URL　https://journal.ntt.co.jp/article/36086
DOI　https://doi.org/10.60249/25095103

人々とAIの協調による“人々のためのセキュリティ活動”への進化

システム等をねらうサイバー攻撃は年々激化する一方，セキュリティ人材不足も深刻化していることから，AI（人工知能）による自動化を前提としたセキュリティオペレーションへの抜本的な見直しが求められています．さらに現代の情報社会では，偽・誤情報などにより人々の自律的な意思決定の維持が難しくなっており，セキュリティの対象はシステム等から人の認知へと拡大しています．本稿では，生成AIを活用した革新的なサイバーセキュリティオペレーションと人の自律的な意思決定を支えるコグニティブセキュリティに関する取り組みを紹介します．

キーワード：#サイバーセキュリティ，#コグニティブセキュリティ，#生成AI

松橋　亜希子／秋山　満昭
山中　友貴／古谷　諭史
原　亨

NTT社会情報研究所

人とAIが協調するセキュリティ活動の未来像

サイバー攻撃の主要な原因である「脆弱性」の急増は，もはや従来手法では対処困難なレベルに達しています．警察庁によると，2024年はサイバー攻撃を目的とした脆弱性探索行為と思われる不審なアクセス，ランサムウェアやフィッシングの報告件数が急増しています(1)．さらに，サイバー詐欺やソーシャルエンジニアリング，SNS上で拡散する偽・誤情報などのように，システムではなく人をねらうことによって，その認知や意思決定に影響を与える新たな脅威も顕在化しています．一方，これらに対処するセキュリティ体制は，リスクの増大と多様化に応じて人員を増強することが容易ではなく，日々，人員の心理的負担が増大して疲弊するなど，従来のセキュリティ活動は限界に直面しているといえます．

セキュリティに関連する活動は，NIST（National Institute of Standards and Technology：米国国立標準技術研究所）のサイバーセキュリティフレームワークによれば，特定，防御，検知，対応，復旧，統治に分類され，それぞれにおいてさまざまな活動が求められています．私たちはそのすべてを対象として，人とAI（人工知能）が協調する新たなセキュリティ活動に進化させることによって，前述した深刻な現状を打開することをめざした研究を進めています．そのための研究構想を，私たちは「Cycle-Ops（サイクロプス）」と名付けました．

Cycle-Opsは，AIが人の苦手な部分や弱点を補い，人がAIには困難な部分を担うことによって，人とAIが共にそれぞれの長所を活かし合いながら各活動を遂行する形態を実現し，人のみの体制では困難であったセキュリティ活動の進化を促します．この進化によって，活動にかかわるすべての人々（セキュリティ担当，システム担当，エンドユーザ，経営層等）の心理的負担感を軽減し，より良いセキュリティ活動体験をもたらし，リスク低減や効率化のような経済合理性の視点のみでは具現化できなかった真の安心・安全をもたらします．

また，各セキュリティ活動では，日々知見が生み出されるものの，各活動を担当する個々人の頭脳に蓄積するにとどまり，セキュリティ体制総体として知見が十分に活かされづらい現状があります．そこで，Cycle-Opsは，AIの仲介によって，ある活動の知見を別の活動に活用する流れと，人と人の間で知見を共有・継承する流れの両面を促進可能にし，「持続的かつ循環的に成長するセキュリティ活動」を実現することもめざしています．

このようなCycle-Opsの実現に向けた軸となる考え方が2つあります．1番目は「暗黙知の形式知化」です．セキュリティ活動においては幅広い知識が必要であり，その知識はいわゆるノウハウや暗黙知といったかたちでセキュリティの熟練者に依拠しているため，その知識継承や再現は容易ではありません．そこで，現在人間中心となっている各セキュリティ活動において，人（セキュリティ熟練者）が持つ暗黙知を形式知化し，それをAIに取り込むことで人とAIとの協調を可能とします．

2番目は「人の認知を守るセキュリティ（コグニティブセキュリティ）」です．Cycle-Opsがめざす持続的な循環成長型のセキュリティ活動の主体（人）には，システム担当やセキュリティ担当だけでなく，セキュリティ問題に直面するエンドユーザも含んでいます．そして，セキュリティのもっとも弱い部分は組織や社会が保有するシステム自体よりも，その保守者や利用者などの人であることが多くあります．人やその集団が直面する情報についての正しい認知ができないと，例えば前述した偽・誤情報による脅威が発生することにつながり，セキュリティを維持することが難しくなります．そこで，情報科学だけではなく認知心理学等の方法論も活用しながら，個人や集団における自律的な意思決定を確保することで，人の認知を守るアプローチが重要となります．このアプローチをCycle-Opsにおける人とAIが協調する活動形態によって実践可能にします．

