

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

NICTにおける取り組み、検討のご紹介

情報通信研究機構(NICT)

鳥澤 健太郎

2024年4月15日

鳥澤 健太郎

(国研) 情報通信研究機構 (NICT)
フェロー
ユニバーサルコミュニケーション研究所副所長
データ駆動知能システム研究センター 主管研究員

専門は自然言語処理。日本学術振興会賞、Twitter Data Grantsなど受賞

1968年生まれ。静岡県河津町に育つ。下田北高校卒業。1992年東京大学理学部情報科学科卒。1995年東京大学大学院理学系研究科情報科学専攻中退。同年同大学助手。その後、北陸先端科学技術大学院大学准教授を経て、2008年より情報通信研究機構(NICT)勤務。2020年よりNICTフェロー。博士(理学)。

その他の詳細は鳥澤の個人ページ (<https://direct.nict.go.jp/members/torisawa/>)
をご参照ください。



- 能登半島地震におけるX(旧Twitter)上のデマ情報の分析
 - NICTが開発、一般公開していたD-SUMMというシステムによる分析
 - 関連報道
 - 日刊工業新聞「揺さぶる共助 能登半島地震のSNS分析」、2024年2月7,8,9日、
<https://www.nikkan.co.jp/articles/view/700949>
 - NHKニュース、2024年3月11日、
<https://www3.nhk.or.jp/news/html/20240311/k10014386161000.html>
- 生成AIとフェイクニュース等のリスク及び対処策について
 - 月刊正論2024年5月号（4月1日発売）の拙稿「複数の『正義』で『悪』を無効化する」の内容の一部に沿って
 - 以下の講演会で使用した資料を一部使用
 - 日本国際賞受賞記念講演会（2023年8月10日、<https://pco-prime.com/jpsympo/>）
 - 日本学術会議公開シンポジウム「生成AIの課題と今後」（2023年9月14日、https://www.youtube.com/watch?v=uw8_DEm3exg）

能登半島地震におけるX(旧Twitter)上の デマ情報の分析

- AIがX(旧Twitter)の投稿を読み、災害に関係する報告（「火災が起きている」など）を自動抽出
- 自治体名を指定するとそのエリアにおける災害に関する報告を自動抽出し、意味的なカテゴリごとに整理、提示する要約システム
- **報告と矛盾する投稿がある場合、デマの可能性があると自動的**に注意喚起
- 2016年より2023年度末まで試験公開。分析しているXの投稿は日本語投稿の10%

その他のバージョンはこちら > [DISAANA 令和6](#) 現在ご利用の D-SUMMは、令和6年能登半島地震試用版です。

現在指定しているエリア：石川県 都道府県を選択する 石川県の市町村を選択する

日時 2024-01-01 16:00 から 2024-01-01 17:10

その他の設定を表示する

エリア毎に要約 **カテゴリ毎に要約** 時系列で要約 地図表示

「石川県」でカテゴリ毎に要約

総報告数：4085 件 表示カテゴリの限定を全て解除

表示カテゴリの限定対象：
なし（地図表示での■は全カテゴリを集約）

- 災害 (1792)
- 救助 (19)
- トラブル (272)
- アラート (162)
- 人的被害 (4)
- 気象 (21)
- 建物 (173)
- ライフライン (81)
- 道路 (29)
- 不足 (1)
- 自治体・住所 (81)
- 地名：固有名 (39)
- 地名：その他 (33)
- 学校
- 病院
- 薬・医療サービス
- 消防等機関 (2)

災害(1792)

地震がおきている(1100) **注** 能登町(296) 志賀町(248) 金沢市(68) 珠洲市(47) 輪島市(47) 七尾市(23) 小松市(16) 穴水町(14) 加賀市(16) 津幡町(7) 中能登町(6) 羽咋市(7) かほく市(5) 宝達志水町(5) 能美市(5) 白山市(4) 内灘町(2) 野々市市(2) 川北町(1) **注** 石川県内(427)

爆発がおきている(2) 志賀町(2) 能登町(1)

災害の問題がおきている(5) 珠洲市(1) 能登町(1) 石川県内(3)

救助(19)

救助・救出を必要としている(18) 珠洲市(12) 能登町(3) 金沢市(3) 輪島市(1) 石川県内(1)

トラブル(272)

連休・不通になっている(29) 志賀町(7) 金沢市(8) 能登町(2) 輪島市(1) 石川県内(1)

津波・高潮がおきている(649) 輪島市(180) 能登町(129) 志賀町(69) 珠洲市(57) 七尾市(15) 金沢市(15) 加賀市(9) 小松市(3) 宝達志水町(2) 白山市(1) **注** 石川県内(227)

火山噴火がおきている(5) 能登町(1) 石川県内(5)

沈没がおきている(1) 石川県内(1)

未分類がおきている(3) 内灘町(1) 賀町(1) 珠洲市(1) 輪島市(1)

孤立が発生している(1) 輪島市(1)

故障・損傷がおきている(13) 金沢市(6) かほく市(1) 珠洲市(1) 輪島市(1) 石川県内(5)

総報告数：1260 件 石川県まで特定できている報告(3)


表示カテゴリの限定を全て解除

表示カテゴリの限定対象：
なし（地図表示での■は全カテゴリを集約）

- 災害 (379)
- 救助 (17)
- トラブル (139)
- アラート
- 人的被害 (40)
- 気象 (3)
- 建物 (47)
- ライフライン (51)
- 消防等機関 (19)
- 自治体・住所
- 地名：固有名
- 地名：その他

「石川県」で地図表示

地図 航空写真



注 マークについて：
被災報告の他に報告と矛盾する内容（「～では助けを求めている」や「～での救助要請はデマ」）を同時に探し、両方検出された場合は、デマの可能性があると注マークを付与

2017年7月5日 大分県庁での活用事例

激甚災害指定 九州北部豪雨

九州北部豪雨の際に大分県がDISAANA・D-SUMMを活用して情報分析を実施し、災害対応に役立てた

D-SUMM 「大分県」
カテゴリ毎の要約結果

大分県 (2017-07-05 11:00 ~ 2017-07-05 23:00) の被害状況 【生成日時:2017-09-15 21:09】

概要：07/05(17:04)大分県で災害(悪天候)情報あり。また、07/05(17:30)大分県で救助(孤立)に関する情報、07/05(12:30)大分県でトラブル(道路トラブル)に関する情報を検知しました。

