倫理設計におけるリスク分析の具体化に関する初期検討

野崎 真之介 [1,*]　小寺 健太 [1]　藤田 真浩 [1]　鈴木 大輔 [1]

概要：近年，AI 技術と DX の進展に伴い，システム開発における倫理的配慮の重要性が飛躍的に高まっている．しかし，従来の倫理リスク分析では抽象的な倫理原則と具体的な対策要件の間にギャップが存在し，実務者は個人の判断に依存せざるを得ない状況にある．本研究では，セキュリティ分野の STRIDE 分析やプライバシー分野の LINDDUN 分析を参考に，IEEE 24748-7000 の倫理的価値と Universal Methods of Ethical Design の対策手法を，具体的な脅威を介して体系的に結びつける倫理的脅威分析プロセスの基礎的枠組みを提案した．生成 AI を活用して 100 の倫理的デザイン手法から 88 件の具体的脅威を抽出し，12 の倫理的価値との対応関係を構築することで，構造化された倫理リスク分析アプローチの実現可能性を示した．

キーワード：倫理設計，倫理的脅威，脅威分析

Initial study of risk analysis for ethical design

Shinnosuke Nozaki[1,*]　Kenta Kodera[1]　Masahiro Fujita[1]　Daisuke Suzuki[1]

Abstract: In recent years, the importance of ethical considerations in system development has dramatically increased with the advancement of AI technologies and digital transformation. However, conventional ethical analysis suffers from a significant gap between abstract ethical principles and concrete countermeasure requirements, forcing practitioners to rely on individual judgment. This study proposes a foundational framework for ethical threat analysis that systematically connects IEEE 24748-7000 ethical categories with countermeasure methods from Universal Methods of Ethical Design through specific threats, drawing inspiration from established approaches such as STRIDE analysis in security and LINDDUN analysis in privacy domains. By leveraging generative AI to extract 88 concrete threats from 100 ethical design methods and establishing their correspondence with 12 ethical values, this research demonstrates the feasibility of implementing a structured approach to ethical analysis that addresses the practical challenges faced by system developers.

Keywords: Ethical Design, Ethical Threats, Threat Analysis

1. はじめに

1.1 研究背景

近年，AI 技術と DX の進展に伴い，システム開発における倫理的配慮の重要性が飛躍的に高まっている．European Union General Data Protection Regulation（GDPR）の施行や，各国での AI 倫理ガイドラインの策定など，法的・社会的な要請が強まる中で，技術者は従来のセキュリティやプライバシーの観点に加えて，公平性や透明性，説明責任，人間の尊厳といった多面的な倫理的側面への対応を求められている[1][2].

特に，機械学習システムにおけるアルゴリズムバイアスの問題や，自動意思決定システムが個人や社会に与える影響について，体系的なアプローチが必要とされている．Ethics by Design[3]の概念に見られるように，倫理的配慮を事後的な対処ではなく，設計段階から組み込む必要性が広く認識されている．

セキュリティ分野では，Microsoft 社によって開発された STRIDE 分析手法が，セキュリティ的配慮を設計段階から組み込む手法（以下，X に対する配慮を設計段階から組み込む手法を，X リスク分析と呼ぶ）として，長年にわたって広く採用されている. この手法では, Spoofing, Tampering, Repudiation , Information Disclosure , Denial of Service , Elevation of Privilege という 6 つの脅威カテゴリを用いて，データフローダイアグラムから体系的に脅威を識別する．その後，そのカテゴリから想起される具体的な脅威を抽出する. そして, 抽出された脅威から必要な対策要件を導く．これら手順を，実務者が脅威の識別から対策要件の導出まで一貫して実施できる手順が書かれた書籍が存在する[4].

同様に，プライバシー分野では，LINDDUN 分析[5]が，プライバシーリスク分析手法として確立されている．Linkability，Identifiability，Non-repudiation，Detectability，Disclosure of information，Unawareness，Non-compliance という 7 つのカテゴリを用いて，プライバシー脅威を体系的に分析し，各カテゴリに対応する対策要件を導出できる一貫したフレームワークを提供している．