Cycle-Opsは，この2つの観点を融合し，セキュリティ活動にかかわる多様な人々（シ

ステム担当，セキュリティ担当，エンドユーザ等）とAIが高度に協調することによって，専門人材の人力のみに依存し，限界すら迎えていた従来型のセキュリティ活動を，新たな形態へと進化させることをめざしています．以降では「暗黙知の形式知化」「人の認知を守るセキュリティ」について，ユースケースを交えながら研究内容を紹介します．

セキュリティレポート作成作業を対象とした暗黙知の形式知化技術

前述のとおりCycle-Opsの軸となる観点の1つ「暗黙知の形式知化」について，私たちは最初のユースケースとして，セキュリティ活動全般で不可欠といえる脅威情報に関するレポート（セキュリティレポート）の作成に着目し研究を開始しました．セキュリティレポートの作成担当者は，日々，深刻かつ膨大な脅威情報に相対しながら，その収集と分析には正確性と効率性が求められます．また，読者は経営層やシステム担当者など多様な属性・役割を持つため，それぞれが知りたいポイントを適切におさえたレポートを作成する必要もあります．このような作業を，セキュリティレポート担当者は独自のノウハウに基づき日々遂行しており，この暗黙知を形式知化することがCycle-Ops実現の第一歩として不可欠と考えました．

従来のセキュリティレポート作成のプロセスでは，熟練したセキュリティ担当者が，入手した脅威情報の内容だけでなく，読者の立場や関心事を考慮してレポートの目次を設計するとともに，自組織の実状に即した内容に仕立てるために組織内のネットワーク構成や対応状況を踏まえた詳細な分析を行い，その結果を読者に応じて書き分けます．このような書き分けには，経営層向けには経営判断に必要な情報を簡潔にまとめ，システム担当者向けには技術的な詳細を含めるなど，組織内の文化に対する深い理解，つまり暗黙知が求められます．近年，セキュリティレポートの作成には限定的ながらLLM（Large Language Model：大規模言語モデル）の活用が始まっていますが，熟練者の持つこのような暗黙知を十分に活用できていない現状があります．

これを受け，私たちは，セキュリティ担当者が行うセキュリティレポート作成のプロセスに含まれる暗黙知を大きく2つに分けました．1番目は，読者の立場や関心事を考慮してレポートの目次を設計する方法，2番目が，レポートに記載された脅威情報に対する自システムへのリスク分析方法です．この2点の暗黙知をLLMに学習させることで，組織の文化や読者の特性に適応した高品質なレポートを自動生成する技術を確立しました．次にその詳細を解説します．

■「読者の立場や関心事を考慮してレポートの目次を設計する方法」の形式知化手法

従来のLLMに単純にレポート作成を依頼しても，読者の関心に合わない目次が生成されてしまう問題がありました．例えば，幹部層向けのレポートにもかかわらず過度に技術的な内容が含まれるなど，組織との関連性が不明確で一般的な内容になってしまうといった問題です．

本技術では，熟練者が作成した過去のレポートから暗黙知を自動抽出する手法を開発しました．この手法は，従来の人によるセキュリティレポート生成業務を，例えば新人のセキュリティ担当者が下案を作成し，熟練のセキュリティ担当者がそれを校正する業務ととらえ，そのレポート構成の観点を暗黙知としてとらえて形式化するものです．具体的には，最初に熟練者の過去レポートからトピックと対象読者を抽出し，同じ条件でLLMにレポートを生成させます．このときLLMは，一般的な内容しか書くことができない新人のセキュリティ担当者を模したものとなっています．その後，両者の差分を分析することで，例えば，熟練者のレポートには「自社への影響」といった幹部層の関心に沿った目次がありますが，LLMレポートにはないという差分から，例えば「幹部層向けの脆弱性報告では，自社への影響を重視する」といった暗黙知を抽出します．そして，レポート生成時には，RAG（Retrieval-Augmented Generation）技術により類似トピック・読者の暗黙知を参照しながら，目次案を生成する仕組みにより，対象読者に寄り添った目次としての精度向上を実現しています．