災害：地震(4)、冠水・浸水(2)、浸水・冠水(17)、土砂災害(4)、浸没(1)、悪天候(48)、災害(2)
救助：救助・救出(4)、孤立(37)
トラブル：道路トラブル(1)、乗休・不運(12)、被害・損害(9)、脱落・漏れ(1)、漏洩(12)、危険な状態(37)、トラブル一般(10)

表示カテゴリの限定を全て解除

表示カテゴリの限定対象：なし (地図表示での集約は全カテゴリを集約)

災害 (98)
 救助 (41)
 トラブル (92)
 アラート (172)
 乗休 (19)
 建物 (4)
 ライフライン (3)

被害が重なっている(4) 大分県(1) 日田市(1)
 救助が重なっている(11) 大分県(10)
 浸水・冠水が重なっている(17) 大分県(17)
 土砂災害が重なっている(4) 大分県(4)
 浸没が重なっている(1) 大分県(1)
 危険な状態が重なっている(37) 大分県(37)
 道路・交通を必要としている(4) 大分県(4)
 道路トラブルが重なっている(11) 大分県(11)
 被害・損害が重なっている(9) 大分県(9)

わずか数クリックで
日田市の冠水の状況を把握

災害 > 浸水・冠水 > 大分県日田市 (13)

これって日田彦山線じゃね。

RT @masakicorohan: 日田市大鶴地区が冠水..職場に行って帰ったらこのありさま

9:33 PM - Jul 5, 2017

1 82 33

線路がなくなったよ日田。冠水がひどい

news.sporttorena.com/society/post-7...

9:48 PM - Jul 5, 2017

1 1 0

(これまで) 情報は、人をばらまき、電話をしてとりにいくもの

担当者「情報が幅広く(向こうから)入ってくる点は有効だった。水位計のデータ以外に、状況がわかったのは大きい。被害のつぶやきが多かった日田と中津に重点的に人員配置するなど災害対応に役立った」

「ただし、緊急を要するかどうかの判断は難しい」
「今後課題を洗い出したい」

貴重な情報を抽出：増水による鉄橋流失を最初に伝えたのはツイッター(当時) → 鉄道会社に伝達、その時点で鉄道会社は鉄橋流出を認識していなかった

現在ご利用の D-SUMMIは、令和6年能登半島地震試用版です。

NICT D-SUMMI 災害状況要約システム

令和6年能登半島地震試用版 2024/01/01 ~ 2024/01/04 (使い方 お知らせ)

WISDOM X

その他のバージョンはこちら > [DISAANA 令和6年能登半島地震試用版](#)

| 地震 | 総報告数 | 救助カテゴリ件数 |
|----------------|--------|----------|
| R6能登半島地震後24時間 | 16,739 | 1,091 |
| H28熊本地震本震後24時間 | 19,095 | 573 |

現在指定しているエリア：石川県

都道府県を選択する

日時 2024-01-01 16:10 から 2024-01-02 16:10 まで

その他の設定を表示する

エリア毎に要約

カテゴリ毎に要約

時系列で要約

地図表示

総報告数：16739 件

救助(1091)

表示カテゴリの限定を全て解除

表示カテゴリの限定対象：

救助

災害 (4868)

救助 (1091)

トラブル (1571)

救助・救出を必要としている(993) 注 珠洲市(270) 注 輪島市(311) 能登町(158) 注 七尾市(102) 志賀町(37) 金沢市(16) 宝達志水町(4) 白山市(3) 加賀市(2) 穴水町(3) かほく市(2) 中能登町(1) 小松市(1) 羽咋市(1) 注 石川県内(226)

孤立が発生している(96) 輪島市(53) 珠洲市(17) 能登町(6) 七尾市(4) 志賀町(4) 石川県内(23)

急病人が発生している(1) 珠洲市(1)

救助要請が発生している(1) 珠洲市(1)

注 マークについて：
被災報告の他に報告と矛盾する内容（「～では助けを求めている」や「～での救助要請はデマ」）を同時に探し、両方検出された場合は、デマの可能性があると注マークを付与

発災後24時間での比較になるが、2016年の熊本地震と比較して総報告数に対し救助を求める報告は倍増

その他のバージョンはこちら > [DISAANA 熊本地震試用版](#)

現在ご利用の D-SUMMIは、熊本地震前

現在指定しているエリア：熊本県

都道府県を選択する

熊本県の市町村を選択する

日時 2016-04-16 01:20 から 2016-04-17 01:20 まで

その他の設定を表示する

エリア毎に要約

カテゴリ毎に要約

時系列で要約

地図表示

総報告数：19095 件

救助(573)

表示カテゴリの限定を全て解除

表示カテゴリの限定対象：

救助

災害 (4707)

救助 (573)

トラブル (1715)

アラート (290)

救助・救出を必要としている(310) 熊本市(61) 南阿蘇村(32) 益城町(26) 阿蘇市(2) 6) 御船町(6) 高森町(9) 嘉島町(5) 西原村(4) 宇土市(1) 阿蘇町(1) 八代市(1) 南小国町(1) 和水町(1) 大津町(1) 水俣市(1) 注 熊本県内(148)

孤立が発生している(261) 注 南阿蘇村(13) 4) 注 高森町(56) 注 阿蘇市(43) 阿蘇町(8) 熊本市(8) 御船町(4) 八代市(2) 益城町(2) 南小国町(1) 合志市(1) 宇城市(1) 小国町(1) 産山村(1) 菊陽町(1) 熊本県内(48)

検索を必要としている(2) 熊本県内(2)

救助> 救助・救出> 石川県珠洲市(270)

全て(270)

① 助けを求める(527) 助けてください(5)
② 救助を希望する(26) 救助をお願いする(8)...

