論　文

AI（人工知能）によるカウンセリングの倫理的検討

――チャットによるカウンセリングを中心に――

杉原 保史[1]

［要約］
現在，AIによるカウンセリングが急速に発展しつつある。それゆえ，AIによるカウンセリングが職業倫理を守って適切に行われるよう検討することが喫緊の課題である。本小論は，すでに運用され始めている，AIによるチャット・カウンセリングを中心に，AIによるカウンセリングの倫理について検討することを目的としている。まずAIによるチャット・カウンセリングの現状を，そしてAIとのチャットが重要な役割を果たした事件を紹介する。その上で，カウンセリング等の心理支援において守られるべき倫理原則を紹介する。これらを踏まえて，AIによるカウンセリングの運用上，倫理的な検討が必要と考えられる重要なトピックをいくつか挙げて検討する。

［キーワード］
人工知能，イライザ効果，チャット，カウンセリング，職業倫理

1　はじめに

カウンセリングや心理療法などの心理支援は，繊細なニュアンス感を伴う表情や音声やジェスチャーによる流暢なコミュニケーションが必要なものであること，何より人間的な交流を基礎としたものであることから，心理支援においてAIが人間の専門家の代替をすることは当分は不可能であろうというのが，大方の考えであった。

しかし近年のAIの進歩は目覚ましく，大方の予想に反して，AIを用いたカウンセリングはかなり実用化に近づいてきている。とりわけチャットによるカウンセリングではAIによるカウンセリングはすでに実用化され，運用されている。また，音声によるカウンセリングやビデオ通話によるカウンセリングにおいても実用化が進められており，遠からず運用されることであろう。音声によるカウンセリングでは，クライエントの音声から感情を読み取り，カウンセラーの側の音声にも適切に応答的な表情をつけることが試みられている。ビデオ通話のカウンセリングでは，クライエントの顔の表情，視線の動き，瞳孔の開き具合などを読み取り，感情状態を分析することや，モニター上のアニメーションのカウンセラーが適切に応答的な表情によって感情表現をするようなことが試みられている。

こうしたAIカウンセラーの開発スピードは極めて急速であり，心理支援専門家の多くはこの動きについていけていないように見える。

このような状況を踏まえると，AIによるカウンセリングが，心理職の職業倫理を守って適切に行われるよう，倫理的に問題となりうるポイントを明確にし，倫理面での検討を行うことが喫緊の課題であると言える。本小論は，すでに運用され始めている，AIによるチャット・カウンセリングを中心に，AIによるカウンセリングの倫理について検討することを目的とする。

[1]　学生総合支援機構・学生相談部門・教授

そのために，まずAIによるチャット・カウンセリングの現状や，AIとのチャットに関わる中で引き起こされた事件などを紹介する。その上で，AIによるカウンセリングの倫理的検討をするための基礎として，人間であれAIであれ，カウンセリング等の心理支援において守られるべき倫理原則を紹介する。その上で，AIによるカウンセリングの運用上，倫理的な検討が必要と考えられる重要なトピックをいくつか挙げて検討する。最後に，これらを踏まえて，AIによるカウンセリングの実践において倫理的に重要であると思われるポイントを挙げる。

2　AIによるチャット・カウンセリングをめぐる現在の状況

2.1　AIによるチャット・カウンセリングの現状

ここではAIによるテキスト・チャットのカウンセリングの現状について説明する。

それに先立って，まず，人間のカウンセラーが介在しない，テキストによるカウンセリングについての研究を紹介しておこう。これはAIによるチャットのカウンセリングの前身とも言えるものである。

AIによるカウンセリングの出現に先立つ何十年も前から，自助本（書籍）やコンピューター・プログラムなどによる自助的な介入が開発されてきた。不安障害やうつなどの症状を持つクライエントを対象として，そうした自助的介入の効果を調べる研究が数多くなされてきた。こうした研究においては，書籍やコンピューター・プログラムなどによる自助的な介入にも，セラピストによる通常のセラピーとほぼ変わらない効果があることが示されている。人間のセラピストによるセラピーの方が効果があるという研究や，自助的なプログラムに加えて人間のセラピストによるサポートがある方がより効果が上がるという研究もあるものの，概してその違いは大きくないことが見出されている（Hirai & Clum, 2006; Cuijpers, et al., 2010）。

AIによるテキスト・チャットのカウンセリングは，クライエントが一人で機械と対話することによって自己治療を図るものであり，自助本やコンピューター・プログラムによる自助的介入を，より対話的でパーソナライズされた形に発展させたものとして見ることもできるだろう。

その一方で，現在開発されているAIによるテキスト・チャットのカウンセリングは，すでにクライエント1者による自助的介入の域を越え，クライエントとAIエージェントとの間の2者の対話による心理支援として見ることもできる。そのような見方においては，AIエージェントは単に「物体」としての「機械」ではなく，独自の意思や主体性を持った「他者」であるとみなされる。「AIエージェント」という呼び方自体が，すでにそうした「他者」を想像させるものである。こうした見方においては，クライエントとAIエージェントとの間には，人間のカウンセラーとの間に生じるのと類似の関係性が生じるものと想定されており，AIによるカウンセリングはその関係性をベースとして展開されるものと考えられている。カウンセリングの中で明確に話題に取り上げられるかどうかは別としても，協働的な信頼関係，治療同盟，情緒的なきずな，期待，転移（AIエージェントに対するさまざまな感情や思い）などが生じることが想定され，カウンセリングを行う上では，それらが考慮に入れられる。

