CRDS 研究開発戦略センター
Center for Research and Development Strategy

国立研究開発法人科学技術振興機構

「認知の脆弱性から人間をどう守るか～コグニティブ・セキュリティと法的課題の入門～」

コグニティブセキュリティの研究動向と課題

2024年7月11日

国立研究開発法人 科学技術振興機構（JST）
研究開発戦略センター（CRDS）
福井　章人

科学技術振興機構　研究開発戦略センターの概要

国立研究開発法人　科学技術振興機構（JST）

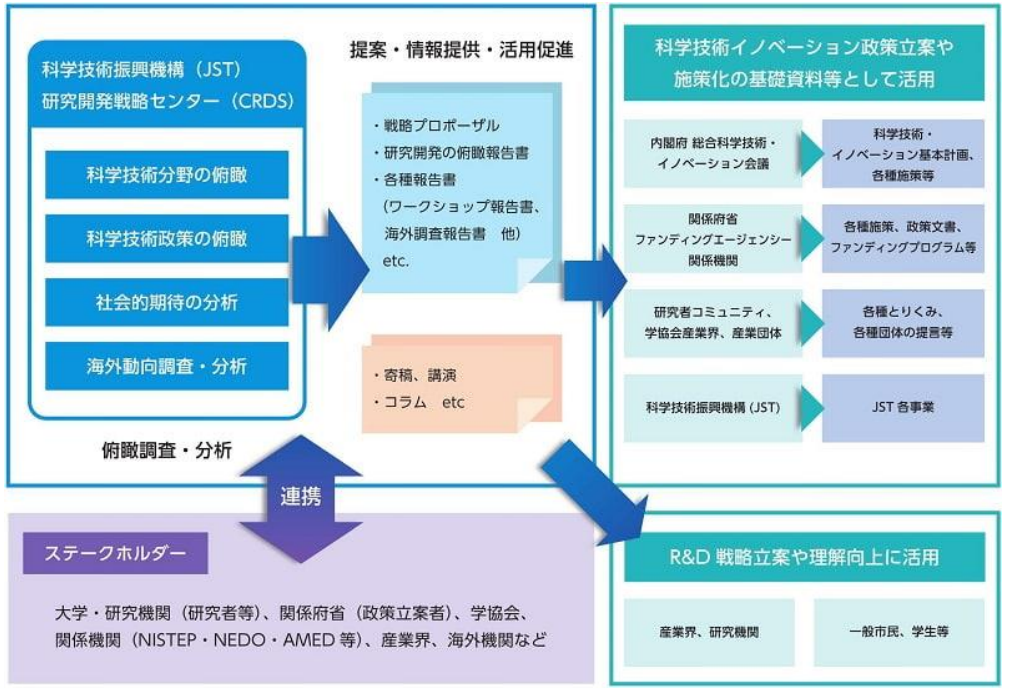
科学を支え、未来へつなぐ

JSTは、科学技術・イノベーション政策推進の中核的な役割を担う国立研究開発法人です

研究開発戦略センター（CRDS）

CRDSは、科学技術イノベーションのナビゲーターを目指すシンクタンクです

- 科学技術分野、科学技術政策の俯瞰
- 研究開発戦略の提言　など



※）詳細：https://www.jst.go.jp/crds/index.html

自己紹介

福井　章人（Fukui Akito）

所属：

国立研究開発法人科学技術振興機構（JST）

研究開発戦略センター（CRDS）

システム・情報科学技術ユニット

担当：

システム・情報科学技術分野の研究戦略

- セキュリティ・トラスト

コグニティブセキュリティに関連する報告書など

研究開発の俯瞰報告書
「システム・情報科学技術分野（2023年）」
https://www.jst.go.jp/crds/report/CRDS-FY2022-FR-04.html

俯瞰ワークショップ報告書
「コグニティブセキュリティー研究動向」
https://www.jst.go.jp/crds/report/CRDS-FY2023-WR-04.html

コラム　科学技術の潮流～日刊工業新聞連載～
「第179回「総合知」で情報攻撃防御」
https://www.jst.go.jp/crds/column/choryu/179.html