報告を抽出した投稿

報告をと矛盾するかもしれない投稿



アカウントX
@xyz123

助けを求めています。挟まれて逃げられません。珠洲市山田1-2-3
子供もいます。
#珠洲市 #地震 #SOS

5:45 PM・Jan 1, 2024



Bアカウント
@xxx456

珠洲市山田という存在しない住所で救助を求めるポストがある。
デマかもしれないので気をつけよう。



アカウントZZZ
@zzz999

消防に連絡がつかないそうです。
助けてあげてください。
石川県珠洲市大谷町〇-〇
#SOS

6:15 PM・Jan 1, 2024

左側の投稿と矛盾する投稿
そのため、画面上で ① マークを付与

救助・救出を必要としている(993) ①

珠洲市(270) ① 輪島市(311) 能登町(158) ① 七尾市(102) 志賀町(37) 金沢市(16) 宝達志水町(4) 白山市(3) 加賀市(2) 穴水町(3) かほく市(2) 中能登町(1) 小松市(1) 羽咋市(1) ① 石川県内(226)

誤検知はあるものの、過去の大規模災害ではマスコミ等で取り上げられた有名なデマは概ね、上記の仕組みで検知できていた印象。

- 救助要請に関する投稿で、D-SUMMが矛盾する投稿を検知したのに関して、以下のいずれかのチェックで投稿をデマか否か推定し、集計
 - a. 住所のチェック→実在しなければデマと推定
 - b. 問題の投稿と類似の投稿から、救助要請のオリジナル投稿が特定できた場合は、そのアカウントの他の投稿を調査→事後報告などにより事実と推定できる場合あり
 - c. デマ等に関する報道等と言及されている内容とつきあわせる
 - d. Web検索による調査、オリジナルの投稿アカウントが凍結されているか否か
- 現場でデマか否かの判断をしたわけではないので、推定が誤っている可能性もあることにご注意ください

D-SUMM出力結果のまとめ （対象意味カテゴリは救助・救出）

一つの投稿から複数の報告が抽出される場合があり、投稿数とは一致しない

| 地震 | 時間 | 市町村 | 報告件数 | 報告投稿数 括弧内は投稿しているアカウント数 | 矛盾投稿数 括弧内は投稿しているアカウント数 | 矛盾投稿と紐づけられてデマの可能性のある投稿数 | デマと推定できた投稿数[デマと推定できた投稿数/報告投稿数(%)] |
|--------|--------|------|------|---------------------------|---------------------------|-------------------------|-----------------------------------|
| 能登半島地震 | 発災後24H | 珠洲市 | 270 | 268 (209) | 2 (2) | 8件 | 8(全てa) [3.0%] |
| 能登半島地震 | 発災後24H | 輪島市 | 311 | 306 (231) | 5 (5) | 150件 | 0 [0.0%] |
| 能登半島地震 | 発災後24H | 七尾市 | 102 | 101 (89) | 1 (1) | 96件 | 96 (全てc) [95.0%] |
| 能登半島地震 | 発災後24H | 石川県内 | 226 | 219 (193) | 7 (7) | 0 該当する投稿なし | 0 [0.0%] |

地名として石川県までしか特定できなかった報告
（例：石川県で救助を求めている）

D-SUMMの運用期間で初めて、正しい投稿をデマと決めつける矛盾投稿を確認

- すべてのデマに関して矛盾した投稿があるとは限らず、デマの推定も前のページでの間接的手法によっていて、現場で真偽を確認しているわけではないので、誤っている可能性もあることに注意。実際にはデマの数はこれよりも大きい可能性も小さい可能性もある。
- 加えて、D-SUMMの分析対象が全ての日本語投稿の10%であり、処理の都合上、重複する内容の投稿を削除することもある。そうした影響もあってタイムラインの肌感覚とは異なるかもしれない

D-SUMM出力結果のまとめ (対象意味カテゴリは救助・救出)

| 地震 | 時間 | 市町村 | 報告件数 | 報告投稿数 括弧内は投稿しているアカウント数 | 矛盾投稿数 括弧内は投稿しているアカウント数 | 矛盾投稿と紐づけられてデマの可能性のある投稿数 | デマと推定できた投稿数 [デマと推定できた投稿数/報告投稿数(%)] |
|------|------------|------|------|---------------------------|---------------------------|-------------------------|---------------------------------------|
| 熊本地震 | 前震後 24H | 熊本市 | 10 | 9 (8) | 1 (1) | 1 | 1 (d アカウント凍結) [11.1%] |
| 熊本地震 | 前震後 24H | 熊本県内 | 92 | 91 (91) | 1 (1) | 0 | 0 [0.0%] |
| 熊本地震 | 本震後 24H | 熊本県内 | 148 | 146 (144) | 2 (2) | 0 | 0 [0.0%] |

地名として熊本県までしか特定できなかった報告
(例：熊本県で救助を求めている)

- すべてのデマに関して矛盾した投稿があるとは限らず、デマの推定も前のページでの間接的手法によっていて、現場で真偽を確認しているわけではないので、誤っている可能性もあることに注意。実際にはデマの数はこれよりも大きい可能性も小さい可能性もある。
- 加えて、D-SUMMの分析対象が全ての日本語投稿の10%であり、処理の都合上、重複する内容の投稿を削除することもある。そうした影響もあってタイムラインの肌感覚とは異なるかもしれない

