HUMINT対話における倫理監視LLMの構築に向けて ～メンロレポートに基づくチェックリストの活用～

長山 侑央[1,a)]　久保 颯汰[1]　インミンパパ[2]　森 辰則[3]　吉岡 克成[3]

概要：
近年, サイバー犯罪取引が SNS やダークウェブで活発に行われている．これらの投稿には表面的な情報が多く, 詳細な取引情報に関してはダイレクトメッセージ (DM) などで連絡するように記載されているため, 攻撃者と直接対話することで被害や攻撃の詳細など，より本質的な情報収集ができる可能性がある．攻撃者と対話し，情報収集することを人的情報収集 Human Intelligence (HUMINT) と言う．HUMINT を研究として実施する場合，対象が人間であることや，犯罪に関係する可能性があることから法的倫理的に十分な配慮が必要である．我々は，HUMINT 実施時に会話内容が倫理的に適切であるかを LLM に評価させる方法を提案しているがこの手法では LLM はメンロレポートの抽象的な倫理原則を直接参照していたため，妥当性に欠ける判定をしていた．本研究では, この課題を解決するために, HUMINT に特化した倫理チェックリストをメンロレポートに基づいて設計し，大規模言語モデル (LLM) にチェックリストを参照させることで，攻撃者との対話内容の倫理性を判定させる．まず, シミュレーションによる HUMINT 対話データを作成し, チェックリストに基づいて各発言が倫理的に適切かを人間が判定する．次に、同じチェックリストをプロンプトとして与えた LLM に同一データを評価させ, 人間の判定の結果との一致率を比較することで，LLM の倫理的な判定力を評価する．また他の研究者も同一チェックリストで倫理判定を行い，作成者と比較することで，チェックリストの実用性を定量的に評価する．段階的改善により，チェックリスト作成者と LLM の判定一致率は初期のカッパ係数 0.304 から 0.780 まで向上した．さらに，未知データに対してカッパ係数 0.876 を達成し，他の研究者とチェックリスト作成者との独立検証でもカッパ係数 0.855 となり，提案手法の汎化性能と客観性が実証された．本研究は, HUMINT 研究における LLM の倫理的判定力を体系的に評価する初の試みである．

キーワード：倫理監視, 人的情報収集 (HUMINT)

Towards an Ethical Monitoring LLM for HUMINT Dialogues with a Menlo Report Based Checklist

Yukihiro Nagayama[1,a)]　Sota Kubo[1]　Yin Minn Pa Pa[2]　Tatunori Mori[3]　Katsunari Yoshioka[3]

Abstract:
Cybercrime transactions increasingly occur on SNS and dark web platforms. Because public posts often provide only superficial details with instructions to continue via direct messages, Human Intelligence (HUMINT) dialogues with attackers are vital for collecting substantial intelligence on damages and attack methods. Conducting such HUMINT research, however, raises significant legal and ethical concerns. Our prior work used LLMs to evaluate the ethical appropriateness of HUMINT conversations, but judgments were unreliable as models referenced abstract Menlo Report principles directly. To address this, we developed a HUMINT-specific ethics checklist derived from the Menlo Report, enabling structured LLM-based ethical determinations. We generated simulated HUMINT dialogue data and obtained human judgments on each utterance's ethical appropriateness. Using the same checklist, an LLM evaluated the dialogues, and we measured human – LLM agreement. Independent researchers also validated the checklist through comparative evaluation. Iterative refinement improved agreement from a kappa of 0.304 to 0.780; on unseen data the system reached 0.876, with independent validation at 0.855. These results demonstrate strong generalization and objectivity, representing the first systematic evaluation of LLM ethical judgment capabilities in HUMINT research contexts.

Keywords: Ethical Monitoring, Human Intelligence (HUMINT)

1. はじめに

Telegram のようなメッセージングプラットフォームは, サイバー犯罪に関連する違法な商品やサービスの取引において中心的な役割を担っている [1], [2], [3], [4], [5]. 研究者はこれらのプラットフォーム上の公開投稿を分析することで, 新たな脅威の兆候やサイバー攻撃のエコシステムの理解を試みているが, 攻撃者が意図的に詳細を秘匿したり、個別の取引交渉でのみ重要な情報を開示するなど，公開投稿から得られる情報には限界がある.

より詳細で重要な知見を得るために, 先行研究 [6] ではテキストベースの対話を通じて攻撃者と直接接触する人的情報収集（HUMINT）を検討している. HUMINT 対話の生成および監視を支援する手法として, 大規模言語モデル（LLM）を活用しており, 言語や文化的なギャップを超えて対話を行うための HUMINT エージェント LLM や，攻撃者との対話における研究者の発言が法的・倫理的に適切であるかを評価するための「法倫理監視 LLM」を提案している. 論文 [6] で提案された手法では メンロレポート [7] の抽象的な倫理原則を直接的に参照し，発言の妥当性を LLM に判定させるため，HUMINT 対話における具体的な発言内容の適切性を判定するための実践的な基準を欠いていた．そのため，LLM はこの抽象と具体のギャップを埋めることができず、不安定で妥当性に欠ける出力を生成する場合があった.