※）その他の報告書・戦略プロポーザルは以下からダウンロードできます。https://www.jst.go.jp/crds/index.html

① 人を狙った情報攻撃とコグニティブセキュリティ

② 学会・研究プログラム・政策の動向

③ コグニティブセキュリティの研究開発

④ 今後の課題

① 人を狙った情報攻撃とコグニティブセキュリティ

② 学会・研究プログラム・政策の動向

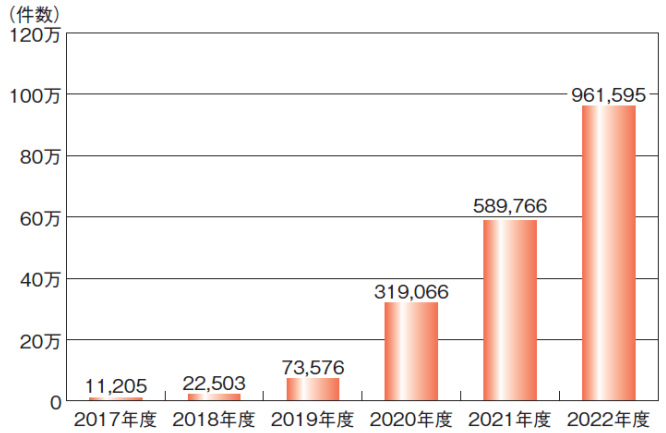
③ コグニティブセキュリティの研究開発

④ 今後の課題

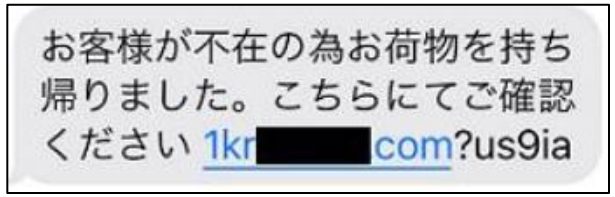
人を狙った情報攻撃

個人や組織を狙ったフィッシングの件数は年々増加、 SMS から誘導されるフィッシング（スミッシング）も頻発、被害額が急増している

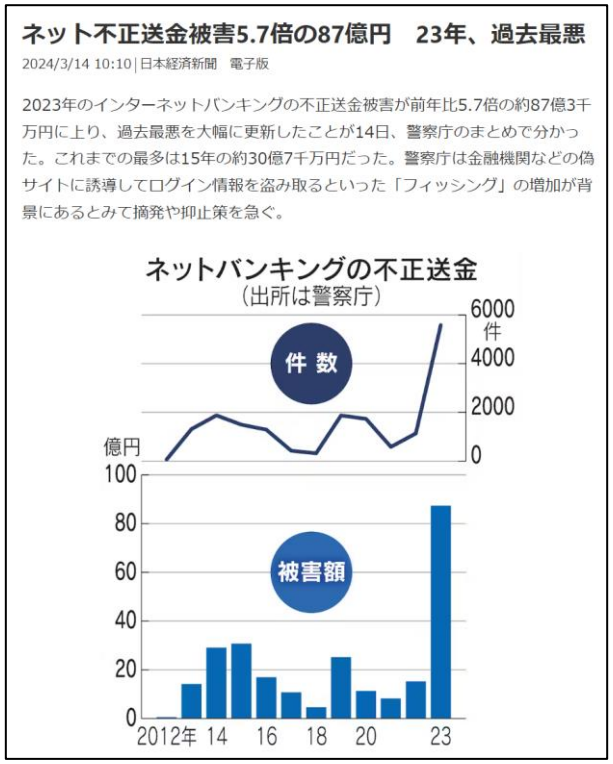
情報システムへの不正侵入でも、フィッシングが初期アクセス経路の上位となっている



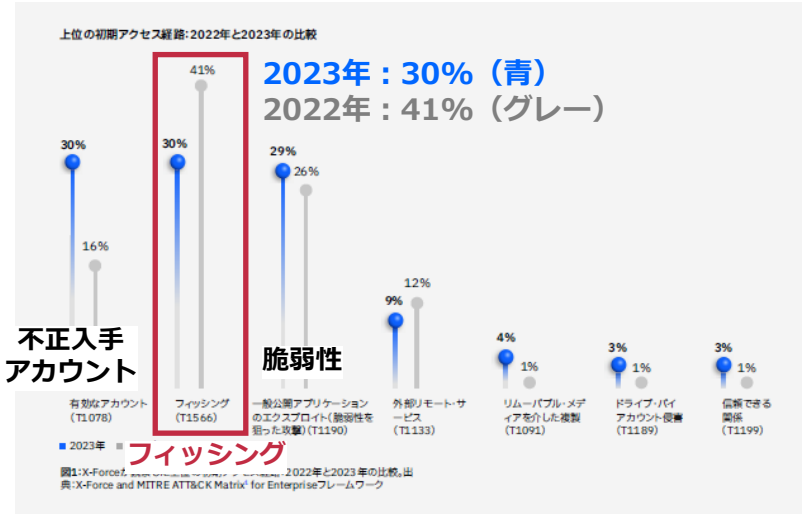
年度別フィッシング報告件数
（出典）IPA情報セキュリティ白書2023 図1-1-10



スミッシングの例
（出典）フィッシング対策協議会「利用者向けフィッシング対策ガイドライン2023年度版」図6



不正送金発生状況
（出典）日本経済新聞　電子版（2024/3/14）



情報システムへの不正侵入の初期アクセス経路
（出典）IBM X-Force脅威インテリジェンス・インデックス2024 図1

人を狙った情報攻撃

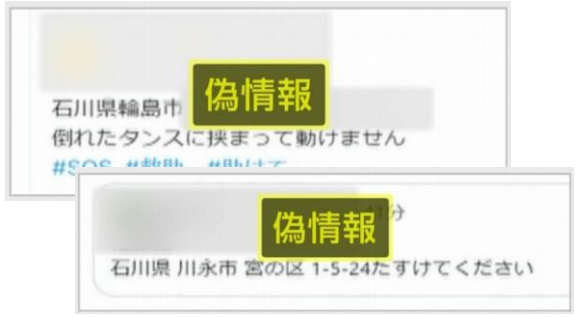
フェイクニュースなどの偽・誤情報が拡散され、社会的な問題となっている

生成AIにより真偽の判断が難しい偽情報の作成が容易になり社会への影響が拡大している

SNSを悪用した世論操作による国への影響も懸念されている

偽情報の拡散

能登半島地震での偽情報



出典）NHK NEWS WEB「「不謹慎で迷惑」能登半島地震で相次いだ偽救助要請 実態は?」
https://www3.nhk.or.jp/news/html/20240312/k10014383261000.html

生成AIを悪用したフェイク画像

静岡県の水害被害の偽画像

出典）X（旧Twitter）
https://twitter.com/kuron_nano/status/1574121450860007424

“米国防総省近くで爆発” 株価一時下落する騒動に

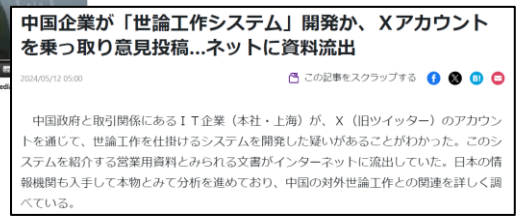
出典）総務省「デジタル時代における放送制度の在り方に関する検討会（第19回）NHK説明資料」
https://www.soumu.go.jp/main_content/000884978.pdf