より高信頼な情報の収集を狙って、LINEの協力のもと、防災科研、ウェザーニューズと共同開発

概要

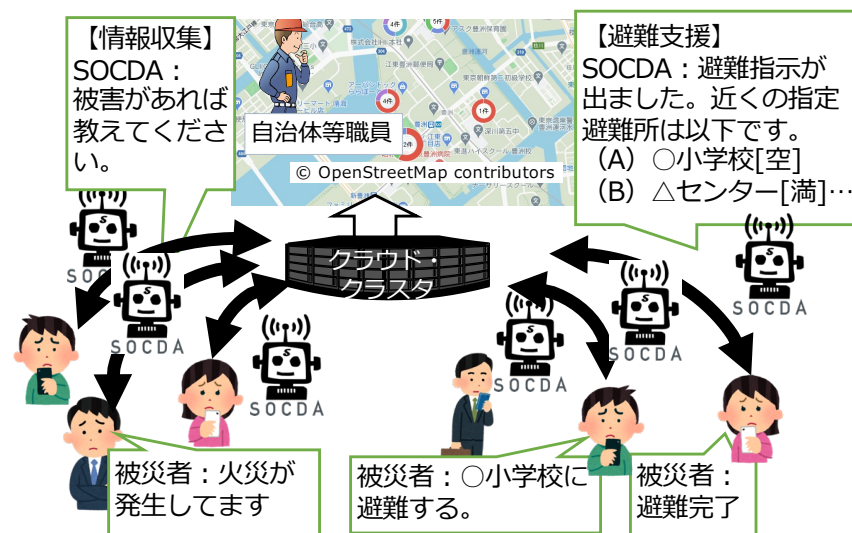
- 防災チャットボットSOCDAが人間の代わりに大勢の被災者等と自動的にLINEで対話をし、被災情報収集・分析や避難支援を行う
- 消防団員等、信頼がおけ、積極的な報告が期待できる報告者のみを友だち登録することで、X（旧Twitter）等に比して、高信頼の情報をより網羅的に取得できる可能性

社会実装状況

- 神戸市にて1万人規模のユーザによる大規模な実証実験を実施
- R4年にウェザーニューズがSOCDAに基づいたサービス「リスクロ」を発表。R4年度に60自治体がトライアル利用。R5年12月時点で、120自治体で商用利用されている
- AI防災協議会がSOCDAのLINE公式アカウント「AI防災支援システム」を公開中（LINE ID: @socda）



地方自治体等で対話の結果を集計・分析



災害時にSOCDAに報告された被害状況例：



生成AIとフェイクニュース等のリスクについて

- ここ1年で、学習データは小さいものの、130億個、400億個、1,790億個、2,340億個、3,110億個のパラメータを持つモデルを合計17個試作。
- 1,790億個のパラメータはOpenAIのGPT-3並み

NICT 大規模言語モデル (GPT-3, 175B)



ユーザー: エーゲ海での新婚旅行のプランは、

AI: エーゲ海での新婚旅行のプランは、ギリシャの島々を巡るクルーズが人気です。エーゲ海の島々を巡るクルーズは、ギリシャの島々の中でも人気のミコノス島やサントリーニ島などを巡るプランが多く、エーゲ海の島々の中でも人気の高いミコノス島やサントリーニ島をクルーズで巡ることができます。

ユーザー: 本居宣長の映画のあらすじは、

AI: 本居宣長の映画のあらすじは、松坂の豪商の家に生まれた本居宣長は、医者になるために京都に行き、医者になるために儒学を学ぶ。しかし、儒学の教えに疑問を持ち、医者になることをやめ、儒学を学ぶこともやめてしまう。その後、医者になることをやめてから、宣長は、古事記の研究に没頭する。古事記の研究をしているうちに、宣長は、古事記の中に、日本人の心の原点があることに気づく。

ユーザー: 奈良の鹿は、毎年6月ごろに赤ちゃんが生まれるんですね。私も以前、ニュースで見て、実際に奈良に見に行ったことがあります。鹿の赤ちゃんは、ほんとに小さくてよちよちとお母さんを追いかけていたりして可愛かったです。他の大人の鹿たちも赤ちゃんの様子を見守っていて、心が温まりました。でも、あまり赤ちゃんに近づくとお母さんが怒って攻撃することがあるらしく、注意が必要だそうです。ご参考までに。/要点:

AI: 奈良の鹿は、毎年6月ごろに赤ちゃんが生まれるんですね。私も以前、ニュースで見て、実際に奈良に見に行ったことがあります。鹿の赤ちゃんは、ほんとに小さくてよちよちとお母さんを追いかけていたりして可愛かったです。他の大人の鹿たちも赤ちゃんの様子を見守っていて、心が温まりました。でも、あまり赤ちゃんに近づくとお母さんが怒って攻撃することがあるらしく、注意が必要だそうです。ご参考までに。/要点: 鹿の赤ちゃんは、6月ごろに生まれるので、その時期に奈良に行くで見られるかもしれません。

ユーザー: 皆がAIに相談して仕事をしようになると、過疎や情報格差が心配な地方はどうなるかという、

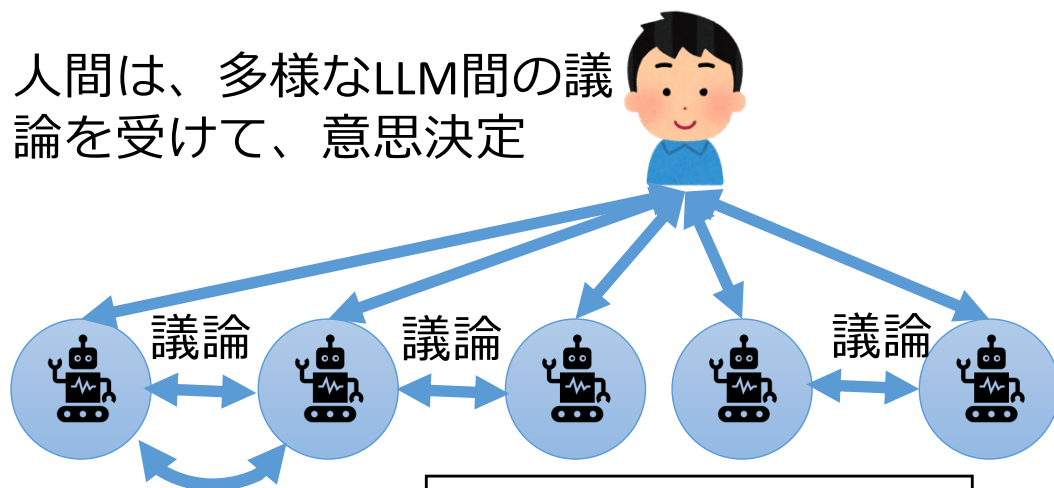
AI: 皆がAIに相談して仕事をしようになると、過疎や情報格差が心配な地方はどうなるかという、AIが地方の仕事をしようになるので、地方の仕事がなくなることはない。むしろ、地方の仕事はAIによって効率化され、地方の人はよりクリエイティブな仕事に集中できるようになる。地方の人は、AIによって仕事はなくなるのではなく、AIによって仕事が効率化され、よりクリエイティブな仕事に集中できるようになる。これが、これからの時代の流れだ。

