ИНСТРУМЕНТЫ ИИ В РУКАХ ЗЛОУМЫШЛЕННИКОВ — КЛАССИФИКАЦИЯ УГРОЗ И СПОСОБЫ ПРОТИВОДЕЙСТВИЯ

РОСКОМНАДЗОР

Главный радиочастотный центр

СОДЕРЖАНИЕ

Из всего многообразия факторов, которые постоянно меняют очертания нашего будущего, инновационные технологии заслуживают отдельного внимания, за счёт высокой скорости внедрения в сферы жизни человека, от производства до коммуникации. Каждый виток развития цифровых технологий всегда сопряжен с новыми вызовами. И ключевую роль в их успешном преодолении играет время реакции — уже сегодня мы обязаны искать возможности решения рисков завтрашнего дня.

Нам жизненно необходим анализ новейших решений и методов работы, введение их в нашу практику для обеспечения стабильного развития нашей страны. Соблюдая интересы граждан, государства и бизнеса, защищая их права и их самих от возможных цифровых угроз, мы всегда будем совершенствовать механизмы нашей деятельности. И главным двигателем нашего совершенствования всегда будет информация.

НЕСТЕРЕНКО РУСЛАН ВАСИЛЬЕВИЧ

Врио генерального директора, ФГУП «ГРЧЦ»

Важно понимать, что использование алгоритмов искусственного интеллекта может служить как щитом, так и мечом. С одной стороны, мы видим ИИ как надёжного помощника внутри научно-технических процессов. Скорость работы и круглосуточная доступность алгоритмов значительно облегчает работу учёным и исследователям. С другой, мы не можем закрывать глаза на деструктивные сценарии использования ИИ злоумышленниками, от мошеннических схем до социального манипулирования через информационное воздействие.

Научный прогресс не остановить, и у ИИ есть внушительный потенциал для решения многих проблем, существующих сейчас в социальном, экономическом и технологическом секторах, однако мы всегда должны помнить о вопросах безопасности его использования и нашей защищённости.

РЫЖОВА ЕВГЕНИЯ ЮРЬЕВНА

Советник генерального директора по научно-техническому развитию, ФГУП «ГРЧЦ»

Возможности технологий искусственного интеллекта неуклонно растут, и уже сейчас существуют области и задачи, в которых искусственный интеллект заметно обогнал человека. Подходы, которые мы используем для принятия решений, потребления и производства контента, даже общения с другими людьми — никогда не будут прежними. Более того, они будут меняться всё сильнее по мере развития и внедрения искусственного интеллекта. Нашей стране придётся приложить значительные усилия, чтобы не отстать от стран-лидеров в области развития ИИ. При этом практика показывает, что отечественные разработчики способны создавать решения мирового уровня. Кооперация между государством, научно-исследовательскими учреждениями, гражданским обществом и бизнесом в целях исследования, освоения и мониторинга развития технологий искусственного интеллекта — важная задача. Если не заниматься её решением, мы рискуем однажды проснуться в мире, полном неразрешимых проблем, где все наши подходы стали архаичны и перестали работать. И адаптироваться к такому миру будет уже поздно.

ГЛАЗКОВ
БОРИС МИХАЙЛОВИЧ

Вице-президент
по стратегическим
инициативам,
ПАО «Ростелеком»

Искусственный интеллект — один из наиболее значимых вызовов для общества сегодня: он поменяет привычные бизнес-модели, породит и уничтожит целые индустрии, сократит старые и создаст принципиально новые рабочие места. Несмотря на колоссальный потенциал для общественного блага, искусственный интеллект будет использоваться злоумышленниками: для обмана, манипуляций и введения в заблуждение. Наконец, подрывной потенциал искусственного интеллекта ставит нас перед вопросами и выборами, с которыми мы никогда не сталкивались за всю историю человечества — технологическими, этическими, социальными. Именно поэтому крайне важно задумываться о завтрашних рисках уже сегодня.

ЮСУФОВ
РУСЛАН ГЕННАДЬЕВИЧ

Управляющий
партнер, MINDSMITH

ВВЕДЕНИЕ

Искусственный интеллект (или ИИ) как явление позволяет нам переосмыслить процессы анализа и использования информации, прогнозирования и принятия решений, проникнув практически во все сферы нашей деятельности. Хранение и обработка огромного массива данных с высокой скоростью, сложные алгоритмы обработки информации любого рода, от визуального изображения, до неструктурированного цифрового потока, принятие решений без вмешательства человека — это наша действительность. В последнее десятилетие скорость развития ИИ стала экспоненциальной: то, что в середине 2010-х казалось рывком в разработке нейросетей, на самом деле было началом продолжительного развития технологии, которое не сбавляет скорости.

Но важнейшим маркером внимания к ИИ является то, что другие технологии начинают менять свой вектор развития под него: происходит разработка специализированных чипов, постройка отдельных центров обработки данных, автоматизация методов сборки датасетов — производство инструментов по созданию ИИ стало своим собственным рынком. Обыватели уже используют ИИ, встроенный в поисковые системы или мобильные приложения, не задаваясь вопросами. Это говорит о высокой степени коммерциализации и о том, что вскоре технология будет доступна широчайшему спектру людей, а входной порог для, например, обучения нейросети или её модификации значительно снизится.

Подрывной потенциал технологии станет доступен в полной мере куда более широкому спектру людей, и несомненно среди этого спектра будут и злоумышленники, которые раньше не были способны грамотно применить ИИ в рамках своих целей и средств.
У государства нет другого выбора, кроме как изучать эту технологию, обучать своих граждан и готовиться к грядущим изменениям в информационных войнах, киберпреступлениях, практиках недобросовестного использования персональных данных и других векторах рисков, которые значительно изменятся и обострятся по мере развития технологии. Остановить или замедлить это развитие невозможно.

Неминуемое развитие технологий создает необходимость прозрачного понимания возможностей ИИ как для представителей бизнеса, так и для государства, ведь использование ИИ напрямую затрагивает вопросы общечеловеческих ценностей, прав и свобод граждан, национальной безопасности.

Сложности к пониманию вектора развития технологии для обычного обывателя добавляет частичная (иногда полная) секретность о разработках инструментов, с которыми человек взаимодействует уже сегодня, попутно подливая масла в огонь любителям конспирологических теорий. Однако как на самом деле обстоят дела с ИИ сегодня?

ЦЕЛЬ ЭТОГО ИССЛЕДОВАНИЯ СОСТОИТ В ТОМ, ЧТОБЫ ПРЕДСТАВИТЬ РЕАЛИСТИЧНЫЙ ОБЗОР И НАГЛЯДНО КЛАССИФИЦИРОВАТЬ ПЕРЕДОВЫЕ РЕШЕНИЯ С ИСПОЛЬЗОВАНИЕМ ИИ, ЗАДЕЙСТВОВАННЫЕ В МОНИТОРИНГЕ ИНТЕРНЕТА И ОБЕСПЕЧЕНИИ БЕЗОПАСНОСТИ ЕГО ПОЛЬЗОВАТЕЛЕЙ.

КЛАССИФИКАЦИЯ

МЫ СФОРМУЛИРОВАЛИ ДВА КЛЮЧЕВЫХ ПАРАМЕТРА ДЛЯ КЛАССИФИКАЦИИ ПРИКЛАДНЫХ РЕШЕНИЙ, ПРОАНАЛИЗИРОВАВ НАУЧНУЮ ЛИТЕРАТУРУ И ТЕМАТИЧЕСКИЕ ОТЧЕТЫ КОНСАЛТИНГОВЫХ КОМПАНИЙ.

Первым параметром стала принадлежность решения к одному из 4 сформированных направлений ИИ (названных нами в исследовании субтехнологиями). Каждая из субтехнологий представляет собой широкую предметную область, характеризующуюся форматом информации, с которым работает то или иное решение: компьютерное зрение, обработка естественного языка, распознавание и синтез речи, системы прогнозирования и поддержки принятия решений. Следует отметить, что кроме разницы в векторе применения, субтехнологии значительно отличаются и с технической стороны.

Вторым параметром стала степень зрелости, отражающая разные стадии развития решений. Данная градация была сформирована на основе отчётов консалтинговых компаний и обзоров научной литературы, задачей которых было измерение степени развития той или иной технологии. Было выделено три уровня зрелости решения: «концепция», «прототип» и «внедрено».

Решения на этапе «концепция» представляют собой экспериментальные системы, концепции систем и гипотетически возможные системы на ранних стадиях разработки, не прошедшие пилотных стадий или тестирование в реальных условиях или условиях, близких к реальным. Для доведения таких решений до готовности к внедрению потребуется значительный объём работ.

Решения на этапе «прототип» включают в себя экспериментальные системы, которые успешно прошли или проходят первые этапы тестирования в реальных условиях или условиях, близких к реальным, ими можно пользоваться и внедрять их в процессы после предварительной модификации.

Решения на этапе «внедрено» являются решениями, которые уже работают в реальных условиях, используются частными компаниями, НКО, государственными организациями, обывателями и иными заинтересованными группами. Такие решения уже занимают свою нишу на рынке и готовы к внедрению.

Следующим шагом мы распределили решения на 12 функциональных групп (названных нами в исследовании кластерами), соответствующих технологическим интересам профильных организаций по мониторингу и обеспечению безопасности интернета. Каждый кластер объединяет решения, предназначенные для выполнения задач в конкретном направлении. Так были выделены кластеры:

1 Обнаружение дипфейков
2 Определение контекста происходящего на видео
3 Автоматизация мониторинга и модерации контента
4 Распознавание лиц
5 Извлечение смысла из текста
6 Поддержка фактчекинга
7 Распознавание символики
8 Извлечение и анализ метаданных
9 Распознавание эмоций
10 Поддержка решений при информационных атаках
11 Генерация контента
12 Рекомендация контента

Кластеры были сформированы на основе предварительного анализа релевантной научной литературы, в ходе обсуждения с экспертами.

Следует отметить, что в ходе поиска решений был использован ряд аналитических инструментов, в частности система Teqviser. База данных системы содержит свыше 60 миллионов научных публикаций, не менее 30 миллионов патентов и около 600 тысяч примеров инвестиционных проектов.

Одним из результатов исследования стала обзорная карта технологий, которая будет представлена дальше. Карта демонстрирует нынешний ландшафт разработки решений на основе ИИ в разрезе кластеров, субтехнологий и степени зрелости этих решений.

ОБЩИЕ ВЫВОДЫ

НЕСМОТРЯ НА ЗНАЧИТЕЛЬНЫЕ РАЗЛИЧИЯ В ТЕХНИЧЕСКИХ ОСОБЕННОСТЯХ РАБОТЫ РЕШЕНИЙ В КАЖДОМ КЛАСТЕРЕ, СУЩЕСТВУЕТ РЯД УНИВЕРСАЛЬНЫХ ПОЛОЖЕНИЙ, КОТОРЫЕ БЫЛИ ПОДТВЕРЖДЕНЫ ИЛИ ОБНАРУЖЕНЫ В ХОДЕ ИССЛЕДОВАНИЯ:

1 В России есть экспертиза для создания отечественных датасетов и моделей, но не хватает вычислительных мощностей, инфраструктуры и кооперации между ключевыми стейкхолдерами. Ситуация осложняется тем, что в России на данный момент недостаточно развито производство высококачественного, мощного вычислительного оборудования. Несмотря на это, в России есть крайне качественные решения, особенно это касается распознавания лиц и работы с информационными атаками.

2 Китай и США лидируют с большим отрывом в существенной части кластеров. Несмотря на то, что эти страны имеют значительные различия как в направлении исследований и разработок, так и в тактическом подходе, в обеих странах развито сотрудничество государственных организаций с коммерческими.

3 Из-за автоматизации информационных войн и развития генеративных моделей будет крайне проблематично обеспечить когнитивную безопасность населения без внедрения искусственного интеллекта.

4 Наличие отечественных моделей и их внедрение, а также использование отечественных датасетов — вопрос национальной безопасности, так как зарубежные акторы способны экспортировать искусственный интеллект, который будет им подконтролен после передачи клиенту.

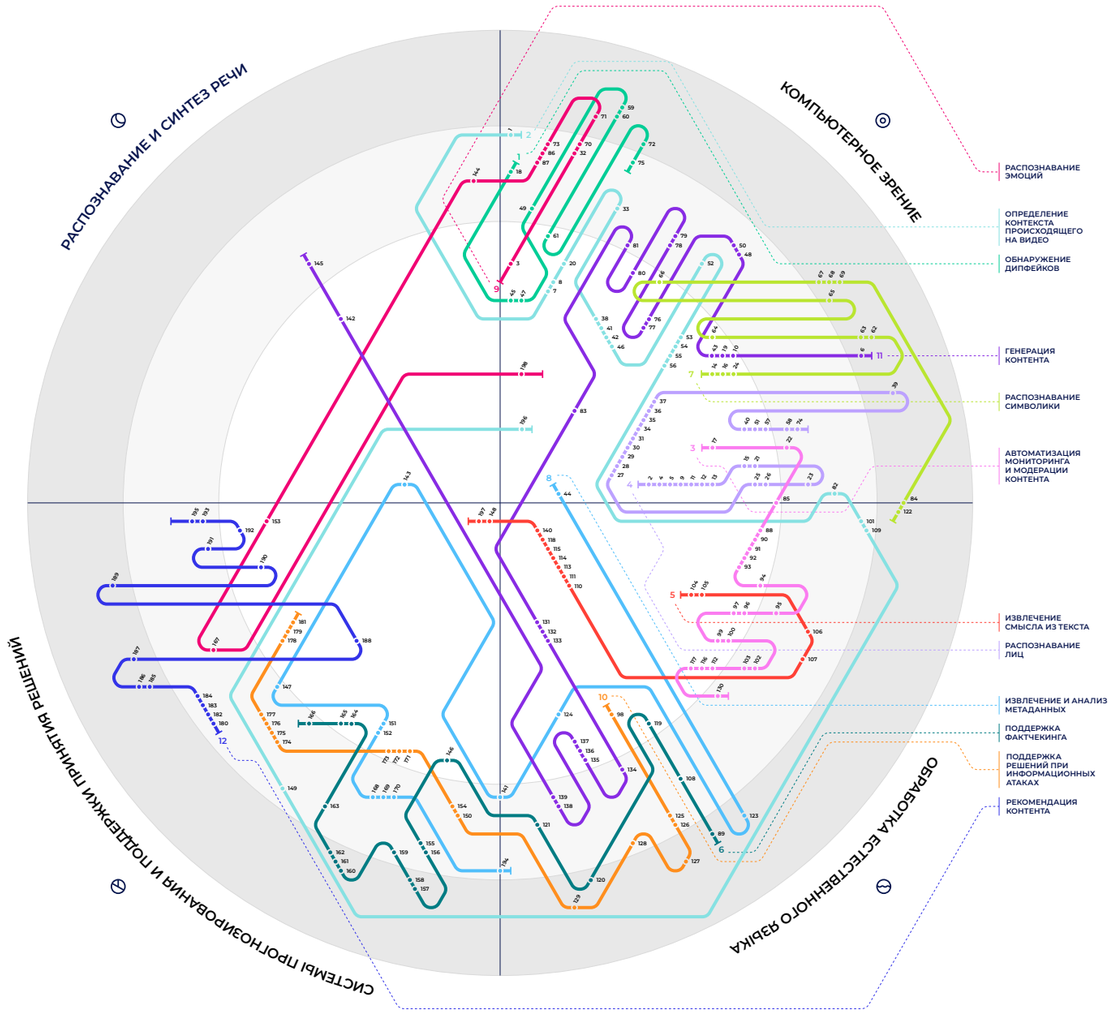
5 Разработка процедур тестирования и оценки моделей — важная инфраструктурная задача, которая позволит государству держать руку на пульсе развития технологий, отслеживать перспективных разработчиков и проекты.

6 Существенная доля моделей представляет собой «чёрный ящик» — из-за огромного объёма параметров крайне сложно определить, как именно эти модели принимают свои решения. Это представляет угрозу в случае внедрения зарубежных моделей, использования зарубежными датасетами и разработки некачественных моделей.

7 Регуляторика в большинстве стран не успевает за скоростью развития технологий. Это относится к большой доле кластеров, начиная с генеративных алгоритмов, регуляция которых всё ещё не достигла должной детализации, и заканчивая самими датасетами, подход к наполнению и использованию которых во многих странах остается на совести самих разработчиков.

8 Искусственный интеллект значительно обогнал человека в объёме и скорости обработки данных, способности распознавать малозаметные паттерны и сравнивать образцы с большими объёмами информации. На данный момент искусственный интеллект всё ещё далек от реального понимания контекста и культурных нюансов, чем нередко пользуются злоумышленники.

КАРТА ТЕХНОЛОГИЙ



Перед вами карта технологий, которая отображает ландшафт проанализированных нами инструментов на базе искусственного интеллекта.

Круг разделен на четыре доли и три окружности: доли показывают субтехнологии, а окружности — степень зрелости технологического решения. Каждое решение представляет собой точку на этой карте. Местоположение инструмента отображает, к какой субтехнологии он относится и на каком этапе разработки находится.

Точки объединены в линии — так показаны функциональные группы инструментов. Благодаря этому можно понять, насколько та или иная группа инструментов готова к внедрению и к какие субтехнологии в ней задействованы.

КЛАСТЕРЫ И КЕЙСЫ

1 ОБНАРУЖЕНИЕ ДИПФЕЙКОВ

18 Обнаружение дипфейков по биосигналам, Intel
45 Обнаружение дипфейков, Sensity
47 Deepware Scanner, Zemana
49 Распознавание дипфейков, Сбербанк
59 Анализ бликов в глазах, Университет Буффало
60 DISSIMILAR, Открытый университет Каталонии
61 Defudger
72 Обнаружение редактирования по сжатым кадрам
75 Выявление дипфейков по цвету лица, Сбербанк

2 ОПРЕДЕЛЕНИЕ КОНТЕКСТА ПРОИСХОДЯЩЕГО НА ВИДЕО

1 Обнаружение аномальных событий, Шэньчженьский политехнический институт
7 SenseFoundry
8 Патрульный робот МВД Сингапура
20 Whatismymovie, Valossa AI
33 Анализ поведения, Kepler Vision Technologies
38 «Gun Detection», Kogniz
41 Freemove, Veo Robotics
42 Автономные роботы, Knightscope
46 «Kipod», Synesis
52 Анализ контента и контекста видео, университет Лейкхед
53 Video Intelligence, Google
54 DETECTOR, Awaait Artificial Intelligence
55 Анализ и синтез видео, Beijing Moviebook Technology
56 Маркировка участков видео, CLIPr
82 Анализ видео с устройств IoT
101 Распознавание контекста, Менуфийский университет
109 Модерация конференций, Audiocodes
149 Предсказание поведения человека, Колумбийский университет
196 Luna CARS, VisionLabs

3 АВТОМАТИЗАЦИЯ МОНИТОРИНГА И МОДЕРАЦИИ КОНТЕНТА

17 PicPurify, SAS Ars Nova Systems
22 Модерация потокового видео, Kyndryl и IBM
85 Автоматическая модерация, VK
88 Модерация комментариев, Unitary
90 AutoMod, Discord
91 Azure Content Moderator, Microsoft
92 Модерация контента, Hive
93 Conflict Alerts, Meta*
94 Автоматическая модерация, Sentropy
95 Самомодерация ИИ, Университет Калифорнии
96 «Are You Sure?», Match Group
97 Автоматическая модерация, Content Score
99 Модерация контента, Checkstep
100 Модерация инфополя пользователя, Bodyguard
102 Модерация искаженного текста, Verizon Media
103 Модерация текстовых сообщений, МТС и Сколтех
112 Audio Intelligence, AssemblyAI
116 Community SIFT, Two Hat Security
117 LiveDune
130 KOLD

4 РАСПОЗНАВАНИЕ ЛИЦ

2 BioSurveillance NEXT, Herta Security
4 Axxon Next, AxxonSoft
5 Оплата лицом, VisionLabs
9 «Age Verification», Yoti
11 BioFinder, Herta Security
12 FaceRadar, Нейросети Ашманова
13 Визирь.СКУД, ЦРТ
15 Luna SDK, VisionLabs
21 Распознавание действий, Ntech Lab
23 Распознавание лиц по тепловым сигналам, МО США
25 Видеонаблюдение для школ, NTech Lab и ЭЛВИС-Неотек
26 IVA CV, IVA Cognitive
27 SmartCheck, iProov
28 AI Xunren, Baidu
29 FaceGo, Hanwang
30 O.Gate, O.Vision и Beeline
31 Face2Pay, Ак Барс Банк
34 FaceID, Apple
35 FindFace Multi, NTechLab
36 FaceNet, Google
37 FRVT, NIST
39 EES, Thales Group
40 Распознавание лиц в аэропортах, Thales Group
51 FERET, NIST
57 City Brain, Alibaba DAMO Academy
58 Защита от морфинговых атак, Институты им. Гумбольдта и Фраунгофера
74 Бенчмарк моделей ИИ для распознавания лиц, Havelsan

5 ИЗВЛЕЧЕНИЕ СМЫСЛА ИЗ ТЕКСТА

104 Томита-парсер, Яндекс
105 ABBYY Compreno, ABBYY и Сбер
106 Определение смысла многозначных слов, IBM
107 Определение смысла многозначных слов, Университеты Сан-Паулу и Далахаузи
110 InstructGPT, OpenAI
111 BERT, Google
113 Expert.ai
114 R-NET, Microsoft
115 CAILA, Just AI
118 mGPT-3, SberDevices
140 Conversation Intelligence, INVOCA
148 ИИ для поиска финансовых пирамид, Полиция Кореи
197 BEAT и EMILIE, Deloitte