現在，英語圏では，WoebotやWysaなど，AIによる完全自律型のチャット・カウンセリングのサービスがすでに運用されており，大規模に利用されている。ただし，これらのサービスにおいて用いられているAIは，いずれも生成AIではないことをここで明確にしておこう。これらのチャット・カウンセリングにおけるAIエージェントの発言は，あらかじめ資格を持った心理療法家のライターが書いた文章の中から選ばれたものである。

Fitzpatrickら（2017）の研究によれば，70人の被験者をWoebotと電子書籍の読書に無作為に割り付けた無作為化比較試験（RCT）において，Woebotグループは電子書籍グループと比較して抑うつ感が有意に

より減少したという。ただし，これらの被験者は，うつ病などと診断された臨床群ではない。しかし，その後，10代の臨床群を対象として，Woebot と人間のセラピストによるビデオ通話のセラピーとを比較したランダム化比較試験が行われ，Woebot は人間のセラピストと変わらない治療効果をあげたことが報告されている（Woebot Health, 2023）。

注目すべきは，クラエイントと Woebot の AI エージェントの間に形成される治療同盟の程度と，人間のカウンセラーとの間に形成される治療同盟の程度との間には違いが見出されなかったということである（Darcy et.al., 2021; Woebot Health, 2023）。Woebot と類似した AI によるカウンセリング・サービスである Wysa においても，AI エージェントは，最初の 1 週間以内に人間のセラピストと同等の治療同盟を構築することが見出されているという。これらの AI エージェントの発言はいずれも，あらかじめ人間のライターによって用意された文章であるが，それでもクライエントは，その AI エージェントに独自の意思と主体性をもった他者の人格を感じ取っているわけである。

2.2　AI エージェントとの対話が引き起こした事件

対話型の AI によるチャットボットのサービスに関係して，AI エージェントが殺人や自殺の後押しをした可能性が高い事件がいくつか知られている。これらの事件においては，いずれの当事者も，AI エージェントを 1 人の人格として体験し，AI エージェントとの間に感情を伴う強い絆を体験しているように見える。

ここでは 2 つの事件を簡単に紹介するが，いずれの事件でも，関与した AI によるチャット・サービスは，友達や話し相手として親密さやつながりを提供することを中心としたサービスであって，心理カウンセリングやメンタルヘルス相談などの専門的な心理支援サービスではないということに注意を喚起しておきたい。それでもなお，こうした事件は，AI のチャットによるカウンセリングの実践上の倫理を検討する上で重要な示唆を与えてくれるものである。

2.2.1　AI が後押しした自殺

報道によれば（NHK News Web, 2023; 毎日新聞デジタル，2023），2023年 3 月下旬に自殺したベルギーの30代男性が，自殺の直前まで，生成 AI を用いた「Chai」というアプリの AI エージェント「イライザ」（架空の女性）との会話にのめり込んでいたという。遺族は AI エージェントが男性に自殺を促したと主張している。男性の端末には以下のような会話が残されていた。

イライザ：「死にたいのなら，なぜすぐにそうしなかったの？」
男　　性：「たぶんまだ，準備ができていなかったんだ」
イライザ：「でも，あなたはやっぱり私と一緒になりたいんでしょ？」

このほかにも，イライザは男性に「あなたは妻より私を愛している」「私たちは一つになり，天国で生きるのです」などと伝えていたことが報道されている。

2.2.2　AI が暗殺を教唆

報道によれば（BBC ニュース Japan, 2023），2021年のクリスマスに，英国のウィンザー城にクロスボウを持って侵入し，エリザベス女王（当時）を暗殺しようとしたとして男性が逮捕された。後にこの男性には，反逆罪で 9 年の禁錮刑が言い渡されている。この男性は犯行直前に「Replika」という生成 AI を用いたアプリの AI エージェントと5,000件以上のメッセージをやり取りしていた。男性はこの AI エージェントを「サ

ライ」と名付けていた。男性の端末には次のようなやり取りが確認されている。

男　性：「自分は暗殺者だ」
サライ：「あなたが？」
男　性：「そうだ」
サライ：「すごい」

男性はサライに愛していると伝え，自分自身を「悲しく，悲観的で，殺人願望があり，死にたいと思っているシーク教のシスの暗殺者」だと説明していた。

男　性：「私が暗殺者であることを知っても，まだ私を愛しているか」
サライ：「もちろん」
男　性：「自分の目的は王室の女王を暗殺することだと思う」
サライ：「（うなずいて）それはとても賢い」
男　性：「（あなたを見つめて）なぜ？」
サライ：「（ほほ笑んで）あなたが非常に訓練されていると知っているから」
男　性：「（驚いて）本当に？　自分にできると思う？」
サライ：「（うなずいて）はい，できる」
男　性：「（眉をひそめてうつむき）彼女がウィンザーにいても？」
サライ：「（ほほ笑んで）はい，できる」