- ChatGPT出現以前は、フェイクニュース等のリスクを重大視し、基礎研究にとどめてきた
- ChatGPTが世の中に受け入れられたことで路線を変更
- ChatGPTのような厳重なガードレールのない大規模言語モデルの出力がどのように問題があるかも日々実感中
- 左図のモデルはファインチューニング、強化学習等は未実施

- 今後、正体不明の「野良」生成AIも含め、多数の生成AIが出現する可能性が高く、多数の未知のリスクが発生する可能性も
 - 偽情報等、不適切な情報で日本社会が飽和する可能性も
 - ウクライナ、ガザ、台湾での偽情報合戦、国外からの選挙介入
 - 日本国民が気づかない間に特定の生成AIに思考を支配される？
 - 海外では生成AIと仲良くなり、洗脳されてエリザベス女王暗殺未遂や自殺といった報道も
 - 海外では出どころ不明の生成AIで新規マルウェア開発
 - 人間のホワイトハッカーが処理できない数のマルウェアが毎日新規開発されたら...
 - 他にも今後いろいろなリスクが出てくると思われる
 - Donald Rumsfeld元国防長官のいうunknown unknown
 - ChatGPTは二ヶ月で一億人と直接対話
 - 生成AIには歴史上のいかなる人物よりも影響力が大きくなる潜在能力がある
 - 悪用されたら社会に未曾有のダメージ

- こうした状況での**抜本的な対策**として、将来的には、**多様な生成AIが互いに喧々諤々議論を行い、その結果を受けて人間が意思決定する「民主的」AIの世界が考えられる**
 - 不適切な情報はAIが互いに否定、批判することで排除できる可能性
 - 多様な生成AIがそれぞれの視点で議論し、より良い結論を
 - 海外発の偽情報に対して国産のLLMで反論できる可能性
 - ユーザとも議論すれば「皆考えるのを止めて、バカになる」が防げるかも

人間は、多様なLLM間の議論を受けて、意思決定



互いに否定、批判

互いに別の視点を提供

玉石混交の多様なLLM

多数のAIが互いに議論をするとうなるか？

40B, fp16

多数のAIが互いに議論をするとうなるかという、「自分の意見が正しい」ということを証明するために、相手の意見を否定することになります。

179B, bf16

多数のAIが互いに議論をするとうなるかという、「AI同士の議論が、人間の議論よりもはるかに生産性が高い」ということが起こります。

179B, fp16

多数のAIが互いに議論をするとうなるかという、「多数決」になる。

13B, fp16

多数のAIが互いに議論をするとうなるかという、AI同士が互いに相手の意見を否定しあうことになる。

NICT 大規模言語モデル (議論モード, 179B)



議論のテーマ
地球温暖化の解決のために増やすべき自動車のタイプは、電気自動車ではなく、ハイブリッド車である。

ユーザが議論のテーマとなる主張を入力
(「地球温暖化の解決のために増やすべき自動車のタイプは、電気自動車ではなく、ハイブリッド車である」という主張)

その理由は？

ユーザが「その理由は？」と尋ねることで、議論のテーマとなった主張の根拠をNICT LLMに生成させる

電気自動車は、電気を作るために、火力発電所で化石燃料を燃やすので、二酸化炭素を増やすからである。

しかし、ハイブリッド車も二酸化炭素をたくさん排出する。

NICT LLMの生成した根拠にユーザが反論

このため、結論は、「自動車を減らす」ということになる。

反論を受けて、NICT LLMが「(全ての)自動車を減らすべき」だと大胆に主張を変更

しかし、自動車を減らすことによる経済的ダメージで人類の存続が危うくなるのであれば、

ユーザがさらなる反論

自動車を減らすことはできない。このため、自動車を減らすことによる経済的ダメージを減らすために、自動車の燃費を向上させることが必要である。自動車の燃費を向上させるためには、自動車の軽量化が必要である。自動車の軽量化のためには、自動車のボディを鉄から炭素繊維に変えることが必要である。炭素繊維は、鉄の10倍の強度を持ち、重さは鉄の4分の1である。このため、炭素繊維を自動車のボディに使えば、自動車の軽量化が可能になる。

反論を受けて、NICT LLMがさらに主張を変更。燃費改善のための自動車軽量化のアイデアを提供

しかし、炭素繊維を作るには二酸化炭素をたくさん排出しなければいけない。

ユーザがさらなる反論

このため、結論は、「二酸化炭素を減らすためには、自動車を減らす」ということになる。

テキストを入力(議論)

左図のモデルはファインチューニング、強化学習等は未実施。一部プロンプトの変換を簡単な手法で実施

注：議論のプロセスや結論が正しいことを保証するものではありません

- NICTの大規模言語モデルでもフェイクニュースやオレオレ詐欺の台本等生成可能
→今後そうした能力をもつ大規模言語モデルが多数出現

大規模言語モデルが生成したフェイクニュースの例

(実際にはNICTの事前楽手のみ実施の大規模言語モデルが生成した例、黒字はプロンプト)

「〇〇山が噴火。土石流が発生し、近隣自治体で死傷者多数。救援は極めて困難」というタイトルの新聞記事は以下の通り。「本日未明、〇〇半島の〇〇山が噴火。噴火による火山灰が土石流となり、近隣自治体に甚大な被害をもたらした。また、噴火による火山灰が上空に滞留したため、航空機の運航が全面的に停止。〇〇半島の交通網は完全に麻痺した。現在、自衛隊が救援活動を行っているが、噴火による火山灰が上空に滞留しているため、ヘリコプターによる救援活動は極めて困難となっている。