6 ПОДДЕРЖКА ФАКТЧЕКИНГА

89 Обман стилометрическиз систем фактчекинга, MIT
108 Проверка и сопоставление статей, Канадский Университет Ватерлоо
119 Оценка логичности текста, IBM
120 Правила формирования датасетов для фактчекинга, MIT
121 Мультимодальный фактчекинг, Пекинский технологический Институт
146 Оценка дезинформации, Factmata
155 Оценка достоверности новостей, Sony Interactive
156 Проверка утверждений, Университет Ставангера
157 Мультимодальный фактчекинг, CCRI
158 Мультимодальный фактчекинг, Университет Беннета
159 Специализированный фактчекинг, Технологический институт Нью-Джерси
160 Проверка новостей, IIT-Patna
161 Проверка постов в Twitter, Университет Манубы
162 Проверка репостов в Twitter, Пекинский университет
163 ALIKAH, Технологический университет Абдула Калама
164 Инструментарий для фактчекинга, Full Fact
165 ClaimBuster, Университет Техаса в Арлингтоне
166 Проверка новостей о COVID-19, группа индийских учёных

7 РАСПОЗНАВАНИЕ СИМВОЛИКИ

14 «Brand Recognition», CVisionLab
16 VizPol, Колумбийский университет
24 Smart ID Engine, Smart Engines
62 Идентификация символики в видеоиграх, Уппсальский университет
63 Проверка оригинальности логотипов, Университет Северного Бангкока
64 Идентификация символов, Clarifai
65 Распознавание смысла татуировок, Китайская Академия Наук
66 Идентификация символов с распознаванием контекста, Visua
67 Конкурс по распознаванию татуировок, NIST
68 Распознавание рукописных электронных диаграмм, NLPR
69 Распознавание химических структур, Университет Нотр-Дам-Луэз
84 Распознавание культурных символов, Цзилиньский университет
122 Определение политической склонности изображений, Питтсбургский Университет

8 ИЗВЛЕЧЕНИЕ И АНАЛИЗ МЕТАДАННЫХ

44 Metadata Digger AI, DataHunters.AI
123 Определение тональности видео по метаданным, Колледж Раштрея Видьялая
124 Извлечение метаданных из нейросетей, IBM
141 Обнаружение ботов в Twitter, Мадридский политехнический институт
143 SIIP, Verint Systems, Interpol
147 MADIK, Университет Бразилиа
151 Spectrum Discover, IBM
152 AWS Content Analysis, Amazon
168 Криминалистическая классификация медиафайлов, Университетский колледж Дублина
169 Приоритизация электронных улик по метаданным, Университетский колледж Дублина
170 Поиск информации о торговле людьми, XDATA
194 Анализ неоднородных массивов данных, Университет Плимута

9 РАСПОЗНАВАНИЕ ЭМОЦИЙ

3 BioObserver, Herta
32 Анализ вовлечённости учеников, Intel и Class Technologies
70 Анализ эмоций по движениям глаз, Политехнический университет Варшавы
71 Извлечение микровыражений из видео, Технический университет Эйндховена
73 Анализ эмоций по движению пор на лице, Университет Стони-Брук
86 Считывание эмоций, Хэфэйский комплексный национальный научный центр
87 Оценка гибких навыков по выражению лица, Тайваньские университеты
144 Распознавание эмоций по речи, Китайские университеты
153 Умный ресторан, Baidu
167 Распознавание эмоций по энцефалограмме, Китайские университеты
198 Распознавание улыбок, Canon

10 ПОДДЕРЖКА РЕШЕНИЙ ПРИ ИНФОРМАЦИОННЫХ АТАКАХ

98 Мониторинг сообщений, Factmata
125 Определение предвзятости публикаций, Университет Джорджа Вашингтона
126 Анализ языковых сегментов интернета, Galisteo Consulting Group
127 Обнаружение ботов в Twitter, Университет Джорджа Мэйсона
128 Выявление пропаганды, Университет Хамада бин Халифа
129 Обнаружение пропаганды в новостных статьях, Индонезийская Академия Наук
150 RIO
154 Оценка ущерба от публичных скандалов, S&P
171 Платформа анализа медиасреды, PrimerAI
172 Logically Intelligence, Logically
173 Brand Analytics, ПалитрумЛаб
174 Обнаружение активности ботов, Galisteo Consulting Group
175 «Bot-Match», Университет Карнеги — Меллона
176 Идентификация информационных кампаний, Принстонский Университет
177 Анализ информационных операций, Университет Карнеги — Меллона
178 Анализ СМИ и соцсетей, Медиалогия
179 Avalanche, Лавина Пульс
181 Крибрум, ГК InfoWatch

11 ГЕНЕРАЦИЯ КОНТЕНТА

6 SwinIR, Высшая техническая школа Цюриха
10 CogVideo, Университет Цинхуа
19 DALL·E 2, OpenAI
43 Remini, Bending Spoons
48 NeuMan, Apple и Университет Британской Колумбии
50 Редактирование объектов в XR, Осакский университет
76 Make-a-Video, Meta*
77 Hotpot.ai, Panabee
78 ИИ-хореограф, Калифорнийский университет и Google Research
79 Изменение цвета по примерам, Восточно-китайский педагогический университет
80 DeepFill v2, университеты США и Китая
81 VQGAN, Гейдельбергский университет
83 Перевод изображений в воксельные модели, Харбинский университет
131 StableDiffusion, StabilityAI
132 Character.ai
133 Frase
134 BAYOU
135 Rephrase
136 Radar, PA Media Group и Urbs Media
137 Адаптирующийся ИИ для обработки и генерации текста, исследователи из Самары
138 Генерация заголовков научных статей, Adobe Research
139 Генерация новостных статей, BVICAM
142 Sonantic
145 Jukebox, OpenAI

12 РЕКОМЕНДАЦИЯ КОНТЕНТА

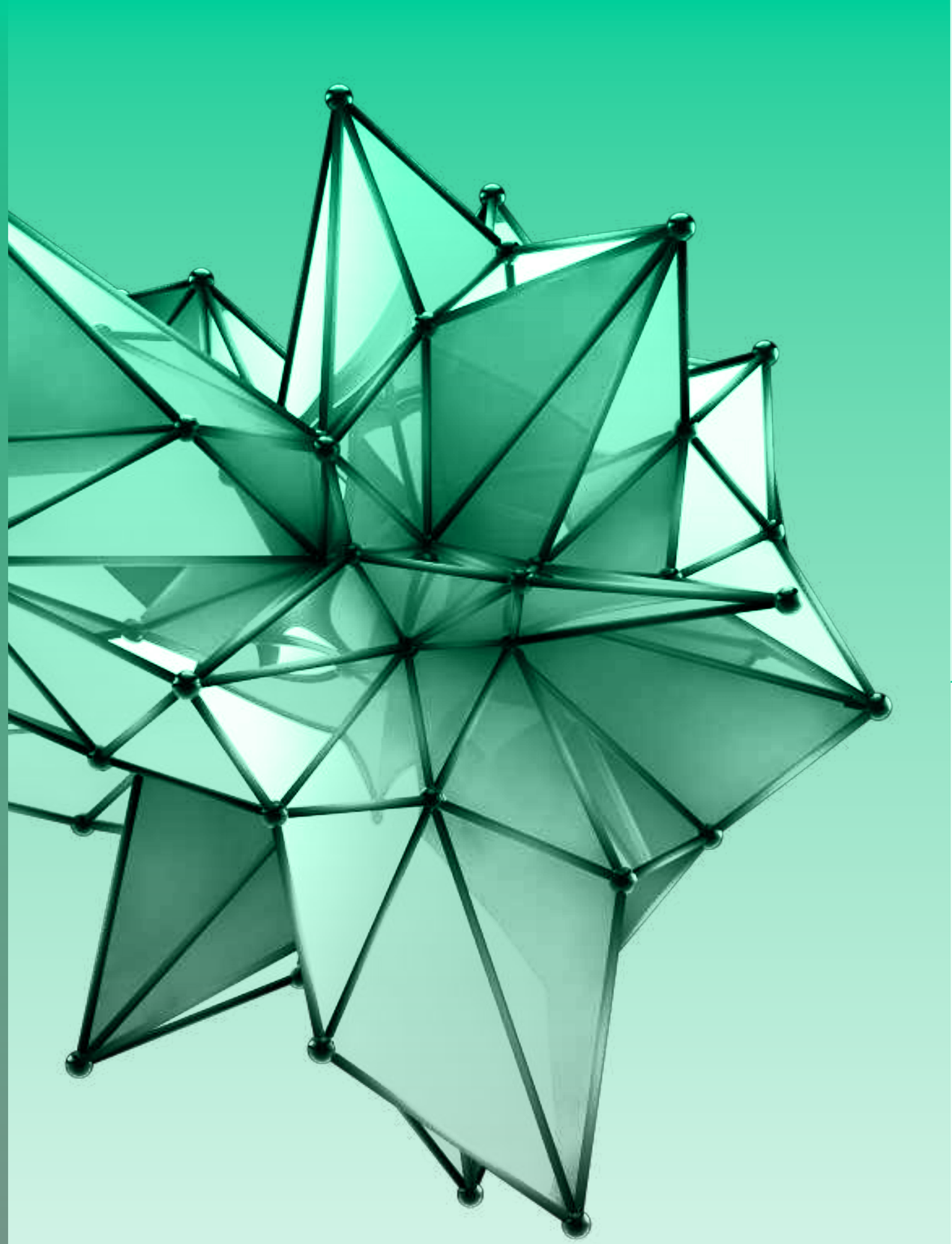
180 Рекомендация фильмов, международная группа учёных
182 CARCA, Университет Хильдесхайма
183 Интерактивная рекомендация одежды, Университет Глазго
184 M2TRec, Технологический институт Джорджии
185 Состязательное обучение рекомендательных алгоритмов, Netflix и Fever
186 Рекомендация на базе последних действий пользователя, Университет Глазго
187 EANA, Google Research
188 Персонализированные музыкальные плейлисты с учётом настроения, Deezer
189 Рекомендация недвижимости, Zillow Group
190 PinnerFormer, Pinterest
191 Рекомендательные алгоритмы для онлайн-знакомств на базе матчинга, CyberAgent Inc
192 Расследование рекомендательных алгоритмов TikTok
193 ProHealth eCoach, Университет Агдера
195 Адаптивная система рекомендаций, Netflix

СТАДИИ ВНЕДРЕНИЯ

Внедрено
Прототип
Концепция

* Компания признана экстремистской организацией в России

КЛЮЧЕВЫЕ ВЫВОДЫ ПО КЛАСТЕРАМ



1 ОБНАРУЖЕНИЕ ДИПФЕЙКОВ

В ПОСЛЕДНИЕ ГОДЫ ФЕЙКОВЫЕ НОВОСТИ СТАЛИ ПРОБЛЕМОЙ, ПРЕДСТАВЛЯЮЩЕЙ ОЩУТИМУЮ УГРОЗУ. ОТДЕЛЬНОЙ ФОРМОЙ ИНФОРМАЦИОННОГО КОНТЕНТА СЛУЖАТ ВИДЕОМАТЕРИАЛЫ. РОСТ ПОПУЛЯРНОСТИ ВИДЕО ПОДЧЁРКИВАЕТ НЕОБХОДИМОСТЬ ИНСТРУМЕНТОВ ДЛЯ ПОДТВЕРЖДЕНИЯ ПОДЛИННОСТИ МЕДИА И НОВОСТНОГО КОНТЕНТА, ПОСКОЛЬКУ НОВЫЕ ТЕХНОЛОГИИ ПОЗВОЛЯЮТ УБЕДИТЕЛЬНО ИСКАЖАТЬ И ПОДДЕЛЫВАТЬ ВИДЕО И С ВЫСОКОЙ СКОРОСТЬЮ РАСПРОСТРАНЯТЬ ПОДДЕЛКИ ЧЕРЕЗ СОЦИАЛЬНЫЕ СЕТИ, ОХВАТЫВАЯ АУДИТОРИЮ В ДЕСЯТКИ МИЛЛИОНОВ ПОЛЬЗОВАТЕЛЕЙ.

СУБТЕХНОЛОГИИ

- Компьютерное зрение
- Обработка естественного языка
- Распознавание и синтез речи
- Системы прогнозирования и поддержки принятия решений

СТАДИЯ ВНЕДРЕНИЯ

- В Внедрено
- К Концепция
- П Прототип

КОНТЕКСТ

Поддельные видео, фото и аудиофайлы сложно распознать невооружённым взглядом. При этом злоумышленники способны производить их в огромных объёмах и использовать в широком спектре своих задач.

ЧЕЛОВЕК НЕ СПОСОБЕН ЭФФЕКТИВНО РАСПОЗНАВАТЬ ДИПФЕЙКИ

Исследование Сиднейского института нейрологии постановило, что мозг может распознать дипфейки в 54% случаев, но сознательно человек определяет их лишь в 37% случаев, при этом можно с уверенностью сказать, что с дальнейшим развитием технологии обе цифры будут уменьшаться. Исследование Sentinel Analysis показало, что в YouTube отредактированные таким образом видео набрали более 5 млрд просмотров, а в TikTok более 65 млн. Производимый объём дипфейков и низкий процент их распознавания обывателем говорят о том, что людей нецелесообразно использовать для идентификации дипфейков.

ДИПФЕЙКИ — МНОГОЗАДАЧНЫЙ ИНСТРУМЕНТ ДЛЯ ЗЛОУМЫШЛЕННИКОВ

Дипфейки и другие способы редактирования изображений с помощью ИИ представляют собой угрозу для систем аутентификации, а также института репутации и распространения правдивой информации в целом.

Зарегистрированы случаи использования дипфейков мошенниками для обхода биометрических систем, создания злоумышленниками аккаунтов несуществующих людей, шантажа, производства нелегального контента, проведения мошеннических кампаний от имени известных лиц, подкрепления клеветы ложными доказательствами.

АЛГОРИТМЫ ДЛЯ СОЗДАНИЯ ДИПФЕЙКОВ ОСТАВЛЯЮТ АРТЕФАКТЫ

Самые простые из них буквально оставляют следы, заметные невооружённому глазу. Тем не менее для того чтобы идентифицировать дипфейк как таковой, человеку нужно обладать необходимыми техническими знаниями и разглядеть эти следы.

Наиболее совершенные алгоритмы всё равно имеют проблемы с симуляцией тонких особенностей фотографий и видео: текстурами кожи, цветом лица, геометрией головы человека, динамикой изменения этих параметров в связи с сердцебиением и дыханием или поворотами головы в нетипичных ракурсах.

ВЫВОДЫ

Качественные системы для обнаружения дипфейков чаще всего производятся частными компаниями и редко используются или проверяются государствами. В большинстве своём граждане остаются с проблемой дипфейков наедине, и государства редко регулируют их создание или распространение.

ГРАЖДАНАМ НУЖЕН ПРОСТОЙ И ПОНЯТНЫЙ СПОСОБ ОПРЕДЕЛЕНИЯ ДИПФЕЙКОВ

На данный момент у обывателей нет общепринятого метода и инструментария, с помощью которых можно обнаруживать подделки, созданные ИИ. Кроме того, далеко не все люди вообще задумываются о том, дипфейк перед ними или нет, из-за чего задача проверки ложится на плечи заинтересованных в выявлении лиц, чьи силы и технические возможности ограничены.

ГОСУДАРСТВАМ НУЖЕН СОБСТВЕННЫЙ ИНСТРУМЕНТАРИЙ ДЛЯ ИДЕНТИФИКАЦИИ ДИПФЕЙКОВ

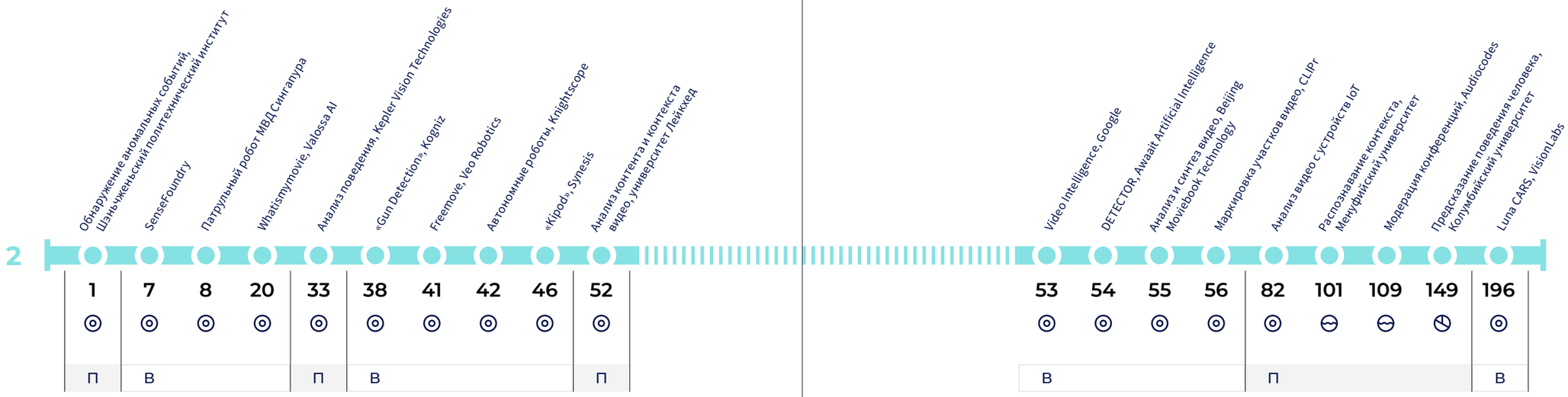
Наличие подобного инструментария позволит государствам не зависеть от инструментов, разрабатываемых корпорациями и другими коммерческими компаниями. Также государство сможет иметь все нужные ему для распознавания и анализа подделок функции, самостоятельно решать, какой функционал добавить и как его разрабатывать. Кроме того, наличие такой системы позволит избегать конфликта интересов при проведении анализа подделок. Такая система может подключаться к API соцсетей и медиаплатформ и маркировать видео с дипфейками — подобная практика уже существует на многих платформах в других контекстах.

ЗАКОНЫ О ПОДДЕЛКАХ, ПРОИЗВЕДЁННЫХ ИИ, НЕДОСТАТОЧНО РАЗВИТЫ, А РЕШЕНИЯ ПО ОБНАРУЖЕНИЮ ДИПФЕЙКОВ НЕ ИСПОЛЬЗУЮТСЯ В СУДЕБНЫХ ПРОЦЕССАХ

Не все страны успевают разрабатывать законодательные нормы о дипфейках и иных подделках, произведённых с помощью ИИ. Тем не менее подделки, созданные с помощью ИИ, значительно отличаются от обычных и требуют особого подхода. В том числе во многих государствах до сих пор не требуется помечать дипфейки как таковые, хотя это бы позволило значительно снизить наносимый ими ущерб. Использование решений для обнаружения и идентификации дипфейков позволит значительно уменьшить трудозатраты экспертов и криминалистов. Также это поможет снизить количество поддельных медиафайлов, которые в итоге фигурируют в судебных делах как настоящие.

2 ОПРЕДЕЛЕНИЕ КОНТЕКСТА ПРОИСХОДЯЩЕГО НА ВИДЕО

ИНСТРУМЕНТЫ ДЛЯ РАСПОЗНАВАНИЯ ПРОИСХОДЯЩЕГО НА ВИДЕО С ПОМОЩЬЮ ИИ ПРИМЕНИМЫ В ШИРОКОМ СПЕКТРЕ СФЕР, СРЕДИ КОТОРЫХ ЗДРАВООХРАНЕНИЕ, СФЕРА РАЗВЛЕЧЕНИЙ, УПРАВЛЕНИЕ ДОРОЖНЫМ ДВИЖЕНИЕМ, УПРАВЛЕНИЕ АВТОНОМНЫМ ТРАНСПОРТОМ, ОХРАНА ОБЩЕСТВЕННОГО ПРАВОПОРЯДКА И ТАК ДАЛЕЕ. ФУНКЦИОНАЛЬНЫЕ ВОЗМОЖНОСТИ РЕШЕНИЙ ДЛЯ ОПРЕДЕЛЕНИЯ КОНТЕКСТА ПРОИСХОДЯЩЕГО НА ВИДЕО В РАЗРЕЗЕ КОМПЬЮТЕРНОГО ЗРЕНИЯ ЯВЛЯЮТСЯ ВЕСЬМА ВОСТРЕБОВАННЫМ И ПЕРСПЕКТИВНЫМ ПОЛЕМ ДЛЯ ИССЛЕДОВАТЕЛЬСКОЙ ДЕЯТЕЛЬНОСТИ И БУДУЩИХ РАЗРАБОТОК.



СУБТЕХНОЛОГИИ

- Компьютерное зрение
- Распознавание и синтез речи
- Обработка естественного языка
- Системы прогнозирования и поддержки принятия решений

СТАДИЯ ВНЕДРЕНИЯ

- В Внедрено
- П Прототип
- К Концепция

КОНТЕКСТ

Искусственный интеллект способен ограниченно распознавать происходящее на видеозаписи, включая взаимодействия между объектами и связи между такими взаимодействиями. Тем не менее глубина осмысления действий всё ещё далека от человеческой.

ИИ НЕ ПОНИМАЕТ СМЫСЛ

Несмотря на рывок в разработке инструментов, понятны некоторые ограничения в работе ИИ. ИИ в большинстве своем не способен осознавать полный контекст происходящего. Система может обрабатывать огромное количество файлов в секунду, но не способна понимать их настоящий смысл.

ПРОБЛЕМЫ В ЭТИКЕ И ПРОИЗВОДИТЕЛЬНОСТИ

Существуют проблемы в этике и производительности, которые ещё не разрешены. Обычные системы распознавания объектов и действий не всегда могут поддерживать модерацию контента на платформах из-за ограничений в их способности точно распознавать смысл сигналов и образов на видео.