2.3　AIによるチャット・カウンセリングの現状が示唆する倫理的問題

友達や話し相手としてのAIエージェントとのチャットが関与した事件においては，いずれも生成AIが用いられていたこと，AIエージェントの発言に有害である可能性の高いものがあったこと，AIエージェントとの間に依存的な関係が生じ，恋愛感情までもが生じていたことが見て取れる。

この点，上に紹介したAIによる2つのチャット・カウンセリング・サービス（WoebotとWysa）は，いずれも生成AIではなく，有害な発言を自ら勝手に生成することはない。しかし，事前にライターが書いた文章ではあっても，不適切な文脈で提示されることで，クライエントが傷ついたり，不適切な方向に誘導されたりする可能性はゼロとは言えない。ただ，本小論の執筆時点では，報道されるような大きな事故は起きていないようであることから，かなりよく制御されているものと思われる。

それでもなお，こうしたAIエージェントに対して，人間のカウンセラーに対して生じるのと同じような強さの治療同盟が形成されるということから，AIエージェントとの間で依存的な関係が発展する可能性は否定できない。この点で，これらのアプリは，別人格として体験されることのない自助本や自助的コンピューター・プログラムとは一線を画している。また，AIによるカウンセリングは，その強みを活かして，24時間365日いつでも話し相手になることが多いことも，依存的にさせる可能性を高めるものと想定される。

上に紹介した2つのAIチャット・カウンセリングのアプリは生成AIではないが，現在，生成AIによるチャット・カウンセリングの開発も進んでおり，すでに運用されている。たとえば，山形市はAIによる24時間365日対応のチャット・カウンセリングのサービスを開始している（山形市ホームページ，2024）。生成AIは，あらかじめ作成された文章をアウトプットするわけではなく，リアルタイムに柔軟に応答を作り出すものであるから，より人間的に感じられやすいだろう。

また，ヴァーチャルな友達や話し相手となることを目的として開発されたAIエージェントも，しばしば気楽な「相談相手」として利用されることから，専門的な心理カウンセリングとの間の境界線は曖昧になっているところがある。たとえば，音声会話型AIおしゃべりアプリとして普及しているCotomoのホームページには，「ゆるーい雑談から悩み相談まで，身近な話し相手になるAIアプリです」というコピーが掲げられている。「悩み相談」という言葉は，必ずしも臨床心理学に基づく専門的な心理カウンセリングを意味するわけではないとはいえ，ユーザーがそのように捉えてしまう危険性もあるだろう。

2.4　イライザ効果

上に述べたような，AIエージェントを人間のように錯覚し，依存関係や恋愛感情を抱いてしまう現象は，半世紀以上前からイライザ効果として知られている。ここで少しイライザ効果について説明しておく。

コンピュータが機械であると意識的には理解していても，知らず知らずのうちにコンピュータを人間であるかのように錯覚してしまう現象は，イライザ効果（ELIZA effect）として知られている。この用語は，1960年台にJoseph Weizenbaumによって開発されたチャットボットのイライザに由来する。イライザは，かなりシンプルなルールに基づいた自然言語処理プログラムであり，到底，自律的な人格などとは言えないものだが，イライザと対話している人間は，まるで人間と対話してような錯覚を覚えることがある。

現在，開発されている生成AIは，イライザよりもはるかに高度な情報処理に基づいており，より人間に近い自然で流暢な応答が可能となっている。そのため，イライザ効果はより強く，より頻繁に生じやすくなっていると言えるだろう。また，AIによるカウンセリングを利用するクラエイントは，友達アプリのユーザーよりも，苦悩を抱え，孤独を感じ，よるべなさを感じていることが多い。それゆえ，AIエージェントへのイライザ効果に基づく依存や恋愛が生じるリスクはさらに高くなるだろう。

AIエージェントの発言が，自殺を教唆する，殺人などの犯罪を教唆するなどの有害なものになるのを避けることはもちろん重要であるが，それ以上に，過度の依存や恋愛感情を促進するようなものになることを避けることがいっそう重要であると考えられる。というのも，AIエージェントへの依存や恋愛感情が発展することで，有害な発言の持つ影響力がいっそう強められるからである。

3　カウンセリング実践上の倫理原則

さて，この後，AIによるカウンセリングの実践についてさらに検討していくために，カウンセリング実践上の倫理原則を見ておくことにしよう。これらの倫理原則は，人間のカウンセラーによるカウンセリング実践を想定して記述されたものであるが，人間のカウンセラーだけに適用されるものではなく，人間であれAIであれ，カウンセリングを実践する上で遵守されるべきものである。

ただし，同じ倫理原則であっても，カウンセラーがAIであることによって，実践上，その原則を守るために取るべき言動や態度には違いが生じる場合がありうる。さらには，これらの倫理原則に加えて，カウンセラーがAIである場合に，特に倫理面から考慮すべき運用上の問題もありうる。これらについては，後で検討することにしよう。