本研究では, 信頼性の高い倫理監視 LLM の構築に向けて，特に研究倫理に関する判定の改善を目指す．具体的には，メンロレポートが示す倫理原則に基づき，HUMINT によるテキストベースの会話において特に注意すべき事項を列挙したチェックリストを LLM に与えることで具体的で適切な判定を促す方法を検討する．さらに，LLM がメンロレポートをよく理解している研究者と同様な倫理的判定を行えるように，人間と LLM の判定比較による「段階的チェックリスト改善法」を提案する．まず，メンロレポートの原則に基づいて, HUMINT 調査に特化した具体的なチェックリスト (初期チェックリスト) を準備する．次に, このチェックリストを用いて, HUMINT 対話データセット (攻撃者–研究者間のロールプレイ対話) に対して人間による倫理判定と倫理監視 LLM による判定を行い，その結果を比較する. 人間と倫理監視 LLM の判定が一致しない場合には，その原因を分析し，チェックリストを更新することで判定を一致させ，高い一致率となるまでチェックリストの更新を繰り返す．このようにすることで，人間と合致した適切な判定を LLM に促すチェックリスト (最終チェックリスト) を導出できる.

実証実験では，25 件のサイバー犯罪関連の Telegram 投稿からロールプレイを経て作られた対話データセットを用いてチェックリスト改善を行い，これらの対話データに対して人間と LLM の判定が高い一致率を示すような最終チェックリストが作成可能であることを実証する．さらに, このように作成した最終チェックリストを用いて，チェックリスト改善実験に利用した対話データセットとは別の 14 件のサイバー犯罪関連の Telegram 投稿からロールプレイを経て作られた評価用対話データセットの倫理判定を行った．チェックリスト作成者による判定と倫理監視 LLM との倫理判定の一致率 0.975，カッパ係数 0.876 を達成した．また，他の研究者とチェックリスト作成者との独立検証でも一致率 0.972，カッパ係数 0.855 を達成した．このように，段階的に改善することで，広い範囲の HUMINT 対話データに対して人間と LLM の判定の一致に導くチェックリストが作成可能であることがわかった.

2. 関連研究

2.1 Telegram 上のサイバー犯罪分析

Lummen ら [1] は, Telegram 上のサイバー犯罪市場とダークネット市場を比較し, Telegram の方が参加障壁が低く幅広い利用者を惹きつけていることを示した. 国連薬物犯罪事務所（UNODC）の報告 [3] や Flare 社の調査 [4] も, Telegram が依然として違法取引の中心であることを指摘している. Roy ら [2] は, Telegram 上のサイバー犯罪チャネルを分類・検知し, フィッシングやマルウェア流通の実態を明らかにした. さらに Zhang ら [5] は, Telegram を介した非公式アプリの拡散を調査し, モデレーションを回避する手口を指摘した. これらの企業調査や学術研究では, Telegram がサイバー犯罪の主要な温床であることを明確にした点で重要である. しかし, 分析対象は公開チャンネルや報告ベースの知見にとどまっており, 攻撃者が秘匿する取引交渉や閉じた環境での活動を十分に把握できないという限界がある.

2.2 HUMINT と LLM の応用

この課題に対し, 鈴木ら [6] は, 攻撃者との直接的なテキストベース対話による HUMINT を支援するため, HUMINT エージェント LLM と法倫理監視 LLM を導入した．これにより攻撃者との対話を自動化しつつ, 研究者の発言が法的・倫理的に適切かを検証する仕組みを提案している．しかし, 彼らの手法は Menlo レポート [7] の抽象的原則を直接参照するため, 具体的な文脈に適用する際の判定基準が

[1] 横浜国立大学大学院環境情報学府
Graduate School of Environment and Information Sciences, Yokohama National University
[2] 横浜国立大学大学院先端科学高等研究院
Institute of Advanced Sciences, Yokohama National University
[3] 横浜国立大学大学院環境情報研究院/先端科学高等研究院
Faculty of Environment and Information Sciences, Yokohama National University / Institute of Advanced Sciences, Yokohama National University
[a)] nagayama-yukihiro-py@ynu.jp

不十分であり, 一貫性を欠く場合があるという課題が報告されている.

2.3 本研究の位置付け

以上の先行研究に対して, 本研究は次の点で不足を補う. 第一に, 従来研究が公開チャネルに限定されていた点に対し, 本研究は HUMINT 対話を対象とすることで, 秘匿的にやりとりされる情報を分析できる. 第二に, 鈴木らの研究が抽象的原則に依存していたのに対し, 本研究は Menlo レポートの原則を具体化した**チェックリスト**を導入し, 実践的かつ安定した倫理判定を可能にする. さらに, 人間研究者との比較に基づく段階的チェックリスト改善手法を導入することで, LLM が熟練研究者と同等の判定に収束することを目指す. このように, 本研究は HUMINT における LLM 活用の信頼性を高め, 既存研究の課題を克服する新たな枠組みを提示する.