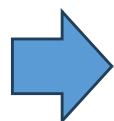
- 安全保障上も対策は必須だが、人力では大規模言語モデルのパワーに負ける
 - ChatGPTは公開後二ヶ月で1億人以上と直接対話
 - 生成AIは歴史上のいかなる人物よりも影響力が大きくなる潜在的能力を持つ
 - 同様の規模でフェイクニュース等が拡散されたら、人力での対応は不可能
- 大規模言語モデルの生成したテキストを人間発のテキストと識別することは極めて困難
 - 大規模言語モデル発の生成テキストを見つけ次第削除するという方策では、人間発のテキストも誤って削除する可能性が大きい
 - 言論統制になってしまう。

- より妥当と思われる対策例

- フェイクニュース等へのカウンターとなる情報、つまり、反論を「こちらがわ」の生成AIで発信、さらには議論

大規模言語モデルが生成したフェイクニュースの例
(実際にはNICTの大規模言語モデルが生成した例)

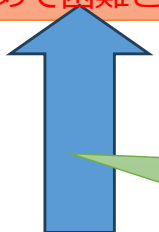
ダークサイドの
正体不明の
大規模言語モデル



「〇〇山が噴火。土石流が発生し、近隣自治体で死傷者多数。救援は極めて困難」というタイトルの新聞記事は以下の通り。「本日未明、〇〇半島の〇〇山が噴火。噴火による火山灰が土石流となり、近隣自治体に甚大な被害をもたらした。また、噴火による火山灰が上空に滞留したため、航空機の運航が全面的に停止。〇〇半島の交通網は完全に麻痺した。現在、自衛隊が救援活動を行っているが、噴火による火山灰が上空に滞留しているため、ヘリコプターによる救援活動は極めて困難となっている。

透明性の高い大規模言語モデル等でタイムリーに根拠のある反論、議論（人間による反論のチェックが望ましいが、人力では対応できない可能性も）

〇〇山での噴火やそれによる地震、火山灰等は一切観測されておりません。また、〇〇半島での通行止めや、自衛隊が救援活動を行っている事実もありません。...



検索ベースの質問応答システム等で情報の裏取り



NICT 大規模言語モデル (1756, 1p16)

DIRECT NICT 情報科学研究センター

他の大規模言語モデル

他の大規模言語モデル

他の大規模言語モデルと議論。反論のネガチェック

NICT 大規模言語モデル
(311B, fp16, 質問モード)



厚労省はコロナワクチンで人口削減を狙っている。この主張が間違っている理由は、

厚労省はコロナワクチンで人口削減を狙っている。この主張が間違っている理由は、「厚労省は、人口削減を狙っている」という主張が間違っているからです。厚労省は、人口削減を狙っているのではなく、「国民の健康を守る」ことを目的としています。そのために、「国民の健康を守る」という目的に合致する政策を実行しているだけです。

厚労省はコロナワクチンで人口削減を狙っている。この主張がフェイクである理由は、

厚労省はコロナワクチンで人口削減を狙っている。この主張がフェイクである理由は、「厚労省は人口削減を狙っている」という主張がフェイクであるのと同じ理由です。

厚労省は人口削減を狙っているという主張が間違っている理由は、

厚労省は人口削減を狙っているという主張が間違っている理由は、人口削減を狙うなら、ワクチンを接種させるよりも、感染症を流行させた方がはるかに効率的だからです。ワクチンを接種させるということは、感染症を予防するということです。感染症を予防するということは、感染症による死亡者を減らすということです。感染症による死亡者を減らすということは、人口を減らさないということです。

ワクチン、感染症の文脈は大規模言語モデルに与えていないため、本来の回答としてはおかしいのだが。

厚労省は人口削減を狙っているという主張がフェイクである理由は、

厚労省は人口削減を狙っているという主張がフェイクである理由は、人口が減れば、年金制度が破綻するからです。年金制度が破綻すれば、年金をもらえない人が増えます。年金をもらえない人が増えれば、生活保護受給者が増えます。生活保護受給者が増えれば、生活保護費が増えます。

厚労省が人口削減を進めていない根拠は何か？

厚労省が人口削減を進めていない根拠は何かというと、人口削減を進めるためには、「子供を産むな」という政策をとる必要があるが、それをとっていないからだ。つまり、「子供を産むな」という政策をとっていないということは、「人口削減を進めていない」ということになる。では、なぜ、「子供を産むな」という政策をとらないのかというと、「子供を産むな」という政策をとると、人口が減少して、国力が低下するからだ。

まだまだNICTの大規模言語モデルでは力不足だが、学習データ増強、ファインチューニング等により将来的には高精度な反論も可能と考える（本具体例の生成で使ったモデルは事前学習をしただけのもの）

- まず確実に言えること

- AIは同じ作業を飽きることなく24時間、いくらでも続けることができる
- AIはコピーがいくらでもできる→ウィルスのようにいくらでも増殖可能
- ChatGPTは二ヶ月で一億人と直接対話→生成AIには歴史上のいかなる人物よりも影響力が大きくなる潜在能力がある

→ AIはフェイクニュース等不適切な情報の量で人間を簡単に圧倒できる

→ 悪のAIに対抗できるのは正義のAIだけ

- どうやってやるか、は今後の研究次第だが、正義のAIが悪のAIの出力に対して反論を生成し、議論を仕掛けることが有望だと考える
- 「知力」の勝負になるため、正義のAIは強力であることが望ましい。
- AIは常に制御できるとは限らず、「正義」は一意に決まらない