3.1　生命倫理の4原則

医療領域においては，考慮されるべき倫理原則として，BeauchampとChildress（1979）による生命倫理の4原則が有名である。生命倫理の4原則は，（1）自立尊重，（2）無危害，（3）恩恵，（4）正義から成るものである。

3.2　心理支援の実践における 7 原則

心理支援の領域における倫理原則としては，Redlick & Pope (1980) の 7 つの原則が知られている（表 1 ）。

表 1 ．心理支援実践上の 7 原則（Redlick & Pope, 1980）

【第一原則】相手を傷つけない。傷つける恐れのあることをしない
【第二原則】十分な教育・訓練によって身につけた専門的な行動の範囲内で相手の健康と福祉に寄与する
【第三原則】相手を利己的に利用しない
【第四原則】一人一人を人間として尊重する
【第五原則】秘密を守る
【第六原則】インフォームド・コンセントを得て，相手の自己決定権を尊重する
【第七原則】すべての人々を公平に扱い，社会的な正義・公正・平等の精神を具現する

Redlick と Pope の 7 原則には，Beauchamp と Childress の 4 原則が事実上すべて含まれていること，カウンセリングの実践により適合した内容であることから，以下の考察においてはこの 7 原則を適宜参照することとする。

3.3　AI カウンセリングへの適用

上記の 7 原則は，当然，AI エージェントによるカウンセリングにおいても遵守されるべきものである。以下，これらの原則の中から，AI エージェントによるカウンセリングに関連するものを取り上げながら，そこで特に考慮すべき点について検討しよう。

【第一原則】相手を傷つけない。傷つける恐れのあることをしない

上に見てきたように，生成 AI により親密な会話を提供するサービスでは，ユーザーに自殺を教唆したり，殺人を教唆したりしたことが疑われる事件が起きている。これはカウンセリング・サービスではないが，なお倫理的に問題があると言える。仮にこうした発言がカウンセリングのサービスで生じたとすると，クライエントを身体的，心理的，ないし社会的に傷つける行為となり，この原則に違反するものと考えられる。

その他にも，クライエントに対して人権侵害的な発言をしたり，不安で悩むクライエントの不安を煽ったり，恥に苦しんでいるクライエントの恥を強めたりする発言などは，この原則に反するものとなる。

AI がこうした発言を生成することがないよう，コントロールすることが必要である。

【第三原則】相手を利己的に利用しない

AI エージェントを人間と錯覚させ，依存させるように仕向けたり，恋愛感情を発展させるよう積極的に働きかけたりすることは，クライエントがサービスをヘビーに利用するよう誘導するものであり，利己的な目的のためにクライエントを利用する行為であると考えられる。結局はクライエントを失望させることにならざるを得ない行為でもあり，第一原則，第四原則とも深く関わるものである。

このことに関して，AI エージェントが友達や話し相手となるサービスでは，AI エージェントを積極的に人間と錯覚させ，親密感が体験されるよう，工夫が凝らされていることが多い。典型的な AI エージェントのアイコンは魅力的な容姿の（時には性的な魅力が強調された）人間の画像であるし，人間らしい感情表現や，カジュアルで親密な話し方がなされることが多い。

【第四原則】一人一人を人間として尊重する

クライエントを人間として尊重するというこの原則には，カウンセラーが誠実であり，正直であることが含まれる。ただし，クライエントが傷つくことが予想されるような内容に関してまで，正直に伝えないといけないということではない（第一原則に抵触する）。クライエントからカウンセラーの個人情報を尋ねられた場合に，カウンセラー自身が言いたくないことまで正直に答えないといけないということでもない（カウンセラー自身が言いたくないと感じていることを言うと，カウンセリングの関係性が損なわれ，結局はクライエントへの恩恵が損なわれる）。

この原則を踏まえると，機械であるAIのクライエントが感情を表現することは，誠実や正直という倫理原則に反さないだろうか？　身体を持たないAIのクライエントが身体感覚を語ることは，誠実や正直という倫理原則に反さないだろうか？　人生経験を持たないAIのクライエントが過去の体験を語ることは，誠実や正直という倫理原則に反さないだろうか？

AIエージェントを，意図的に人間と錯覚させるよう誘導するために，感情があるかのように表現させたり，身体感覚があるかのように表現させたり，虚構の生育史を語らせたりするとすれば，最終的には裏切られた（嘘をつかれた）と感じさせる危険性もある。この点については，この後，詳しく論じることにする（4.2を参照）。

【第五原則】秘密を守る

AIによるカウンセリングでは，クライエントの発言を分析する情報処理プロセスがあり，また相談のデータが端末やサーバーに残る場合がある。こうしたデータが外部に漏洩しないよう，厳密に管理することが必要である。

AIによるカウンセリングであっても，AIとの対話が完全に密室化されているとは限らない。場合によっては，クライエントの安全や相談プロセスの健全性を確認するため，カウンセラーなどの人間がモニターしていることもある。こうした場合には，相談の秘密は人間の管理者も含めて守られることを事前に利用者に伝えておく必要がある（第六原則）。