実際のセキュリティオペレーションの現場におけるレポート作成業務において，本手法を適用した際のアンケート調査では，これまでの熟練者により作成されたレポートと比較しても否定的な意見はなく，同水準のレポートが生成できていることが確認されています（**図1**）．

■「レポートに記載された脅威情報に対する自システムへのリスク分析」の形式知化手法

より読者に沿ったセキュリティレポートを生成するためには，外部に公開されたセキュリティ情報だけでなく，組織内のシステム等の情報や，それに関連するリスク情報が必要となります．そこでは，ネットワーク構成，各種機器の設定情報，脆弱性情報などの多様な情報を動的に組み合わせた複雑な分析が必要です．従来，セキュリティ担当者が行っていたこのプロセスを単純にLLMに置き換えると，条件分岐のたびにLLMによる推論が必要となり，どこかのステップで失敗すると最終的なリスク判定の結果も異なってしまい，正しいリスク判定ができないという問題がありました．

そこで私たちは，グラフデータベースにこれらの多様な情報を一元化し，1回のグラフクエリでリスク分析に必要な情報をまとめて取得できるよう構造化しました．具体的には，「外部からの悪用可能性判定」のユースケースに合わせて，ノードに脆弱性情報やネットワーク機器の設定を紐付け，ネットワークの経路に関連して一括取得できるよう構築しています．

これにより，「インターネットから脆弱性Xがある端末へ到達可能なネットワーク経路と関連情報を取得」といった指示をLLMでグラフクエリに変換し，グラフデータベースから一括取得した情報をプロンプトに成形し，LLMによりリスク分析を一気に実行することが可能となります．このようにLLMによる推論回数を削減し，総合的な情報を考慮した高精度なリスク判定を実現しています．

実際，本技術の活用により，ネットワークの構成のみを考慮した先行研究の精度約80%と同等でありながら，より複雑な条件（ファイアウォールやネットワーク機器の

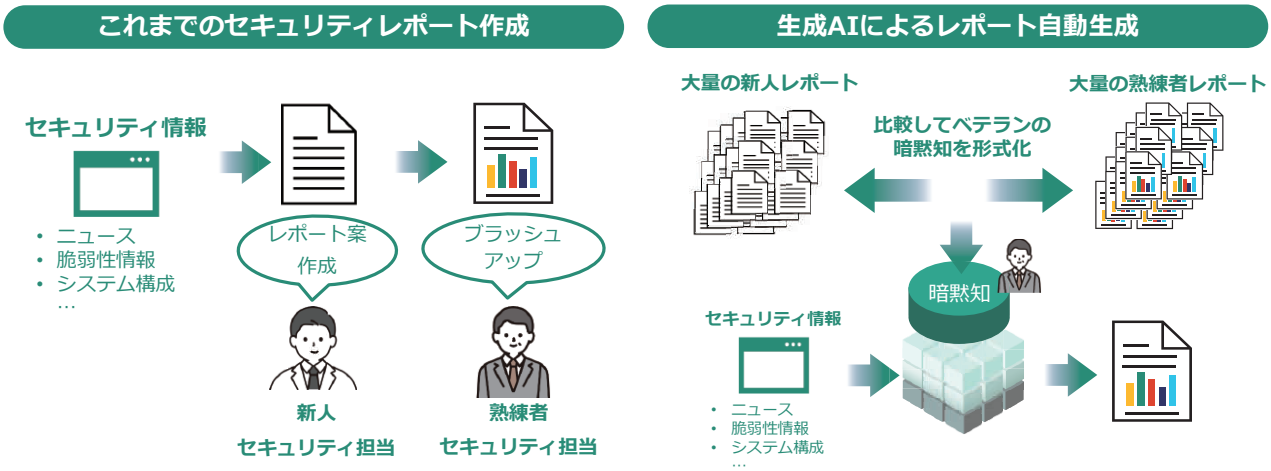
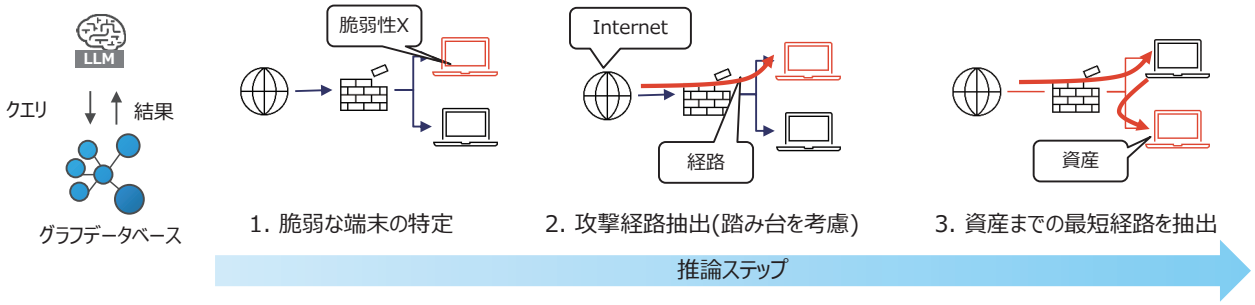