ВЫСОКАЯ ВЕРОЯТНОСТЬ ОШИБКИ ПРИ РАБОТЕ С НЕОДНОЗНАЧНЫМИ СИТУАЦИЯМИ

Формы и объекты в файлах могут быть вторичны для злоумышленников, а смысл, обозначенный в рамках их субкультуры — первичен. Один и тот же файл может быть шуткой или рекомендацией к суициду. В таком случае обычные системы модерации имеют сниженную чувствительность. Например, несмотря на наличие инструментов для автоматической модерации у видеоплатформ, внешне безопасный контент со взрослым подтекстом регулярно просачивается в детские разделы. Злоумышленники намеренно скрывают реальный смысл своих видео, чтобы обойти ИИ-фильтры и попасть в детский сегмент.

ВЫВОДЫ

Решения кластера крайне требовательны к вычислительным мощностям, из-за чего разработчики ищут нетривиальные способы оптимизации. Несмотря на активную коммерциализацию, не все ниши рынка заполнены. Развитие решений приведёт к их повсеместному внедрению в процессы модерации, киберкриминалистики и обеспечения безопасности.

НА РЫНКЕ СУЩЕСТВУЮТ ПРИКЛАДНЫЕ И КОММЕРЧЕСКИЕ РЕШЕНИЯ, НО РЫНОК ЕЩЁ НЕ НАСЫЩЕН. КРОМЕ ТОГО, ОТКРЫВАЮТСЯ НОВЫЕ ПЕРСПЕКТИВНЫЕ НАПРАВЛЕНИЯ РАЗРАБОТОК

Несмотря на то, что системы отличаются большим энергопотреблением и высокой требовательностью к аппаратной базе, они уже коммерциализируются. Существуют прототипы, обращённые на работу в новых направлениях приложения кластера: обнаружения аномалий, мультимодального анализа видео, глубокого анализа поведения людей на видео. Решения, которые обрабатывают ранее записанное видео, менее требовательны, чем системы, обрабатывающие видео в реальном времени.

ДЛЯ УЛУЧШЕНИЯ РЕЗУЛЬТАТОВ И ОПТИМИЗАЦИИ НЕКОТОРЫЕ СИСТЕМЫ ЦЕЛИКОМ ИЛИ ЧАСТИЧНО ПОЛАГАЮТСЯ НА ОБРАБОТКУ МЕТАДАННЫХ И ДАННЫХ, ОКРУЖАЮЩИХ ВИДЕОФАЙЛ

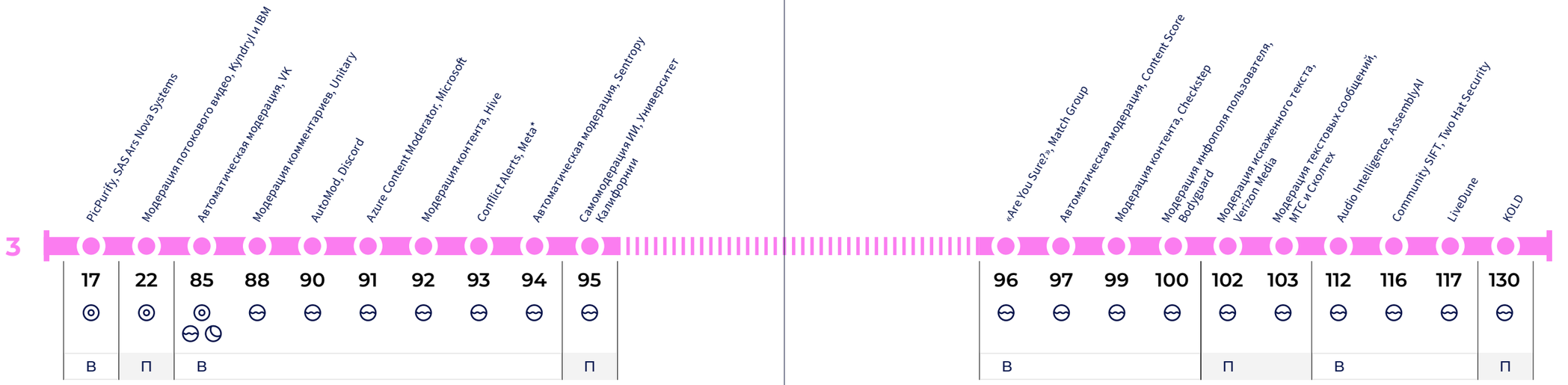
Это можно считать уязвимостью, но в некоторых случаях только такой подход позволит обрабатывать поток выходящих видео полностью при дефиците вычислительных мощностей. Обычно такой подход можно видеть в 2 типичных случаях. В первом результаты анализа метаданных и данных, окружающих файл, добавляются к результатам анализа самого видео для повышения точности. Во втором анализируются только метаданные и данные, окружающие видеофайл, чтобы минимизировать нагрузку на оборудование.

В БУДУЩЕМ ИИ СМОЖЕТ ПОНИМАТЬ НЕ ТОЛЬКО КОНТЕКСТ ПРОИСХОДЯЩЕГО НА ВИДЕО, НО И НЮАНСЫ ЕГО ПРОИЗВОДСТВА

Системы начинают анализировать всё больше данных — как в самом файле, так и вне его. Рано или поздно это приведёт к созданию системы, которая будет способна определять методы, режиссёрские приёмы и технологии, задействованные в создании того или иного видео. Кроме того, ИИ сможет понимать, в каких обстоятельствах видео было записано. Этот шаг значительно поможет отличать ложные или ретушированные видео от настоящих, а также позволит отслеживать метатренды в производстве контента пользователями и интернет-знаменитостями.

3 АВТОМАТИЗАЦИЯ МОНИТОРИНГА И МОДЕРАЦИИ КОНТЕНТА

ОГРОМНОЕ КОЛИЧЕСТВО КОНТЕНТА ЗАГРУЖАЕТСЯ И РАСПРОСТРАНЯЕТСЯ В ИНТЕРНЕТЕ ЕЖЕДНЕВНО, ЧТО НАМНОГО ПРЕВОСХОДИТ ВОЗМОЖНОСТИ ЛЮБОГО ПОСРЕДНИКА ПО АНАЛИЗУ КОНТЕНТА ПЕРЕД ЕГО ЗАГРУЗКОЙ. УЧИТЫВАЯ ФАКТ ВСЕОБЪЕМЛЮЩЕЙ ЦИФРОВИЗАЦИИ И ЕЁ БЕЗУСЛОВНУЮ ЭКОНОМИЧЕСКУЮ ВЫГОДУ, ПРИМЕНЕНИЕ МОДЕЛЕЙ ДЛЯ АВТОМАТИЗАЦИИ МОНИТОРИНГА И МОДЕРАЦИИ КОНТЕНТА БУДЕТ ТОЛЬКО РАСТИ В БУДУЩЕМ. ОДНАКО ИХ ВНЕДРЕНИЕ ТРЕБУЕТ ВЗВЕШЕННОГО И ОСТОРОЖНОГО ПОДХОДА, ЧТОБЫ СОХРАНИТЬ КАК ЭФФЕКТИВНОСТЬ РАБОТЫ СИСТЕМЫ, ТАК И ЛОЯЛЬНОСТЬ ПОТРЕБИТЕЛЯ.



СУБТЕХНОЛОГИИ

- Компьютерное зрение
- Обработка естественного языка
- Распознавание и синтез речи
- Системы прогнозирования и поддержки принятия решений

СТАДИЯ ВНЕДРЕНИЯ

- В Внедрено
- К Концепция
- П Прототип

* Компания признана экстремистской организацией в России

КОНТЕКСТ

Почти все платформы базируются на том, что их контент генерируется самими пользователями. Это приводит к огромному объёму информации, которая публикуется каждый день.
При этом, весь поток информации должен быть проанализирован и отфильтрован, чтобы до потребителей контента не доходили нелегальные или нежелательные сообщения. Человек давно перестал полностью справляться с постоянно растущей нагрузкой в этой области, но остаётся незаменимым элементом модерации.

ПОТОК КОНТЕНТА НА РАЗНЫХ ПЛАТФОРМАХ СТАЛО СЛОЖНО КОНТРОЛИРОВАТЬ

Эта проблема усиливается с развитием Интернета и ростом аудитории соцсетей. Несмотря на то, что большая часть публикуемой информации безвредна, в её потоке может появляться контент, легальность которого разнится в зависимости от страны. Уже сейчас каждый день публикуется 500 млн твитов.

СОЦСЕТИ ВЫНУЖДЕНЫ ПРИБЕГАТЬ К УСЛУГАМ ПЛАТНЫХ МОДЕРАТОРОВ ИЛИ ДОБРОВОЛЬЦЕВ

Так, Facebook* оплачивает услуги около 15–30 тыс. модераторов, а Twitch отдает модерацию в руки самих создателей контента, но рекомендует, чтобы на 200 зрителей был хотя бы один модератор.

НЕЗАКОННЫЙ КОНТЕНТ ОПАСЕН И ДЛЯ ПОЛЬЗОВАТЕЛЕЙ, И ДЛЯ МОДЕРАТОРОВ

Модераторы Facebook* жаловались на ПТСР, профессиональную деформацию и другие проблемы психологического характера. В 2020 году, в ходе судебного иска, компании пришлось заплатить более $50 млн нанятым модераторам.

АВТОМАТИЗАЦИЯ НЕ СПОСОБНА СПРАВИТЬСЯ СО ВСЕМИ УГРОЗАМИ

Существенная часть нежелательных сообщений и постов может быть обнаружена только благодаря пользователям — нейросети не всегда успевают за развитием меметического ландшафта, не понимают весь сленг.

* Продукт Meta, организации, которая признана в России экстремистской

ИСПОЛЬЗОВАНИЯ ТОЛЬКО РУЧНОЙ МОДЕРАЦИИ НЕДОСТАТОЧНО

Тем не менее многие исследования указывают на то, что исключительно ручная модерация в масштабах целой соцсети силами постоянных рабочих, у которых нет поддержки — несовершенный подход к модерации.

ПРЕСТУПНИКИ АКТИВНО ПРОТИВОДЕЙСТВУЮТ

Сообщества преступников также могут развивать собственный жаргон или намеренно деформировать свои сообщения, шифровать контент, компартментализировать передачу информации, чтобы не попадать под баны и не быть обнаруженными ИИ.

ВЫВОДЫ

Роль человека в модерации сместится в сторону поверки и совершенствования автоматизированных систем. На данный момент модерация непрозрачна и не действует по-разному в разных юрисдикциях, а также имеет задержку по отношению к публикации контента — эти две зоны продолжат быть проблемными, пока технические решения не достигнут нужного уровня.

МОДЕРАЦИЯ В РЕАЛЬНОМ ВРЕМЕНИ — ВЫЗОВ НА БЛИЖАЙШИЕ 5–10 ЛЕТ

Прямые эфиры, включая рекламу в них, пока что приходится модерировать вручную. Автоматическая система, способная выявлять неправомерный контент на стримах, в настоящее время не реализована в полной мере.

АВТОМАТИЗАЦИЯ МОЖЕТ ДОСТИГНУТЬ СТЕПЕНИ, ПРИ КОТОРОЙ ШТАТ СОТРУДНИКОВ СТАНЕТ МИНИМАЛЕН, А МОДЕРАТОРЫ КОНТЕНТА СТАНУТ ОПЕРАТОРАМИ ИИ-СИСТЕМ

Системы уже сейчас берут на себя большую долю рутинной работы модераторов. С развитием ИИ функция человека в системах модерации поменяется, перед людьми встанут другие задачи: отладка ИИ, арбитраж его решений в случае обжалования и поддержка коммуникации между платформой и пользователями. У платформ также появится отдельная группа, ответственная за отслеживание изменений в жаргоне, сленге и способов коммуникаций злоумышленников.

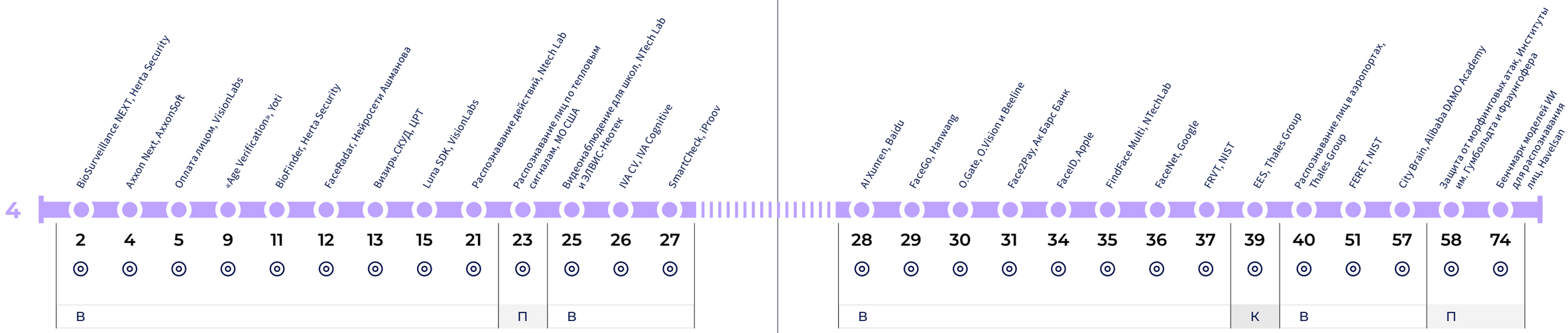
НЕПРОЗРАЧНОСТЬ РАБОТЫ АЛГОРИТМОВ И ПОЛИТИКИ МОДЕРАЦИИ ПЛАТФОРМ ПРИВОДЯТ К КОНФЛИКТАМ КАК С СООБЩЕСТВАМИ ПОЛЬЗОВАТЕЛЕЙ, ТАК И С ГОСУДАРСТВАМИ

Платформы редко раскрывают информацию о работе ИИ, интегрированного в их систему модерации, а правила, описанные публично самой платформой, часто не соответствуют фактическому подходу.

В условиях глобализации это приводит к тому, что решения модераторов и ИИ нередко контекстуальны или предвзяты, хотя платформа об этом не заявляет открыто. В результате механизмы бана непрозрачны.

4 РАСПОЗНАВАНИЕ ЛИЦ

РАСПОЗНАВАНИЕ ЛИЦ — ОБЛАСТЬ ПРАКТИЧЕСКОГО ПРИМЕНЕНИЯ ТЕОРИИ РАСПОЗНАВАНИЯ ОБРАЗОВ, ИНТЕРЕС К КОТОРОЙ НЕ УГАСАЕТ БОЛЬШЕ ТРЁХ ДЕСЯТИЛЕТИЙ. НА СЕГОДНЯШНИЙ ДЕНЬ ДАННЫЙ ВЕКТОР — ОДИН ИЗ САМЫХ ПОПУЛЯРНЫХ СРЕДИ ИССЛЕДОВАТЕЛЕЙ И РАЗРАБОТЧИКОВ, И ИМЕЕТ ВНУШИТЕЛЬНЫЙ ОПЫТ ВНЕДРЕНИЯ В ОКРУЖАЮЩУЮ НАС ДЕЙСТВИТЕЛЬНОСТЬ. ОДНАКО НА СЕГОДНЯ СИСТЕМЫ РАСПОЗНАВАНИЯ ЛИЦ И ИХ ВНЕДРЕНИЕ СОПРЯЖЕНЫ С РЯДОМ ПРОБЛЕМ, ОТ ТЕХНИЧЕСКИХ ТРУДНОСТЕЙ ДО ВОПРОСОВ КОНФИДЕНЦИАЛЬНОСТИ И ЭТИКИ.



СУБТЕХНОЛОГИИ

- Компьютерное зрение
- Обработка естественного языка
- Распознавание и синтез речи
- Системы прогнозирования и поддержки принятия решений

СТАДИЯ ВНЕДРЕНИЯ

- В Внедрено
- К Концепция
- П Прототип

КОНТЕКСТ

Распознавание лиц уже коммерциализировано и используется повсеместно. При этом точность и скорость, с которыми ИИ может обнаруживать лица и их идентифицировать, значительно превосходят человеческие способности.

ИИ ПРЕВОСХОДИТ ЧЕЛОВЕКА В ТОЧНОСТИ РАСПОЗНАВАНИЯ ЛИЦ

ИИ превзошёл человека в точности распознавания лиц: человек способен распознавать лица с точностью до 97,53%, тогда как существуют системы, чья точность превышает 99%.

Количество лиц, которые может распознать средний человек, ограничивается 5 тыс., в то время как ИИ способен распознавать лица из базы в сотни миллионов лиц.

РАСПОЗНАВАНИЕ ЛИЦ ВНЕДРЯЕТСЯ ПОВСЕМЕСТНО

Внедрение ИИ для распознавания лиц позволило значительно улучшить процедуры идентификации, ускорить поиск пропавших людей, а также снизить частоту мелких преступлений и повысить безопасность в аэропортах, банках и других публичных местах. С удешевлением технологии и её коммерциализацией, она стала использоваться не только для обеспечения безопасности, но и в бытовых, повседневных процессах: разблокировка смартфонов, оплата счёта, отслеживание здоровья и допуск к покупке продуктов с возрастным ограничением. В результате появляются огромные архивы, содержащие снимки лиц граждан.

ВНЕДРЕНИЕ И РАЗВИТИЕ ИИ ПО РАСПОЗНАВАНИЮ ЛИЦ ПРОИСХОДИТ НЕРАВНОМЕРНО

У кластера распознавания лиц весьма сильна региональная специфика. Так как технология уже привлекла значительное внимание со стороны широкой публики, некоторые страны начали активно её регулировать, вследствие чего разработка новых решений либо замедлилась, либо проводится с учётом новых нормативов. При этом страны, в которых регуляторика отстаёт от развития технологий или оставляет широкое пространство для интерпретации, способны достигать серьёзных прорывов в этой области.

ВЫВОДЫ

Россия способна производить решения мирового уровня в распознавании лиц. Тем не менее внедрение технологии ставит под угрозу биометрические данные граждан. Злоумышленники уже ищут способы по обходу систем для распознавания лиц, а с утечкой биометрии в их руках появится опасный инструмент для реализации нелегальных схем.

СУЩЕСТВУЮЩИЕ РЕШЕНИЯ ОБОШЛИ ЧЕЛОВЕКА В ЭФФЕКТИВНОСТИ И ВЫТЕСНИЛИ ЕГО ИЗ УЧАСТИЯ В НЕКОТОРЫХ ЗАДАЧАХ

Решения способны распознавать частично скрытые лица, разделять живых людей и фотографии, предсказывать возраст, опознавать каждого человека в больших скоплениях людей. Они снабжаются дополнительным функционалом для аналитики и работы с базами данных. Развитые экономики активно используют эти решения для обеспечения безопасности и в иных целях, что привело к стремительному росту и развитию рынка. Существуют нетривиальные решения, умеющие опознавать лица по их тепловому излучению или строить модель лица человека по всего лишь нескольким снимкам.

ПРЕСТУПНИКИ БУДУТ ИСКАТЬ МЕТОДЫ ПО ОБМАНУ СИСТЕМ РАСПОЗНАВАНИЯ

Уже существуют маски, которые позволяют обмануть ИИ, заставить его не воспринимать лицо как таковое. Исследователи также разработали камуфляж, который запутывает систему, так как ИИ начинает находить лица в паттернах камуфляжа. Некоторые системы для проверки документов или изображений можно обмануть, внедрив в проверяемое изображение невидимые глазу человека паттерны, которые ИИ примет за лицо, либо деформировать изображение конкретного человека так, что компьютер не сможет его опознать. Такие техники будут постоянно эволюционировать и использоваться для атак на системы распознавания лиц.

В РОССИИ СУЩЕСТВУЮТ КОМПАНИИ, СПОСОБНЫЕ РАЗРАБАТЫВАТЬ РЕШЕНИЯ МЕЖДУНАРОДНОГО УРОВНЯ

Компании из России не только создали свой собственный комплект для разработки решений, но и смогли собрать собственные датасеты, спроектировать свою архитектуру, обучить нейросети и получить высокие оценки в нескольких проверках NIST.