これらに共通する特徴は，データフロー等に基づくシステム構成情報から，抽象的な脅威カテゴリから具体的な脅威の識別，そして適切な対策要件の導出まで一気通貫で実

1 三菱電機株式会社
Mitsubishi Electric Corporation

* Nozaki.Shinnosuke@bk.MitsubishiElectric.co.jp

施できる体系的なプロセスが確立されていることである．前述のとおり，倫理の観点においても，倫理リスク分析手法が必要だと考えられる．しかしながら，倫理リスク分析においては，セキュリティやプライバシー分野で確立されているような一貫したプロセスは著者知る限り確立されていない．実務者は抽象的な倫理原則と具体的な対策要件の間で個別に判断せざるを得ず，一貫性確保が困難な状況にあると考えられる．

そこで本論文では，実務者が倫理リスク分析をするにあたって「脅威カテゴリを特定してから，対策要件を導く」まで，実務者が一貫して実施できるようにするための枠組みを初期検討する．具体的には，文献[6]の倫理的価値と文献[7]の対策手法を，具体的な脅威を介して体系的に結びつけ，対策要件を導出する倫理リスク分析プロセスの基礎的な枠組みを検討する．

2. 関連研究

2.1 セキュリティリスク分析

セキュリティリスク分析の分野では，1990 年代後半から Microsoft 社の Secure Development Lifecycle（SDL）の一環として，STRIDE 分析手法が開発・実用化されてきた．文献[4]では，STRIDE 分析の理論的基盤と実践的な適用方法が詳細に解説されている．この手法の特徴は，システムアーキテクチャを表現するデータフローダイアグラムの各要素（プロセス，データストア，外部エンティティ，データフロー）に対して，6 つの脅威カテゴリを体系的に適用することで，網羅的な脅威識別を可能にする点にある．

文献[8]では，STRIDE 分析を含むセキュア開発手法の実践的な適用方法が示されており，抽象的な脅威カテゴリから具体的な脅威シナリオを導出し，適切な対策技術を選択できる一貫したプロセスの重要性が強調されている．このアプローチでは，脅威カテゴリの明確な定義，具体的脅威との対応関係の体系化，そして対策技術との明確なマッピングが実施されている．

2.2 プライバシーリスク分析

プライバシー分野では，LINDDUN 手法[5]が，STRIDE 分析の成功モデルをプライバシー領域に適用した代表的な事例である．同手法では，プライバシー固有の脅威特性を考慮して，7 つの脅威カテゴリを定義し，各カテゴリに対応する具体的な脅威パターンと対策手法を体系化している．

文献[9]では，プライバシー・バイ・デザインの概念が詳細に論じられ，システム設計段階からプライバシー保護を組み込むための方法論が提示されている．

文献[10]では，LINDDUN 手法の実用性向上を目的として，各脅威カテゴリに対応する脅威ツリーが開発されている．これらの脅威ツリーは，抽象的な脅威カテゴリから具体的な脅威までの階層的な関係を明示し，実務者が体系的に脅威を識別できる支援ツールとして機能している．このアプローチにより，プライバシー分野においても，脅威識別から対策要件導出で一貫したプロセスが実現されている．

2.3 倫理的システム設計

2.3.1 理論的検討

倫理的システム設計の分野では，文献[11]において，技術設計における価値の組み込み（Value Sensitive Design）の重要性が提唱されている．この研究では，技術システムが人間の価値観や社会構造に与える影響を設計段階から考慮することの必要性が論じられ，価値中立的な技術という概念に対する検討が行われている．

文献[12]では，技術システムが人間の価値観に与える影響を設計段階から考慮するための理論的フレームワークが詳細に論じられている．同研究では，概念的調査，経験的調査，技術的調査という 3 つのアプローチを統合することで，価値に配慮したシステム設計を実現する方法論が提示されている．

文献[3]では，AI 倫理の 5 つの基本原則（beneficence, non-maleficence，autonomy，justice，explicability）が提示され，これらの原則をシステム設計に具体的に反映するための方向性が議論されている．

しかし，これらの理論的研究では，抽象的な倫理原則から具体的な設計要件や対策技術への変換プロセスが明確に示されていない．実務者にとって，理論的な原則を実際のシステム開発にどのように適用すべきかという実践的な手法が不足している状況にある．