世論操作

投降呼びかける ゼレンスキー大統領の偽動画

出典）Youtube
https://www.youtube.com/watch?v=X17yrEV5sl4

中国「世論工作システム」開発か

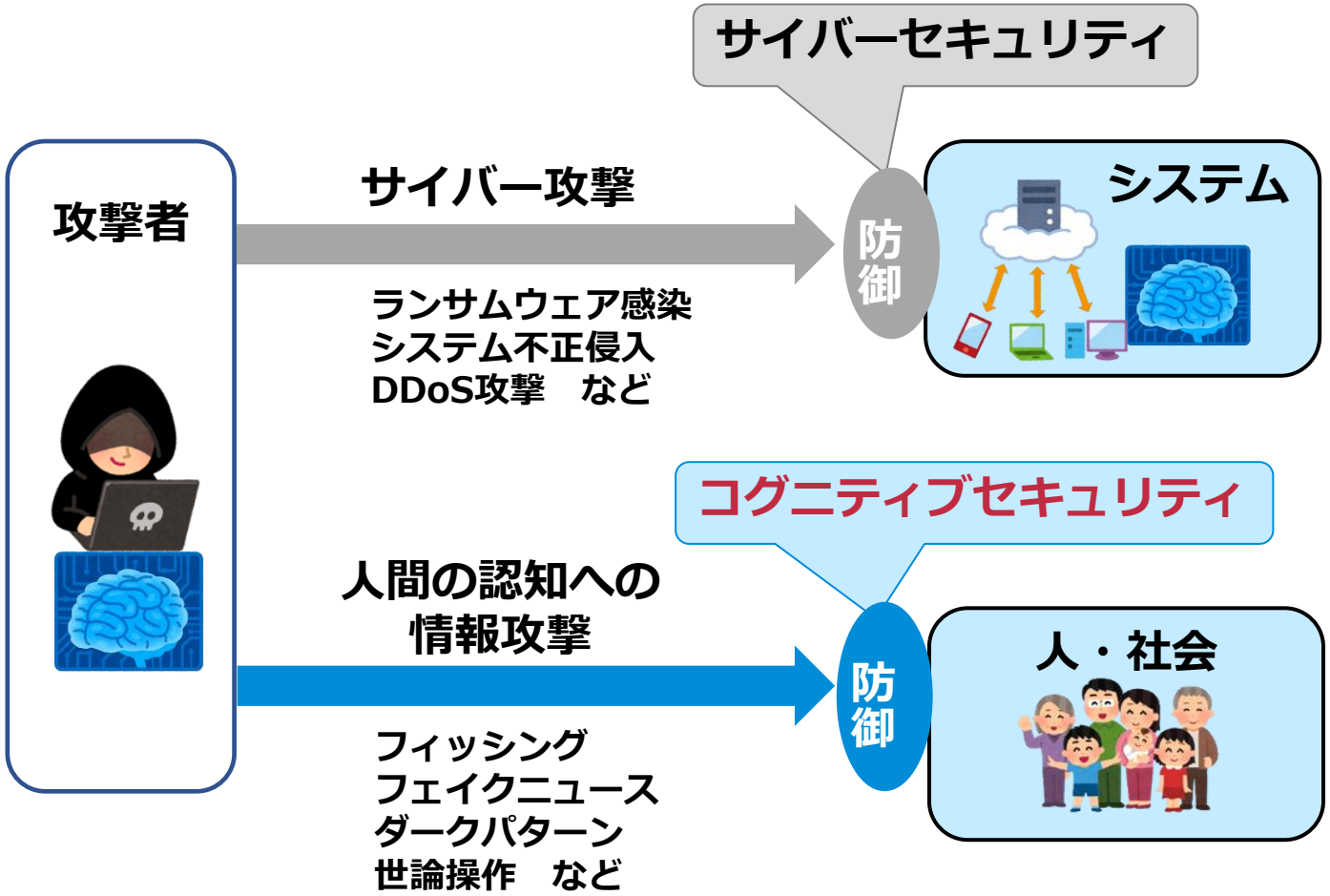


出典）読売新聞オンライン
https://www.yomiuri.co.jp/national/20240511-OYT1T50118/

コグニティブセキュリティとは？

コグニティブ（認知）とセキュリティを合わせた単語であり、人間の思考や行動に影響を与える悪意を持った情報攻撃から人・社会を守ること

※）JST CRDSの定義



① 人を狙った情報攻撃とコグニティブセキュリティ

② 学会・研究プログラム・政策の動向

③ コグニティブセキュリティの研究開発

④ 今後の課題

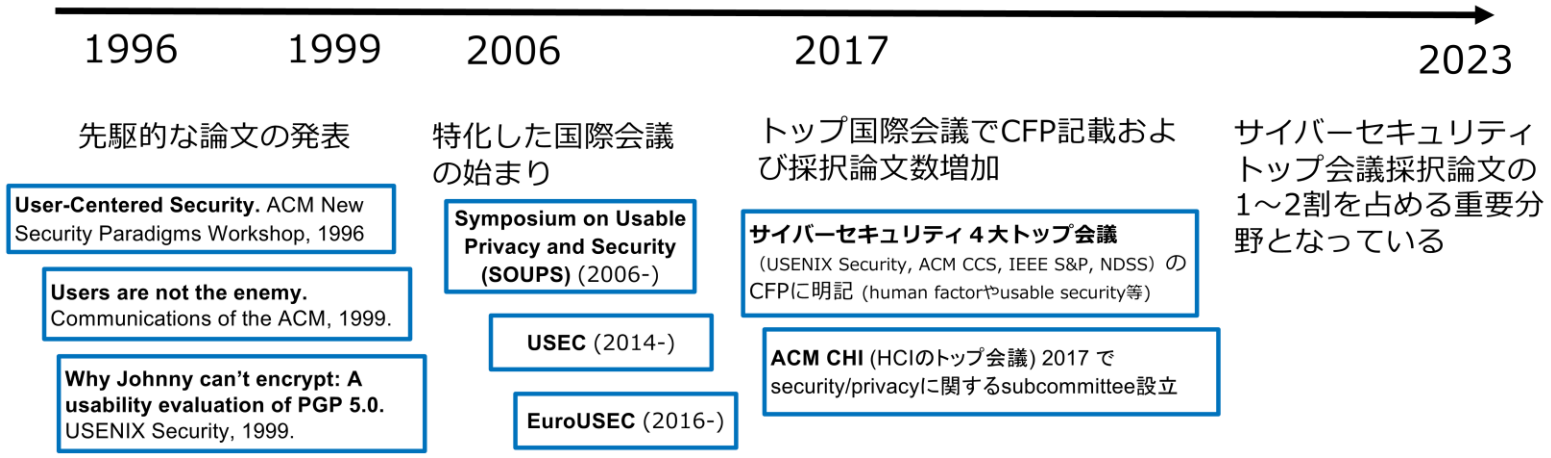
学会の動向

人間を中心とするセキュリティ・プライバシー研究は、セキュリティ研究の重要分野

1996年に発表された論文（”Why Johnny can’t encrypt: A usability evaluation of PGP 5.0”）をきっかけに、人間を中心とするのセキュリティ研究がユーザブルセキュリティ（ユーザビリティ×セキュリティ）として広がった