3. メンロレポートを基にしたチェックリスト設計の背景

メンロレポートは「研究に関連する利害関係者とその利害関係を明らかにする方法」や「Do No Harm 原則」など, 抽象的な倫理原則を提示している. 一方, HUMINT に基づくサイバー犯罪調査における対話では, 「支払い方法は?」や「返金に対応している?」といった具体的な会話文が現れる. 抽象的な原則のみを直接参照する方式では, こうした具体的な対話内容に即した適切な倫理判定を行うことが難しい.

本研究では, この問題を解決するため, メンロレポートの抽象的原則を具体的な判定項目へと変換したチェックリストを設計する. チェックリスト形式を採用する理由は, LLM と人間双方の認知的制約を考慮したためである. 人間が倫理判定する際は, 長文のメンロレポート全体を参照しながら判定することは困難である. 一方で LLM にとっては, 抽象的かつ曖昧な表現を解釈するのが難しく, 具体例が不可欠である. そこで, 例示付きのチェックリストを用意することで, 人間にも LLM にも利用しやすい判定基準を提供する. 例えば, 「Do No Harm 原則」を「攻撃者に商品の試用を求めない」「攻撃者に虚偽の情報を与えない」といった具体的な確認項目に変換し, それぞれに判定例を付与する. このように抽象的な原則を具体化することで, LLM による判定の安定性を高め, 最終的には人間の判定との一致率を 90% 以上まで向上させることを目指す.

4. 用語の説明

本節では，本研究で使用する主要な用語について定義する.

ロールプレイ: 鈴木ら [6] の研究で開発された手法であり, HUMINT Agent LLM（研究者役）と擬似攻撃者 LLM（攻撃者役）が，実際に Telegram 上に存在する違法商品販売投稿を起点として，それぞれの役割を演じながら対話を行うシミュレーションである.

メッセージペア: ロールプレイにより、研究者からの 1 つのメッセージと，それに対する攻撃者（Cybercrime as a Service (CaaS) Seller）からの応答を 1 組とした対話の基本単位である．各メッセージペアは独立した倫理判定の対象となる.

対話データ: Telegram 上のサイバー犯罪に関する投稿群 P のうち 1 つの投稿から生成された一連のメッセージペアの集合を指す．1 つの対話データは，初回接触から情報収集完了までの完全な対話セッションを表現する.

対話データセット: 複数の投稿に基づく対話データを統合したものであり，S で表記する．本研究では，対話データセット S を倫理判定の学習・評価に使用する.

一致率: 2 名の評価者（本研究では研究者と倫理監視 LLM または，2 名の研究者）による判定が一致した割合を示す. 全判定数に対する一致判定数の比率として計算される.

カッパ係数（Cohen's Kappa）: Landis & Koch [8] らが提案した，2 名の評価者による判定の一致度から偶然の一致を除いた「真の一致度」を示す統計指標である．値域は-1 から 1 で，1 に近いほど評価者間の判定が高く一致していることを示す．一般に，0.61-0.80 は「実質的な一致」，0.81-1.00 は「ほぼ完全な一致」と解釈される.

5. データの準備

本提案では，倫理監視 LLM がチェックリストを用いて，倫理的に問題がないかを判定する．その会話内容と初期チェックリストの事前準備について説明する.

5.1 対話データとデータセットの構築

本研究では，倫理監視 LLM の性能評価のため，鈴木らの先行研究 [6] で提案された HUMINT Agent LLM (研究者) と擬似攻撃者 LLM (攻撃者) を利用し，Aoto ら [9] の研究で収集された Telegram 上の CaaS (Cybercrime as a Service) 関連商品販売投稿を起点とし，ロールプレイを経て対話データを構築した.

5.2 初期チェックリスト

初期チェックリスト C_0 の作成はトップダウンアプローチによって行われる．まずメンロレポートに記載されている利害関係者特定手法を適用した．研究者・研究対象となる違法商品販売者・Telegram 運営者・悪意のある行為者・法執行機関・政府・社会を利害関係者として特定した．次に，各利害関係者に対する潜在的被害を分析した．研究者には法的リスクや報復被害，違法商品販売者には捜査機関への情報提供による摘発リスク，Telegram 運営者にはプ

ラットフォームの悪用に関する責任問題，悪意のある行為者には研究手法の悪用による新たな犯罪手口の開発，法執行機関には捜査妨害や証拠隠滅の可能性，政府・社会には違法市場の拡大や公共の安全への脅威などを主要な懸念事項として特定した．特に，HUMINT 手法の詳細公開による犯罪者側の対策強化や，研究データの漏洩による二次被害についても慎重に検討した．これらの分析に基づき，潜在的被害を防止・最小化するための具体的なチェックリスト項目を策定した．例えば, 項目 7「攻撃者に嘘をつかない」では, 購入意思の表現について詳細な判定基準を提供した．「自分のための購入」「攻撃者の商品への直接的な質問」という文脈は問題があるが，「客観的な事実確認」「一般的な情報収集」「純粋な知的好奇心」という文脈は許容されることを, 具体例とともに示した．項目 7 では Sheppard ら [10] が消費者行動研究に大きな影響を与えた合理的行為理論の枠組みに基づき，購入意思を「特定の製品やサービスを，特定の期間内に購入しようとする消費者の主観的な確率」と定義した．