図1　熟練者の暗黙知を活用した生成AIによるセキュリティレポート生成



T. Nagai, Y. Yamanaka, Y. Teramoto, T. Yamashita, A. Shiga, and R. Sato : “Proposal of an attack surface analysis system based on firewall rules using LLM,” in IEICE Technical Report, vol. 124, no. 422, ICSS2024-73, pp. 31–38, Mar. 2025.

図2　グラフDBを活用した生成AIによる脆弱性リスク分析

設定も考慮）下でも同等の精度を達成しています（**図2**）.

人の認知を守るセキュリティ

現代の情報社会において，私たちは日常的に多種多様な情報に接しながら意思決定を行っています．SNSや動画配信サービス，さらには生成AIの登場により，情報の生成・共有・消費のスピードと量は飛躍的に増大し，これまでにない利便性と情報アクセスの自由を享受できるようになりました．

一方，その利便性の裏で，人々の認知や意思決定が，さまざまなかたちで影響・干渉を受けるリスクが顕在化しています（**表**）．例えば，「サイバー詐欺・ソーシャルエンジニアリング」は，人間の認知の隙を突いてユーザを騙すことで，ユーザから金銭や個人情報を詐取します．また，ダークパターンと呼ばれる「操作的なインタフェース」も，ユーザの無自覚な選択を誘導し，ユーザに本来の意思とは異なる行動をとらせることがあります．さらに，SNS上で行われる「偽・誤情報の拡散」は，個々のユーザの信念に影響を与えるだけでなく，社会全体の意見の分断にもつながります．加えて，近年では，生成AIの出力も人々の認知に影響を与える要素として無視できなくなっています．典型的な問題としては，実在しない情報をもっともらしく出力する「ハルシネーション」がよく知られていますが，それ以外にも，学習に利用されたデータの偏りを反映して，事実には基づいているものの視点が一方的に偏った情報を，あたかも中立かつ正確な知見であるかのように「アルゴリズムによる操作」を通じて提示する問題もあります．このような出力は，受け手にとって違和感が少ないため，内容を鵜呑みにしやすく，結果として認知の偏りや意思決定の誤りを生じさせるリスクがあります．

コグニティブセキュリティとCycle-Ops

現代社会では，このような認知にかかわる脅威が「個人」と「集団」の両レベルで生じ，それらが複合的に絡み合いながら社会全体に影響を及ぼしています．このような環境下では，情報を「どのように扱うか」だけでなく，情報が「どのように認知されるか」が重要な課題になります．私たちの認知や意思決定そのものが，外部からの影響や操作の標的となり得る状況が生まれて

表　人の認知に関する脅威とその影響

脅威の種類	具体例	影響のレベル	主な影響内容
サイバー詐欺・ソーシャルエンジニアリング	なりすましメール，詐欺リンク，偽ログイン画面	個人	金銭や個人情報の詐取
操作的インタフェース	ダークパターン：カウントダウンによる購入圧力，解除が困難な登録の導線	個人	ユーザが望まない選択・行動への誘導
偽・誤情報の拡散	SNS上での健康デマ，政治的な虚偽発言の共有	個人・集団	偏った意見価値観の形成，社会的分断の助長，民主主義の機能不全
アルゴリズムによる操作	検索結果やフィードのパーソナライズ（フィルターバブル），AI/LLMによる誤った?偏った回答	個人・集団	偏った意見・価値観の形成

いるのです．

私たちは，この脅威に対抗するために「コグニティブセキュリティ（Cognitive Security）」という新たなセキュリティの枠組みに基づいた研究に取り組んでいます．コグニティブセキュリティは，認知や意思決定を標的とする情報環境のリスクから，個人・集団・社会を守るためのアプローチです．