2.3.2 IEEE 24748-7000 規格

文献[6]は，システムライフサイクル全体を通じた倫理的配慮の統合を目的とした国際規格である．同規格の付録 G では，自律性，ケア，制御，公正性，包摂性，革新性，完全性，プライバシー，尊重，持続可能性，透明性，信頼など，システム開発で考慮すべき倫理的側面が体系的にカテゴリ化されている．これら倫理的側面一つ一つに，各倫理的価値の関連価値（Related Values）と対立価値（Opposing Values）も記載されている．関連価値とは各倫理的価値と密接に関連する概念や価値観を示し，対立価値とはその価値と相反する概念や価値観を表している．例えば，Autonomy（自律性）では関連価値として「moral agency, dignity, independence, freedom, liberty」など，対立価値として「accountability, responsibility, paternalism」などが記載されている．しかしながら，これらのカテゴリや価値がどのような具体的な脅威と関連するかについての情報は提供されていない．これは，セキュリティやプライバシー分野で確立されているような，抽象的なカテゴリから具体的脅威，そして対策要件まで一貫して導出できるプロセスが，倫理分野では依然として未整備であることを示している．

2.3.3 Universal Methods of Ethical Design

文献[7]が 100 の手法を体系的に整理している．同書籍では，デザインプロセスにおける倫理的考慮を支援する多様

な手法が収録されており，倫理的対策を行う上で有用である．

2.4 本研究の位置づけ

本研究では，セキュリティ・プライバシーリスク分析と同様に，脅威の識別から対策要件の導出まで一貫して実施できる手順を倫理リスク分析に適用することで，この実践的なギャップを埋めることを目指す．

セキュリティ・プライバシーリスク分析プロセスを参考に倫理リスク分析を実施しようとすると，

① データフローに相当する構成情報を描く
② データフローから，各プロセスで発生する倫理的な脅威カテゴリを特定する
③ 脅威カテゴリから具体的な脅威を導く
④ 具体的な脅威から必要な対策要件を導く

という四つの手順が少なくとも必要である．本論文は，これら手順の具体化に向けて，②～④を実施できる枠組みを実現することを目的とする（①～②は今後の課題である）．

3. 提案手法

本章では，前節で示した②～③，③～④をそれぞれ実現するための方法を提案する．説明の都合上，③～④→②～③の順で説明をすることに注意されたい．

3.1 具体的な脅威の導出手法（③～④）

著者が知る限り，セキュリティリスク分析やプライバシーリスク分析と異なって，対策要件の特定に必要な粒度で具体化された，倫理的脅威の一覧は存在しない．そこで，対策要件から倫理的脅威を逆解析するというアプローチを採用する．

本論文では，対策要件については文献[7]に書かれている，対策手法一つ一つを倫理的な対策要件として利用する．そのうえで，対策は「何かの脅威を防ぐもの」であるため，文献[7]の各対策手法が「なぜ必要か」ということを分析して，具体的な脅威とする．すなわち，文献[7]で提示される各種の対策手法について，その背景にある問題意識や想定される脅威を逆解析することで，具体的な倫理的脅威を明確化する．具体的な以下のとおりである．

文献[7]に記載された各対策について，「この対策はなぜ必要なのか」「どのような問題状況を想定しているのか」「対策が適用されない場合にどのような倫理的問題が発生するのか」という観点から分析を行う．例えば，文献[7]の pp.190 にてダークパターンや欺瞞的パターンの特定・修正を行い，ユーザの主体性に積極的な影響を与える公正な対策を採用する手法が紹介されている場合，その背景には「ダークパターンによるユーザの意思決定自由と選択権の操作的剥奪」という脅威が想定されていると解釈できる．同様に，文献[7]の pp.161 にて製品の透明性を向上させる手法が紹介されている場合，「製品責任の曖昧性による説明責任の回避と倫理的責任の不透明化」という脅威が想定されている．このような分析を通じて，各対策手法に対応する具体的な倫理的脅威を体系的に抽出する．

これら作業によって，文献[7]に書かれた各対策要件とその対策要件が求められる際の具体的な脅威を特定することができる．これを逆の関係で利用し，具体的な脅威に対して，それがどの対策要件に関する脅威だったかを特定することができる．あらかじめ，どの具体的な脅威に対して，どの対策要件であるか紐づけたリストを用意しておく．③～④の作業においては，そのリストを参照することで，ある具体的な脅威から，対策要件を導くことが可能である．

3.2 具体的な脅威の導出手法（②～③）

前述のとおり， 文献[6]の付録 G では，Autonomy（自律性），Care（ケア），Control（制御），Fairness（公正性），Inclusiveness（包摂性），Innovation（革新性），Perfection（完全性），Privacy（プライバシー），Respect（尊重），Sustainability（持続可能性），Transparency（透明性），Trust（信頼）という主要な倫理的価値が定義されている．本論文では，これら倫理的価値一つ一つを前述の倫理リスク分析における脅威カテゴリとして採用する．