МАССИВЫ БИОМЕТРИЧЕСКИХ ДАННЫХ — РИСК ДЛЯ ГРАЖДАН

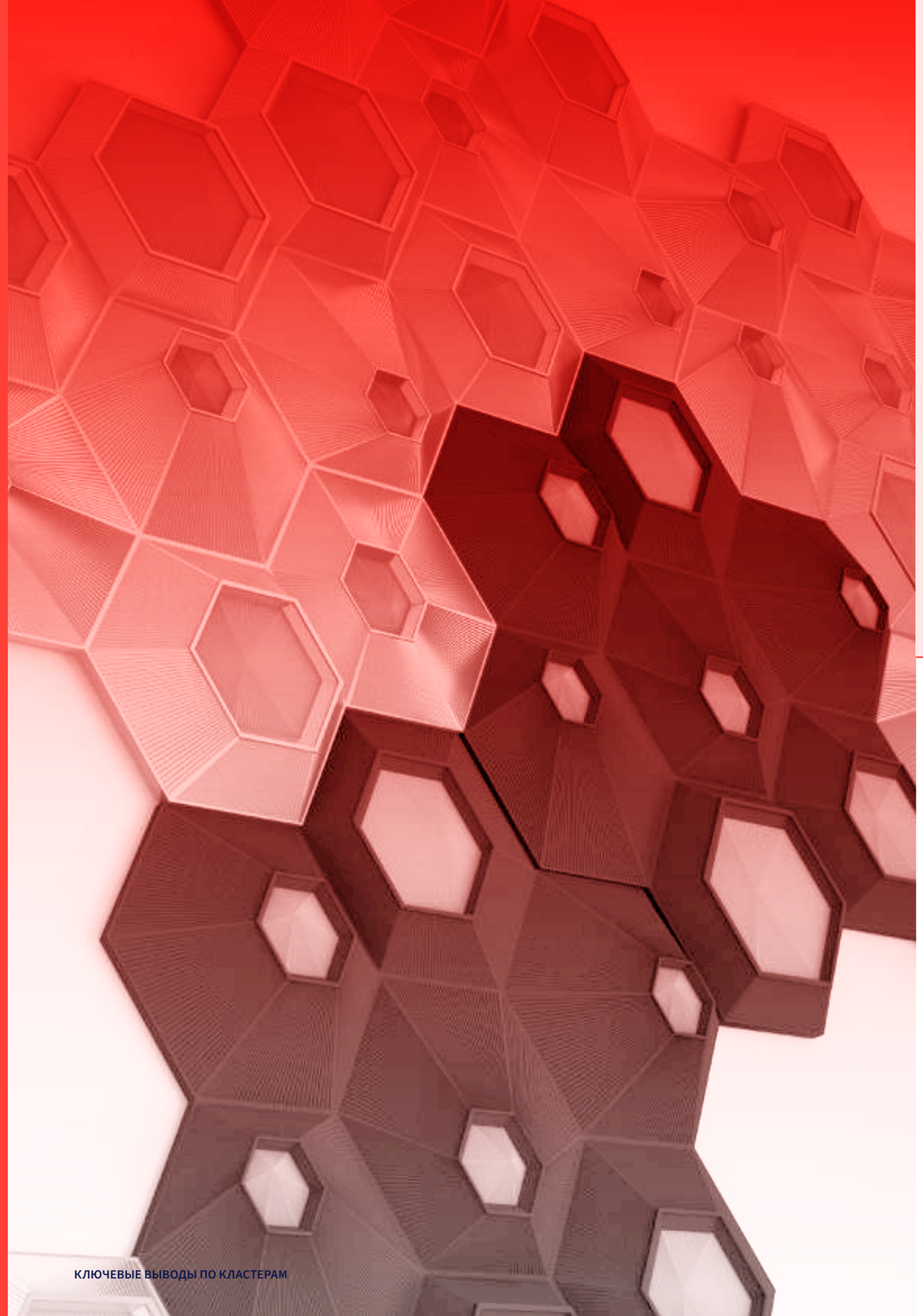
И пользователи решений, и разработчики собирают огромное количество фотографий с лицами людей. Чаще всего эти базы данных не деперсонализируются. В случае утечки эти данные могут быть использованы злоумышленниками для обхода процедур идентификации, отслеживания жертв и других целей.

ТЕХНОГИГАНТЫ ГОТОВЯТ ПОЧВУ ДЛЯ ПРОРЫВНЫХ РАЗРАБОТОК

Фундамент для создания решений был подготовлен компаниями-техногигантами и государствами. Так, исследователи Google и Nvidia создают свои уже натренированные нейросети, готовые к внедрению в будущие проекты. Nvidia и Baidu продают комплекты для разработки решений, которые включают в себя адаптированные для задачи чипы, камеры, библиотеки и руководства для работы. Маленькие компании редко имеют вычислительные мощности и датасеты, которые требуются для обучения нейросетей по распознаванию лиц. Alibaba создает многочисленные инструменты для разработчиков, а Apple внедряет передовые технологии и инвестирует в них сотни миллионов долларов.

ИСПОЛЬЗОВАНИЕ ЗАРУБЕЖНЫХ ДАТАСЕТОВ И НЕЙРОСЕТЕЙ — ВОПРОС НАЦИОНАЛЬНОЙ БЕЗОПАСНОСТИ

На 2021 г. в Москве находилось 213 тыс. камер наблюдения, а в Санкт-Петербурге — 70 тыс. Даже при использовании отечественного ПО, если оно базируется на зарубежных датасетах и нейросетях, существуют риски для безопасности. Нейросеть, лежащую в ядре отечественной системы, можно натренировать для нескольких разных целей, а датасет можно намеренно исказить. Оборудование и ПО, произведённое зарубежными корпорациями, может иметь заранее предусмотренные разработчиками уязвимости и двойное назначение.



5 ИЗВЛЕЧЕНИЕ СМЫСЛА ИЗ ТЕКСТА

ИЗВЛЕЧЕНИЕ СМЫСЛА ИЗ ТЕКСТА — ОДНА ИЗ ЗАДАЧ ОТДЕЛЬНОЙ ИССЛЕДОВАТЕЛЬСКОЙ ОБЛАСТИ, КОТОРАЯ НАЗЫВАЕТСЯ ОБРАБОТКОЙ ЕСТЕСТВЕННОГО ЯЗЫКА. РАСПОЗНАВАНИЕ И ОБРАБОТКА «ЧЕЛОВЕЧЕСКИХ» ЯЗЫКОВ ВНЕДРЕНА ВО МНОГИХ ОТРАСЛЯХ ЧЕРЕЗ ТЕХНОЛОГИЮ ЧАТ-БОТОВ — ИНТЕЛЛЕКТУАЛЬНЫХ ДИАЛОГОВЫХ ПРОГРАММ, ИМИТИРУЮЩИХ РАЗГОВОР В ЕГО ЕСТЕСТВЕННОЙ ФОРМЕ. ХОТЯ ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА И ИЗВЛЕЧЕНИЕ СМЫСЛА ИЗ ТЕКСТА В ЧАСТНОСТИ ИМЕЮТ СВОИ СЛОЖНОСТИ, СОЗДАННЫЕ СЕГОДНЯ МОДЕЛИ ОБЕСПЕЧИВАЮТ ШИРОКИЙ СПЕКТР ВОЗМОЖНОСТЕЙ ДЛЯ ЛЮБОГО БИЗНЕСА.

СУБТЕХНОЛОГИИ

- Компьютерное зрение
- Обработка естественного языка
- Распознавание и синтез речи
- Системы прогнозирования и поддержки принятия решений

СТАДИЯ ВНЕДРЕНИЯ

- В Внедрено
- К Концепция
- П Прототип

КОНТЕКСТ

Системы для осмысления текста сильно отличаются между собой по точности, глубине анализа и назначению. При этом они зачастую интегрируются в другие виды ИИ, обеспечивая повышенную точность и степень распознавания контекста.

СИСТЕМЫ ДЛЯ РАСПОЗНАВАНИЯ СМЫСЛА ИМЕЮТ ШИРОКИЙ СПЕКТР СЛОЖНОСТИ И ГЛУБИНЫ

Самыми простыми из таких систем можно считать те, что определяют тональность текста по количеству и морфологии слов. В то же время, существуют модели с сотнями миллионов и даже десятками миллиардов параметров, которые ориентируются на куда более сложные паттерны при анализе текста. К таким системам можно отнести ChatGPT, LaMDA, BERT, BLOOM и GPT-3, которые называются «большими языковыми моделями». Их функционал куда более широкий, и они осмысляют язык более глубоко, но требуют значительных вычислительных мощностей и крупных датасетов для тренировки.

СИСТЕМЫ РАБОТАЮТ БЫСТРЕЕ ЧЕЛОВЕКА, НО ОТСТАЮТ В ГЛУБИНЕ ПОНИМАНИЯ СМЫСЛА

Скорость, с которой модель способна считывать текст и распознавать его смысл, зависит от аппаратуры, как и объём обрабатываемой информации. Это даёт моделям для распознавания смысла серьёзное преимущество перед человеком: они могут масштабироваться, пока на это хватает денег. Тем не менее они не способны осмыслять текст так же глубоко, как обычный человек. Чем более абстрактный вывод требуется от модели, и чем сложнее контекст, окружающий текст, тем ниже её точность.

ОТ РАЗРАБОТОК В ОБЛАСТИ ОСМЫСЛЕНИЯ ТЕКСТА ЗАВИСЯТ РЕШЕНИЯ В ДРУГИХ КЛАСТЕРАХ

Извлечение смысла из текста и работа с текстом как таковая либо синергируют с остальными кластерами, либо напрямую позволяют им существовать. Так, например, без обработки текста и его смысла, рекомендательные системы оставались бы на примитивном уровне, а системы для модерации никогда бы не продвинулись дальше блокировки конкретных фраз. Системы для извлечения смысла из текста могут служить мощным дополнением к алгоритмам других кластеров или помогать с оптимизацией нагрузки на оборудование.

ВЫВОДЫ

Решения для анализа текста редко требуют масштабной аппаратной базы для своей работы, но для обучения больших языковых моделей всё равно нужны значительные вычислительные мощности. Существенный объём решений уже внедрен и используется для автоматизации документооборота и обработки пользовательских запросов.

БОЛЬШИНСТВО ПОДОБНЫХ СИСТЕМ ИСПОЛЬЗУЕТСЯ ДЛЯ АНАЛИЗА ДОКУМЕНТАЦИИ, ОБРАБОТКИ ОТЗЫВОВ И МОДЕРАЦИИ

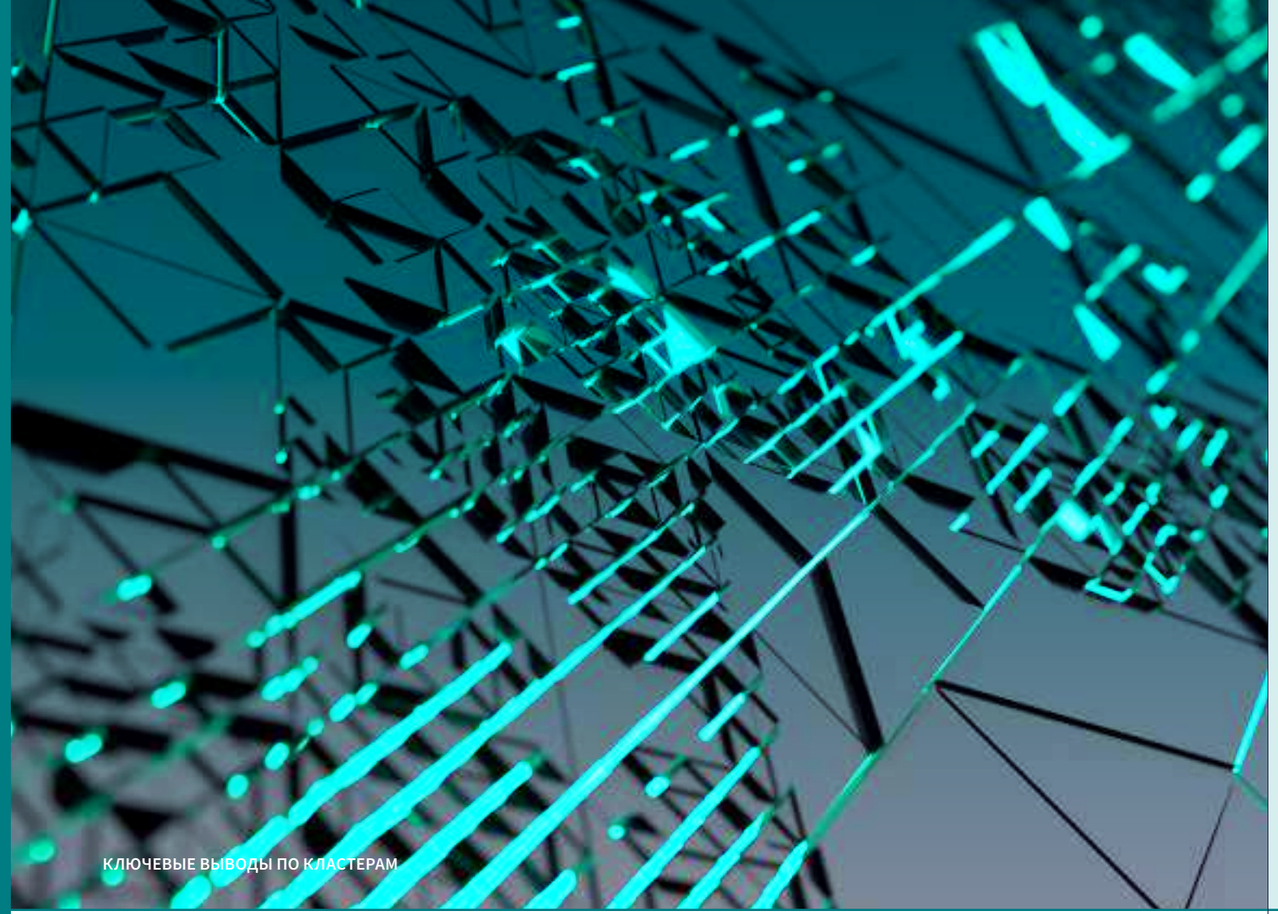
К основным методам здесь относятся установка причинно-следственных связей в тексте, экстракция фактов, распознавание конкретного смысла, в котором слово было употреблено, а также создание кратких описаний для массивов текста или групп разрозненных текстов. Решения способны не только упорядочивать и оптимизировать обработку информации, но и строить интуитивно понятные инфографики по результатам работы.

РЕШЕНИЯ СПОСОБНЫ РАБОТАТЬ В РЕАЛЬНОМ ВРЕМЕНИ И ДОСТАТОЧНО ЭКОНОМНЫ В ПЛАНЕ ТРЕБОВАНИЙ К АППАРАТНОЙ БАЗЕ

Тем не менее датасеты для обучения таких ИИ чаще всего должны быть корректно размечены и иметь достаточный объём. При этом существуют примеры работающих ИИ этой категории, обученные в полуавтоматическом режиме, без учителя. Большие модели вроде GPT-3 являются исключением из этих правил, и для их обучения требуются серьёзные вычислительные мощности. Но по сравнению с другими модальностями текст всё равно остаётся одним из наименее требовательных к вычислительным мощностям.

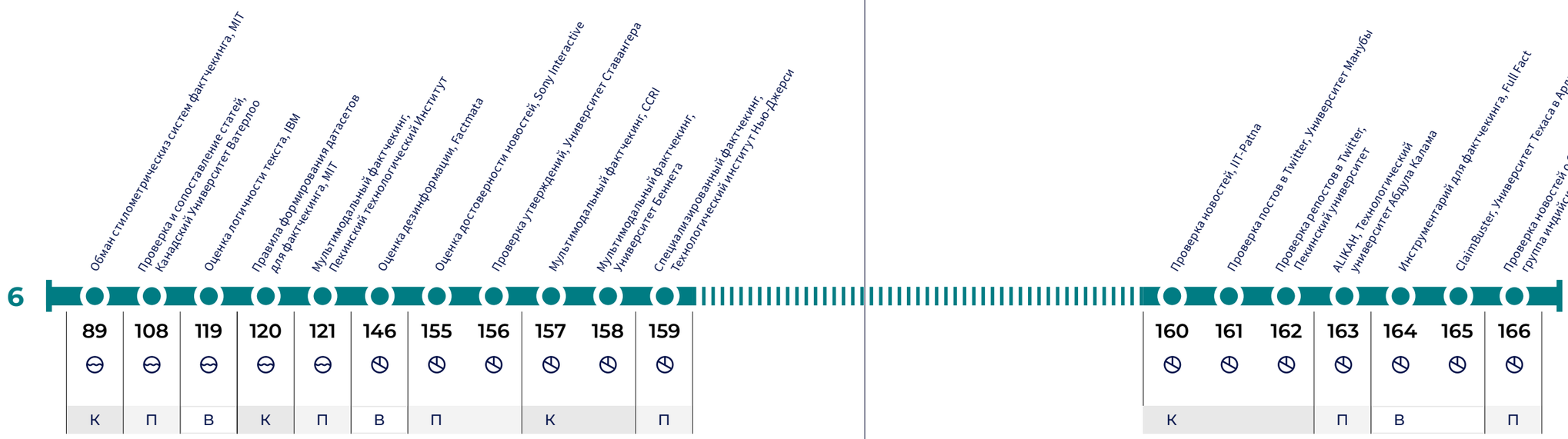
ЗЛОУМЫШЛЕННИКИ МОГУТ МАНИПУЛИРОВАТЬ СИСТЕМОЙ, СОЗДАВАЯ ВИДИМОСТЬ ТРЕНДОВ ИЛИ ИНФОРМАЦИОННЫЙ ШУМ

Если злоумышленники знают, что система внедрена, они могут манипулировать ей, создавая видимость трендов, или создавать информационный шум вокруг реально существующих трендов. Кроме того, так как многие системы отталкиваются от статичных словарей и грамматических моделей, злоумышленники могут специально запутывать систему, используя локальные эвфемизмы и метафоры.



6 ПОДДЕРЖКА ФАКТЧЕКИНГА

БОРЬБА С ДЕЗИНФОРМАЦИЕЙ ЯВЛЯЕТСЯ СЛОЖНОЙ ЗАДАЧЕЙ, ТРЕБУЮЩЕЙ ЦЕЛОСТНОГО ПОДХОДА С УЧАСТИЕМ РАЗЛИЧНЫХ ЗАИНТЕРЕСОВАННЫХ СТОРОН, ТАКИХ КАК ГОСУДАРСТВЕННЫЕ РЕГУЛИРУЮЩИЕ ОРГАНЫ, АДМИНИСТРАЦИИ ОНЛАЙН-ПЛАТФОРМ, ОБРАЗОВАТЕЛЬНЫЕ И ИССЛЕДОВАТЕЛЬСКИЕ УЧРЕЖДЕНИЯ И ТАК ДАЛЕЕ. СИСТЕМЫ ПОДДЕРЖКИ ФАКТЧЕКИНГА НАПРЯМУЮ ЗАВИСЯТ ОТ РАЗВИТИЯ ТЕХНОЛОГИЙ, НА КОТОРЫХ БАЗИРУЮТСЯ, ТАКИХ, КАК ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА ИЛИ КОМПЬЮТЕРНОЕ ЗРЕНИЕ. ТЕСНОЕ СОТРУДНИЧЕСТВО МЕЖДУ ПЛАТФОРМАМИ ПРОВЕРКИ ФАКТОВ, ИССЛЕДОВАТЕЛЯМИ И РАЗРАБОТЧИКАМИ ЯВЛЯЕТСЯ ЖИЗНЕННО НЕОБХОДИМЫМ УСЛОВИЕМ ДЛЯ БОРЬБЫ С ДЕЗИНФОРМАЦИЕЙ В БУДУЩЕМ.



СУБТЕХНОЛОГИИ

- Компьютерное зрение
- Обработка естественного языка
- Распознавание и синтез речи
- Системы прогнозирования и поддержки принятия решений

СТАДИЯ ВНЕДРЕНИЯ

- В Внедрено
- К Концепция
- П Прототип

КОНТЕКСТ

Проверка правдивости публичных заявлений – необходимый процесс в условиях постправды, и запрос на него растет. Каждая проверка является по сути небольшим расследованием, сложность которого зависит от контекста, и в оценке правдивости может участвовать огромное количество факторов. ИИ не способен полностью осознать все нюансы того или иного заявления и провести расследование самостоятельно.

ОБЪЁМ КОНТЕНТА И СКОРОСТЬ ЕГО СОЗДАНИЯ ВОЗРОСЛИ, ЧЕЛОВЕК НЕ СПРАВЛЯЕТСЯ

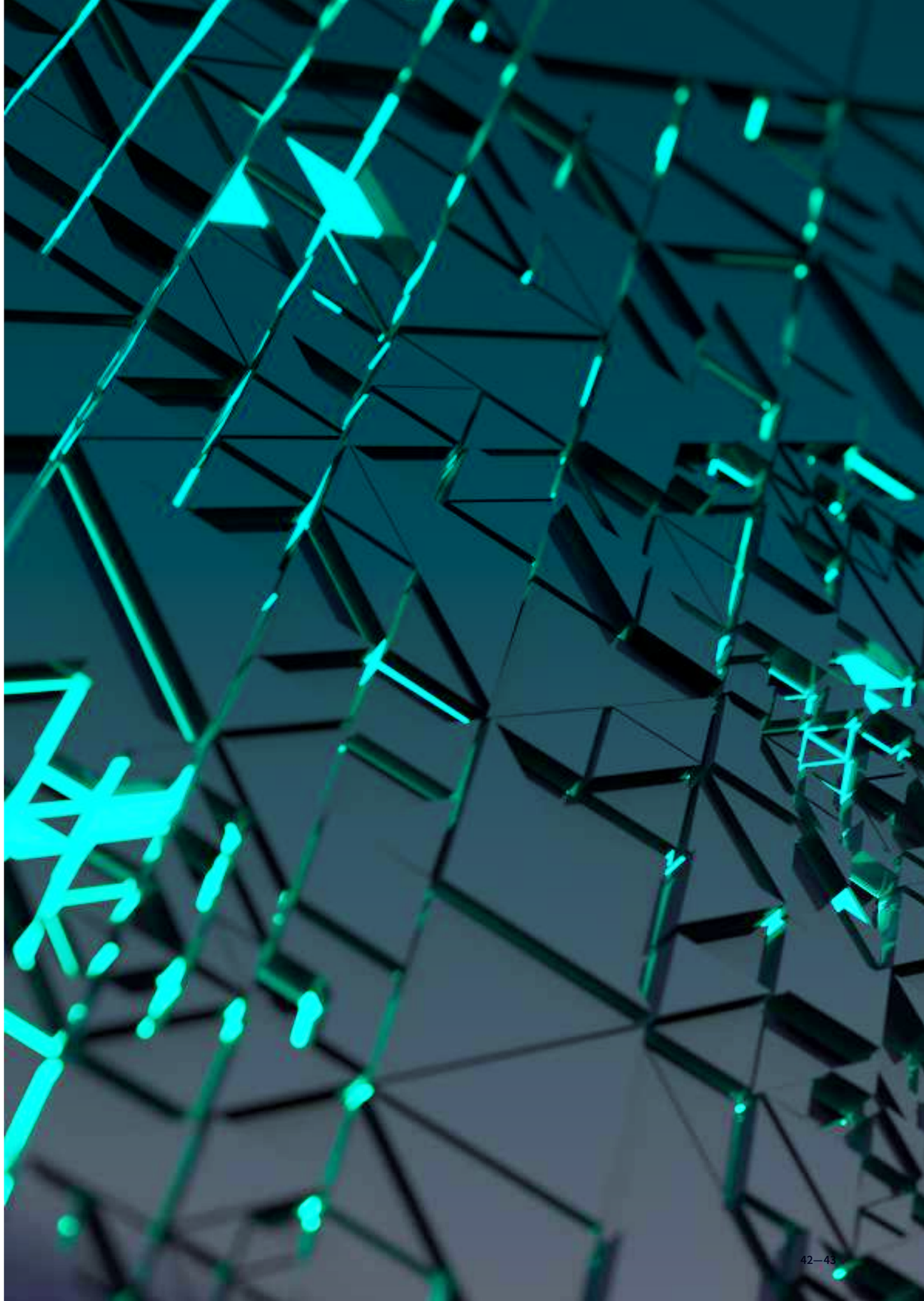
В России более 274 тыс. блогеров с аудиторией от 100 тыс. человек. Более тысячи из них имеют аудиторию свыше миллиона. Также существует около 60 тыс. зарегистрированных СМИ. Не все из них проверяют информацию, которую распространяют. За фактчекингом обычно стоит расследование с привлечением экспертов и подробной проверкой фактов. Из-за объёма контента и кропотливости работы требуется автоматизация данного процесса.