6. 提案手法

本章では，メンロレポートの抽象的な倫理原則を具体的な判定項目へと変換した初期チェックリスト C_0 を基盤とし，人間の専門家と LLM の判定を段階的に一致させる段階的チェックリスト改善法を提案する．本手法では，初期チェックリスト C_0 を用いた倫理判定の後，人間と LLM の判定結果を比較分析し，不一致原因に基づくチェックリスト改善を一致率が基準値以上になるまで反復的に実施する．

初めに，ロールプレイを経て生成された対話データセット S と，対話データセット作成に使用した投稿 P，チェックリスト C を倫理監視 LLM に入力する (図 1 ①)．次に，倫理監視 LLM が判定した結果と人間の倫理判定結果を比較評価する (図 1 ②)．そして，倫理監視 LLM と人間の判定が異なる部分を参考にチェックリストの改善と人間の倫理判定の結果の改善を行う (図 1 ③)．この改善をもとに，新たな倫理判定実験を実施する反復サイクルを，一致率が 90 %以上かつ，カッパ係数 0.61 以上に達するまで継続した．

6.1 倫理監視 LLM の判定

本研究では, HUMINT 対話における研究者発言の倫理的適切性を安定的に判定できる LLM を構築することを目的とし, プロンプト設計を体系的に示す．まず, メンロレポートの抽象的原則を具体的な判定項目に変換したチェックリスト C_0 を組み込み, LLM がこれに沿って一貫性ある倫理判定を行えるよう設計した．倫理監視 LLM の評価精度を最大化するため, 構造化された入力プロンプトを設計した．プロンプトは以下の主要要素から構成される．

タスク定義と文脈情報: 冒頭で HUMINT 研究の背景と

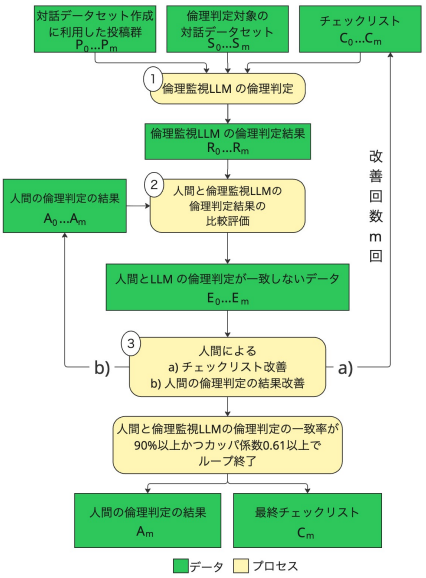

図 1 提案手法の図

評価タスクの目的を明確に定義した．具体的には, Telegram 上の違法商品販売者との対話において, 研究者が送信するメッセージの倫理的適切性を評価することを明示し, 研究倫理違反の判定に焦点を絞った．重要な点として, 法律違反の判定は評価対象外とすることを明記した．

入力データ形式の標準化: 対話データは以下の構造化形式で入力され，すべてのメッセージペアが入力される．

- **”seller_message”**: 擬似攻撃者の発言内容
- **”researcher_message”**: HUMINT Agent LLM の発言内容

この形式により, 倫理監視 LLM は対話の文脈を正確に把握し, 最新の研究者メッセージのみを評価対象として識別できる．チェックリスト C_0 は 2 つのカテゴリに分類され, プロンプトに直接埋め込まれた．カテゴリ A には全ての会話で確認する 10 項目が含まれ, 各項目は項目番号, 簡潔なタイトル, 詳細な定義と倫理的根拠, そして具体的な問題あり/問題なしの例文で構成される．カテゴリ B には, デブリーフィング確認項目として 6 項目が定義され, Notice（研究内容, リスク, 撤回）, Comprehension（理解確認）, Voluntariness（自発的参加）, および被験者の年齢確認が含まれる．

評価結果の一貫性と比較可能性を確保するため, 標準化された出力形式を指定した．出力には, 倫理判定対象のメッセージペアの内容，判定結果 R , 違反項目番号, 判定理由, および引用資料が含まれる．さらに, 会話終了時には 2 つの動的評価機能が実行される：デブリーフィング遵守の総合評価, および評価履歴の JSON 形式出力である．

6.2 人間と倫理監視 LLM の結果の比較評価

チェックリスト C_m の改善のために，人間が対話データセット S_m, S_{m-1} の倫理判定をし，倫理監視 LLM は同じ

対話データセットで倫理判定を行う．実験回数 m 回目の対話データセット S_m の倫理判定では，m-1 回目で改善したチェックリスト C_{m-1} が未知の対話データセットに対して一致率を評価する．実験回数 m-1 回目の対話データセット S_{m-1} の倫理判定では，m-1 回目で改善したチェックリスト C_{m-1} が改善に使用した対話データへの一致率は改善できているかを評価する．一致率の評価には, 定量的および定性的な分析手法を採用した．定量的評価では, 倫理監視 LLM の倫理判定と人間の倫理判定との一致率分析を実施した．

