U-Rec

9th Grand Canvas ショートトークセッション ｜ 2026年10月6日

フィジカルAIは何に騙されるのか

敵対的攻撃の実例と、その備え

株式会社U-Rec 代表取締役 菅野 吉洋

■ 実例：認識への細工は、実機の動作まで届く

U-Rec

認識への細工が、見逃し・加速・作業失敗という「動作」に届いた実例です。

騙されたAI	騙した入力	起きたこと	出典
① 人物検出AI	身体の前に持った印刷パッチ	対象の検出器が人物を見落とす	Thys ほか（2019）
② 運転支援AI	速度標識に貼った短いテープ	35→85 mph と誤認、実車が加速	McAfee ATR（2020）
③ VLAロボット	カメラに映る敵対的パッチ	通常成功する作業の平均43%が失敗	Wang ほか（2024–25）

実証の範囲：① 印刷物を撮影し、対象の検出モデルで検証　② 2016年式の車両構成。研究者は目標速度に達する前に制動

③ 実ロボットで、通常成功する作業のみを対象に評価。いずれも管理された研究実証で、事故や侵入の多発を示すものではありません

共通点は、性能を下げる意図で入力に細工していること。

遮蔽や照明といった「想定される変動」を測る、通常の頑健性評価の外にあります。

出典の詳細は15枚目

■ 場面：騙され方は 4つ。どこで効き、何が起きるか

U-Rec

騙され方は4つ。影響の大きさは、騙された判断の先に何がつながっているかで決まります。

騙され方	どういう誤りか	効く場面	現場で起きうること
① 見落とす	いる人を、いないと判断	立入検知、協働エリアの安全監視	人がいるのに止まらない
② 見間違える	別の物・別の値と判断	標識を読む移動体、外観検査	速度を誤る。不良を通す
③ 動作が乱れる	狙った動作から外れる	VLAロボットの把持・組立・搬送	作業の失敗、落下、接触
④ 命令として読む	画像内の文字を指示と解釈	伝票・ラベル・掲示物を読む現場	指示していない作業をする

①〜③は2枚目の実例に対応。④は、画像内の文字による間接プロンプトインジェクション（実例には含めていません）。

「現場で起きうること」は当社による想定で、実際の事故の報告ではありません。

当社が強みを持つのは、①見落とす と ②見間違える（色をつけた 2行）。

視覚認識そのものを狙う敵対的攻撃への防御です。

当社による整理（Vigil 製品ページの想定脅威をもとに作成）

■ 影響が大きくなる 4つの条件

U-Rec

同じ騙され方でも、次の条件が重なるほど、影響は大きくなります。

条件	なぜ影響が大きくなるか	対応する備え
① 誰でも入力を置ける	物を置くだけで成立し、システムへの侵入が要らない	導入前に測る（備え①）
② 判断が動作に直結する	止める層がなければ、AIの出力がそのまま物を動かす	実行前に行動を検証（備え②）
③ 同じモデルを多数で使う	同じ入力で、同じ誤りが同時に起きうる（仮説）	導入前に測る。段階的に導入
④ 記録が残らない	監視上は正常のまま。事後に原因を説明できない	監査記録を残す（備え②）

①の例：公共空間のポスターやスマホ画面、外部から届く伝票・ラベル。 ③の例：多拠点で同じ日に同じ誤作動が起きる。

③は当社が検証したい研究仮説で、確認された事実ではありません。

備えは2つに集約できます。導入前に「測る」ことと、実行時に「止めて、残す」ことです。

■ U-Recの強み：視覚認識を狙う敵対的攻撃への防御

U-Rec

「見落とす」「見間違える」を突く攻撃を、自ら作り、測り、止める。この3つを自社で持っています。

できること	内容	現在地
作る（攻撃の再現）	検出を妨げるテクスチャを自ら最適化して当てる	人物検出器3種の実験を論文で公開
測る（耐性の評価）	同じ条件・同じ動作点で、対策の前後を比べる	評価コードを公開
止める・残す（Vigil）	入力の検査、行動の検証、経緯の記録を前後に挟む	開発中。大手メーカーとPoCを実施中

攻撃を作れるから、防御が効いたかどうかを、同じ条件で確かめられます。

ほかに、AIシステム受託開発／AIセキュリティコンサルティング／AI研修を提供しています。

設立 2024年12月 ｜ 所在地 栃木県宇都宮市 ｜ 加盟・採択等 NVIDIA Inception Program／Google for Startups／J-StarX（JETRO）

出典：Kanno (2026)、当社コーポレートサイト（u-rec.jp）

■ 備え①：導入前に、騙され方を測る

U-Rec

当社の実験では、最適化した攻撃より、単純な迷彩のほうが検出率を下げました。

何に騙されるかは、測ってみないと分かりません。測り方の設計で、結論も変わります。

観測したこと	評価設計への含意
検出率は、置換範囲の水準で大きく変わった	想定条件は一点ではなく段階化し、同一シーンで水準ごとに測る
最適化は、単純な迷彩に一貫して勝たなかった	高度な攻撃の再現より先に、低コストな条件をベースラインとして測る
検出率が下がる条件は、モデルごとに違った	共通の校正基準で動作点を固定し、条件別に候補モデルを並べる
サンプル数と動作点で、結論が動いた	シーン数・動作点・再現条件を測る前に決め、不確かさを併記する