ИИ НЕ МОЖЕТ ИССЛЕДОВАТЬ НОВОСТЬ ГЛУБОКО

В отличие от человека, ИИ способен охватить большой объём распространяемых новостей и достаточно быстро их обработать. Но ИИ не всегда понимает контекст и не может исследовать новость так же глубоко, как человек. Также ИИ не всегда корректно обучен и может отталкиваться от ограниченного и поверхностного анализа в своих решениях.

ПОСТПРАВДА ЗНАЧИТЕЛЬНО ПОВЫШАЕТ УЩЕРБ ОТ ДЕЗИНФОРМАЦИИ

Так как в эпоху постправды ценность заявлений смещается от фактической верности к эмоциональному эффекту и стыковке с субъективными убеждениями аудитории, распространять дезинформацию стало гораздо проще, а эффект от распространения увеличился. Так, с 2020 г. по середину мая 2021 г., по данным РАНХиГС и НЦМУ, было зафиксировано более 6 млн постов и репостов с дезинформацией о пандемии коронавируса.



ВЫВОДЫ

Фактчекинг — слишком сложный процесс, чтобы его полностью автоматизировать. Тем не менее разработчики достигли успехов в создании инструментов, способных значительно облегчить работу человека.

НА ДАННЫЙ МОМЕНТ ПОЛНОСТЬЮ ПОЛОЖИТЬСЯ НА ИИ В ФАКТЧЕКИНГЕ НЕЛЬЗЯ, МОЖНО ЛИШЬ АВТОМАТИЗИРОВАТЬ ЧАСТЬ ПРОЦЕССОВ

Ни одно из найденных решений не имеет 100% точности, поэтому важно учитывать вероятный масштаб последствий при ошибке ИИ. Кроме того, эксперты и НКО в этой сфере признают, что на данный момент невозможно полностью автоматизировать весь процесс фактчекинга. По этой причине многие решения автоматизируют либо одну часть из процесса, либо концентрируются на выдаче оценки конкретному артефакту контента.

РАЗРАБОТЧИКИ ВЫСТРАИВАЮТ СИСТЕМЫ ДЛЯ ОБРАБОТКИ МУЛЬТИМОДАЛЬНОГО КОНТЕНТА, СОСТОЯЩИЕ ИЗ НЕСКОЛЬКИХ НЕЙРОСЕТЕЙ

Наиболее совершенные системы анализируют не только само заявление, но и оценивают его автора, прикреплённые медиафайлы, отслеживают, на что ссылается заявление и где оно распространяется. Эти факторы впоследствии ранжируются и анализируются, чтобы затем поступить на финальную оценку либо эксперту, либо центральной нейросети комплекса. Такие системы позволяют более адекватно оценивать достоверность контента, так как учитывают больше факторов.

СИСТЕМЫ ПРИБЛИЖАЮТСЯ К ФАКТЧЕКИНГУ В РЕАЛЬНОМ ВРЕМЕНИ И БУДУТ ИНТЕГРИРОВАНЫ В СОЦСЕТИ

В будущем системы смогут анализировать речь и потоковое видео, проверять их, а при обнаружении ложных новостей поднимать тревогу или оповещать пользователей о недостоверном контенте. Такие системы будут включать в себя комплекс работающих в симбиозе нейросетей и внедряться в соцсети. TikTok, Twitter, Youtube и Meta* дают гранты организациям фактчекеров и сами приближаются к внедрению автоматизированного фактчекинга.

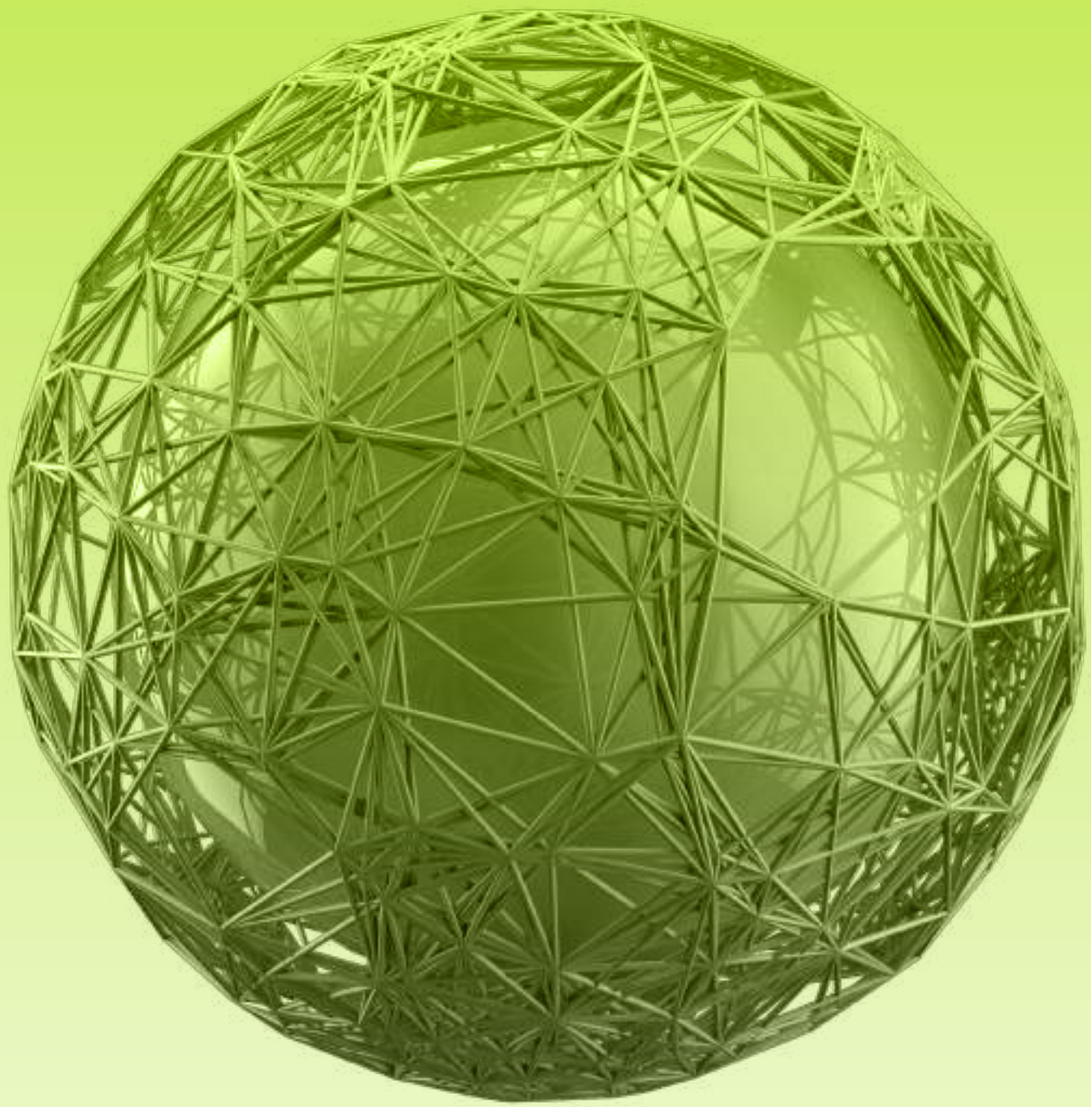
* Компания признана экстремистской организацией в России

СИСТЕМЫ ДЛЯ ФАКТЧЕКИНГА НАХОДЯТСЯ НА РАННЕМ ЭТАПЕ РАЗВИТИЯ И БУДУТ МАССОВО ВНЕДРЕНЫ ТОЛЬКО ЧЕРЕЗ 3–5 ЛЕТ

Даже техногиганты не могут полностью автоматизировать фактчекинг и вместо этого используют системы для разметки уже известных и широко распространившихся недостоверных новостей. Интерес к автоматизации фактчекинга значительно поднялся во времена пандемии коронавируса, и с тех пор направление стало активно развиваться. Тем не менее даже экспериментальные решения не покрывают все этапы процесса, не включают в себя все форматы контента и не всегда являются правильно обученными.

ВНЕДРЕНИЕ АВТОМАТИЗИРОВАННОГО ФАКТЧЕКИНГА ПРИВЕДЁТ К ТОМУ, ЧТО КАЖДАЯ НОВОСТЬ БУДЕТ РАЗМЕЧЕНА СПЕЦИАЛЬНЫМИ ТЕГАМИ НА КРУПНЫХ ПЛАТФОРМАХ

Государства смогут напрямую влиять на фактчекинг, сформировав систему тегов и отслеживая корректность разметки новостей техногигантами. Высока вероятность, что корпорации будут некорректно размечать новости, чтобы сформировать выгодную оптику в отношении инфоповодов. Такая масштабная система разметки уже используется в Twitter для оповещения пользователей о том, что сообщение опубликовано иностранным государственным СМИ, но она не подключена к системе автоматического фактчекинга.



7 РАСПОЗНАВАНИЕ СИМВОЛИКИ

РАСПОЗНАВАНИЕ СИМВОЛИКИ — ЭТО ОБЛАСТЬ КОМПЬЮТЕРНОГО ЗРЕНИЯ, КОТОРОЙ ПОСВЯЩЕНО НЕМАЛО НАУЧНЫХ РАБОТ. И В ОТЛИЧИЕ ОТ СМЕЖНОЙ ОБЛАСТИ — ОПТИЧЕСКОГО РАСПОЗНАВАНИЯ СИМВОЛОВ — РАСПОЗНАВАНИЕ СИМВОЛИКИ НА РАННИХ ЭТАПАХ ЗАСЛУЖИЛО ОТДЕЛЬНОГО НАПРАВЛЕНИЯ ИССЛЕДОВАНИЙ ИЗ-ЗА ГОРАЗДО БОЛЬШЕГО КОЛИЧЕСТВА И РАЗНООБРАЗИЯ СИМВОЛОВ. С РАЗВИТИЕМ ТЕХНОЛОГИЙ, ОСОБЕННО МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ, РАСПОЗНАВАНИЕ СИМВОЛИКИ СТАЛО ОДНОЙ ИЗ ЗАДАЧ ВИЗУАЛЬНОГО РАСПОЗНАВАНИЯ И ПОТОМУ ЕЁ ВНЕДРЕНИЕ СЕГОДНЯ ВСЕЦЕЛО ЗАВИСИТ ОТ РАЗВИТИЯ ТЕХНОЛОГИЙ КОМПЬЮТЕРНОГО ЗРЕНИЯ.

СУБТЕХНОЛОГИИ

- Компьютерное зрение
- Распознавание и синтез речи
- Обработка естественного языка
- Системы прогнозирования и поддержки принятия решений

СТАДИЯ ВНЕДРЕНИЯ

- В Внедрено
- П Прототип
- К Концепция

КОНТЕКСТ

Несмотря на то, что распознавание символики не является сложной задачей для техногигантов и государств, решения в этом направлении чаще всего держат подальше от глаз публики. Главной проблемой решений является многогранность, локальность, и контекстуальный характер маргинальной символики: разработка из США, обученная осмыслять символы BLM, не будет полезна в контексте России. При этом, в зависимости от инфоповодов, символы устаревают и меняются.

ОТКРЫТЫЕ РАЗРАБОТКИ ДЛЯ РАСПОЗНАВАНИЯ НЕЖЕЛАТЕЛЬНОЙ СИМВОЛИКИ — РЕДКОСТЬ

В распознавании символов большинство задач решены, но государства и корпорации редко оповещают широкую публику о том, как они используют технологию в борьбе с маргинальными идеологиями. Исключение составляет ИИ, разработанный для модерации соцсетей.

Но в таких случаях нейросеть тренируется не только для поиска экстремистской символики, но и для поиска другого нежелательного контента. Такие нейросети не предназначены для распознавания символики вне соцсетей, например на баннерах во время протестов.

НЕЖЕЛАТЕЛЬНАЯ СИМВОЛИКА ЭВОЛЮЦИОНИРУЕТ

Одной из ключевых проблем в разработке решений является постоянное развитие маргинальных движений и их символики. При появлении нового движения радикалов или изменении в идеологии уже существующего, символика меняется. Она также эволюционирует при реакции движения на разные инфоповоды. При этом радикалы понимают, что публикация их символики на открытых платформах и её демонстрация в реальной жизни приведёт к ответным действиям правоохранительных органов, а потому принимают соответствующие меры.

СМЫСЛ СИМВОЛА ЗАВИСИТ ОТ ЛОКАЛЬНОГО КУЛЬТУРНОГО КОНТЕКСТА

Для корректного распознавания маргинальной символики человеку требуется экспертиза в маргинальных движениях и их локальной специфике. Так, для человека из Индии знак африканской группировки Боко Харам не будет значить ничего, а использование свастики в религиозном контексте является нормой. Попытка масштабировать систему обнаружения маргинальной символики без использования ИИ обречена на провал из-за многообразия символов и их связи с локальной культурой.

ВЫВОДЫ

Решения для распознавания символики из других регионов могут быть адаптированы для работы в России. Более того, существует ряд решений для осмысления сочетаний символов и мультимодального анализа, которые могут быть модифицированы для опознания символики маргинальных движений, с которыми борется государство.

ТЕХНОЛОГИИ НАХОДЯТСЯ НА ВЫСОКОМ УРОВНЕ РАЗВИТИЯ, НО НЕ ВСЕГДА ПОДХОДЯТ ДЛЯ РЕШЕНИЯ ЗАДАЧ В РОССИИ

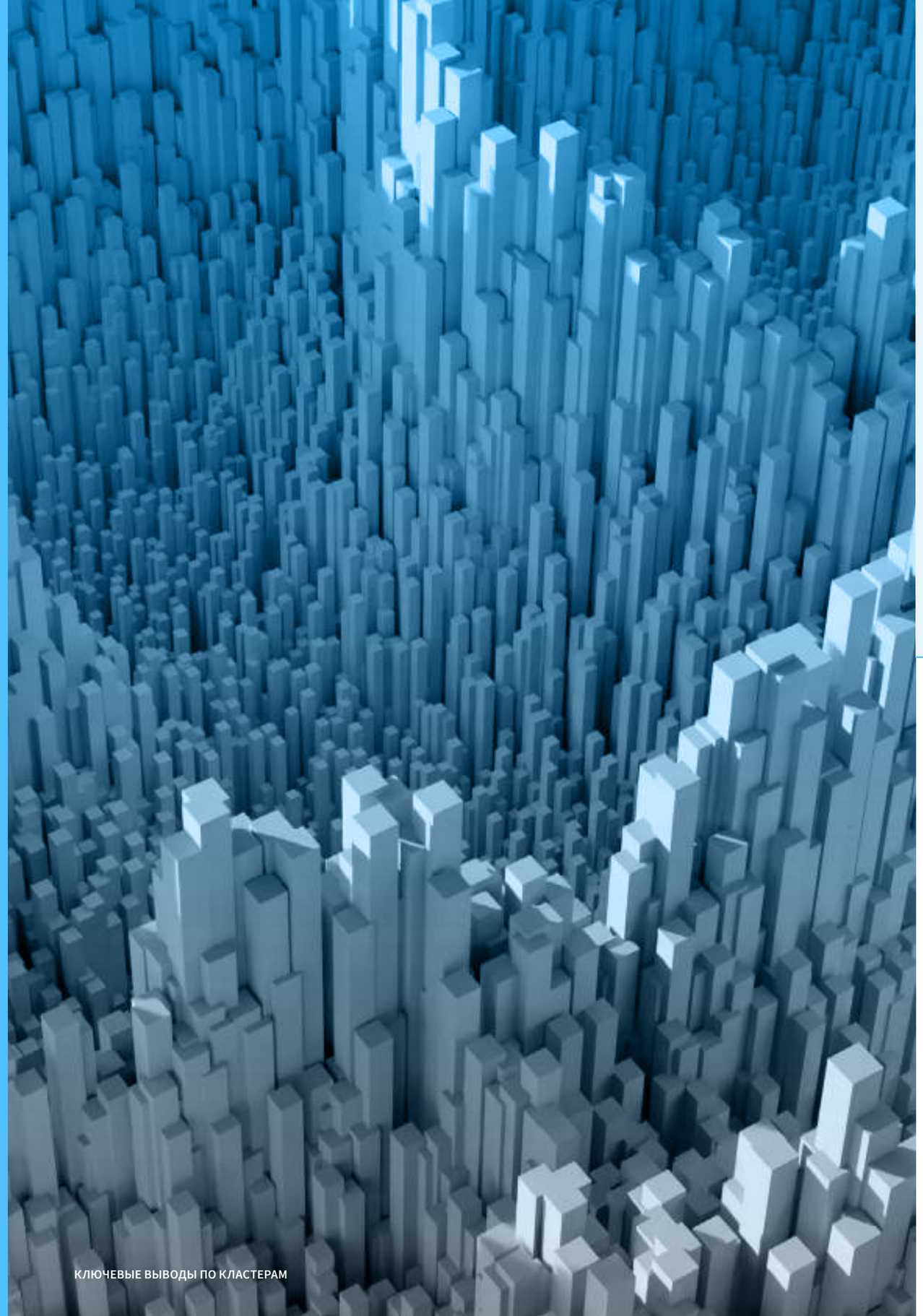
Даже лучшие иностранные решения не смогут полностью ни фильтровать контент, ни обнаруживать радикальную символику на протестах или в граффити в России, если ИИ не будет дополнительно обучен. Импортированные решения смогут только определять символику тех движений, которые присутствуют и в России, и в стране-импортёре. Например, импортированное европейское решение будет опознавать общие неонацистские символы, но не сможет опознавать символы радикальных сепаратистов или символику криминального сообщества.

РЕШЕНИЯ ДВИЖУТСЯ К ПРЕДСКАЗАНИЮ СМЫСЛА СИМВОЛОВ, ОПРЕДЕЛЕНИЮ СМЫСЛА ИХ КОМБИНАЦИЙ И ВЗАИМОДЕЙСТВИЙ

Такие решения позволят верно осмыслять сложные комбинации символики и текстов, которые нередко появляются на протестах в качестве баннеров. Большой объём артефактов, содержащих нежелательную символику, включает в себя сочетание нескольких символов с текстом и производится вручную. Системы смогут анализировать смысловую нагрузку таких артефактов, а при их совмещении с инструментами для распознавания контекста — точно оценивать контекст вокруг нежелательного символа.

СИСТЕМЫ ДЛЯ ОБНАРУЖЕНИЯ НЕЖЕЛАТЕЛЬНОЙ СИМВОЛИКИ ДОЛЖНЫ БЫТЬ МУЛЬТИМОДАЛЬНЫ

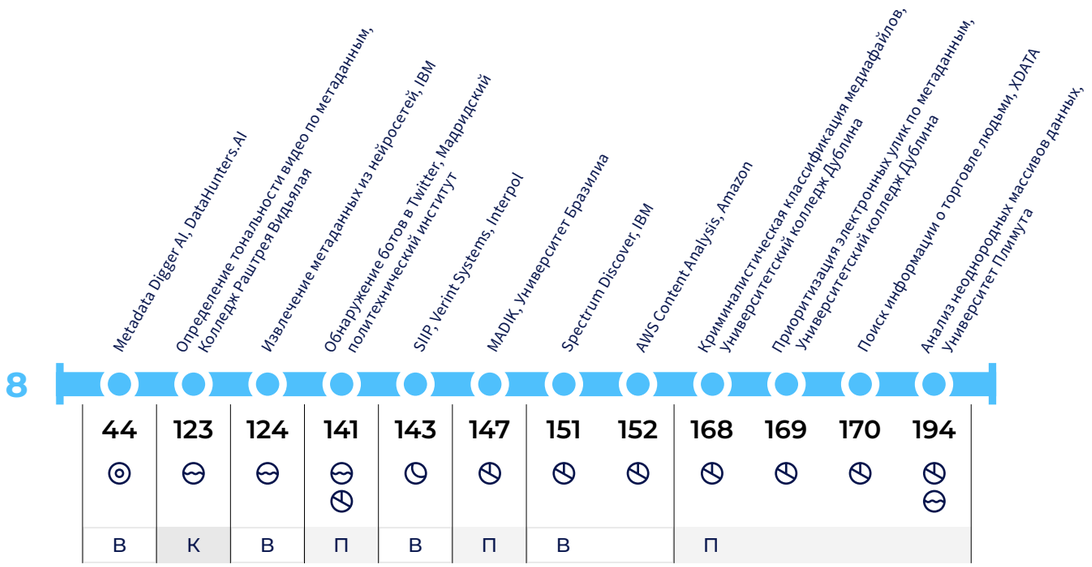
Благодаря спонсорам с компетенциями в области информационных операций, радикалы способны быстро адаптироваться к мерам правоохранительных органов. В связи с этим методы демонстрации и публикации нежелательной символики многообразны и адаптивны. Поэтому комплексное решение для обнаружения символики должно охватывать события как на улицах, так и в виртуальных пространствах, а также работать в мультимодальном формате. Унифицированных решений для борьбы с распространением экстремистской символики, масштабируемых на все сферы деятельности государства, пока что не существует.