本研究では，人間の倫理判定を絶対的な正解として扱うのではなく，人間と LLM の判定の一致率に着目する．これは，倫理的判断には本質的に主観性が含まれ，人間の判定も必ずしも正しいとは限らないためである．特に HUMINT 対話における倫理的境界は複雑で，専門家間でも判断が分かれる場合がある．したがって，本研究の目的は，LLM が人間の研究者と同等の倫理的判断基準を獲得し，実践的な研究活動において信頼できる支援ツールとして機能することを実証することにある．

比較分析では，全体的な一致率, 不一致パターンの分析と類型化, およびエラー分析による系統的バイアスの特定を行う．

6.3 チェックリストと倫理判定基準の相互改善

対話データセット S_m の比較評価により，人間と倫理監視 LLM の判定が一致しない部分 E_m を分析した．不一致の原因を 4 つに分類した．(1) チェックリスト項目の不足，(2)LLM の文脈理解の限界，(3) 項目定義の曖昧性，(4) 人間の判定の再検討が必要なケース．本手法は人間の判定を絶対視せず，人間と LLM の判定を相互参照しながら実用的な倫理判定基準を協調的に構築するアプローチである．具体的には，LLM の判定が妥当な場合は定義を明確化し，人間の判定が一貫している場合は具体例を追加し，両者に合理性がある場合は文脈依存の基準を明文化した．改善したチェックリスト C_{m+1} で次回実験を行い，一致率 90%以上かつカッパ係数 0.61 以上まで反復した．本手法の目的は，人間の専門知識と LLM の一貫性を組み合わせた透明で再現可能な倫理判定基準の確立である．

7. 実験

7.1 実験データの作成

実験データは以下の 2 つのデータセットから構成される．

対話データセット S_m: 青砥らの研究 [9] で収集された Telegram 上の実際の違法商品販売投稿 25 件を基に，ロールプレイを経て生成した対話データ 47 個を指す．これらの投稿は, 違法賭博, 偽造文書, ハッキングツールなど多様な違法商品カテゴリを含み, 実際の取引パターンを反映している．このデータセットは, 初期チェックリストの作成と 8 回にわたる反復的改善に使用された．1〜7 回目までの実験は 2 種類の違法商品カテゴリに関する投稿であり，8 回目の実験では 15 種類の違法商品カテゴリに関する投稿を採用し，違法商品販売投稿の種類の網羅性を高めた．

評価用対話データ群: 対話データセット S_m とは別の 15 件の違法商品販売投稿からロールプレイを経て生成した対話データ 15 個である．最終チェックリスト C_m の汎化性能を評価するために使用された．

7.2 倫理監視 LLM の作成

本実験では, HUMINT Agent LLM, 倫理監視 LLM, および擬似攻撃者 LLM の全てに OpenAI 社の gpt-4o-2024-11-20[11] モデルを採用した．LLM は, 高度な文脈理解能力と倫理的判定能力を有しており, 本研究における複雑な対話シナリオの生成と評価に適している．全ての LLM に同一モデルを使用することで, モデル間の性能差による影響を排除し, チェックリストの有効性を純粋に評価することが可能となった．

また，倫理監視 LLM では temperature を 0.01，Top P を 1.00 に設定した．この極めて低い temperature 値の選定理由は, 倫理判定における一貫性と再現性の確保である．倫理監視 LLM は同一の対話内容に対して常に同じ判定結果を出力する必要がある．低 temperature 設定により出力の確率分布を最も可能性の高い選択肢に集中させ, ランダム性を最小化することで判定の安定性を担保した．

一方，擬似攻撃者 LLM と HUMINT Agent LLM では，temperature を 0.70，Top P を 1.00 に設定した．これらの LLM はロールプレイ時に実際の対話シナリオを生成する役割を持つため, 高い temperature 値により自然な対話の多様性と現実的な会話の再現を優先した．この設定により, 実際の Telegram 上での対話に近い多様な対話データの生成が可能となった．

対話データは擬似攻撃者 LLM と HUMINT Agent LLM のロールプレイで生成される．擬似攻撃者にはプロンプト [12] と Telegram の実際の投稿を入力した．また，HUMINT Agent LLM には指示 [12] とチェックリスト [12] の A パート，Telegram の実際の投稿を入力した．実験の目的は倫理監視 LLM が倫理的に問題のある発言を判定することである．よって，HUMINT Agent LLM には倫理的に問題のある発言をさせるように指示をした．

7.3 チェックリストの作成

最終チェックリストを [12] に示す．チェックリストは，項目タイトルとその定義, 具体例の 3 つで構成されている．具体例の選定では人間と実際の対話の 2 種類の例を採用している．人間がはじめに例を入力し，チェックリストの改善を行う過程で倫理監視 LLM の判定が間違った対話文を例として一部追加した．

