference on Recommender Systems*, ACM, New York, NY, (2016)

[9] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, EdH. Chi, Top-k off-policy correction for a REINFORCE recommender system, *ACM International Conference on Web Search and Data Mining*, (2019)