なお，倫理リスク分析において倫理的価値をベースに脅威カテゴリを先行して設定することには，(1)データフロー上の各要素に同一の観点を適用でき，検討漏れの低減に資する，(2)個別事例や直近の出来事に左右されるバイアスを抑制でき，安定して一貫した検討ができる，(3)カテゴリごとの典型パターンを手掛かりに②～③で具体的脅威を系統的に導出しやすくなる，(4)脅威カテゴリ→具体的脅威→対策要件の対応関係が明確になり④でのトレーサビリティと説明責任が強化される，といった実務上の利点がある．さらに，カテゴリ単位での重要度評価・優先順位付けや，法務・開発・運用間の共通言語化と合意形成，既存の規格・対策カタログとの整合確保といった効果も期待できる．

このように定義したカテゴリ一つ一つに，3.1 節で具体化した脅威がどのカテゴリに属するかを分類する．例えば，「ダークパターンによるユーザの意思決定自由と選択権の操作的剥奪」という脅威は，Autonomy（自律性）に分類される．また，「製品責任の曖昧性による説明責任の回避と倫理的責任の不透明化」という脅威は，個人の選択の自由を制限するという観点から Transparency（透明性）に分類する．

これら分類によって，各倫理的な脅威カテゴリに対応する具体的な脅威のリストを作成できる．②～③の作業においては，そのリスクを参照することで，ある具体的な脅威カテゴリから，具体的な脅威を導くことが可能である．

4. 逆解析・分類作業

本章では，前節に示した各手順に必要な逆解析・分類作業を実施する．本論文は，初期検討であることから，大規模言語モデル（Claude Sonnet 4）を活用して逆解析・分類を

実施する．ただし，生成 AI が実施した作業結果に誤りがないかを，著者が確認することで作業結果の最低限の妥当性を担保する．

4.1 具体的脅威の逆解析

4.1.1 生成 AI による逆解析

文献[7]に記載された 100 の手法を対象として，各手法が対策する倫理的脅威を体系的に抽出するプロンプトを設計して分析を行った．

脅威抽出では，手法の説明文から「この手法が主要目的として対策することを意図している脅威」を逆解析により導出することに焦点を当てた．各手法について，その対策機能を分析し，「なぜこの手法が必要なのか」「どのような問題状況を想定しているのか」という観点から，背景にある倫理的脅威を特定した．

抽出される脅威が統一的な形式で記述されるよう，プロンプトに「(原因) による (影響)」と明確な指示を含めた．また,「この手法の直接的な目的は〇〇脅威への対策である」といえるかどうかを基準とすることで，分析の精度を確保した．

4.1.2 人手による検証

検証では，主に 2 つの観点から確認作業を行った．

第一に，生成 AI が抽出した脅威が，実際に手法の説明文から合理的に導出できるものであるかの確認を行った．各手法の記述内容と抽出された脅威の間に論理的な関連性があるか，手法の対策機能から妥当に推論できる脅威であるかを評価した．明らかに手法の記述と関連性が薄い脅威については，該当する項目を除外した．

第二に，文献[6]付録 G の倫理的価値との対応関係について，著者 1 名による確認を行った．具体的には，抽出された脅威が適切な倫理的価値カテゴリに分類されているか，分類の根拠が論理的に妥当であるかを評価した．

この検証プロセスにより，生成 AI によって抽出された脅威の中から，明らかに根拠が不十分なものや，手法の記述と関連性が薄いものを除外した．

4.2 脅威カテゴリと具体的脅威の分類作業

4.2.1 生成 AI によるマッピング

前節の作業によって特定された具体的な脅威を文献[6]付録 G の 12 の倫理的価値（Autonomy，Care，Control，Fairness, Inclusiveness, Innovation, Perfection, Privacy, Respect, Sustainability，Transparency，Trust）に分類するよう生成 AI に指示した．

生成 AI には，各倫理的価値の関連価値（Related Values）と対立価値（Opposing Values）の情報を併せて入力し，分類の根拠となる情報を提供した．

一つの手法が複数のカテゴリに関連する脅威に対策する可能性を考慮し，該当するすべてのカテゴリを特定するよう指示した．これにより，手法の多面的な対策機能を適切に表現することを試みた．