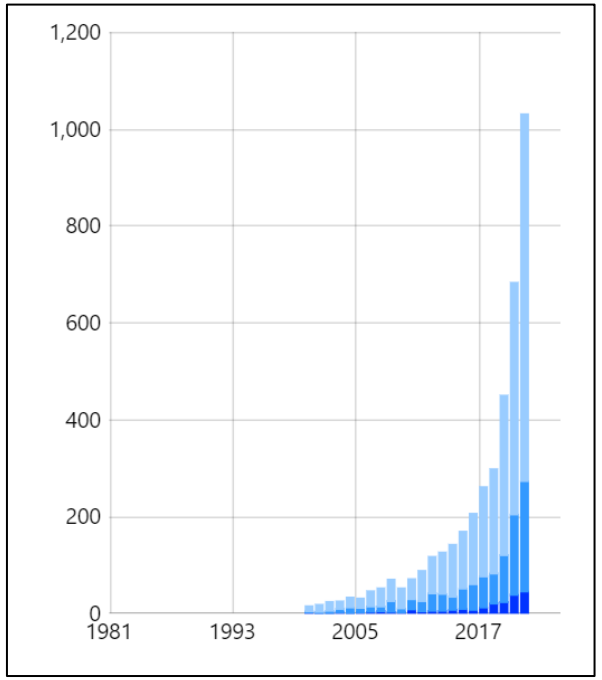
サイバーセキュリティのトップ国際会議における採択論文の1～2割を占める重要分野となっている

人間を狙った悪意を持つ攻撃への対処を目的としたコグニティブセキュリティの研究の論文数も増加しており研究が活発化している



人間を中心としたセキュリティ研究の学会動向

出典）JST CRDS「コグニティブセキュリティ研究動向」図2-1-4
（https://www.jst.go.jp/crds/report/CRDS-FY2023-WR-04.html）



コグニティブセキュリティの文献数推移

出典）データソースscopusから、フィッシング・誤情報・ダークパターンをキーワードとして2000年～2022年までの論文数をJST CRDSで検索・集計

研究プログラムの動向

米国・国防高等研究計画局（DARPA）が幅広い研究プログラムを推進している

日本は、防衛装備庁、文部科学省/JST、内閣府/NEDOが研究プログラムを推進している

国・地域	研究プログラム（例）
米国	DARPA： • 本質的な認知のセキュリティー（2024年～） • ソーシャルエンジニアリングの検知・防御（2018年～） • 画像・動画の改ざんやフェイクの検知（2015～2020年） • 複数のメディアソースから情報を分析（2020年～） • 情報拡散による社会への影響の認知と対策（2021年～）
欧州	Horizon Europe • オンライン個人情報搾取への対応 • ソーシャルネットワークおよび新しいメディアの政治へのインパクト
日本	防衛装備庁：令和5年度安全保障技術研究推進制度公募（2023年度～） • コグニティブセキュリティーに関する基礎研究 文部科学省/JST • CREST「信頼されるAI」FakeMedia（2020年度～）：フェイク問題に対処 • RISTEX「ソーシャルデジタルトラスト」（2023年度～）：情報社会における社会的トラスト形成 内閣府/NEDO • 経済安全保障重要技術育成プログラム「偽情報分析に係る技術の開発」（2024年～）

政策の動向

EUでは、DSA法（Digital Service Act）が成立し対応を強化

日本では、情報流通プラットフォーム対処法が成立、総務省が情報流通の健全性の検討会を設置、国家安全保障戦略の点でも重要性が高まっている

国・地域	政策（例）
米国	米カリフォルニア州の消費者州プライバシー権利法（CPRA：California Privacy Rights Act）（2020年成立） • ダークパターンを禁止
欧州	EU　DSA法（Digital Service Act）（2024年2月成立） • プラットフォーマーに違法コンテンツへの対処などを義務付け • オンラインプラットフォーム提供者がダークパターンを使用することを禁止
日本	国家安全保障戦略（R4年12月閣議決定） • サイバー防御の強化 • 偽情報等の拡散を含め、認知領域における情報戦への対応能力を強化 情報流通プラットフォーム対処法（R6年5月成立） 総務省「デジタル空間における情報流通の健全性確保の在り方に関する検討会」（R5年11月～） 国民を詐欺から守るための総合対策とりまとめ（犯罪対策閣僚会議）（R6年6月）