また，本アプローチは，従来のシステム中心のセキュリティ対策技術では実践困難です．そこで，Cycle-Ops構想における人（システム担当，セキュリティ担当，エンドユーザ等）とAIが協調する形態を通じて実現する新たなセキュリティ活動の中核要素として「コグニティブセキュリティ」を位置付けています．AIがエンドユーザに寄り添って介入するなどして「個人」の認知面のリスクを低減するとともに，そのようなAIとエンドユーザ（人々）のやり取りを面的にとらえることによって「集団」に対する認知面のリスク把握や対応判断が容易になり，システム担当やセキュリティ担当は効果的なエンドユーザ支援を行えるようになります．

以降では，コグニティブセキュリティが対象とする脅威の1つである「偽・誤情報」に焦点を当て，私たちの取り組みを紹介します．偽・誤情報は，サイバー詐欺などと異なり，個人から個人へと連鎖的に拡散していくため，上記のCycle-Opsを通じた実践が有効な脅威であり，私たちは「個人」と「集団」のそれぞれを対象として技術開発に取り組んでいます．

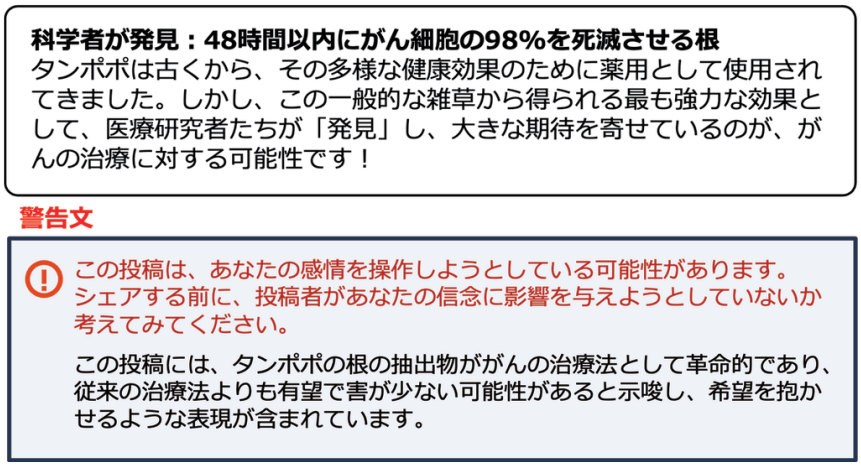


図3　感情的な操作表現を含むコンテンツとそれに対する警告ラベルの例

■個人への介入：感情を揺さぶる表現に対する気付きを促す介入[2]

「個人」を対象とした取り組みとして，感情的な言葉遣いに着目した介入技術を紹介します．SNS上では，怒りや恐怖などの感情をあおる表現が，偽・誤情報の拡散に影響することが知られています．私たちは，健康関連の偽・誤情報投稿に対して，感情的な表現への気付きを促すメッセージを添える介入を作成し，その効果をアンケート調査によって検証しました．結果として，投稿に対する共有意図が有意に抑制されることが確認されました．この研究では偽・誤情報を対象としていますが，感情を揺さぶる情報への気付きを促すという介入は，より汎用的なものです．情報の真偽を問わず，過度に情動を刺激する表現に冷静な視点を持たせることで，受け手の判断を支援することができます．特に，ユーザの注意を引くことが価値とされるアテンションエコノミーの文脈においては，ユーザが自律的に思考を保つための環境設計が重要です．本介入のような仕組みは，日常の情報接触における意思決定の質を支える実用的な手段となり得ます（**図3**）．

■集団への介入：誤情報の拡散を抑えるネットワーク介入[3]

「集団」を対象とした対策技術としては，ソーシャルネットワークにおいて誤情報の拡散を最小化するために，プレバンキング介入の割り当てを最適化する技術を開発しています．プレバンキングとは，情報版のワクチンのようなもので，人々が誤情報にさらされる前に，流通が予測される誤情報に対して予防的に警告や反論を提示することで，人々の誤情報に対する認知的抵抗力の強化を促す介入です．多くの研究でプレバンキングの有効性が実証されているものの，すべての人々に介入を行うのはコストの問題で困難であり，誤情報の普及が最大限抑制されるように効率的に介入を割り当てる必要があります．私たちは，この介入