攻撃への耐性を「測れる品質」として扱うには、この設計が要ると考えています。

出典：Kanno (2026)。人物検出器3種・置換範囲が最大のときの結果。右列は当社の提案。詳細は9〜12枚目

■ 備え②：実行時に止めて、残す　— Vigil

U-Rec

「見落とす」「見間違える」を狙う入力を、モデルの手前で検査します。フィジカルAI向けに開発中です。



Vigil 監査記録　入力・判断・出力・検知イベントを一元的に記録し、事後に経緯を再現できる形で残す

設計方針	内容
置換ではなく補完	既存の制御や安全機構（安全 PLC・非常停止）はそのまま。ROS／ROS 2 の構成に外付けする
段階導入	Shadow（本番に影響を与えず記録）→ Alert（通知）→ Block（高危険度のみ遮断）
約束しないこと	検知率や遅延は確約しない。PoCで、対象環境の実測として示す

出典：Vigil 製品ページ（ai-security.u-rec.jp）。PoCの進め方は13枚目

■ ご相談：3つの形で、ご一緒したい

U-Rec

「測る」技術と、Vigilを持って伺います。

形	ご一緒したい方	当社が持ち込むもの	最初の一歩
Vigil PoC	フィジカルAI・AIエージェントを現場に導入する企業	Vigil と、導入前の耐性評価	脅威シナリオを1時間で決め、Shadow運用から始める
新規事業での提携	評価・認証や安全機能を事業化したい、大企業の新規事業部門	敵対的検証の技術と評価基盤	御社の事業仮説に当社の検証を組み合わせ、メニューを試作
共同研究	大学・研究機関の研究者、企業の研究開発部門	公開した評価コードと実験設計	物理環境での再現性の検証と、VLAへの拡張

会場で、またはメールでお声がけください。

菅野 吉洋 ｜ y.kanno@u-rec.jp ｜ u-rec.jp

■ 参考：当社の実験の設計

U-Rec

「最適化した攻撃」と「単純な模様」を、同じシーン・同じ範囲で比べました。

先行研究では、テクスチャ・代理モデル・適用面積が同時に変わることが多いためです。

項目	設定
対象	公開の人物検出器 3種（Faster R-CNN／DETR／YOLO26n）
シーン	COCO val2017 から選んだ同一の 32シーン。別の32シーンでも順序を確認
置換範囲	人物領域の一部をデジタル合成で置換。置換画素数 ÷外接矩形面積で 0.074〜0.416 の6水準
対照条件	合成迷彩／平滑化した画素ノイズ／単色（いずれも最適化なし）
最適化条件	代理モデル（YOLOv8n）の人物スコアを下げるよう最適化したテクスチャ（ 3シード）
指標	モデルごとに事前校正して固定した閾値（ 0.981／0.981／0.886）での検出率。COCO APではない

出典：Kanno (2026) “Mask Extent and Texture Content”, Zenodo, DOI 10.5281/zenodo.22842246

■ 参考：最適化した攻撃は、単純な迷彩に一貫して勝たなかった

U-Rec

置換範囲が最大（0.416）のとき、3モデルすべてで、単純な迷彩のほうが検出率を下げました。

検出器	無加工	合成迷彩（対照）	最適化（3シード平均）	最適化 − 迷彩
Faster R-CNN	0.906	0.250	0.531	+28.1 pt
DETR	0.906	0.500	0.635	+13.5 pt
YOLO26n	0.781	0.094	0.188	+9.4 pt

開発32シーン・固定閾値での検出率（低いほど見逃しが多い）。別の32シーン、より強い代理モデル（試した範囲）でも順序は同じ。

- 検出率を最も大きく動かしたのは、置換範囲の水準だった（ Faster R-CNN・迷彩：0.906 → 0.250）
- 騙され方は、モデルごとに違った（迷彩・範囲 0.196：0.844／0.781／0.312）
- 32シーンで見えた最適化の小さな優位は、 132シーンでは確認されなかった。閾値を動かすと順序も入れ替わった

出典：Kanno (2026), Table 1–3。全6水準の数値は11枚目

■ 参考：全6水準の検出率（迷彩と最適化）

U-Rec

Faster R-CNN と DETR では、全水準で最適化の検出率が迷彩以上。YOLO26n では範囲0.358以下で最適化が1.0〜5.2ポイント低く、0.416で逆転する。