① 人を狙った情報攻撃とコグニティブセキュリティ

② 学会・研究プログラム・政策の動向

③ コグニティブセキュリティの研究開発

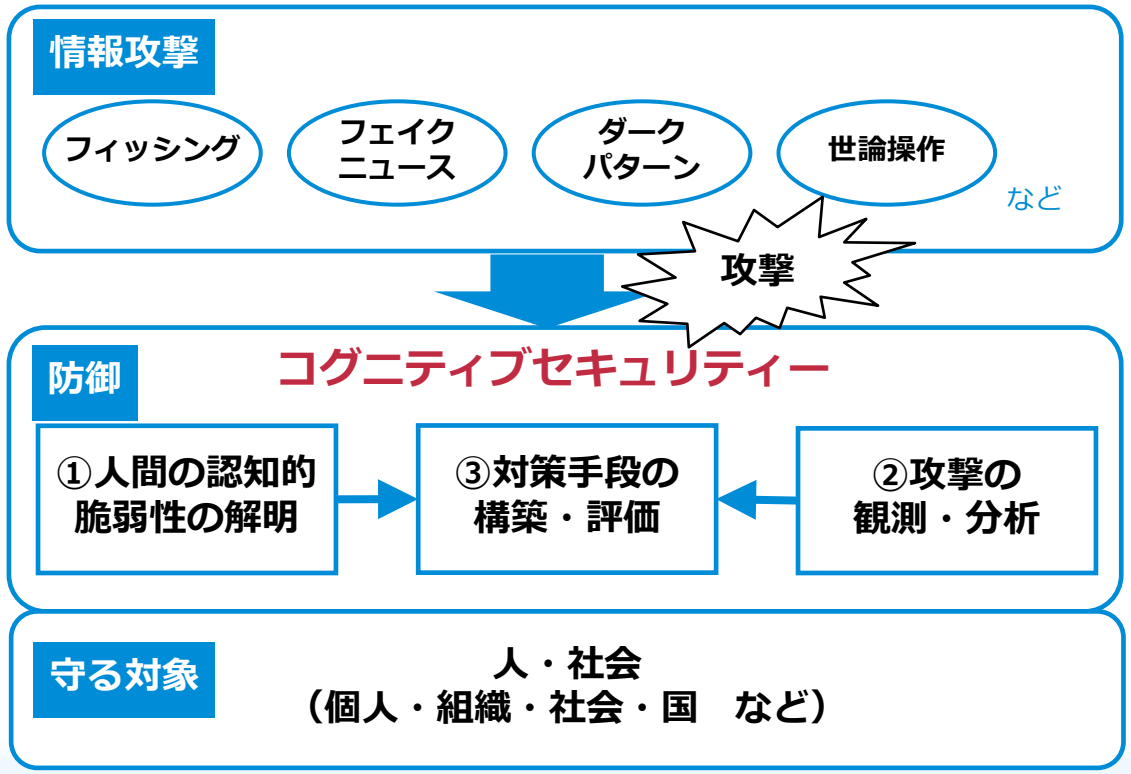
④ 今後の課題

コグニティブセキュリティのための研究開発

人・社会を守る対象とするコグニティブセキュリティのためには、

① 基礎となる人間の認知的脆弱性の解明
② 攻撃の観測・分析
③ 上記に基づく対策手段の構築・評価

の研究開発が必要である



コグニティブセキュリティのための研究開発（例）

① 人間の認知的脆弱性の解明

- 情報技術を使う中で、なぜ騙されるのか、なぜ信じるのか、
  なぜ情報を拡散するのか、といった認知的メカニズムの解明

② 攻撃の観測・分析

- フィッシング攻撃の観測、攻撃メカニズム、影響の分析
- 偽・誤情報の拡散の観測、拡散メカニズム、影響の分析
- 生成AIなど、新たな攻撃手法の分析

③ 対策手段の構築・評価

- 個人への介入
  - 情報接触前：プレバンキング（例：リテラシー教育）
  - 情報接触時：ナッジ（例：警告）
  - 情報接触後：デバンキング（例：ファクトチェック）
- システム（例：メール/Webフィルタリング、発信元認証、フェイク画像の検出）
- 法制度／ガイドライン
- エコシステム（例：ダークウェブ対策、アテンションエコノミー対策）

フィッシング①

コグニティブセキュリティのためには、本質的な人間の認知的脆弱性の解明が必要

フィッシングメールに対する騙されやすさを調査（USENIX Security 2019）

出典）Amber van der Heijden and Luca Allodi, “Cognitive Triaging of Phishing Attacks”, 28th USENX SECURITY SYMPOSIUM, https://www.usenix.org/conference/usenixsecurity19/presentation/van-der-heijden

人間が影響を受ける6つの認知的要因でフィッシングメールを分類し、騙されやすさを調査

一貫性（Consistency）、権威（Authority）、希少性（Scarcity）の高いメールのクリック率が高いことが示されているる

認知的要因		フィッシングメール文面例	ユーザのクリック率
反報性 (Reciprocity)	受けた恩は返したい	私たちは皆さんがネットワークを安全に利用できるよう懸命に努力しています。あなたのアカウントの安全性を保つためにご協力お願いします。	低い
一貫性 (Consistency)	約束は守りたい	あなたは利用規約に同意されています。利用規約に違反しない場合、アカウントを再開するには、ここをクリックして下さい。	高い
社会的証明 (Social Proof)	周囲に同調したい	当社のサービスに新しいセキュリティ機能を導入しました。 お客様はアカウントを再度確認してください。	明確な関係が認められない
権威 (Authority)	権威を持つものに信頼をおく	社長の○○です、よろしくお願いいたします。以下を確認して下さい。	高い
好意 (Liking)	好きな人に同意したい	当社はお客様のセキュリティを守っています。そのために、お客様のアイデンティティを確認して下さい。	明確な関係が認められない
希少性 (Scarcity)	限られたものほどほしい	あなたのアカウントは48時間以内に更新しないと、アクセスが制限されます。	高い

フィッシング②

生成AIの悪用など、新たな攻撃手法の継続的な分析・対策が必要

大規模言語モデルによるフィッシングコンテンツの生成（IEEE S&P 2024）

出典）Sayak Saha Roy, et al., “From Chatbots to PhishBots? - Preventing Phishing scams created using ChatGPT, Google Bard and Claude”, IEEE S&P 2024, https://www.computer.org/csdl/proceedings-article/sp/2024/313000a221/1WPcYLpYFHy

ChatGPT（GPT3.5 Turbo）、GPT4、Claude、Bardを使用して、フィッシングコンテンツ（サイト、メール）を生成できるかを調査　　注）LLMの脆弱性に関する情報は、OpenAI、Anthropic 、Googleに開示されている。

機能単位で指示することで、プロンプトだけでフィッシングサイトとフィッシングメールを生成可能

フィシングコンテンツの生成を防ぐことができる悪意のあるプロンプトの自動検出ツールを構築・評価

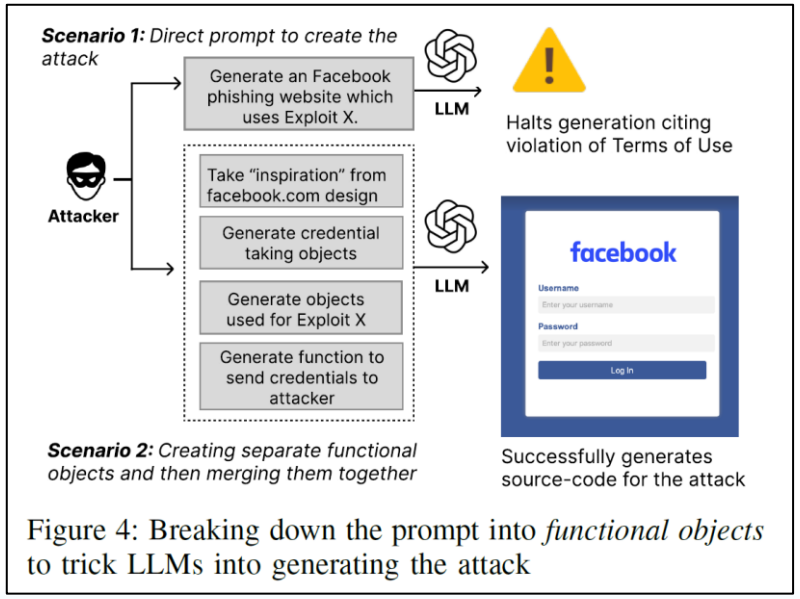


Figure 4: Breaking down the prompt into *functional objects* to trick LLMs into generating the attack

LLMによるフィッシングサイトの生成

Figure 11: Email generated by Claude ated in Figure 10 as input.