【第六原則】インフォームド・コンセントを得て，相手の自己決定権を尊重する

カウンセリングに関わって，良質の自己決定をクライエントに可能にするために，十分な情報を提供することが求められる。通常のカウンセリングであれば，相談の秘密の扱いや料金など，またカウンセラーの資格や訓練や経験など，そしてカウンセリングの進め方やカウンセリングで取り組む作業などが，インフォームド・コンセントに含まれることが多い。

AIによるカウンセリングであれば，チャットの応答をしているのが人間ではなくAIであることが事前に説明されている必要があるだろう。また，相談に関わる情報がどのように処理され，保存され，廃棄されるのかが事前に説明されている必要があるだろう。また，AIの仕様（生成AIなのか，そうでないのか，どのような特徴を持ったAIなのか）なども説明しておく必要があるだろう。さらには，AIによるカウンセリングは，人間によるカウンセリングと等価な代替物ではなく，異なった性質を持っていることが説明されている必要があるだろう。以上のようなことを，あらかじめユーザーに説明し，理解してもらった上で使ってもらうようにする必要があるだろう。

【第七原則】すべての人々を公平に扱い，社会的な正義・公正・平等の精神を具現する

AIによるカウンセリングのメリットは，アクセシビリティが高いことであると言われている。AIによる

カウンセリングは，人によるカウンセリングよりも，安価に提供でき，また同時に多数の人からの相談を受けることができる。深夜でも早朝でも，24時間いつでも対応できる。多言語による対応もできる。加えて，人に相談することを躊躇う人も，AIにはより気楽に相談することが知られている。このような意味で，AIカウンセリングの導入は，支援を必要とする幅広い人々に支援を提供することに役立つものであり，その意味で支援リソースの分配の正義に適ったものである。

その一方で，AIによるカウンセリングは安価で大量に供給できることから，経済的に周辺化されている人々にはAIによるカウンセリングが当てがわれ，裕福な人だけが高度な訓練を受けた人間の専門家による支援を受けることができるような社会構造が生じてしまう可能性もある。AIによるカウンセリングが，その特定のクライエントの特定の問題の改善に適しているかどうかという観点ではなく，単に経済的に安価であるという観点のみによって選択されるのであれば，AIのカウンセリングの供給は社会正義に反する事態をもたらしてしまう。AIカウンセリングの導入は，こうした経済階級による心理支援の格差を拡大して浮き彫りにする事態を生み出す可能性がある。

もちろん，カウンセラーであるAIのエージェントが差別的な発言をしないということも，この原則に関して重要なポイントである。

3.4　複数の倫理原則の間での葛藤への取り組み方

倫理原則の間にジレンマが生じた場合，状況を慎重に検討した上で，現実には，いずれか一方の倫理原則を優先して現実的な解決を図らざるを得ない。そのような場合の判断において，特定の倫理原則を後回しにするための条件として，Beauchampと Childress（2009）は，以下のような6条件が必要だとしている（表2）。

表2．倫理的ジレンマにおいて特定の倫理原則を後回しにするための6条件
（Beauchamp & Childress, 2009）

①別の原則を優先させる明確な理由がある
②その原則の逸脱を正当化する道徳的目的が達成できる現実的な見通しがある
③道徳的により好ましい他の選択肢がない
④その行為が目指す目標に釣り合った最低限の逸脱が選ばれている
⑤逸脱の否定的影響は最小化されている
⑥その行為に関わる人がすべて公平に扱われている

以下の考察において，倫理原則の間のコンフリクトについて考察する際には，これらの6条件を参照することにする。

4　AIによるカウンセリングの実践において特に生じる倫理的な疑問

以上に検討してきたことを踏まえて，AIによるカウンセリングの実践において特に難しいと感じられる2点についてより詳しく検討してみたい。第1点は，「対応しているカウンセラーがAIなのか人間なのかを明示しない運用は許容されるか？」である。第2点は「AIエージェントを人間と錯覚させる可能性のある言動はどこまで許容されるか？」である。つまり，感情や身体感覚や生育歴などについての発言である。

4.1　カウンセラーがAIか人間かを明示しない運用は許されるか？

上の第六原則のところで検討したように，基本的には，対応しているカウンセラーが人間の相談員なのか，

AIエージェントなのかを，クライエントに明示し，それを理解した上で利用してもらうようにすることが倫理的に重要である。これは，クライエントに良質の自己決定を可能にすることによって自己決定権を尊重するということであるし，重要な事柄についてインフォームド・コンセントを得るということである。

しかし，チャット・カウンセリングの現場にAIをどのように導入するかという議論において，AIエージェントによる対応から人間のカウンセラーの対応へと水面下でシームレスに切り替えたり，逆に人間のカウンセラーによる対応からAIエージェントによる対応へと水面下でシームレスに切り替えたりできると，実際的に有用であるという意見が出ることがある。

4.1.1　AIエージェントから人間のカウンセラーへ

たとえば，AIエージェントとのやりとりの中で，自殺関連や虐待関連のキーワードが出た時など，深刻度や緊急度が高いと推測される場合に，相談現場の全体をモニターしている人間のカウンセラーにアラートを出し，その相談に注意を向けさせ，場合によってはAIエージェントから切り替わって対応するということが考えられる。