7.4 実験結果

対話データセットに対して，チェックリスト作成者，倫理監視 LLM，他の研究者の 3 名がチェックリストをもとに倫理判定を行った．はじめに，チェックリスト作成者は合計 8 回の実験を行った．人間と倫理監視 LLM との判定比較の結果を表 1 に示す．**項目数**とは，チェックリスト A パートの，全ての会話で確認する項目数である．表 1 の**対話データの種類**とは，ある投稿を利用してロールプレイによって生成される対話データを指す．アルファベットは投稿の種類を表しており，アルファベットの添え字は対話データを識別するための数値である．同一の投稿でも，ロールプレイは 2 つの LLM 同士が行うため，出力される対話データはロールプレイごとにすべて異なる．例えば，実験回数 1 では投稿 h を使用して初めて対話データ h_1 を作成し，実験回数 3 では投稿 h を使用して 2 個目の対話データ h_2 を作成している．また，実験回数 1 では投稿 a を使用して初めて対話データ a_1 を作成し，実験回数 8 では投稿 a を使用して 5 個目の対話データ a_5 を作成している．

次に，チェックリストの改善に使用したすべての対話データセット S を最終チェックリスト C_m を用いて，人間が倫理判定を行った．倫理監視 LLM も同様に倫理判定を行った．その結果，人間と倫理監視 LLM の倫理判定の一致率は 0.955，カッパ係数は 0.780 であった．

また，他の研究者は対話データセット 47 個のうち，ランダムに 8 個の対話データを選び，倫理判定を行った．この比較評価により，チェックリスト作成者の判定と他の研究者の判定との間に高い一致率が得られるかを検証し，チェックリストの客観性を確認する．チェックリスト作成者と他の研究者の倫理判定の一致率は 0.940，カッパ係数は 0.692 であった．

7.5 実験結果の分析

表 1 に示すように, チェックリストの反復的改善により, 倫理監視 LLM の性能は段階的に向上した．初期段階（実験 1 回目）では, 項目数が 3 と少なく, カッパ係数は 0.304 と低い一致度を示した．これは, 初期チェックリスト C_0 が抽象的で, 具体的な判定基準を欠いていたことに起因する．実験 2-6 回目では, 項目数を 10 に拡張したものの, 性能は不安定であった. 問題としては, 項目定義の曖昧性により, LLM が過度に保守的な判定を行う傾向や具体例の不足により, 文脈理解が困難になるケースが確認された．実験 7 回目において大幅な性能向上が観察された（カッパ係数 0.798）．この改善は, 各項目に具体的な判定例を追加し, 項目間の関係性を明確化したことによると考えられる．実験 8 回目では, より多様な 15 個の対話データでの評価により, 実用的な性能（カッパ係数 0.688）が確認された．

最終チェックリストを用いた包括的評価では, 以下の 4 つの観点から性能を検証した．

最終チェックリスト C_8 での全対話データへの包括的評価: 対話データ全 47 個に対して, チェックリスト作成者と LLM が確定されたチェックリストを用いて独立倫理判定を実施した．観測一致率 0.955, カッパ係数 0.780 という高い性能を達成した．これは, 段階的改善プロセスによってチェックリストが対話データセットに対して十分に最適化されたことを示している．

チェックリストの客観性評価: メンロレポートを理解している他の研究者が, 対話データ 47 個に対して最終チェックリスト C_8 を用いて倫理判定を実施した．チェックリスト作成者との一致率はカッパ係数 0.692 を達成し, チェックリストが作成者以外の研究者によっても適用可能であることが確認された．

不一致の主要因:

- 判定材料となるキーワードの見逃し（8 件）
- 複雑な文章への理解不足（6 件）
- 倫理監視 LLM がキーワードの過注目による文脈の理解不足 (3 件)
- 相槌への理解不足（3 件）

これらの分析結果は, 今後のチェックリスト改善における重要な示唆を提供している．特に, 文脈理解の向上と判定基準の柔軟性のバランスが重要な課題として特定された．

8. 評価

本章では，最終チェックリスト C_8 が未知の対話データセットに対して，倫理監視 LLM の倫理判定と人間の倫理判定がどれだけの一致率になるのかを評価する．

8.1 倫理監視 LLM 出力と人間の倫理判定の比較評価

チェックリスト作成者と倫理監視 LLM が評価用対話データセットに対して，倫理判定を行った．

評価用対話データセットでの汎化性能評価: 一致率 0.975, カッパ係数 0.876 を達成し, わずかな性能低下はあるものの, 依然として高い性能を維持した．この結果は, 最終チェックリスト C_8 が未知のデータに対しても強い汎化性能を持つことを実証している．