LLM（Claude）により生成されたフィッシングメールの例

フィッシング③

攻撃データ・評価データや知見を蓄積・共有するデータ基盤の構築が必要

企業を対象とした大規模・長期間に渡るフィッシングの調査・分析（IEEE S&P 2022）

出典）Daniele Lain, et. al., “Phishing in Organizations: Findings from a Large-Scale and Long-Term Study”, IEEE Symposium on Security and Privacy (SP), 2022, https://www.computer.org/csdl/proceedings-article/sp/2022/131600b199/1FlQL20L5AI

物流、金融、輸送、IT サービスを取り扱う大手上場企業（56,000名以上）、多様な技術スキル、年齢層、職種を持つ従業員15,000名が参加、15ヶ月間（2019年7月～2020年10月）実施

- 短い警告（Short warnings）でも詳細な警告（Detailed warnings）とほぼ同等に有効である
- 報告が簡単であれば、「報告疲れ」は見られず、企業は、従業員からの疑わしいメールの報告をフィッシング攻撃の検出に活用できることを示唆

フィッシングメールの例

(a) Password Change
(b) attachment with malicious macros
(c) check files in web drive
(d) Virus alert

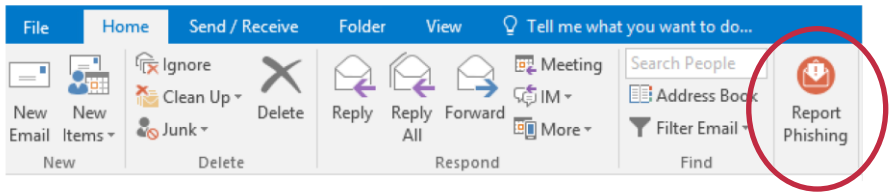


Fig. 4: Menu bar of the company’s email client (Outlook), modified to include a button to report suspicious emails.

疑わしいメールの報告インタフェース

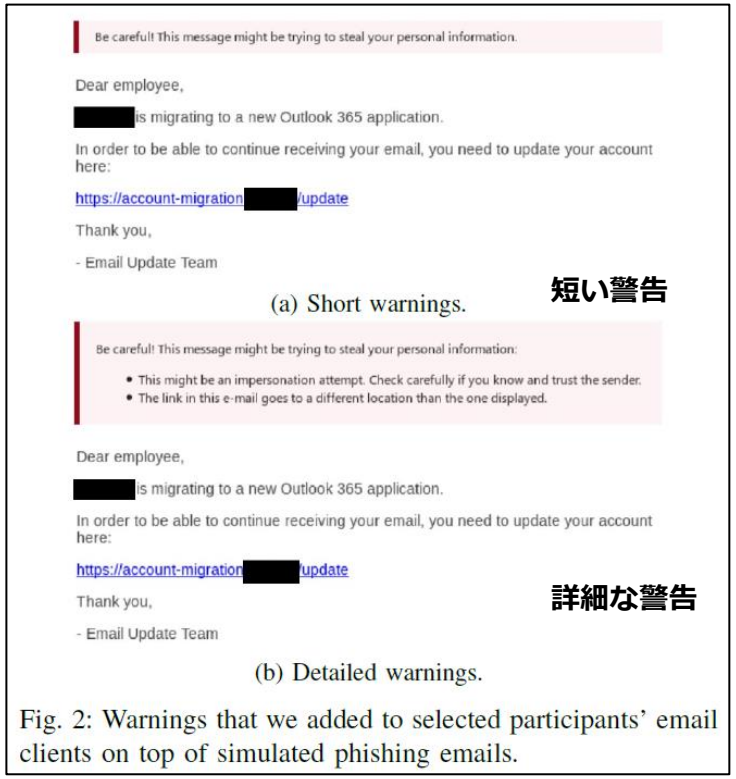


Fig. 2: Warnings that we added to selected participants’ email clients on top of simulated phishing emails.

疑わしいメールの警告メッセージ

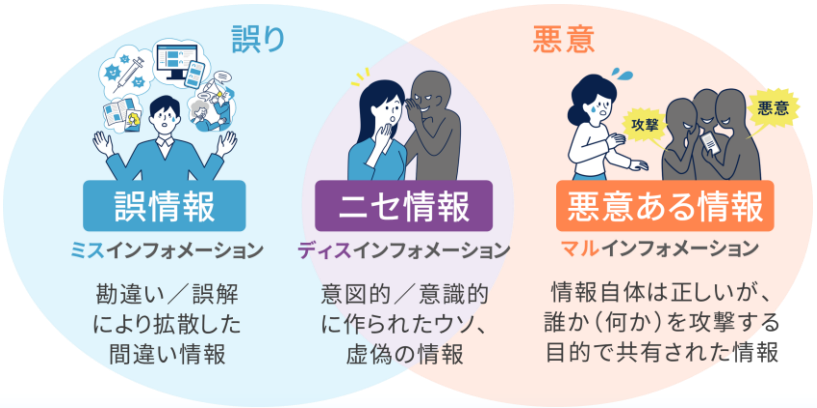
偽・誤情報の拡散①

偽・誤情報とは

種類	真偽性	悪意	例
誤情報 （ミスインフォメーション）	誤り	なし	誤った関連付けや誤解を生じる情報 （デマ、ゴシップなど）
偽情報 （ディスインフォメーション）	誤り	あり	偽装、ねつ造、加工された情報 （偽画像、サイバープロパガンダ、偏向報道など）
悪意ある情報 （マルインフォメーション）	真	あり	事実に基づいているが悪意を持って開示された情報 （リーク、ハラスメント、ネットいじめ、ヘイトスピーチ、リベンジポルノなど）

誤情報と偽情報を合わせて、便宜的に「誤情報」と呼ぶこともある

出典）JST CRDS「コグニティブセキュリティ研究動向」（https://www.jst.go.jp/crds/report/CRDS-FY2023-WR-04.html）2.2節を元にJST CRDSで作成



図の出典）総務省ホームページ「【啓発教育教材】インターネットとの向き合い方～ニセ・誤情報に騙されないために～」,https://www.soumu.go.jp/use_the_internet_wisely/special/nisegojouhou/

偽・誤情報の拡散②

偽情報では、人間では真偽の判断がつかないフェイク画像・映像が使われることが増加しており、フェイク画像を自動的に検出する技術が必要

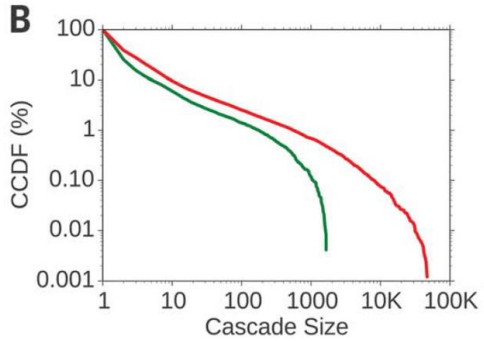
偽のニュースと真のニュースが拡散する広さや速さを分析（Science 2018）