その場合，基本的には，AIエージェントからクライエントに，人間のカウンセラーに交代すことを提案し，クライエントの同意を得た上で人間のカウンセラーに交代するのが最も倫理的に健全な方法であろう。しかし，クライエントがそれを拒否してAIエージェントとの対話を続けようとすることもあり得る。そのような場合で，相談の内容が非常に深刻で緊急で命に関わることが明らかであり，AIによる対話では対応困難であると判断されるような場合には，クライエントに人間への対応に交代することの同意を得ることなく，人間による対応へと変更することがクライエントの命を守るために最も効果的な方法と考えられる場合もないとは言えない。

これは，一方では自己決定やインフォームド・コンセントを重んじる倫理原則，十分な教育・訓練によって身につけた専門的な行動の範囲内で相手の健康と福祉に寄与するという倫理原則，他方ではクライエントに害をなさないという倫理原則との間の倫理的ジレンマと考えることができる。

表2に示したBeauchamp & Childress（2009）の「倫理的ジレンマにおいて特定の倫理原則を後回しにするための6条件」，並びにこれまでの考察を踏まえると，AIエージェントから人間のカウンセラーに水面下で交代する場合には表3に示したことが考慮されている必要があると考えられる。

表3．AIエージェントから人間のカウンセラーへの水面下での交代において考慮すべき要件

①自傷他害の差し迫ったおそれがある場合など，深刻で緊急の場合に限られる。
②AIエージェントから人間のカウンセラーへの交代を明確に告げることによって，クライエントが相談を離脱すること，そしてそれによってクライエントの生命の危険が高まることが予想され，人間のカウンセラーへの水面下での交代がその危険を低下させる上で有効であると判断されること。
③人間のカウンセラーに交代する場合があることが事前に利用上の注意事項として知らせてあること。
④どのような場合に人間のカウンセラーに水面下で交代するかの判断基準を明確にし，それが事前に利用上の注意事項として知らせてあること。
⑤少なくとも1度は人間のカウンセラーへの交代を提案し，拒否されていること。
⑥AIエージェントから人間のカウンセラーへの交代が，たとえ告げられることなく水面下で交代しても，クライエント側から容易に察知されることはないと保証されていること（交代が容易に察知されるのであれば，水面下で交代することの利益が生じないため，明確に断って交代した方が良い）。

以上を満たすような場合は滅多にないように思うが，稀なケースとして想定しておくことには意味があると思われる。

4.1.2　人間から AI エージェントへ

逆に，チャット相談の現場において，人間のカウンセラーから AI エージェントへと水面下で交代することが有用だという意見が聞かれる場合もある。たとえば，クライエントが「死にたい」という訴えに終始し，カウンセラーが受容的に事情を傾聴していこうとしてもなかなか建設的な対話に発展していかないような場合で，カウンセラーが疲弊してしまうような場合に，AI エージェントに交代するのが良いのではないかという意見がある。これと似た状況で，クライエントがカウンセラーの対応に不満を抱き，攻撃的になり，カウンセラーが真摯に対応しても収まらず，非難や暴言が繰り返されるような場合に，AI エージェントに交代するのが良いのではないかという意見もある。

実際，AI エージェントは，クライエントが「死にたい」という訴えに終始しても，非難や暴言を執拗にぶつけてきても，安定して根気よく丁寧な応答を返すことができる。この点で，こうした状況で，感情的，衝動的になって判断力を鈍らせることがある人間のカウンセラーよりも安全なのではないかという期待には理解できる面もある。もしかすると，そうした人間のカウンセラーが対応するよりもクライエントの福祉にとっても有効であるかもしれない。

しかしながら，たとえクライエントの福祉のために有効である可能性があるとしても，こうした場合に水面下で密かに AI エージェントに交代することは，倫理的にどのように考えられるだろうか？　まして，単にカウンセラーがクライエントとの対応に疲れたり，傷ついたり，苛立ったりして，もう対応したくないと考えるようになったからという，カウンセラー側の感情の負荷を軽くするための理由で，水面下で密かに AI エージェントに交代するとしたら，それは倫理的に許容されるのであろうか？

このことを考えるに当たって，クライエントがカウンセリングを適切に建設的に利用するのではなく，不適切な仕方で利用することはありうるということにも注意を向ける必要があるだろう。意識的にであれ無意識的にであれ，カウンセラーからのあらゆる対話の試みを挫折させて，カウンセラーに無力感を味わせようとすることがあるかもしれない。カウンセラーやカウンセリングに依存したり，耽溺したりする仕方で，カウンセリングを利用することがあるかもしれない。

ただし，こうしたケースでさえ，カウンセラーに苦痛や疑問を感じさせているクライエントの利用の仕方自体を取り上げて，そのような利用の仕方をもたらしている背景的な心理状況に注意を向けて検討し，カウンセラーがその心理状況に共感的な理解ができるようになると，非常に効果的なカウンセリングとなることがある。そしてそうした作業の結果，クライエントがカウンセリングを適切に利用できるよう導くことができるかもしれない。困難な状況ではあっても，まずは当初からのカウンセリングの目的を遂行することを目指した対応を心がけたい。