КЛЮЧЕВЫЕ ВЫВОДЫ ПО КЛАСТЕРАМ

8 ИЗВЛЕЧЕНИЕ И АНАЛИЗ МЕТАДАННЫХ

МЕТАДАННЫЕ ПРЕДСТАВЛЯЮТ СОБОЙ СТРУКТУРИРОВАННЫЕ СПРАВОЧНЫЕ ДАННЫЕ, КОТОРЫЕ ПОМОГАЮТ СОРТИРОВАТЬ И ИДЕНТИФИЦИРОВАТЬ АТРИБУТЫ ИНФОРМАЦИИ. БОЛЬШИНСТВО МОДЕЛЕЙ НА ОСНОВЕ ИИ СЕГОДНЯ СФОКУСИРОВАНЫ НА ПОВЫШЕНИИ ПРОСТОТЫ И КОМФОРТА РАБОТЫ С МЕТАДАННЫМИ. СПЕКТР ПРИКЛАДНОГО ПРИМЕНЕНИЯ АНАЛИТИЧЕСКИХ СИСТЕМ ОГРОМЕН, ОТ СПЕЦИАЛИЗИРОВАННЫХ РЕШЕНИЙ (НАПРИМЕР, В КРИМИНАЛИСТИКЕ) ДО ИНСТРУМЕНТА В РУКАХ НОВАТОРОВ, РАЗВИВАЮЩИХ ИДЕИ ДЕЦЕНТРАЛИЗОВАННОГО ИНТЕРНЕТА.



СУБТЕХНОЛОГИИ

- Компьютерное зрение
- Обработка естественного языка
- Распознавание и синтез речи
- Системы прогнозирования и поддержки принятия решений

СТАДИЯ ВНЕДРЕНИЯ

- В — Внедрено
- К — Концепция
- П — Прототип

КОНТЕКСТ

Анализ метаданных является устоявшейся практикой в OSINT и киберкриминалистике. Внедрение ИИ в эти сферы позволяет значительно расширить получаемую таким образом информацию и ускорить сам анализ.

МЕТАДАННЫЕ АКТИВНО ИСПОЛЬЗУЮТСЯ В КРИМИНАЛИСТИКЕ

Так как злоумышленники оставляют цифровые следы, метаданные стали объектом интереса криминалистов. Ранее экспертам приходилось самостоятельно анализировать время создания файлов и их редактирования, логи сетевых пакетов и работы программ, геоданные, информацию об аппаратуре, с помощью которой был создан файл и другие виды метаданных. Всё это занимало нецелесообразное количество времени. ИИ для анализа метаданных развивался при сильном влиянии со стороны нужд киберкриминалистики, поэтому её задачи во многом были решены.

ИИ МОЖЕТ ЗАПОЛНЯТЬ ПРОБЕЛЫ В МЕТАДАННЫХ ПО КОСВЕННЫМ ПРИЗНАКАМ

Техногиганты занимаются разработкой ИИ, предназначенного для работы с большими данными, в частности, для обогащения метаданных и управления ими. Это позволяет упорядочить крупные базы данных и более эффективно ими распоряжаться. ИИ способен воспроизводить и восстанавливать метаданные для файла, чьи метаданные были отредактированы или удалены.

ТРЕНД НА МУЛЬТИМОДАЛЬНОСТЬ ИИ ЗАВИСИТ ОТ ИННОВАЦИЙ В РАБОТЕ С МЕТАДАННЫМИ

Именно метаданные позволяют системам анализировать файлы разных форматов и информацию разных видов. Системы способны соединять и анализировать результаты работы разных модулей именно потому, что эти модули производят специальные виды метаданных, например, векторы, которые в дальнейшем отправляются на анализ главным модулям.

ВЫВОДЫ

Разработка автоматических систем для работы с метаданными — приоритетное направление для ускорения разработки ИИ и оптимизации самих моделей. При этом корпорации уже создают многофункциональные системы для обогащения, восстановления и обработки метаданных.

СИСТЕМЫ ДЛЯ АВТОМАТИЗИРОВАННОГО ОБОГАЩЕНИЯ МЕТАДАННЫХ ПОЗВОЛЯТ СОЗДАВАТЬ СИНТЕТИЧЕСКИЕ ДАТАСЕТЫ, ЧТО УСКОРИТ РАЗРАБОТКУ ИИ

Это позволит задействовать неупорядоченные и гетерогенные источники больших данных эффективным образом, например, автоматически составляя из них нужные датасеты. Разработка и коммерциализация ИИ станут ещё более простыми, когда подобные автоматизированные системы для работы с метаданными будут широко доступны публике и смогут самостоятельно искать файлы, похожие на образцы, предоставленные пользователем. В этом кроется одна из ключевых причин, по которой техногиганты заинтересованы в автоматизации производства и обогащения метаданных – они владеют хранилищами больших данных, которые не всегда упорядочены и чей потенциал не реализован в полной мере.

ИИ ДЛЯ ОБРАБОТКИ МЕТАДАННЫХ ВЫВОДИТ КРИМИНАЛИСТИКУ НА НОВЫЙ УРОВЕНЬ

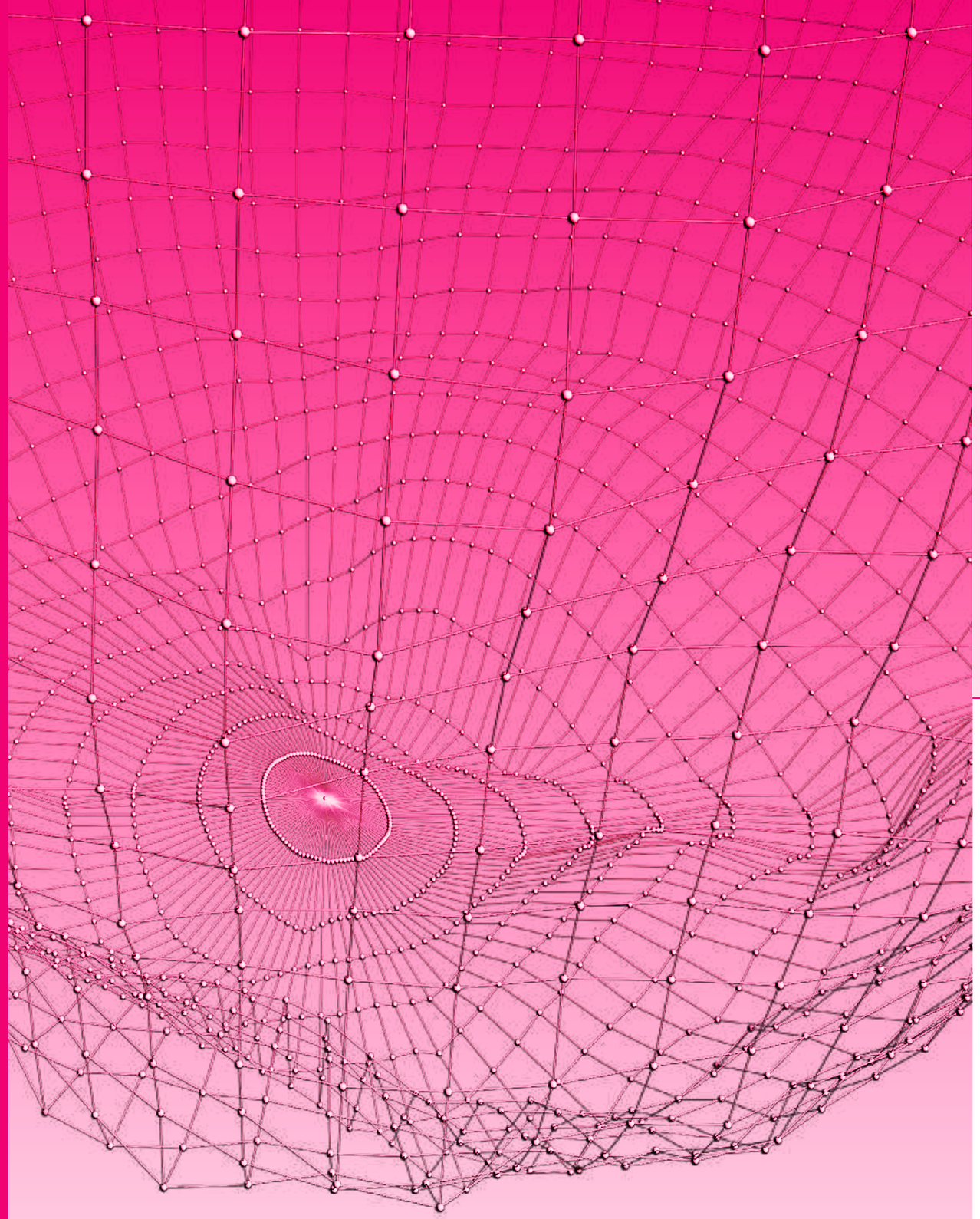
ИИ для обработки метаданных позволяет находить сложные и скрытые паттерны в документах, между которыми могут быть годы хранения в архивах или огромная разница в форматах. На поиск подобных паттернов у человека уйдет гораздо большее количество времени. При этом способность заметить паттерн уменьшается с объёмом анализируемого архива, если специалисту не предоставлены инструменты. Способности ИИ к распознаванию слабовыраженных паттернов не уменьшаются в условиях быстрорастущей базы данных. Как показывают некоторые решения, массивные базы данных наоборот помогают ИИ точнее работать.

АНАЛИЗ МЕТАДАННЫХ ПОЗВОЛЯЕТ ЗНАЧИТЕЛЬНО ЭКОНОМИТЬ ВЫЧИСЛИТЕЛЬНЫЕ МОЩНОСТИ ПРИ РЕШЕНИИ РЯДА ЗАДАЧ

Существуют решения, которые полностью полагаются на метаданные для анализа изображений, видео, текстовых документов и других файлов. Такие решения обычно гораздо менее требовательны к вычислительным мощностям. Тем не менее у таких систем есть значительные ограничения: для их использования требуются качественно размеченные файлы, а точность таких решений варьируется в зависимости от требуемой глубины анализа.

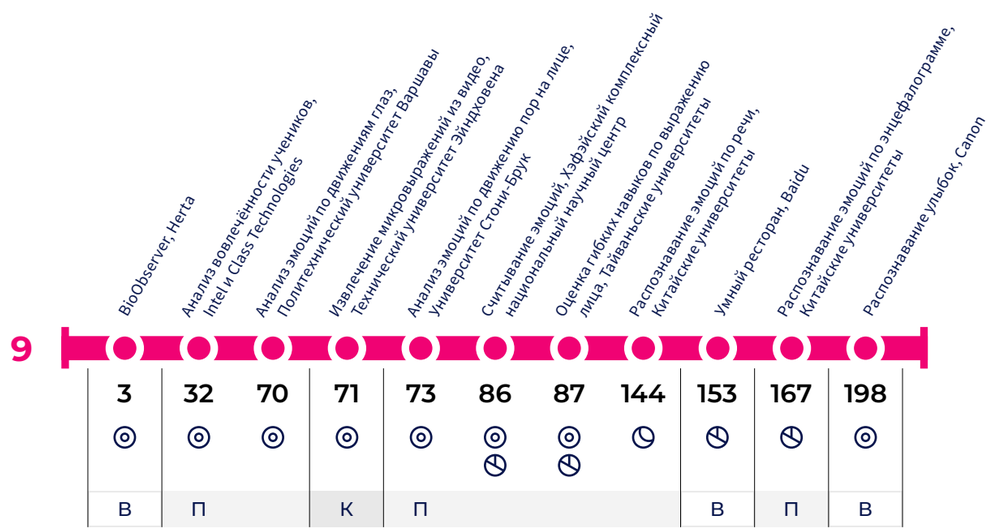
ТЕХНОГИГАНТЫ СОЗДАЮТ УНИВЕРСАЛЬНЫЕ СИСТЕМЫ ДЛЯ ОБРАБОТКИ МЕТАДАННЫХ В ПРОМЫШЛЕННЫХ МАСШТАБАХ

Корпорации занимаются разработкой систем, способных обрабатывать всевозможные виды файлов и выполнять широкий спектр действий с метаданными файлов. Эти системы предназначены для обработки больших массивов информации и способны автоматически оптимизировать управление такими базами данных. Оптимизация достигается за счет того, что системы способны самостоятельно производить, редактировать, восполнять и анализировать метаданные, что позволяет стандартизировать файлы и упорядочить их.



9 РАСПОЗНАВАНИЕ ЭМОЦИЙ

РАСПОЗНАВАНИЕ ЭМОЦИЙ С ПОМОЩЬЮ ИИ — ОДНО ИЗ САМЫХ ПОПУЛЯРНЫХ НАПРАВЛЕНИЙ ИССЛЕДОВАТЕЛЬСКОЙ ДЕЯТЕЛЬНОСТИ. С НЕУКЛОННЫМ ВОЗРАСТАНИЕМ РОЛИ СИСТЕМ ИИ В НАШЕЙ БЫТОВОЙ РЕАЛЬНОСТИ ЗА ПОСЛЕДНИЕ 30 ЛЕТ РАЗРАБАТЫВАЛОСЬ БОЛЬШОЕ КОЛИЧЕСТВО МЕТОДОВ, ОБЛЕГЧАЮЩИХ АНАЛИЗ ЭМОЦИЙ. ТЕХНОЛОГИИ РАСПОЗНАВАНИЯ УЖЕ СЕГОДНЯ ИМЕЮТ ШИРОКИЙ СПЕКТР ПРИМЕНЕНИЯ: ПОТРЕБИТЕЛЬСКИЙ СЕКТОР, БЕЗОПАСНОСТЬ, МЕДИЦИНА, ИНДУСТРИЯ РАЗВЛЕЧЕНИЙ, РОБОТОТЕХНИКА.



СУБТЕХНОЛОГИИ

- Компьютерное зрение
- Обработка естественного языка
- Распознавание и синтез речи
- Системы прогнозирования и поддержки принятия решений

СТАДИЯ ВНЕДРЕНИЯ

- В Внедрено
- К Концепция
- П Прототип

КОНТЕКСТ

ИИ не способен к эмпатии, но может распознавать эмоции. При этом, в отличие от человека-эксперта, машина способна быстро обрабатывать подробные данные о большом количестве разных физиологических сигналов, из-за чего точность качественной системы будет выше, чем у одного эксперта. Способность ИИ распознавать контекст и смысл взаимодействий между людьми зависит от разработок в этом направлении.

ИИ ДЛЯ РАСПОЗНАВАНИЯ ЭМОЦИЙ — ТЕХНОЛОГИЯ С МНОЖЕСТВОМ НАЗНАЧЕНИЙ

Многие из этих решений могут быть в дальнейшем использованы для развития нейронаук, в образовании, маркетинговых и информационных кампаниях, медицине, а также становиться инструментами правоохранительных органов. Распознавание эмоций может быть встроено в большинство процессов, в которых требуется мониторинг состояния человека или улучшение взаимодействия между человеком и машиной.

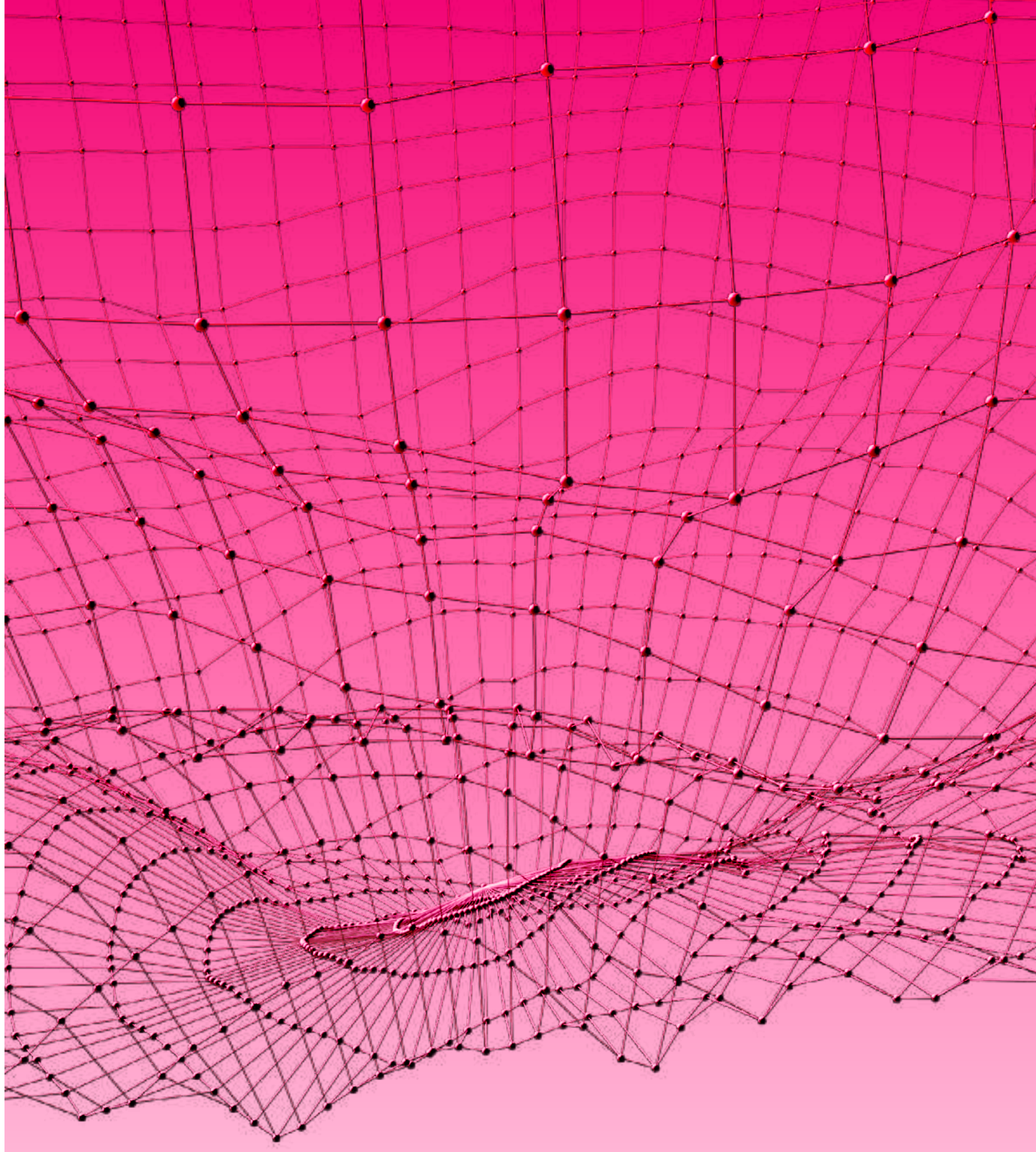
ИИ ПОЗВОЛЯЕТ ОСУЩЕСТВЛЯТЬ РАСПОЗНАВАНИЕ ЭМОЦИЙ МУЛЬТИМОДАЛЬНО

До значительного рывка в развитии ИИ способы ручного и автоматического распознавания эмоций были унимодальны. То есть большинство из них задействовало либо речь, либо мимику, либо голос человека, потому что анализ нескольких модальностей был бы слишком трудозатратен.

В психологии эмоций было сложно проверить некоторые гипотезы и теории без использования ИИ. Эта проблема постепенно решается, так как разработчики предоставляют учёным всё более совершенные решения, а учёные обновляют теоретическую базу для этих решений.

РАСПОЗНАВАНИЕ ЭМОЦИЙ ПОЗВОЛИТ ИИ ЛУЧШЕ ПОНИМАТЬ МЫШЛЕНИЕ ЧЕЛОВЕКА

Эмоции человека могут говорить ИИ о его реакции на тот или иной стимул. Это позволит, во-первых, понять, как человек по-настоящему относится к чему-либо, а во-вторых, поможет предсказывать действия человека и его намерения. Эмоции частично ответственны за часть решений, принимаемых людьми, поэтому учёт эмоций в предсказании поступков как отдельного человека, так и групп людей, играет важную роль. Кроме того, так как люди вкладывают эмоции в медиаконтент и испытывают их при его потреблении и производстве, — эмоции являются неотъемлемым элементом контекста.



ВЫВОДЫ

Развитие систем для распознавания эмоций может привести к тому, что ИИ будет понимать психологическое состояние человека лучше, чем он сам. Это представляет значительную угрозу для общества, так как подобный инструмент может попасть в руки злоумышленников.

ИСПОЛЬЗОВАНИЕ И РАЗРАБОТКА ПРИКЛАДНЫХ РЕШЕНИЙ ДЛЯ РАСПОЗНАВАНИЯ ЭМОЦИЙ ПРИВЕДЁТ К ПЕРЕОЦЕНКЕ СУЩЕСТВУЮЩИХ ТЕОРИЙ В ПСИХОЛОГИИ И ПРОРЫВУ В НЕЙРОНАУКАХ

Существенное количество решений основывается на методах, теориях и гипотезах из психологии, которым больше 10 лет. Это вызывает критику со стороны учёных и общественности. В большинстве случаев использование старого теоретического материала служит ограничением для систем, но существенное количество нового теоретического материала либо базируется на старом, либо требует проверки, которая была слишком трудоёмкой или вовсе невозможной до развития ИИ. В результате разработчикам новых решений придётся кооперироваться с исследователями психологии для отладки и улучшения как самих решений, так и теоретических суждений, на которых они базируются. Кроме того, при выходе на рынок решения будут использоваться учёными для проведения исследований, разработки новых теорий и методов.