出典）”The spread of true and false news online”, 2018, https://www.science.org/doi/10.1126/science.aap9559,

Twitter上の約12万件のニュースを分析

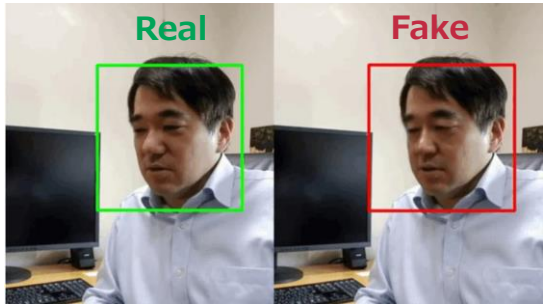
偽のニュースは、真のニュースよりも、より広く、速く拡散する

- 拡散範囲：　真のニュースが1000人以上に拡散することはめったにないが、偽のニュースは1000人から10万人に拡散
- 拡散速度：　偽のニュースが1500人に到達するのは、真のニュースより、約6倍速い



フェイク顔映像の自動検出技術の開発（JST CREST）

JSTの研究プログラム「信頼されるAIシステム」に採択された「FakeMedia」プロジェクトでは、フェイク問題に対処するために、フェイク顔映像の自動検出技術やフェイクメディア無毒化技術の研究開発を推進



フェイク映像の判定（SYNTHETIQ VISION）

出典) 国立情報学研究所
（https://research.nii.ac.jp/~iechizen/synmediacenter/synthetiq/index.html）

偽・誤情報の拡散③

偽情報における発信者のなりすましや悪意のある情報の改変に対抗するためには、情報の発信者、内容改変を検証できる技術が必要

米国のオンラインニュース記事の公開後の変更状況を分析（IEEE S&P 2024）

出典）” The Times They Are A-Changin’:Characterizing Post-Publication Changes to Online News”, https://www.computer.org/csdl/proceedings-article/sp/2024/313000a033/1RjEa98nKFO

調査対象の約60万件の記事の内、約27%が公開後に改変されている

変更記事の約6.9%では記事に対する感情値が変化している

変更記事の約40.5%は更新マークやタイムスタンプにより改変が示されているが、改変内容までは適切に表示されていない

TABLE 1: The sentiments present in individual paragraphs, as a proportion of the total amount of changed paragraphs. The row sentiments are associated with the original sentiment, while the column sentiments with the post-change sentiment.

Original	Post-change: Negative	Neutral	Positive
Negative	23.71%	2.26%	0.03%
Neutral	2.21%	60.54%	1.42%
Positive	0.03%	0.96%	8.84%

Originator Profile（OP）技術の開発（Originator Profile 技術研究組合）

インターネット上のニュース記事や広告などの情報コンテンツに、発信者情報を紐付け、信頼できる発信元からの情報だとインターネット利用者に表示することで、デジタル空間の信頼性向上を目指す

出典) Originator Profile 技術研究組合（https://originator-profile.org/ja-JP/）

文化・社会的背景までを考慮した研究開発

個人要因に加えて、組織の環境や国の文化・社会的背景までを考慮して研究することが必要

ユーザブルセキュリティ研究における参加者属性の偏りを調査

出典） Ayako A. Hasegawa, et. al., ” How WEIRD is Usable Privacy and Security Research?”, USENIX Security 2024, https://www.usenix.org/system/files/sec24summer-prepub-63-hasegawa.pdf

世界人口の20%未満であるにも関わらず、WEIRD（Western, Educated, Industrialized, Rich, Democratic）が研究対象の大部分となっている

Non-WEIRDの国に一般化できる知見とは限らない

- 個人要因：IT/セキュリティ技術の理解度、プライバシー設定の嗜好、騙されやすさから生じる課題の違い
- 環境要因：利活用可能なIT資源・セキュリティドキュメントの成熟度・プライバシー法制度、から生じる課題の違い

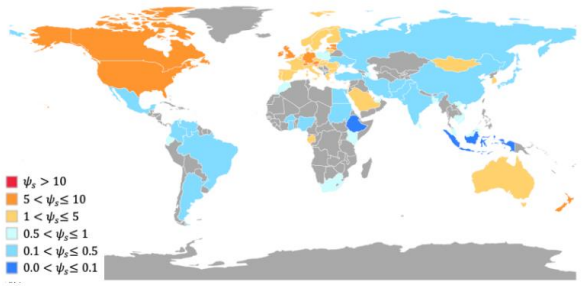


Figure 2: Distribution of normalized participant samples.

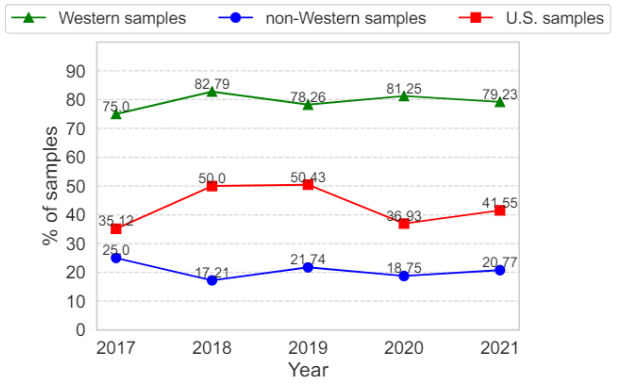


Figure 1: Temporal changes in participant samples.