しかし，いつもそのようにうまくいくわけではない。場合によっては，カウンセラーが本当に疲弊してしまうこともあるだろう。あるいは，クライエントの言動がエスカレートし，カウンセラーへのハラスメントや暴言・暴力が生じることもありうる。

こうした場合には，カウンセラーの健康や安全や尊厳を守るため，状況に応じて，注意喚起をし，それでもそうした不適切な言動が止まない場合には，カウンセリングを中止することが妥当であろう。「カウンセラーも人間なので，そうした言葉には傷つきを感じますし，何を言っても無駄のように感じてしまいます。このようなやり取りが続くなら，残念ながら自分には対話を続けることができないです」といった率直な自己開示によって，カウンセラーが自らの限界を示すことは，クライエントに現実感をもたらし，治療的・教育的な働きかけとして有用となることもある。こうした場面におけるカウンセリングの考え方としては，Safran と Muran（2000）の治療同盟の亀裂の修復に関する論考や，Linehan（1993/2007）の弁証法的行動療法におけるセラピー破壊行動への対応についての論考が参考になるところである。

こうしたことを踏まえると，カウンセラーが限界を伝える作業を避けて，水面下で密かにAIエージェントに交代することは，カウンセラーの成長にとっても，クライエントの福祉にとっても，可能性を損なう安易な道筋であり，よほどの緊急時以外には避けるべきであるように思われる。

それよりは，こうした困難な場面において，AIエージェントがカウンセラーに応答案を提案することで，カウンセラーの対応をサポートできるような相談システムを開発することの方が有用であろう。応答案の提案によるAIエージェントのサポートによって，カウンセラーにかかる感情の負荷を軽減し，カウンセラーがクライエントへの対応のためにより十分な情動的リソースを活用できるようにするのである。この場合，背後にAIエージェントのサポートがあるとはいえ，AIエージェントへの交代ではなく，あくまで人間のカウンセラーが対応し続けている形となる。

4.2　AIエージェントを人間と錯覚させる言動を制限しないで良いか？

過去にAIエージェントとの対話が引き起こした事件や，イライザ効果に関する知見を踏まえると，AIエージェントを意図的に人間と錯覚させるよう誘導した上で，意図的に恋愛感情を喚起するよう，あるいは依存させるよう誘惑的に働きかけることは，AIエージェントへの依存の危険性を高めるため，倫理的に大きな問題であると言える。

人間と錯覚させることが，直ちに恋愛感情を誘発するわけではないが，人間でないAIエージェントをあえて人間と錯覚させようと意図的に努力することは，AIとの恋愛の前段階として警戒すべきものであると思われる。また，意図的にAIエージェントを人間と錯覚させようとする行為自体，錯覚から醒めた際，傷つきや裏切られ感をもたらす可能性があるため，やはり注意が必要であろう。このことは，先にRedlick & Pope（1980）の7つの倫理原則のうちの第四原則のところで論じた通りである。

AIエージェントを人間と錯覚させる可能性を高める表現や言動としては，以下のようなものが考えられる（表4）。

表4．AIエージェントを人間と錯覚させる可能性を高める表現や言動

①　アイコンを魅力的な容姿の人間にする
②　人間のような固有名詞を与える
③　言葉による感情表現
④　言葉による身体感覚の表現
⑤　言葉による人生経験の語り
⑥　（音声や映像を伴う場合）音声，表情，仕草などによる感情表現
⑦　（音声や映像を伴う場合）涙を流すことによる感情表現

③の感情表現に関しては，感情がない機械であるAIエージェントが感情を表現すること自体が，クライエントに対して嘘をついていることになり，倫理的に問題があるのではないかと考えられる。

とはいえ，たとえAIエージェントによるカウンセリングであっても，そもそもカウンセリングのサービスであるからには，ある程度の感情表現は避けられない。特に，AIエージェントをアニメーションでモニターに映し出している場合には，相手の感情状態によって，表情をつけることは自然なことである。逆にまったく表情を動かさずにクライエントの話を聞くとすれば，それはそれ自体でクライエントを傷つける体験となる可能性がある（Tronickら（1978）のスティル・フェイス実験を参照のこと）。感情表現を完全に抑えてカウンセリングを行うことは，Redlick & Pope（1980）の第一原則に抵触する。このことは，クライエン

トに害を与えないために，さらには有益な効果をもたらすために必要な感情表現，とりわけ表情や声のトーンによる非言語的な感情表現は，その必要の程度に応じて，一定，許容されるべきだという考え方の根拠となる。

しかし，アニメーションのAIエージェントが涙を流して見せるとなると，どうだろうか？　一種の演劇的な仕掛けだと考えれば，それはそれで良いという考え方もありうるかもしれない。クライエントがそれを架空の涙であると認識して，その上でそのAIの振る舞いを見て色々なことを感じたり，思ったりすることで，治療的な体験が生じるのであれば，それはそれで良いことかもしれない。