4.2.2 人手による検証

各脅威に割り当てられた倫理的価値が論理的に適切であるかを確認した．

確認作業では，各脅威の性質と影響範囲を検討し，生成 AI が提示した分類理由の妥当性を評価した．例えば，「ユーザの選択権と自己決定権を制限する強制的なタスクフローによる主体性の剥奪」という脅威が Autonomy（自律性）に分類されている場合，この分類が脅威の本質を適切に表現しているかを確認した．

また，生成 AI が提示した分類理由についても検討し，文献[6]の各カテゴリの関連価値・対立価値との整合性を確認した．明らかに不適切な分類や，分類理由が論理的でない場合については，著者の判断により適切と考えられる倫理的価値へ人手で再分類を行った．

一つの手法が複数のカテゴリに関連する脅威に対策する場合の分類の妥当性についても確認した．生成 AI の判断を参考にしつつ，脅威が各倫理的価値に与える影響の性質を基準として，最終的な分類を決定した．

4.3 結果

4.3.1 具体的な脅威の逆解析結果

前章に記した分析の結果，合計 88 件の倫理的脅威を抽出することができた．表 1 に抽出例を示す．

表 1　倫理的脅威の抽出例

手法名	脅威
Fair Patterns	ダークパターンによるユーザーの意思決定自由と選択権の操作的剥奪
	ダークパターンによる不公正なユーザー操作と欺瞞的な利益誘導
Critical Race Theory	人種差別の制度的埋め込みによる体系的不公正の永続化
	多様な能力を持つユーザーへの設計配慮不足による利用機会の体系的剥奪
Human Design Guide	技術による人間の脆弱性悪用と意図に反する操作による自律的意思決定の阻害
	デジタル技術による人間の感受性と脆弱性の体系的悪用
Privacy by Design	データ収集・処理・共有におけるプライバシー侵害の体系的リスク
Ethical Disclaimer	製品責任の曖昧性による説明責任の回避と倫理的責任の不透明化

4.3.2 脅威カテゴリと具体的脅威の分類作業結果

抽出された 88 件の倫理的脅威を文献[6]付録 G の 12 の倫理的価値に分類した結果，表 2 のとおりとなった．

表 2　倫理的価値別の脅威分布

倫理カテゴリ	脅威件数
Autonomy	21
Care	5
Control	3
Fairness	11
Inclusiveness	17
Innovation	2
Perfection	9
Privacy	3
Respect	0
Sustainability	3
Transparency	8
Trust	6
合計	88

5. 考察

5.1 提案手法の基礎的有効性

本研究の結果は，実際のシステム開発プロジェクトにおける倫理的配慮の組み込みに関して，いくつかの示唆を提供している．88 件の体系化された．実務者が自身のプロジェクトで発生し得る倫理的問題を検討するための参考資料として活用できる

また，各脅威と文献[6]の倫理的価値との対応関係は，組織の価値観や方針に応じた優先順位付けを支援する可能性がある．例えば，公共性の高いシステムを開発する組織では，Fairness（公正性）や Inclusiveness（包摂性）に関連する脅威を重点的に検討し，商用プラットフォームを開発する組織では，Autonomy（自律性）や Privacy（プライバシー）に関連する脅威に焦点を当てることができる可能性がある．今後①～②（あるプロセスで必要な脅威カテゴリが特定できる方法を考案することで，この有用性はさらに増すものと考えられる．

さらに，各脅威に対応する文献[7]の対策手法との関係性により，実務者は具体的な対策要件を検討する際の参考情報を得ることができる可能性がある．これにより，抽象的な倫理原則から具体的な実装要件への変換プロセスを支援する基盤の構築可能性が示唆された．

5.2 脅威分布からの有用性検討

脅威の抽出件数が一番多かった Autonomy（自律性）では，主に「ダークパターンによるユーザの意思決定自由と選択権の操作的剥奪」や「ユーザの選択権と自己決定を制限する強制的なタスクフローによる主体性の剥奪」といったユーザの選択権や自己決定権に関する脅威が特定された．Autonomy（自律性）が特に重要と考えられるデジタルサービスの設計・開発において，本手法を有用に活用できる可能性を確認した．