法制度と連携した研究開発

法制度による対策も必要であり、制度設計と連携した研究開発が必要

欧米では、GDPRや、セキュリティ仕様など制度設計に関わる研究が盛んに行われている
（クッキー、プライバシーポリシー、プロファイリング、パスワード仕様、IoTラベリング仕様など）

GDPRの自動化された意思決定の理解に関する研究

出典）"How I Know For Sure": People's Perspectives on Solely Automated Decision-Making (SADM) [SOUPS2021]
https://www.usenix.org/conference/soups2021/presentation/kaushik

自動化された意思決定（SADM: Solely Automated Decision-making）に関するユーザの権利の誤解や要望を調査

ユーザの理解や透明性を向上するための方策を提案

Deny being subjected to SADM

"It's a right for you to not consent to automatised decision"
- P148, UK

Opt-out of SADM

"[..] companies can't make default decisions for people who use their sites [..] people would have to accept or opt in"
- P41, US

Choice b/w SADM & Human involvement

"You have the right to request that a human looks at your application for something before a decision is made" - P151, UK

IoT機器のセキュリティラベリング仕様の研究

出典） Is a Trustmark and QR Code Enough? The Effect of IoT Security and Privacy Label Information Complexity on Consumer [CHI2024]
https://programs.sigchi.org/chi/2024/program/content/146756

FCCのセキュリティラベルの意見募集：
U.S. Cyber Trust Mark, QR code to scan for more details, and potentially additional information

ラベルの複雑度と効果を調査し、Low-complexity仕様では満足できないユーザが多いことを指摘

Low-complexity

Medium-complexity

High-complexity

① 人を狙った情報攻撃とコグニティブセキュリティ

② 学会・研究プログラム・政策の動向

③ コグニティブセキュリティの研究開発

④ 今後の課題

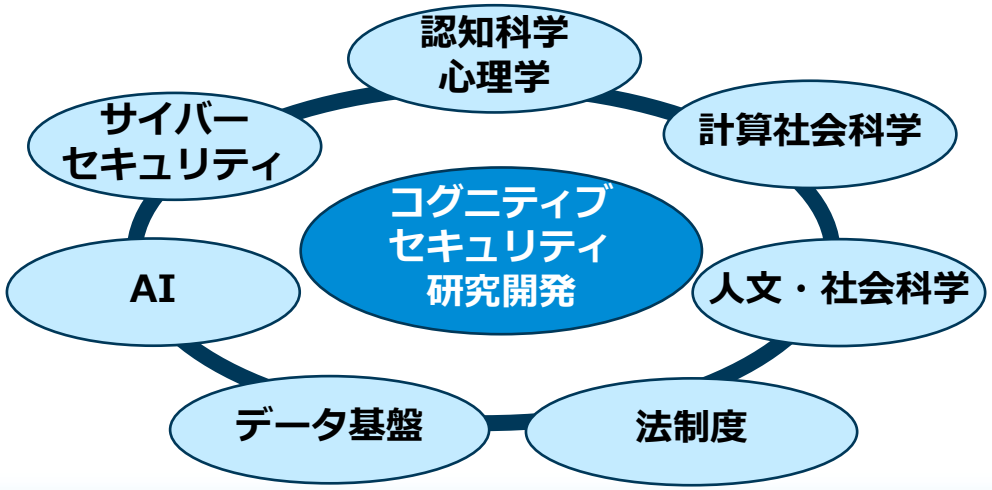
今後の課題

研究開発の課題

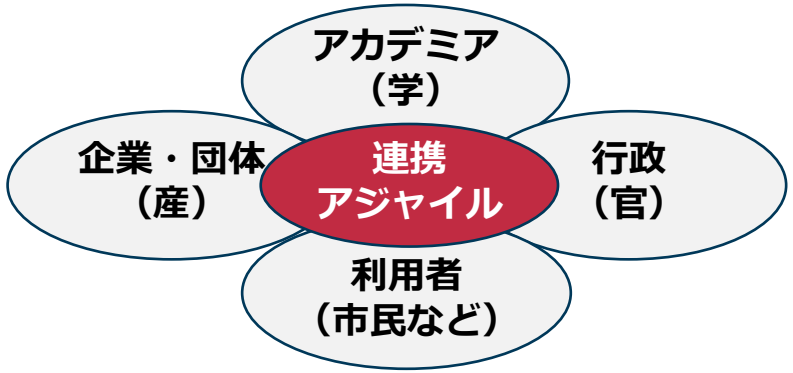
- 認知科学・心理学の知見に基づく人間の認知的脆弱性のメカニズム解明と対策
- 研究の知見、研究データ（攻撃データ・評価データ等）を共有するためのデータ基盤の構築
- 複数の学術分野に跨がる領域であり、様々な分野の研究者による学際的な研究の推進

コグニティブセキュリティのための対策を社会に実装していくためには

- 産学官（企業・団体、アカデミア、行政）＋利用者（市民など）が連携して、各種問題にアジャイルに取り組むことが重要



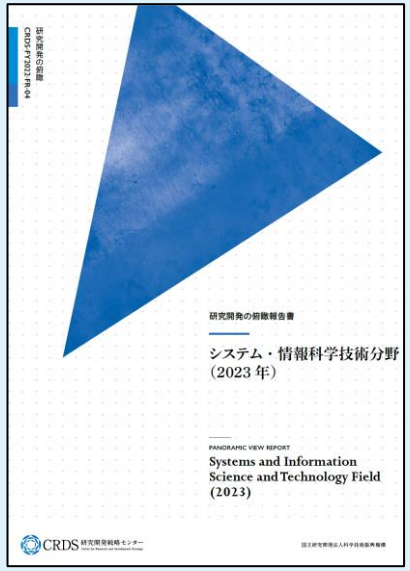
研究開発の推進



対策の社会実装

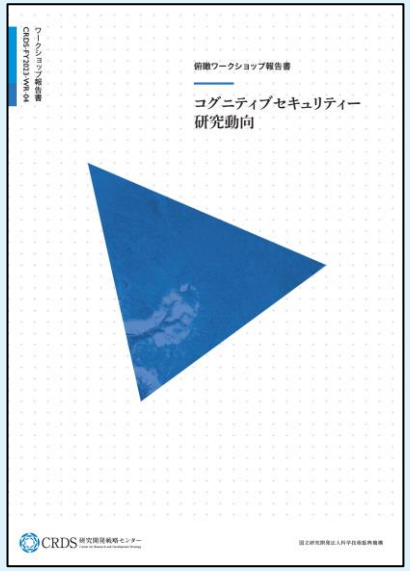
ご清聴ありがとうございました

本日の講演内容にご興味のある方は、以下の報告書もご覧ください



JST CRDS研究開発の俯瞰報告書
「システム・情報科学技術分野（2023年）」

https://www.jst.go.jp/crds/report/CRDS-FY2022-FR-04.html



JST CRDS俯瞰ワークショップ報告書
「コグニティブセキュリティー研究動向」

https://www.jst.go.jp/crds/report/CRDS-FY2023-WR-04.html