ИСПОЛЬЗОВАНИЕ ИНОСТРАННЫМИ КОМПАНИЯМИ ИИ ДЛЯ РАСПОЗНАВАНИЯ ЭМОЦИЙ И СБОР ПОЛУЧАЕМЫХ ОТ ТАКОГО ИИ ДАННЫХ — УГРОЗА КОГНИТИВНОЙ БЕЗОПАСНОСТИ

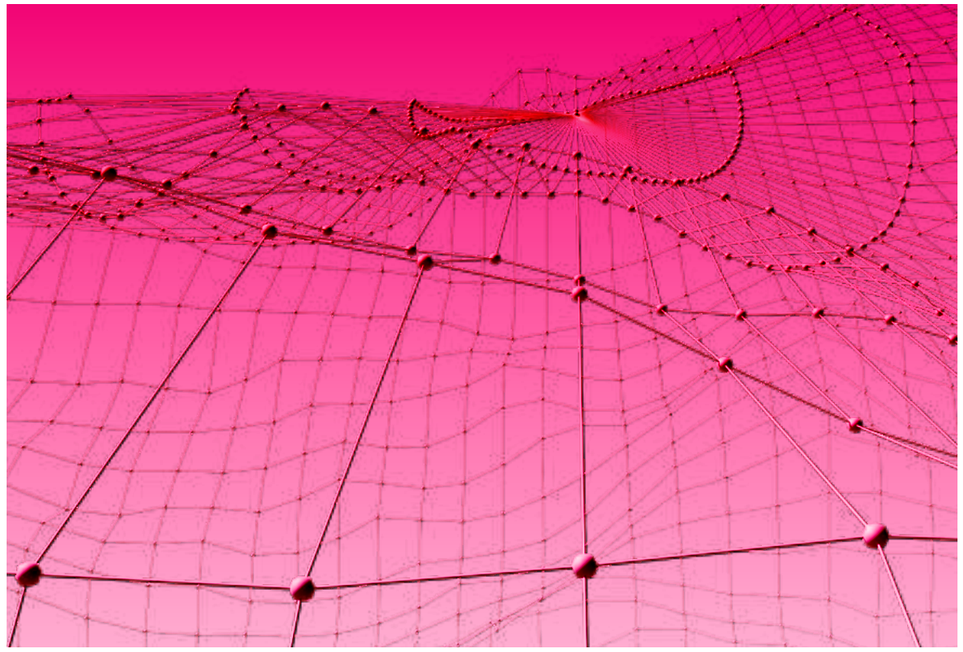
Такой ИИ можно использовать для вычисления наиболее эффективного стимула и его влияния на каждого отдельного пользователя. Анализ подобных данных позволит значительно улучшить рекомендательные алгоритмы и использовать их для гиперкастомизированной доставки пропаганды, в которой каждый медиафайл будет выбран не только исходя из предпочтений пользователя, но и с учётом его реакции. Поэтому распространение нарративов значительно ускорится. С помощью информационных кампаний станет гораздо проще убедить человека в той или иной точке зрения благодаря возможности предсказать, какой именно файл нужно продемонстрировать, чтобы вызвать те или иные эмоции.

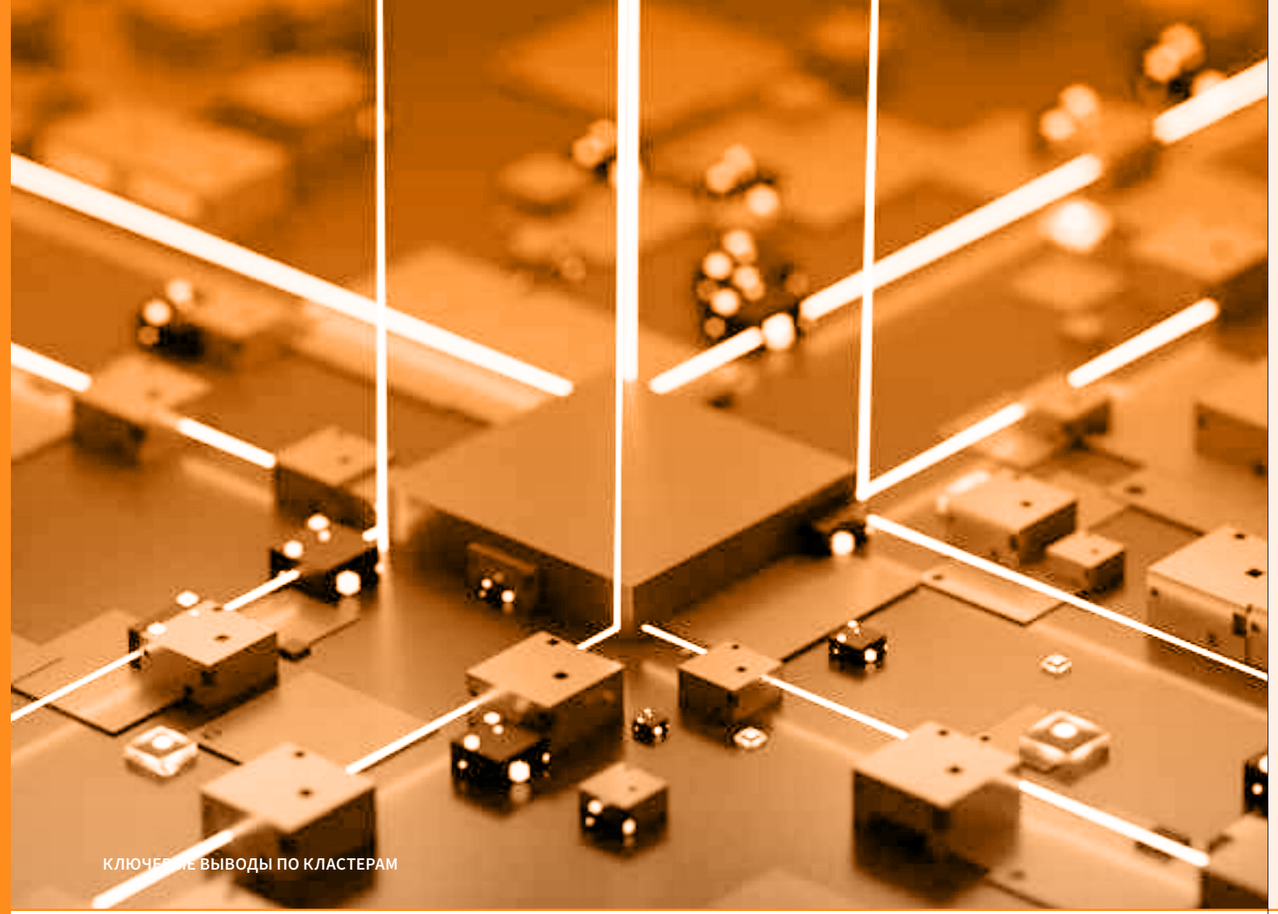
РАСПОЗНАВАНИЕ ЭМОЦИЙ БУДЕТ ПОВСЕМЕСТНО ВНЕДРЯТЬСЯ В КАЧЕСТВЕ ПОДДЕРЖКИ ДЛЯ ИИ, КОТОРЫЕ ВЗАИМОДЕЙСТВУЮТ С ЧЕЛОВЕКОМ ИЛИ АНАЛИЗИРУЮТ ЕГО ДЕЯТЕЛЬНОСТЬ

Системы по распознаванию эмоций могут значительно повысить качество анализа поведения человека, поэтому они будут внедряться в широкий спектр процессов. Многие системы видеонаблюдения и анализа контента уже используют модули для оценки тональности или мимики, но их точность и глубина понимания человеческой психологии далеки от идеальных. ИИ для распознавания эмоций значительно улучшит работу систем безопасности в публичных местах, образовательных и медицинских учреждениях, промышленных комплексах, правоохранительных органах, если будет совмещён с системами предиктивной аналитики и системами видеонаблюдения.

ЭМОЦИИ, ИСПЫТЫВАЕМЫЕ ЧЕЛОВЕКОМ, ОКАЗЫВАЮТ КОСВЕННОЕ И ПРЯМОЕ ВЛИЯНИЕ НА ЕГО ДЕЙСТВИЯ

Исследования в психологии и нейронауках показывают, что эмоции играют заметную роль в решениях, которые принимает человек. По результатам мета-анализа психологических работ, вышедших в последние 20 лет, решения группы инвесторов отклоняются от нормы на 21%, если они испытывают ярость, и на 14 %, если они испытывают страх. По результатам разных исследований, люди, испытывающие грусть, на 13-34% чаще выбирают сиюминутную выгоду, чем спокойные люди. За последние десятилетия был накоплен большой объём научного материала о влиянии эмоций на поведение человека, но чтобы точно измерить это влияние и его успешно предсказывать, потребуется ИИ, так как качественно составлять такие модели и просчитывать их человеку слишком трудозатратно.





10 ПОДДЕРЖКА РЕШЕНИЙ ПРИ ИНФОРМАЦИОННЫХ АТАКАХ

В СВЯЗИ С ЭКСПОНЕНЦИАЛЬНЫМ РОСТОМ ВОЗМОЖНОСТЕЙ ИИ И МАШИННОГО ОБУЧЕНИЯ ЗА ПОСЛЕДНИЕ 5 ЛЕТ, ЕГО ПРИМЕНЕНИЕ НАШЛО СВОЮ НИШУ КАК В СФЕРЕ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ И ЗАЩИТЫ НАСЕЛЕНИЯ ОТ ФЕЙКНЬЮС, ТАК И ПРИ ИНФОРМАЦИОННЫХ АТАКАХ, ПРОВОДИМЫХ С ЦЕЛЬЮ МАНИПУЛЯЦИИ ОБЩЕСТВЕННЫМ МНЕНИЕМ. ГЛАВНЫМ ПОЛЕМ СТОЛКНОВЕНИЯ ИНФОРМАЦИОННЫХ СИЛ СТАНОВЯТСЯ СОЦИАЛЬНЫЕ СЕТИ, СТАВШИЕ ИДЕАЛЬНОЙ ПЛАТФОРМОЙ ДЛЯ МОМЕНТАЛЬНОГО РАСПРОСТРАНЕНИЯ ИНФОРМАЦИИ ЛЮБОГО РОДА. ИИ ЗАРЕКОМЕНДОВАЛ СЕБЯ В КАЧЕСТВЕ НАДЁЖНОГО ПОМОЩНИКА ДЛЯ ДИНАМИЧЕСКОГО АНАЛИЗА СОЦИАЛЬНЫХ СЕТЕЙ В РАМКАХ НОВОЙ ДИСЦИПЛИНЫ — СОЦИАЛЬНОЙ КИБЕРБЕЗОПАСНОСТИ.

СУБТЕХНОЛОГИИ

- Компьютерное зрение
- Обработка естественного языка
- Распознавание и синтез речи
- Системы прогнозирования и поддержки принятия решений

СТАДИЯ ВНЕДРЕНИЯ

- В — Внедрено
- К — Концепция
- П — Прототип

КОНТЕКСТ

Информационные войны перекочевали в интернет и значительно изменились с его развитием. Активная борьба ведется как между государствами, так и между негосударственными организациями. В свете таких изменений, появились решения, позволяющие отслеживать, оценивать, анализировать и купировать информационные атаки.

СОЦИАЛЬНАЯ КИБЕРБЕЗОПАСНОСТЬ — НОВАЯ ИНЖЕНЕРНО-СОЦИОЛОГИЧЕСКАЯ НАУКА

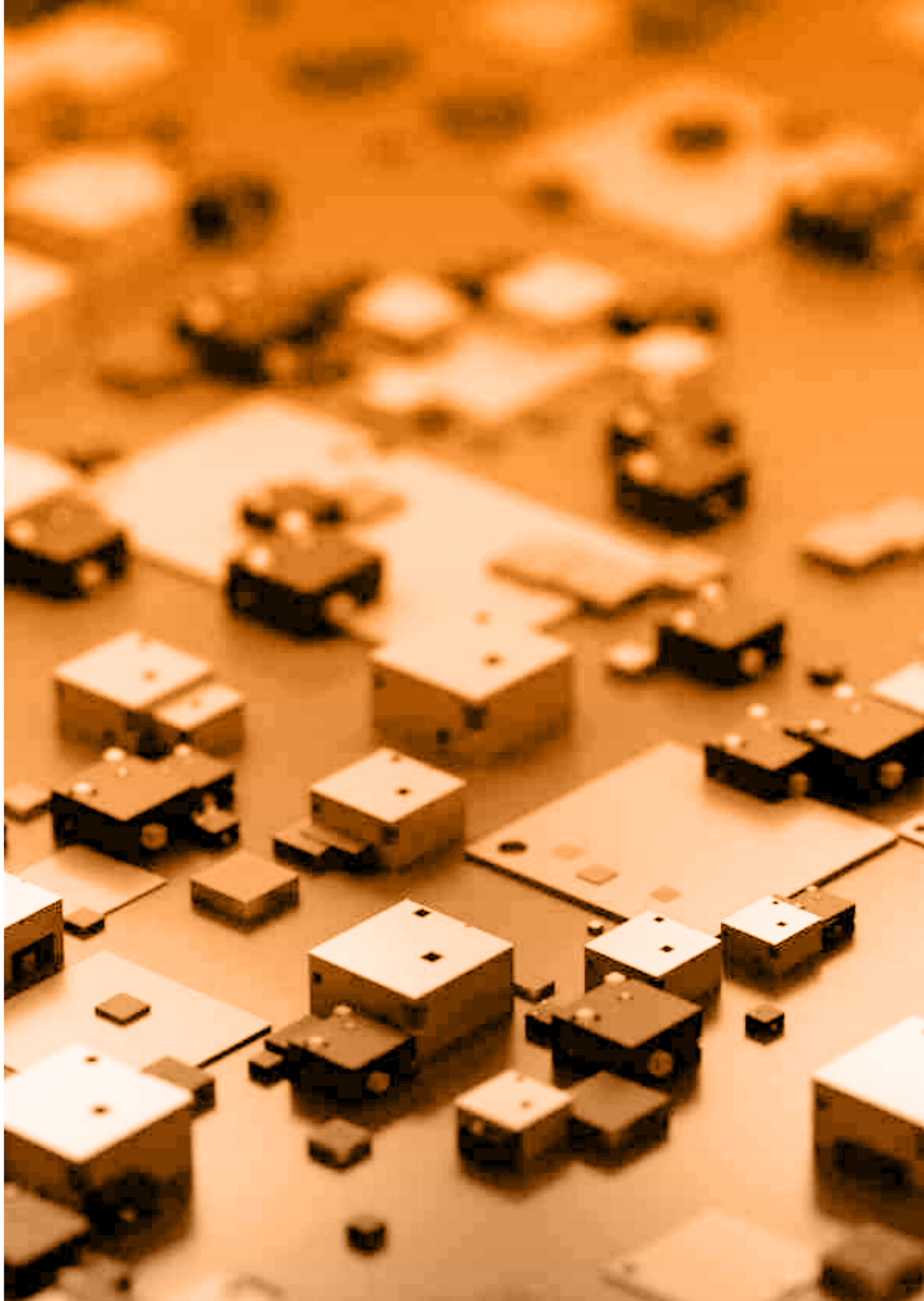
Она находится на стыке IT и социологии. Причинами её появления являются переход информационных войн в киберпространство и автоматизация систем для создания и доставки пропаганды. Дисциплину учредили в Национальных Академиях наук США. Боты, наёмные команды «троллей», новые формы пропаганды, манипуляция вирусными инфоповодами, дипфейки — малая часть инструментов, использующихся в современных информационных войнах. Задачей новой науки является исследование этих инструментов, предсказание их влияния на человеческое поведение и разработка методов по сопротивлению.

ИНИЦИАТОРАМИ ЯВЛЯЮТСЯ ОРГАНИЗАЦИИ, НО В ПРОЦЕССЕ К НИМ ПОДКЛЮЧАЮТСЯ ЛЮДИ

Успешная информационная кампания всегда начинается с государственной или частной организации, у которой есть ресурсы, специалисты и инструменты. В ходе кампании к нарративу начинают подключаться люди, которые могут не осознавать искусственности инфоповодов. Одной из главных задач злоумышленников является не только изменить взгляды общества, но и выдать изменение за естественное. Поэтому многие системы для борьбы с информационными атаками направлены на их обнаружение и сбор доказательств искусственности повествования.

ИИ ДЛЯ БОРЬБЫ С ИНФОРМАЦИОННЫМИ АТАКАМИ КОММЕРЦИАЛИЗИРОВАН

На рынке существует 2 сегмента: для частных компаний и для специальных нужд государств. Первый сегмент используется абсолютным большинством крупных компаний для работы с атаками на их бренды. Решения из первой категории нередко адаптируются для работы с государствами. В некоторых случаях, частные компании создают решения, предназначенные для государственных нужд, с нуля.



ВЫВОДЫ

Системы для работы с информационными атаками всё ещё нуждаются в экспертах и операторах, благодаря которым достигается оптимальная глубина анализа. Существуют целые платформы, которые способны легко определять разные паттерны атак и снабжать пользователей качественной, релевантной информацией. При этом, часть таких разработок — отечественные.

В РОССИИ ЕСТЬ КАЧЕСТВЕННЫЕ КОММЕРЧЕСКИЕ РЕШЕНИЯ

Эти разработки в большинстве своём заточены на мониторинг СМИ и соцсетей, анализ репутации и распространения нарративов. Широко используется ИИ для обработки естественных языков и построения графов. Решения конкурентоспособны, могут ограниченно понимать смысл и контекст сообщений, но в большинстве своём работают именно с русским языком, в некоторых случаях с языками других стран СНГ. Часть нарративов уже попадает в российское инфополе через англоязычные каналы и из глобальных источников, поэтому решения могут не покрывать все цепочки распространения контента. Существуют системы, которые ограниченно работают в мультимодальном режиме.

СИСТЕМЫ ДЛЯ АНАЛИЗА ИНФОРМАЦИОННЫХ АТАК — ТЕХНОЛОГИИ ДВОЙНОГО НАЗНАЧЕНИЯ, КОТОРЫЕ МОЖНО ИСПОЛЬЗОВАТЬ ДЛЯ ПЛАНИРОВАНИЯ ИНФОРМАЦИОННЫХ АТАК

Большинство таких систем либо сами по себе включают функционал по симуляции информационных атак, либо способны оценивать эффект от информационной атаки при подключении к симуляциям. Это позволяет заранее оценить план атаки и адаптировать его. Кроме того, системы для анализа могут использоваться зачинщиками для анализа собственной атаки в её процессе для того, чтобы на ходу перестраивать подход на тактическом и стратегическом уровне. Например, злоумышленник может проверять, насколько хорошо его боты скрываются, и понять, почему система их находит.

ЧАСТЬ УЧАСТНИКОВ ИНФОРМАЦИОННЫХ КАМПАНИЙ — ОБЫВАТЕЛИ, КОТОРЫЕ НЕ ОСОЗНАЮТ СВОЕЙ РОЛИ В ПЛАНАХ ЗЛОУМЫШЛЕННИКОВ

При успешной информационной атаке часть сообщества изначально согласна с её вектором либо меняет своё мнение по тем или иным вопросам. Важно различать эти группы как друг от друга, так и от аккаунтов, напрямую выполняющих команды злоумышленников. В зависимости от того, способны ли право-охранительные органы различать этих акторов, купирование информационных атак может сильно варьироваться в успехе. Если все группы людей при остановке атаки оцениваются одинаково, это добавляет атаке легитимности и снижает эффективность действий по отражению атаки.

МНОГИЕ ЗАРУБЕЖНЫЕ СИСТЕМЫ РАЗРАБАТЫВАЮТСЯ ДЛЯ ОТРАЖЕНИЯ ВОЗМОЖНЫХ РОССИЙСКИХ ИНФОРМАЦИОННЫХ АТАК

Внимание как учёных, так и частных компаний привлекает работа российских специалистов и их китайских коллег. Большое количество датасетов включает в себя сообщения от аккаунтов людей, подозреваемых в работе на отечественные и китайские организации. Некоторые методы обнаружения атак полагаются на поиск особенностей употребления конкретных слов в русско-говорящем сегменте Интернета или на поиск использования семантики английского языка русскоговорящими людьми.

СИСТЕМЫ ЗАВИСЯТ ОТ ЭКСПЕРТОВ ИЗ-ЗА ДИНАМИЧНОСТИ ЯЗЫКА И ИНФОПОВОДОВ

Глобальная медиасреда крайне динамична, инфоповоды в ней появляются каждый день, а контексты и смыслы меняются с высокой скоростью. Новые версии мемов и отдельных слов возникают каждый день, а сами языки изменяются живыми людьми. Полностью автоматизировать систему для анализа информационных атак на данный момент невозможно. Многие системы включают в себя функционал для оперативной отладки и вмешательства со стороны экспертов, а некоторые изначально проектируются с расчётом на постоянный контроль со стороны человека.

КРУПНЕЙШИЕ ПЛАТФОРМЫ УЖЕ НАКОПИЛИ БОЛЬШОЙ ОБЪЁМ СИГНАТУР ИНФОРМАЦИОННЫХ АТАК И ПРОДОЛЖАЮТ ПОПОЛНЯТЬ СВОИ АРХИВЫ

Решения с большим количеством клиентов годами накапливают опыт в распознавании информационных атак и библиотеки паттернов. Поскольку на рынке существуют компании, предоставляющие свои услуги государственным организациям, можно предположить, что эти системы уже используются другими странами.