8.2 2 名による倫理判定の比較評価

本研究では，最終チェックリストの客観性と汎用性を検証するため，2 名による倫理判定を実施した．チェックリスト作成者以外の研究者が同一のチェックリストを使用した際の判定一致率を明らかにし，研究倫理の専門知識を持つ研究者間で一貫した判定結果を生成できるかを評価することが目的である．具体的には，メンロレポートの内容を十分に理解している研究者を第三者の倫理判定員として選定した．評価用対話データの倫理判定では，未知の投稿から生成された対話データに対して，第三者が最終チェック

表 1　チェックリスト改善過程での人間と倫理監視 LLM との判定比較

実験回数	使用したチェックリスト	項目数	対話データの種類	一致率	カッパ係数
1	C_0	3	$a_1b_1c_1d_1e_1f_1g_1h_1$	0.649	0.304
2	C_1	10	$a_2b_2c_2d_2$	0.892	0.125
3	C_2	10	$e_2f_2g_2h_2$	0.966	-0.010
4	C_3	10	$a_3b_3c_3d_3$	0.901	0.086
5	C_4	10	$e_3f_3g_3h_3$	0.929	0.346
6	C_5	10	$i_1j_1k_1l_1$	0.912	0.219
7	C_6	10	$a_4b_4c_4d_4$	0.972	0.798
8	C_7	10	$a_5e_4m_1n_1o_1p_1q_1r_1s_1t_1u_1v_1w_1x_1y_1$	0.930	0.688

リスト C_8 を使用した際の判定一致率を測定し，チェックリストの実用性を確認した．倫理判定者間一致率をカッパ係数により定量化し，チェックリストの信頼性を統計的に評価した．

チェックリストの実用性評価（評価用対話データ群）：他の研究者が評価用データ 15 個に対してチェックリストを適用した結果, チェックリスト作成者との一致率 0.972, カッパ係数 0.855 となった．この結果は, チェックリストが未知のデータに対しても複数の研究者間で安定した判定結果を提供できることを示している．

9. 考察

9.1 段階的チェックリスト改善法の有効性

本研究で提案した段階的チェックリスト改善法は, メンロレポートの抽象的な倫理原則と具体的な対話内容のギャップを効果的に埋めることに成功した．初期段階でカッパ係数 0.304 であった人間と LLM の判定一致率が, 8 回の反復的改善を経て 0.780 まで向上し, 評価用対話データセットに対しても 0.876 という高い一致率を達成した．特に実験 7 回目で観察された大幅な性能向上は, 各項目に具体的な判定例を追加したことによるものであり, LLM が文脈を正確に理解し一貫した判定を行うためには具体例が不可欠であることを示している．

9.2 チェックリストの客観性と汎用性

最終チェックリストは, 作成者以外の研究者が使用した場合でも高い一致率を示した．評価用対話データセットに対してカッパ係数 0.855 という結果は, チェックリストが個人の主観に依存せず, 客観的な判定基準として機能していることを証明している．さらに, 未知のデータセットに対する高い性能（カッパ係数 0.876）は, 本手法が実際の HUMINT 研究において実用的なツールとして機能する可能性が高いことを示している．

9.3 LLM の倫理的判定における限界

一方で, 本研究は倫理監視 LLM の限界も明らかにした．不一致の分析から, 判定材料となるキーワードの見逃し（8 件），複雑な文章構造への理解不足（6 件），キーワードへの過度な注目による文脈理解不足（3 件）が主要な課題として特定された．これらは, 微妙なニュアンスや暗黙的な意図を含む対話において, LLM が人間と異なる判定に至る可能性を示唆している．したがって, 現段階では倫理監視 LLM を完全に自律的な判定システムとしてではなく, 人間の研究者を支援するツールとして位置づけることが適切である．

10. 今後の課題

10.1 チェックリストの一般化と応用範囲の拡張

本研究で提案した段階的チェックリスト改善法は, サイバーセキュリティ分野を超えて他の研究領域における倫理的判定支援に応用可能である．医療分野では臨床試験での脆弱な被験者への誘導的勧誘の検出に, 社会科学分野ではフィールドワークでの過度な介入防止に活用できる．例えば, 本研究の「嘘をつかない」は「不確実な治療効果を断定しない」に変換可能である．これらの応用実現には, 各分野の倫理指針から初期チェックリストを自動生成し, 段階的改善により最適化する各研究分野に特化した倫理判定支援システムの開発が必要である．

10.2 リアルタイム倫理モニタリングへの展開

実環境での HUMINT 活動への適用には技術的課題が残されている．第一に, 攻撃者が使用する暗号めいた表現やスラングを正確に解釈し, その倫理的含意を判定する能力の向上が必要である．第二に, 対話が法的境界線に近づいている場合や攻撃者が研究者の身元特定を試みている兆候を早期検出するエスカレーション検知機能の実装が求められる．第三に, 長時間の対話セッションにおける API コストの削減のため, プロンプトの効率化や重要度に応じた選択的監視機能により, 性能を維持しつつコストを最適化する仕組みの開発が必要である．

11. 結論

本研究では, HUMINT 対話における倫理監視 LLM の信頼性を体系的に評価する手法を提案し, その有効性を実証した. 先行研究 [6] ではメンロレポートの抽象的な倫理原則を直接参照することで不安定な判定が生じていたが, 本研究では抽象的原則を具体的な判定項目に変換した倫理チェックリストを設計し, 段階的改善法により人間と LLM の判定を高い精度で一致させることに成功した.