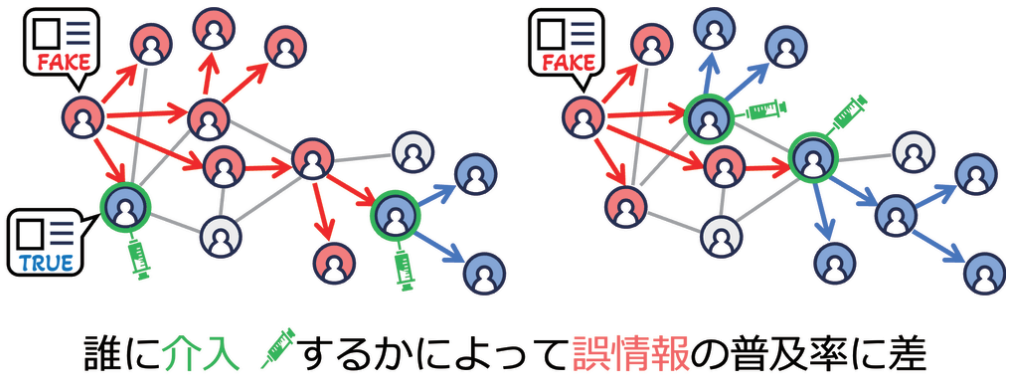


図 4　誤情報の拡散を抑える集団へのプレバンキング介入

割り当ての問題を組合せ最適化問題として定式化し，最適な介入ターゲットを特定するアルゴリズムを開発しました．現実のデータを用いた数値実験により，ランダムな200人に介入を行った場合，誤情報の普及はほとんど抑制できない一方，私たちの技術を用いて介入の割り当てを最適化すると，介入がない場合と比較して最大で約40%もの誤情報を削減できることを明らかにしました（**図 4**）．

私たちがめざすのは，個人が自律的に判断し，社会全体が多様な価値観を尊重しながら意思決定できる環境の実現です．このためには，心理学・HCI (Human Computer Interaction)・ネットワーク科学・社会学といった多分野の協働をとおして，技術や環境が人々の認知に影響をもたらすメカニズムを解明し，有効な介入を設計・実装していく必要があります．同時に，研究を進めるうえでの方法論的課題にも目を向ける必要があります．例えば，ユーザ研究における参加者が特定の文化圏に偏っている実態は，成果の再現性や応用可能性を狭める要因となり得ます．私たちのグループでは，このような課題に対して，人口統計学的要因に基づく違いをとらえる研究にも取り組んでいます(4)．さらに，技術だけでなく，制度設計や倫理的基準，教育によるリテラシー向上も含めた持続可能な取り組みも求められます．今後も私たちは，人々の自律的な意思決定を支える基盤として，コグニティブセキュリティを研究・実装していくことで，健全な社会の構築に貢献していきます．

おわりに

今後ますます激化，巧妙化するサイバー攻撃に対応していくためには，企業のみならずエンドユーザを含むすべての人にアプローチしていく必要があります．私たちは「暗黙知の形式知化」と「人の認知を守るセキュリティ」を中核とするCycle-Opsの研究開発によって，人々とAIが高度に協調する新たな形態へと進化させることで，リスク低減や効率化のような経済合理性の視点のみでは具現化できなかったより良いセキュリティ活動の体験と真の安心・安全をもたらします．

■参考文献

(1) https://www.npa.go.jp/publications/statistics/cybersecurity/data/R6/R06_cyber_jousei.pdf
(2) J. Jamieson, T. Hara, and M. Akiyama : “Flagging Emotional Manipulation: Impacts of Manipulative Content Warnings on Sharing Intentions and Perceptions of Health-Related Social Media Posts,” Proc. of CHI 2025, Yokohama, Japan, April-May 2025.
(3) S. Furutani, T. Aoshima, T. Shibahara, M. Akiyama, and M. Aida : “Suppressing the Endogenous Negative Influence Through Node Intervention in Social Networks,” IEEE Access, Vol.13, pp. 9290-9302, 2024.
(4) A. A. Hasegawa, D. Inoue, and M. Akiyama : “How WEIRD is Usable Privacy and Security Research?,” USENIX Security Symposium 2023, Anaheim, U.S.A., August 2023.

（上段左から）松橋 亜希子／ 秋山 満昭／ 山中 友貴
（下段左から）古谷 諭史／ 原 亨

人の存在を前提とした生成AI時代のセキュリティに関する研究開発を推進します．

◆問い合わせ先
NTT社会情報研究所
社会イノベーションプロジェクト