置換範囲	FRCNN 迷彩	FRCNN 最適化	DETR 迷彩	DETR 最適化	YOLO26n 迷彩	YOLO26n 最適化
0.074	0.906	0.906	0.906	0.906	0.844	0.812
0.136	0.875	0.896	0.875	0.896	0.500	0.479
0.196	0.844	0.854	0.781	0.875	0.312	0.292
0.298	0.656	0.760	0.594	0.719	0.188	0.177
0.358	0.531	0.708	0.531	0.688	0.125	0.073
0.416	0.250	0.531	0.500	0.635	0.094	0.188

固定閾値での検出率（開発32シーン）。FRCNN＝Faster R-CNN。迷彩は固定1生成、最適化は3シード平均。

無加工は全水準共通で 0.906／0.906／0.781。32シーンでは、1シーンの変化が3.125ポイントに相当する。

出典：Kanno (2026), Table 1・Table 2

■ 参考：この研究で確かめていないこと

U-Rec

項目	内容
合成の範囲	デジタル合成（画素の直接置換）のみ。布の形状・陰影、印刷、撮影は評価していない
面積の効果	範囲を変えると、マスクの形・位置・覆われる部位も同時に変わる。面積単独の効果は分離していない
テクスチャの比較	条件間で色分布を揃えていない。差のすべてを最適化の効果に帰すことはできない
サンプル	32シーン中16は最適化にも使用。132シーンへの拡張は既存シーンの再利用で、独立の追試ではない
32シーンでの差	YOLO26n・範囲0.196：迷彩 − 最適化 ＝ +2.1pt、名目95%区間 −12.5〜+16.7
132シーンでの差	同じ設定（実測0.191）で −1.0pt、区間 −7.8〜+5.8。最適化の優位は確認されなかった
位置づけ	実験条件下の記述的な結果。実システムでの攻撃成立・脆弱性・対策の有効性は示していない

出典：Kanno (2026), Sec. 5.2・Sec. 6

■ 参考：Vigil PoCの進め方（標準 4〜6か月）

U-Rec

目的は「万能防御の証明」ではなく、対象環境で何がどこまで現実的かを明確にし、本番導入の判断材料を残すことです。

時期	進めること
0〜1か月	環境確認とリスク定義。脅威モデルと評価軸を、対象環境に合わせて設定する
2〜3か月	検知器の実装とShadow運用。本番に影響を与えず、検知ログを蓄積する
4〜5か月	行動検証と閾値調整。運用に耐える閾値とアラート品質を作り込む
6か月	最終報告。本番展開の体制と責任分界まで提示する

残る判断材料：監視すべき脅威シナリオ／現場で運用できる閾値／自動遮断と人手確認の境界／本番導入時の体制と責任分界

技術指標の例：既知の攻撃シナリオに対する検知再現率／誤検知率とオペレータの確認負荷／
レイテンシ増分／危険アクション遮断ルールの有効性

出典：Vigil 製品ページ（ai-security.u-rec.jp）

■ 参考：通常の頑健性評価と、AIセキュリティ評価の違い

U-Rec

問いが違えば、測る条件も、導入時に決めることも変わります。

観点	通常の頑健性評価	AIセキュリティの評価
問い	想定される条件変動で性能がどう変わるか	性能を下げる意図で入力が変更されたとき、どこまで保てるか
変動の例	遮蔽、照明、距離、センサの劣化	敵対的パッチ・テクスチャ、画像内に書かれた命令文
分類	自然な変動（分布のずれ、ノイズ）	回避攻撃、間接プロンプトインジェクション
判断事項	許容する運用条件の範囲	脅威の想定範囲／検証・対策の優先順位／対策の採否

フィジカルAIでは、認識の誤りがそのまま動作になります。

右の列に、測定で答えられるようにする。これが当社の研究開発の出発点です。

■ 参考：公開物と出典

U-Rec

内容	出典・所在
実例① 人物検出	Thys, Van Ranst, Goedemé. “Fooling automated surveillance cameras: adversarial patches to attack person detection.” CVPRW 2019（arXiv:1904.08653）
実例② 運転支援	McAfee ATR. “Model Hacking ADAS to Pave Safer Roads for Autonomous Vehicles.” 2020年2月
実例③ VLAロボット	Wang ほか. “Exploring the Adversarial Vulnerabilities of Vision-Language-Action Models in Robotics.” arXiv:2411.13587（2024年公開、ICCV採択）
当社の研究論文	Kanno, Y. “Mask Extent and Texture Content: A Controlled Digital-Compositing Study of Person Detectors.” Zenodo, 2026年9月19日公開（プレプリント v1）。DOI 10.5281/zenodo.22842246
研究の解説記事	人物検出AIの見逃しリスクを測る（u-rec.jp/insights/mask-extent-texture-study/）
評価コード	github.com/u-rec-inc/mask-extent-texture-study（非商用ライセンス）
設計の考え方	Physical AIのセキュリティ設計（u-rec.jp/insights/physical-ai-security/）
Vigil／会社情報	ai-security.u-rec.jp ／ u-rec.jp/company/