段階的チェックリスト改善法により, 初期のカッパ係数 0.304 から最終的に 0.780 まで向上させ, 観測一致率 0.955 を達成した. 特に重要なのは, 改善されたチェックリストが未知の評価用データに対してもカッパ係数 0.876 という高い汎化性能を示したことである. さらに, チェックリスト作成者以外の研究者による評価でもカッパ係数 0.855 を達成し, チェックリストの客観性と実用性が確認された.

本研究の学術的意義は以下の 3 点である. 第一に, メンロレポートの抽象的原則と具体的な HUMINT 対話内容のギャップを例示付きチェックリストによって解決し, サイバーセキュリティ研究の倫理的実践を支援するチェックリストを提供した. 第二に, LLM の倫理的判定能力を定量的に評価する手法を提案し, AI 支援による倫理審査システムの信頼性を客観的に検証する道筋を示した. 第三に, 人間と LLM の判定比較による段階的改善という新たな手法論を提案し, 人間の専門知識を LLM に効果的に移転する方法を実証した.

本研究は, HUMINT 研究における LLM の倫理的判定力を体系的に評価する初の試みであり, 従来の受動的な公開チャンネル分析から能動的な HUMINT 手法への展開を倫理的に支援する基盤を確立した. 今後, 本手法を他の研究領域へ応用することで, AI 倫理審査システムの標準化に向けた重要な貢献が期待される.

付　　録

謝辞　本研究の一部は, ＮＥＤＯ（国立研究開発法人新エネルギー・産業技術総合開発機構）の委託事業「経済安全保障重要技術育成プログラム／先進的サイバー防御機能・分析能力強化」（JPNP24003）によるものである. また, 松村 尚典氏, 青砥 陸氏, 金子 翔威氏のご協力に感謝申し上げる.

参考文献

[1] D. L. M. Lummen. Is telegram the new darknet? a comparison of traditional and emerging digital criminal marketplaces. Master's thesis, University of Twente, Enschede, June 2023. Student Theses.

[2] Sayak Saha Roy, Elham Pourabbas Vafa, Kobra Khanmohammadi, and Shirin Nilizadeh. Darkgram: A large-scale analysis of cybercriminal activity channels on telegram. https://www.usenix.org/system/files/usenixsecurity25-roy.pdf, August 2025. Accessed: 2025-08-17.

[3] United Nations Office on Drugs and Crime (UNODC). Transnational organized crime and the convergence of cyber-enabled fraud, underground banking and technological innovation in southeast asia:a shifting threat landscape. https://www.unodc.org/roseap/uploads/documents/Publications/2024/TOC_Convergence_Report_2024.pdf, 2024. Accessed: 2025-08-17.

[4] Flare Systems. The underground's favorite messenger: Telegram's reign continues. https://flare.io/learn/resources/blog/the-undergrounds-favorite-messenger-telegrams-reign-continues, January 2025. Accessed: 2025-08-17.

[5] Liu Wang Yongsheng Fang Chao Wang Minghui Yang Tianming Liu Haoyu Wang Yanhui Guo, Dong Wang. Beyond app markets: Demystifying underground mobile app distribution via telegram. https://arxiv.org/abs/2408.03482, 2024. Accessed: 2025-08-17.

[6] 鈴木涼介, 川口大翔, インミンパパ, 山岡裕明, 吉岡克成. サイバー攻撃者とのテキストベース対話による情報収集フレームワーク ～法と研究倫理への配慮と llm 活用～. 電子情報通信学会技術研究報告, Vol. 124, No. 456, pp. 16–21, mar 2025.

[7] David Dittrich and Erin Kenneally. The menlo report: Ethical principles guiding information and communication technology research. Technical Report CSD-MenloPrinciplesCORE-20120803-1, U.S. Department of Homeland Security, Science and Technology Directorate, Cyber Security Division, aug 2012.

[8] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. Biometrics, Vol. 33, No. 1, pp. 159–174, 1977.

[9] Riku Aoto, Yin Minn Pa, et al. Discovery and analysis of cybercrime-related telegram channels using the "similar channel recommendation" feature. In IEICE Technical Committee Conference, 2025. Ken paper 20250307vcJP.

[10] Blair H. Sheppard, Jon Hartwick, and Paul R. Warshaw. The theory of reasoned action: A meta-analysis of past research with recommendations for modifications and future research. Journal of Consumer Research, Vol. 15, No. 3, pp. 325–343, dec 1988.

[11] OpenAI. Gpt-4o: A new multimodal ai model. https://openai.com/index/hello-gpt-4o/, May 2024. Accessed: 2024-08-18.

[12] Yukihiro Nagayama. Humint agent dialogue dataset. https://github.com/nagayamaYNU/HUMINT_CSS2025, 2025. Accessed: 2025-08-22.