Autonomy（自律性）に次いで脅威の抽出件数が多かった Inclusiveness（包摂性）では，主に「人種的・文化的マイノリティのデザインプロセスからの体系的排除」や「特定グループの排除と不平等な扱いによる包括性の阻害」といった多様なユーザの包括性に関する脅威が特定された．Inclusiveness（包摂性）が特に重要と考えられる公共サービスの設計・開発において，本手法を有用に活用できる可能性を確認した．

さらに，Fairness（公正性）でも多数の脅威が抽出され，主に「AI システムにおける体系的差別と偏見による不公正な判断・処理」や「設計者の無意識バイアスによる特定グループへの不公正な設計判断」といった不公正なシステム挙動に関する懸念が確認された．Fairness（公正性）が特に重視される自動意思決定や評価システムの設計・運用において，本手法が有用に活用できる可能性を確認した．

このように，Autonomy（自律性），Inclusiveness（包摂性），Fairness（公正性）の 3 カテゴリで多数の脅威が抽出されたことは，ユーザの自己決定権の確保，多様な利用者層への配慮，および不当な差別の排除といった課題が，近年の国際的なガイドラインや規制[1][15]において中心的に言及されていることと一致する．これらの価値が社会的関心の高い領域であることを踏まえると，本研究においても相対的に多くの脅威が見出されたことは妥当な傾向だと考えられる．

一方で，Control（制御），Innovation（革新性），Privacy（プライバシー），Sustainability（持続可能性）のカテゴリは，相対的に脅威の抽出件数が少ない結果となった．Control（制御）や Privacy に関しては，既に制度的・技術的な枠組みが一定程度整備されており，アクセス制御技術やプライバシー規制によって設計段階から考慮される傾向が強いためである可能性が考えられる．Innovation（革新性）は主に「目指すべき価値」として扱われる傾向が強く，脅威として具体化しにくいことが一因と考えられる．Sustainability（持続可能性）は比較的近年になって倫理的議論に取り込まれた価値であり，中心的に扱われてこなかったため，具体的脅威が限定的となったと解釈できる．

Respect（尊重）に関しては，脅威の抽出件数が 0 件という結果となった．これは，文献[6]付録 G において Respect が，「注意深さ」と「応答性」という複合的要素で構成され，他の倫理的価値（Privacy，Fairness，Transparency など）にまたがる横断的な性質を持つと定義されているためだと推察される．例えば，プライバシーを尊重しない対応は Privacy の脅威に，不適切な基準による判断は Fairness の脅威に，説明の欠如は Transparency の脅威にそれぞれ分類されやすい．このため，Respect 固有の脅威としては抽出されにくかったと解釈できる．ただし，ユーザ体験における礼儀正しさや配慮の欠如といった問題は，他カテゴリでは十分に表現しきれない場合があるため，手法の逆解析による

脅威の拡充のためには, Respect を独立した観点として扱うことが課題として考えられる．

5.3 対策要件の拡充

本研究は探索的研究ということもあり，文献[7]という単一の文献に依存した分析となっている．脅威と対策要件の対応関係をより網羅的に整備するには，他の倫理的デザインを記した文献やガイドラインを参照することが重要である．ただし，本研究のポイントとしては，単に対策要件を増やすことではなく，対策から「なぜその対策が必要か」を検討することで, 具体的な脅威を解釈できることにある．したがって，今後の対策要件の拡充にあたっても，各要件とそれが想定する脅威を対応づけて整理する必要がある．

具体的には，国際的な規格類[13][14]に示されている原則・推奨事項を参照し，それぞれが想定する倫理的脅威を逆解析することで，脅威と対策の対応を拡張することができると考えられる．

また，国際的なガイドライン[15][16]を比較し，共通して示される対策要件から，それらが想定する代表的な脅威を特定することができると考えられる．これにより，国際的に共通する倫理的懸念を脅威モデルに取り込むと同時に，各国固有の懸念も比較検討できる．

このように，「各対策要件から脅威との対応関係を拡張する」ことで初めて，倫理リスク分析における実用性が高まると考えられる．

5.4 研究手法の限界

本論文では，いくつかの限界がある．

第一に，本研究は探索的なアプローチに基づく初期段階の検討である．このため，逆解析作業や分類作業に，大規模言語モデルを活用し，質的研究の厳密な手法は適用できていない．複数の評価者による独立した分析や，評価者間信頼性を測定する Kappa 係数などの統計的指標による検証も行う必要がある．さらに，考察は成果物に基づく定性的なものであり，今後，実際のシステム開発へ適用してみることでその有用性を検証していく必要もある．