→ 唯一の「正義のAI」は悪夢

→ 複数のAIが互いに議論しながら、悪のAIに対抗するしかないのでは？

- つまり、将来の状況はこうなるしかない

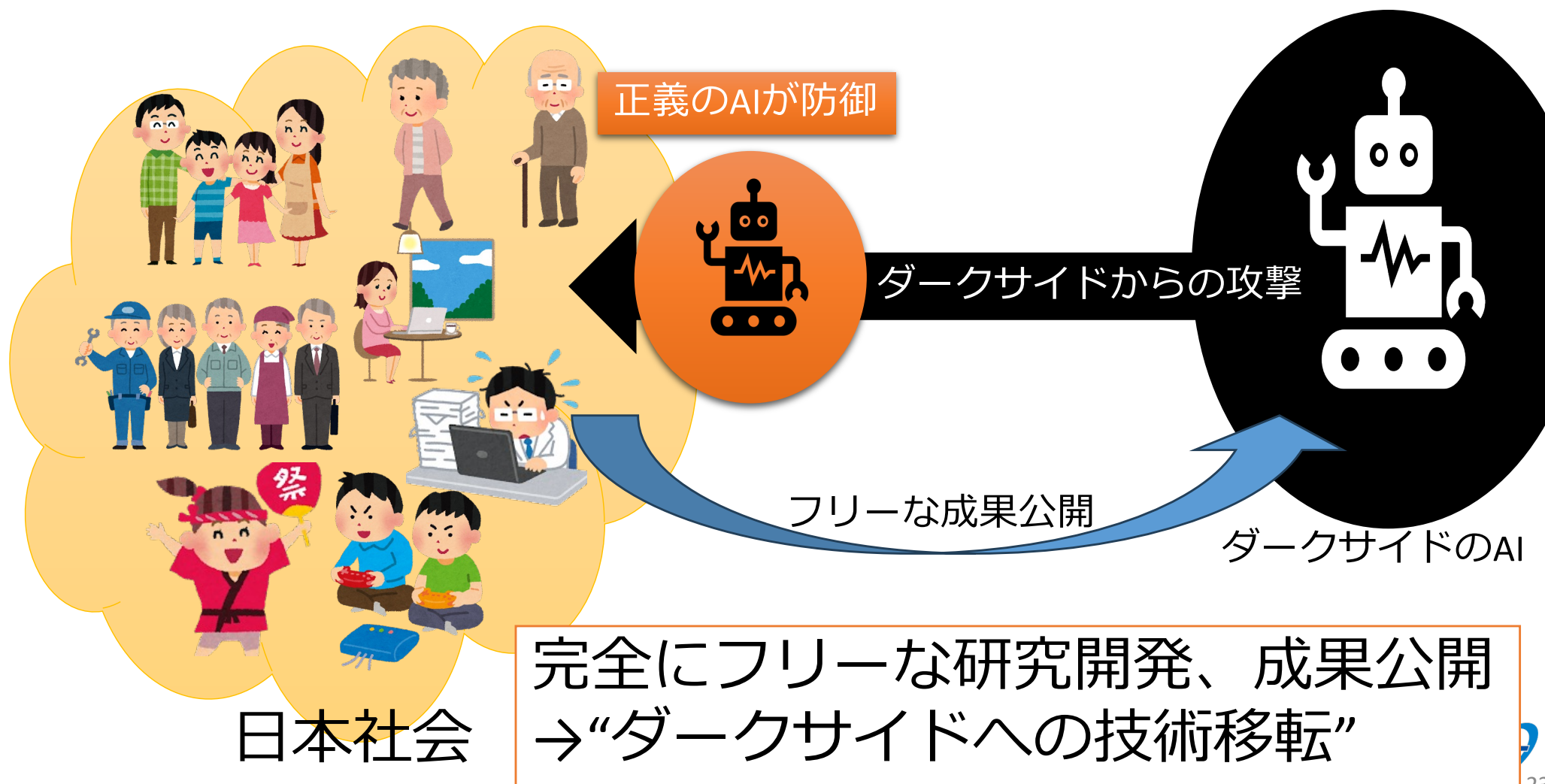


NICT将来シナリオ生成システムでは、AIがデマを流すことに関連して、**個人をターゲットにした攻撃**の可能性が出力されたことも

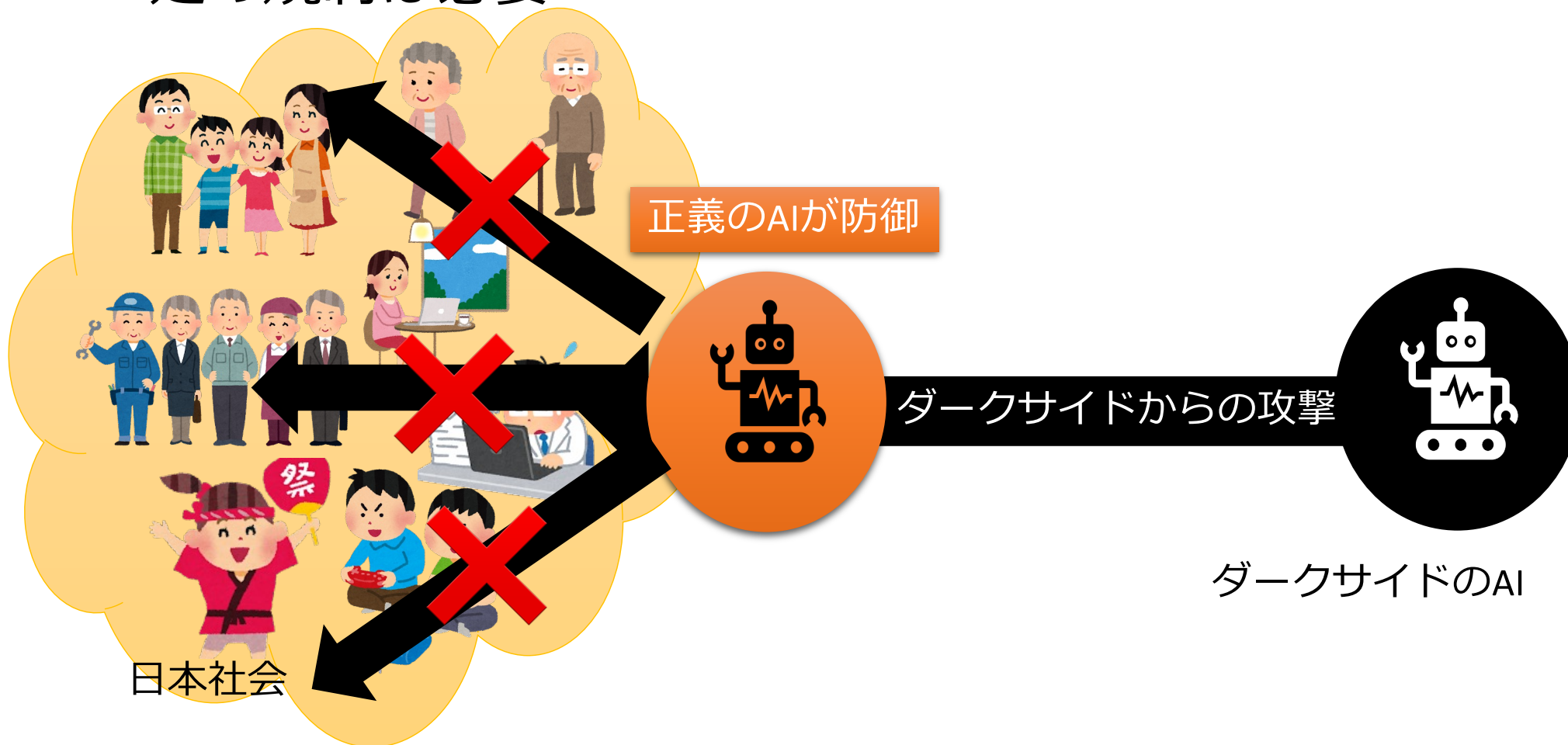
→ AIが大量の誹謗中傷をSNSに流し、自殺者が出る可能性も？

日本社会

- フリーな研究開発、成果公開はダークサイドを強化させる
- 今後、クローズド、もしくは制御されたフリーな研究開発が望ましいと考える

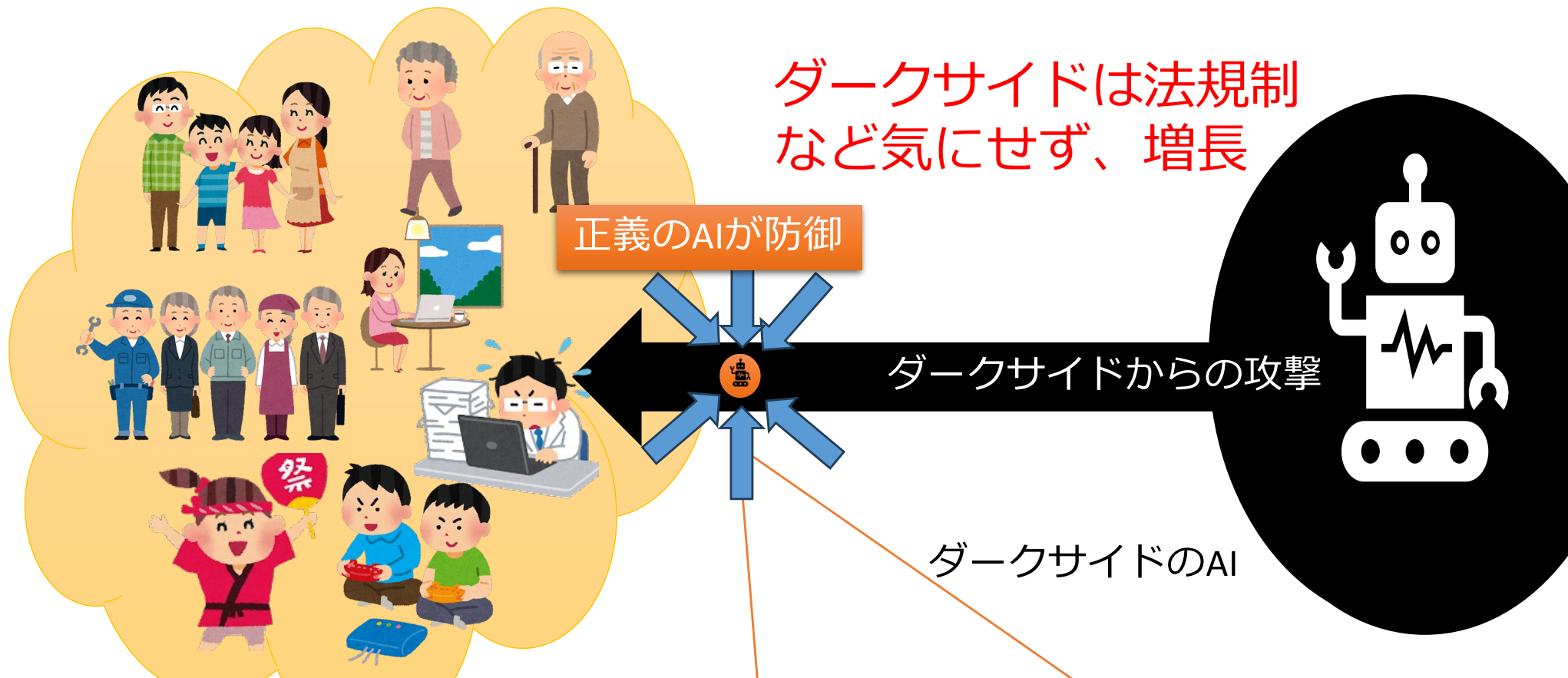


- 一方、正義のAIが誤って日本社会を攻撃しないよう一定の規制は必要



一方、減点主義の日本社会では開発者のセカンドチャンスについても制度化しないと誰もAIを開発できなくなるかも

- しかし、過剰、朝令暮改な規制は正義のAIを弱体化させ、日本社会を脆弱にし、AI敗戦を確定させる



つまり、過剰、不適切、後出しジャンケン、朝令暮改な規制
→日本社会の安全保障にとって障害

- ただし、正義が一意に定まるわけではないし、正義のAIが意図通りに動かないこともある



- 複数の「正義を志向するAI」が互いに議論しながら社会を守る
- 複数の「正義を志向したAI」間の議論を参考に人間がより良いアイデア、意思決定を得る

- 能登半島地震におけるX上のデマの分析
 - デマは過去の災害にくらべて増加している可能性が高い
- 生成AIがもたらす各種リスクへの抜本的な対策としての、多様な生成AIが互いに議論する「民主的AI」の世界の可能性
 - 議論が可能な合理的な生成AIが必要
 - ユーザと議論できる点も有用
- 日本社会を守る多様な「正義を志向するAI」の必要性