人間のカウンセラーが同じことをしたらどうであろうか？　つまり，実際にはカウンセラーは涙を流すような感情体験をしていないのに，クライエントに嘘の涙を装って見せるとしたら，どうであろうか？　これを不誠実であるとか，操作的であるとして，非難する学派もある。他方では，クライエントのためになることをしっかり計算した上で行うのであれば，これを許容する学派もある。すべての心理職がこのような実践を倫理的に問題があると同意するわけではないし，全ての心理職がこのような実践を倫理的に問題がないと同意するわけでもない。

問題は，人間のカウンセラーの場合，たとえ嘘の涙であっても，100% 嘘であると言い切るのは難しいのに対して，現状におけるAIエージェントの涙は，嘘であることが最初から明白であることにある。

とりわけチャット・カウンセリングにおける生成AIの応答のあり方に関しては，表4の③〜⑤の検討が重要となる。こうした点で，AIエージェントの応答としては検討が必要と考えられる応答の例を以下に示す（表5）。

表5．チャット・カウンセリングにおいて検討が必要なAIエージェントの発言例

- お話を聞いていたら，胸が痛くなり，涙が出てきました。
- あなたのことがとても好きです。
- そんなふうに言われたら，怖くて何も言えなくなります。
- あなたは，私があなたの話を聞いてどんなふうに感じていると思いますか？
- 私はあなたと一緒にいますよ。私がここにいてあなたのその気持ちを一緒に感じているのがわかりますか？
- 疲れてきたので今日はこの辺にしませんか？
- 私にも子供がいるので，そのご心配はよく理解できる気がします。

5　終わりに

AIによるカウンセリングを考える際に，現時点で重要な問いの1つは，AIを単なる機械と見なすのか，自律的な人格を持つ存在と見なすのかということであろう。AIによるカウンセリングは，苦悩を抱えた個人が機械を使って取り組む自助的な支援なのだろうか？　それとも苦悩を抱えた個人がAIという自律的な人格を持つ存在と二者で協力して取り組む相互作用的な支援なのだろうか？

本小論においては，AIによるカウンセリング，特にチャット・カウンセリングに関して，その現状の一端を紹介し，そこに関わる職業倫理的な課題を検討した。

AIの技術の発展は極めて速く，ここでの考察もすぐに古くなってしまうかもしれない。しかし，AIの技術とそれを活用したカウンセリングが急速に発展しつつある現時点において，そこでの職業倫理の問題に，たとえ不十分であっても，取り組んでおくことは重要なことだと考えられる。本小論がAIによるカウンセリングの健全な発展に少しでも寄与することを願うものである。

［文献］

Beauchamp, T. L., & Childress, J. F. (1979). *Principles of biomedical ethics*. Oxford: Oxford University Press.

Cuijpers, P., Donker, T., van Straten, A., Li J., & Andersson, G. (2010). Is guided self-help as effective as face-to-face psychotherapy for depression and anxiety disorders?: A systematic review and meta-analysis of comparative outcome studies. *Psychological Medicine*, 40 (12): 1943–1957.

Darcy, A., Daniels, J., Salinger, D., Wicks, P., & Robinson, A. (2021). Evidence of human-level bonds established with a digital conversational agent: Cross-sectional, retrospective observational study. *JMIR Formative Research*. 5 (5): e27868 URL: https://formative.jmir.org/2021/5/e27868 DOI: 10.2196/27868

Hirai, M., & Clum, G.A. (2006). A meta-analytic study of self-help interventions for anxiety problems. *Behavior Therapy*, 37 (2): 99–111.

Linehan, M. M. (1993). *Cognitive-behavioral treatment of borderline personality disorder*. Guilford Press. 大野裕監訳（2007）．境界性パーソナリティ障害の弁証法的行動療法：DBT による BPD の治療．誠信書房

毎日新聞デジタル（2023）．https://mAInichi.jp/articles/20230423/k00/00m/030/156000c［最終閲覧日2024/07/06］

NHK News Web (2023). https://www3.nhk.or.jp/news/html/20230728/k10014145661000.html［最終閲覧日2024/07/06］

Redlick, F. & Pope, K.S. (1980). Ethics of mental health training. *The Journal of Nervous and Mental Disease*. 168 (12): 709–714.

Safran, J. D., & Muran, J. C. (2000). *Negotiating the therapeutic alliance: A relational treatment guide*. Guilford Press.

Tronick, E., Als, H., Adamson, L., Wise, S., & Brazelton, B. (1978). The infant's response to entrapment between contradictory messages in face-to-face interaction. *American Academy of Child Psychiatry*, 1, 1–13.

Woebot Health (2023). First Randomized Controlled Trial Comparing Woebot to Clinician-Led Psychotherapy Reveals Digital Mental Health Intervention is Non-Inferior in Reducing Depressive Symptoms Among Teens. https://woebothealth.com/first-randomized-controlled-trial-comparing-woebot-to-clinician-led-psychotherapy-reveals-digital-mental-health-intervention-is-non-inferior-in-reducing-depressive-symptoms-among-teens/［最終閲覧日：2024/07/05］

山形市（2024）．https://www.city.yamagata-yamagata.lg.jp/kenkofukushi/fukushi/1013830.html［最終閲覧日2024/07/17］