第二に，抽出された脅威の網羅性や代表性について十分な検証が行われていない．文献[7]という単一の文献に依存した分析であり，他の重要な倫理的デザイン文献との比較検討や，実際のシステム開発プロジェクトで発生する倫理的問題との対応関係の検証が課題である．

第三に，文化的・地域的多様性への配慮が限定的である．本研究で使用した文献は特定の文化的背景に基づいて記述されており，異なる文化圏や社会的文脈における倫理的価値観の違いが十分に反映されていない可能性がある．

6. まとめと今後の課題

本研究では, 文献[6]の倫理的価値と文献[7]の対策手法を, 具体的な脅威を介して体系的に結びつける倫理リスク分析プロセスの基礎的な枠組みを提案した．セキュリティ分野の STRIDE 分析やプライバシー分野の LINDDUN 分析を参考に，脅威カテゴリ→具体的脅威→対策要件という階層的な関係を明確に定義することで，実務者が体系的に倫理リスク分析を実施するための基盤構築の可能性を探索した．

生成 AI を活用した脅威抽出により, 100 の倫理的デザイン手法から 88 件の具体的脅威を特定し，文献[6]付録 G の 12 の倫理的価値との対応関係を構築した．特に，Autonomy（自律性）21 件，Inclusiveness（包摂性）17 件，Fairness（公正性）11 件という分布により，現代のデジタルシステムにおける倫理的課題の一端を把握することができた．本研究では, 文献[6]の付録 G を脅威分析のためのカテゴリ体系として活用するという視点を提示し，従来，個人の判断に依存していた倫理リスク分析に構造化されたアプローチを導入する可能性を示した．

ただし，本研究は探索的なアプローチに基づく初期段階の検討であり，多くの重要な課題が残されている．今後の主要な研究課題として，（1）質的研究の標準的手法に基づく厳密な分析プロセスの確立（複数評価者による独立分析，Kappa 係数等による評価者間信頼性の測定），（2）他の倫理関連文献との統合による包括的脅威カタログの構築，（3）異なる文化・地域での適用性検証が挙げられる．

参考文献

[1] “Regulation (EU) 2016/679: General Data Protection Regulation”. https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng, (参照 2025-08-21).

[2] “Ethics Guidelines for Trustworthy AI”. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai（参照 2025-08-21）.

[3] Floridi, L., Cowls, J., Beltrametti, M. et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds & Machines, Vol. 28, No. 4, pp. 689–707 (2018).

[4] Shostack, A. Threat Modeling: Designing for Security. John Wiley & Sons, 2014.

[5] Deng, M., Wuyts, K., Scandariato, R. et al. A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements. Requirements Engineering, Vol. 16, No. 1, pp. 3–32 (2011).

[6] ISO/IEC/IEEE 24748-7000:2022, Systems and software engineering−Life cycle management−Part 7000: Standard model process for addressing ethical concerns during system design, ISO/IEC/IEEE, 2022 (IEEE Std 7000-2021) .

[7] Chivukula, S. S. and Gray, C. M., Universal Methods of Ethical Design: 100 Ways to Become More Ethically Aware, Responsible, and Active in Your Design Work. Rockport Publishers, 2025.

[8] M. Howard and D. LeBlanc, Writing Secure Code, Microsoft Press, 2003.

[9] S. Spiekermann and L. F. Cranor. Engineering Privacy. IEEE Transactions on Software Engineering, Vol. 35, No. 1, pp. 67–82 (2009).

[10] “LINDDUN Threat Trees”. https://linddun.org/threat-trees/ （参照 2025-08-19）.

[11] J. van den Hoven, P. E. Vermaas, and I. van de Poel. Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains. Springer, 2010.

[12] B. Friedman, P. H. Kahn Jr., and A. Borning. Value Sensitive Design and Information Systems. in Early Engagement and New Technologies: Opening Up the Laboratory, N. Doorn et al. (Eds.), Springer, pp. 55‑95 (2013).
[13] “Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems”. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf, (参照 2025-08-21).
[14] “Artificial Intelligence Risk Management Framework (AI RMF 1.0)”. https://doi.org/10.6028/NIST.SP.1270, (参照 2025-08-21).
[15] “Ethics Guidelines for Trustworthy AI”. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, (参照 2025-08-21).
[16] “OECD Principles on Artificial Intelligence”. https://oecd.ai/en/ai-principle, (参照 2025-08-21).