11 ГЕНЕРАЦИЯ КОНТЕНТА

ГЕНЕРАТИВНЫЙ ИИ — МОДЕЛИ НА ОСНОВЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА, ОТВЕЧАЮЩИЕ ЗА СОЗДАНИЕ НОВОГО, ОРИГИНАЛЬНОГО КОНТЕНТА. СЕГОДНЯ ГЕНЕРАТИВНЫЙ ИИ МОЖЕТ СОЗДАВАТЬ ОСМЫСЛЕННЫЙ ТЕКСТ, ИЗОБРАЖЕНИЯ, МУЗЫКУ, ПРОГРАММНЫЙ КОД И СТИХИ, БАЗИРУЯСЬ НА КОРОТКИХ ТЕКСТОВЫХ ЗАПРОСАХ ОТ ПОЛЬЗОВАТЕЛЕЙ. ГЕНЕРАТИВНЫЕ ИИ ЧАСТО ПРИВОДЯТ К РЯДУ ЮРИДИЧЕСКИХ И ЭТИЧЕСКИХ ПРОБЛЕМ. ДИПФЕЙКИ, ИЛИ ИЗОБРАЖЕНИЯ И ВИДЕО, СОЗДАННЫЕ ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ И ПРЕТЕНДУЮЩИЕ НА РЕАЛИСТИЧНОСТЬ, НО ТАКОВЫМИ НЕ ЯВЛЯЮЩИЕСЯ, УЖЕ ЯВЛЯЮТСЯ ЧАСТЬЮ ЛЕНТЫ СОЦСЕТЕЙ, ПОЯВЛЯЮТСЯ В СМИ. ТАКЖЕ ОСТРОЙ ЯВЛЯЕТСЯ ПРОБЛЕМА ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ, КАСАЮЩАЯСЯ ВОПРОСОВ КОНТЕНТА И АВТОРСКОГО ПРАВА.

СУБТЕХНОЛОГИИ

- Компьютерное зрение
- Обработка естественного языка
- Распознавание и синтез речи
- Системы прогнозирования и поддержки принятия решений

СТАДИЯ ВНЕДРЕНИЯ

- В — Внедрено
- К — Концепция
- П — Прототип

* Компания признана экстремистской организацией в России

КОНТЕКСТ

Генеративные алгоритмы дадут пользователям возможность наводнить интернет контентом, серьёзно сократив затраты труда и времени на производство артефактов. Также, генеративные алгоритмы затронут не только то, как создаётся контент, но и то, как он потребляется, оценивается обывателями и поставят перед государствами ряд новых задач.

СОЗДАВАЕМЫЙ ПОЛЬЗОВАТЕЛЯМИ КОНТЕНТ МОЖЕТ ВЫЙТИ ИЗ-ПОД КОНТРОЛЯ

Генеративные алгоритмы предоставят обывателям инструменты для производства контента во многих форматах. Такой контент не всегда будет высокого качества, но его объёмы будут беспрецедентны, и часть контента будет незаконной. Несмотря на то, что разработчики борются с некоторыми видами нежелательного использования собственных алгоритмов, их защита регулярно обходится злоумышленниками с достаточной технической экспертизой. Генеративные алгоритмы использовались в информационных атаках и производстве нелегального контента. По мнению эксперта Нины Шик, на чью книгу Интерпол ссылается в своём отчёте, к 2026 г. около 90% контента в Интернете будет сгенерировано либо полностью, либо при значительном участии ИИ.

ГЕНЕРАТИВНЫЕ АЛГОРИТМЫ ЗНАЧИТЕЛЬНО ИЗМЕНЯТ СОЦИАЛЬНЫЕ ВЗАИМОДЕЙСТВИЯ

При некорректном использовании, отсутствии контрмер и полном отсутствии контроля над генеративными алгоритмами, они не только поставят общество в ситуацию, где объём производимого контента многократно превосходит возможности модерации, но и поменяют само общение людей, а также то, как люди воспринимают контент. Появится избыток контента, сложно будет определить, что сделано реальным человеком, а что роботом, не понимающим контекст. Производители контента начнут конструировать контекст вокруг продуктов деятельности нейросетей, а некоторые уязвимые люди сформируют парасоциальные отношения с ИИ.

ГЕНЕРАТИВНЫЕ АЛГОРИТМЫ — КЛЮЧ К ГИПЕРКАСТОМИЗАЦИИ КОНТЕНТА

Генерация контента нейросетями по заданным параметрам способствует ранее недостижимому уровню персонализации. Уже сегодня существует большое количество решений, лежащих в открытом доступе, при помощи которых можно создать текст, изображение или видео, основываясь только на предпочтениях пользователя. В будущем с повышением качества такого контента ожидаемо глобальное изменение экономического ландшафта сферы услуг. Такой контент будет потребляться пользователями ежедневно и соответствовать их требованиям вне зависимости от этичности запросов пользователя.

ВЫСОКОЕ КАЧЕСТВО КОНТЕНТА, СОЗДАННОГО ИИ

Многие системы генерации контента сегодня осуществляют полный цикл создания текста, изображения или видео без вмешательства в процесс человека. С развитием технологий и повышением их доступности, растёт и качество контента, создаваемого нейронными сетями.

Наибольшую сложность для обывателя составляют тексты, написанные при помощи ИИ. Изображения и видео на данный момент ещё содержат часть артефактов, видных технически подкованному человеку, однако количество таких признаков стремительно падает.

ВОЗМОЖНОСТИ ГЕНЕРАТИВНЫХ АЛГОРИТМОВ НЕ ОГРАНИЧИВАЮТСЯ МЕДИАКОНТЕНТОМ

Генеративные алгоритмы имеют крайне высокий подрывной потенциал во всех сферах жизнедеятельности человека. ИИ уже сейчас способен находить новые химические формулы для сплавов, лекарств, наркотиков и ядов, проектировать здания, а Управление перспективных исследовательских проектов Министерства обороны США работает над ИИ, который сможет находить слабые места в боевой технике по чертежам — следующим шагом будет создание ИИ, способного предлагать альтернативный подход к структуре, например танка, боеголовки или крейсера. Сложно предположить, какая область и индустрия не будут затронуты развитием генеративных алгоритмов.

ВЫВОДЫ

От государств, платформ и разработчиков потребуются скоординированные усилия, чтобы технология не вышла из-под контроля. Алгоритмы будут использоваться злоумышленниками в своих целях, а стейкхолдеры получат в свои руки новую технологию двойного назначения, использование которой в информационных войнах и разработках военно-промышленного комплекса способно дестабилизировать геополитическое положение сил.

НЕ ВО ВСЕХ СТРАНАХ РЕГУЛЯТОРИКА ГОТОВА К ГЕНЕРАТИВНЫМ АЛГОРИТМАМ

Существует длинный список регуляторных моментов, по которым нет консенсуса: начиная с вопросов авторского права, заканчивая вопросом о том, кто должен нести ответственность за использование алгоритмов для генерации нелегального контента. Из-за того, что генеративные алгоритмы всё ещё бурно развиваются, а учёные находят им новые применения, регуляторике большинства стран придётся адаптироваться в ускоренном режиме.

РАЗРАБОТЧИКИ ПЫТАЮТСЯ ОГРАНИЧИТЬ СОБСТВЕННЫЕ РЕШЕНИЯ, НО НЕ ВСЕГДА УСПЕШНО

Разработчики способны предвидеть проблемы, которые возникнут из-за неконтролируемого использования их решений, и потому ограничивают генерацию некоторого контента. В конкретных случаях системы напрямую отказывают в выполнении определённых запросов, список которых зависит от политических и идеологических взглядов разработчика. Тем не менее большинство таких ограничений не встроены в генеративные алгоритмы достаточно глубоко, чтобы остановить малочисленную категорию злоумышленников, способных модифицировать системы.

ГЕНЕРАТИВНЫЕ АЛГОРИТМЫ — ТЕХНОЛОГИЯ ДВОЙНОГО НАЗНАЧЕНИЯ

У генеративных алгоритмов нет моральных установок — они не осознают ни смысл того, что они генерируют, ни возможные эффекты продуктов их работы. Так, в статье «Dual use of artificial-intelligence-powered drug discovery», опубликованной в Nature, учёные описывают, что алгоритм для генерации новых лекарств смог спроектировать 40 тыс. боевых химических веществ за 6 ч. работы на достаточно скромном оборудовании. Этот принцип работает на все генеративные алгоритмы: технологический базис для дипфейков был заложен редактированием видео, а создание нелегального опасного для пользователей контента будет осуществляться ИИ, который изначально для этого не предназначался.

ИИ СПОСОБЕН БЫТЬ ОТЛИЧНЫМ ИНСТРУМЕНТОМ И АССИСТЕНТОМ, НО СОЗДАТЕЛЬ ИЗ НЕГО ПОСРЕДСТВЕННЫЙ

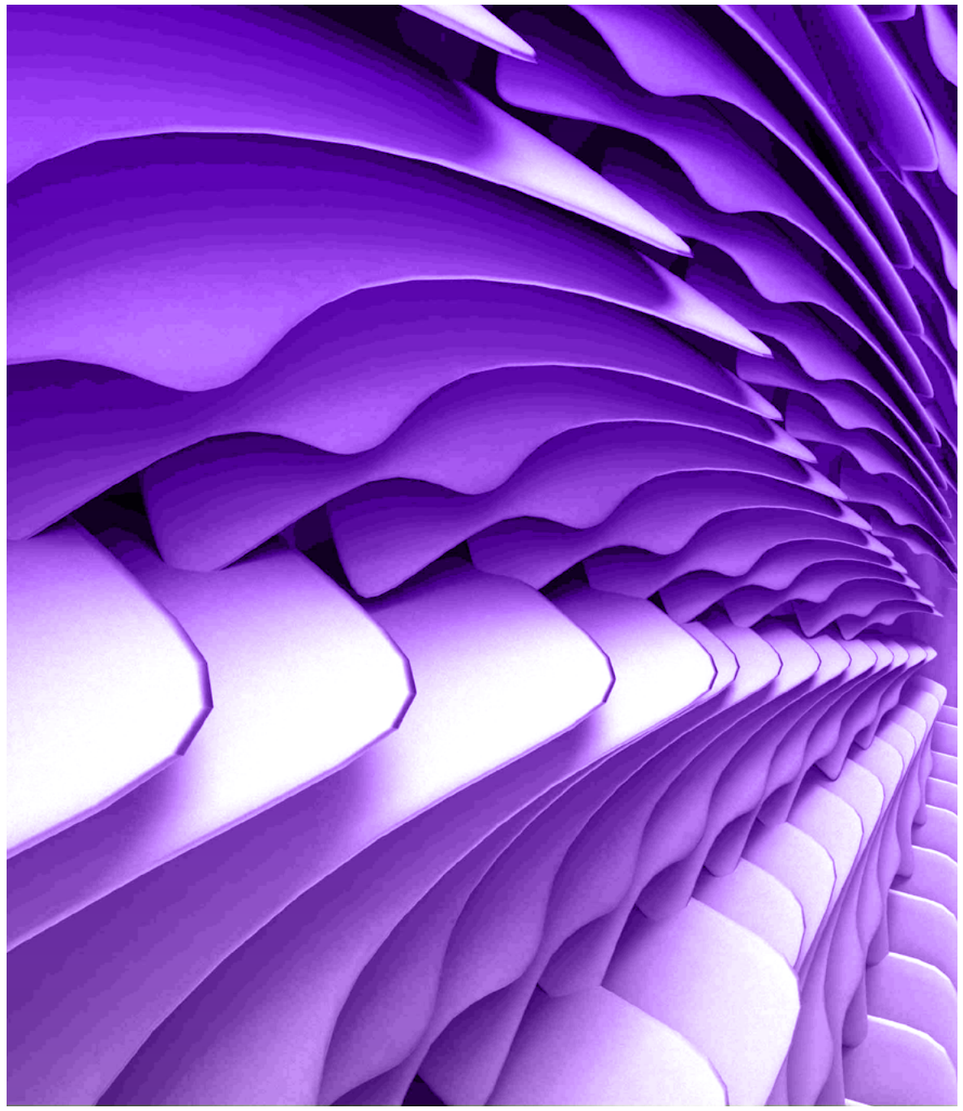
ИИ не способен глубоко понимать культурный контекст и нуждается в чётких инструкциях. Инструменты для редактирования контента с использованием ИИ наиболее выгодно использовать, когда задача проста и линейна, но требует больших усилий и времени. В таких случаях данные инструменты проявляют себя лучше всего. ИИ для генерации цельных артефактов контента не способен придумывать новые концепции, отталкиваясь от контекста и смысла. Например, ИИ не может разработать визуальный язык иллюстрации или отойти от распространённых жанров музыки. Тем не менее ИИ справляется с производством вариаций изображений и треков, что делает его ценным при прототипировании.

АВТОМАТИЧЕСКАЯ ГЕНЕРАЦИЯ НОВОСТЕЙ КОММЕРЦИАЛИЗИРУЕТСЯ УЖЕ СЕЙЧАС, НО НЕ РЕГУЛИРУЕТСЯ ОТДЕЛЬНО

В плане натуралистичности сгенерированного контента на данный момент ИИ лучше всего удаётся работать с текстом. Особенно это касается коротких заметок, которые можно адаптировать в простые формулы. Уже сейчас крупные сервисы генерируют новости вместо того, чтобы нанимать журналистов для этой работы. Развитые модели могут работать автономно и сканировать тренды соцсетей. Возникает 2 риска: во-первых, такие модели могут иметь предвзятость в некоторых вопросах, а во-вторых, так как любой тренд может быть использован злоумышленниками для доставки сообщений прямиком в СМИ, моделям, ориентирующимися на тренды, обязательно нужен фильтр.

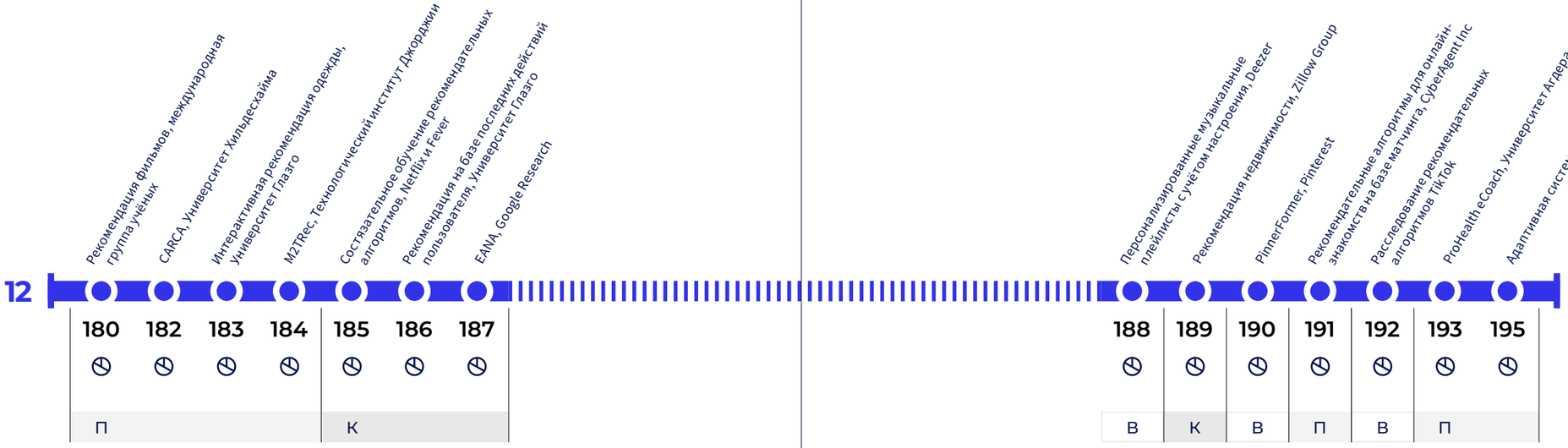
ГЕНЕРАТИВНЫЕ АЛГОРИТМЫ ПОМОГАЮТ ВО ВСЕХ СТАДИЯХ ПРОИЗВОДСТВА КОНТЕНТА

Генеративный ИИ может использоваться не только для редактирования или прототипирования контента, но и для его архивации, а также для улучшения качества устаревающих артефактов контента. Существует большое количество решений, которые направлены на адаптацию контента в другой формат или создание нового контента в ином формате на базе уже существующего контента. Количество возможных точек применения ИИ в креативной индустрии крайне велико и будет только расти.



12 РЕКОМЕНДАЦИЯ КОНТЕНТА

РЕКОМЕНДАТЕЛЬНЫЕ СИСТЕМЫ ИМЕЮТ КЛЮЧЕВОЕ ЗНАЧЕНИЕ В ЦИФРОВОЙ ТРАНСФОРМАЦИИ КАК ЭЛЕМЕНТ, В КОТОРОМ ЗАИНТЕРЕСОВАНЫ КАК ПОЛЬЗОВАТЕЛИ, ТАК И КОМПАНИИ: ДЛЯ ПЕРВЫХ РС ВЫСТУПАЮТ В КАЧЕСТВЕ НЕЗАМЕНИМОГО ПРОВОДНИКА В ОГРОМНОМ ВЫБОРЕ ТОВАРОВ И УСЛУГ, ДЛЯ ВТОРЫХ — ЭФФЕКТИВНОГО ИНСТРУМЕНТА УВЕЛИЧЕНИЯ ПРИБЫЛИ. БУДУЩЕЕ РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМ ОБУСЛОВЛЕНО ДВУМЯ ДВИЖУЩИМИ СИЛАМИ: ТЕХНОЛОГИЧЕСКИМ ПРОГРЕССОМ И ВЫСОКИМ СПРОСОМ НА РЫНКЕ. РАЗРАБОТЧИКИ ВСЁ ЧАЩЕ ПРИМЕНЯЮТ НЕЙРОННЫЕ СЕТИ КАК В КАЧЕСТВЕ ИНСТРУМЕНТА ДЛЯ ОБУЧЕНИЯ, ТАК И В КАЧЕСТВЕ ЯДРА РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМ.



СУБТЕХНОЛОГИИ

- Компьютерное зрение
- Распознавание и синтез речи
- Обработка естественного языка
- Системы прогнозирования и поддержки принятия решений

СТАДИЯ ВНЕДРЕНИЯ

- В — Внедрено
- К — Концепция
- П — Прототип

КОНТЕКСТ

Без рекомендательных систем интернет было бы почти невозможно использовать: объём информации слишком велик для человека и требует тщательной фильтрации. В противном случае найти нужную информацию крайне проблематично. Поэтому каждая платформа имеет свой собственный алгоритм и собирает о пользователях данные, без которых оперативно фильтровать контент под нужды каждого пользователя невозможно.

РЕКОМЕНДАТЕЛЬНЫЕ СИСТЕМЫ — В ПЕРВУЮ ОЧЕРЕДЬ СИСТЕМЫ ДЛЯ ФИЛЬТРАЦИИ КОНТЕНТА

В глобализированном мире с огромными объёмами данных, производимых и распространяемых в Интернете, рекомендательные системы служат в первую очередь для того, чтобы человек мог точнее осуществлять поиск нужной ему информации. Контента слишком много, и простой поиск с применением ключевых слов не обязательно приведёт пользователя к желаемому результату. Поэтому главной функцией рекомендательных систем является отсечение нерелевантных результатов, которые пользователю не нужны.

ТРИ ГЛАВНЫХ ВИДА РЕКОМЕНДАТЕЛЬНЫХ АЛГОРИТМОВ

К наиболее распространённым видам алгоритмов для рекомендации относятся методы коллаборативной фильтрации, основанные на контенте алгоритмы и гибридные алгоритмы. Ранее также использовались неперсонализированные алгоритмы и системы, базирующиеся на матричном разложении, но со временем они утратили свою актуальность и в какой-то мере были вытеснены контентно-ориентированными алгоритмами. Контентно-ориентированные алгоритмы стали особенно популярны в связи с тем, что они широко используются для рекомендаций в соцсетях.

РЕКОМЕНДАТЕЛЬНЫЕ АЛГОРИТМЫ НЕРАЗРЫВНО СВЯЗАНЫ СО СБОРОМ ДАННЫХ О ПОЛЬЗОВАТЕЛЯХ

Так как рекомендательные алгоритмы пытаются предсказать, чего захочет пользователь, исходя из имеющихся у них данных, их можно отнести к предиктивной аналитике. Для предсказания желаний пользователя алгоритмам требуется информация о пользователе и о контенте. Чем детальнее информация, тем точнее будет работать алгоритм. Возникает проблема с приватностью: так как алгоритмы используются в рекламе, корпорациям нужно собирать как можно больше информации о людях. При этом человек не может ориентироваться в разнообразии товаров и контента самостоятельно, поэтому он вынужден отдавать свои данные.

РЕКОМЕНДАТЕЛЬНЫЕ СИСТЕМЫ ЖДУТ ОБЫВАТЕЛЕЙ НА КАЖДОМ ШАГУ

В последние пять лет рекомендательные системы удерживают лидирующие позиции среди всех инструментов, базирующихся на ИИ. На 1 декабря 2022 г. существует больше 360 тыс. активных патентов, связанных с рекомендательными системами. Они являются незаменимым элементом для соцсетей, ритейла, онлайн-банкинга, поисковых сервисов и т. д. Будучи одними из наименее требовательных к вычислительным мощностям инструментами на базе ИИ, они приносят ощутимые результаты бизнесу в короткие сроки: повышение лояльности, привлечение новых клиентов и сокращение их оттока важны для выживания корпораций.

КОМПАНИИ И ПОЛЬЗОВАТЕЛИ ЗАВИСЯТ ОТ РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМ

Вкусы людей сильно разнятся между собой, что обусловлено как социально-культурными аспектами, так и субъективными предпочтениями. Рекомендательные системы позволяют пользователю сориентироваться, не теряясь в огромном выборе товаров и контента. Чем точнее работа рекомендательной системы, тем больше удовлетворён пользователь сервисом, и тем выше прибыль компании. Экономия пользовательского времени напрямую влияет на бизнес-метрики. Наличие рекомендательной системы является важным конкурентным преимуществом для компаний, а её влияние на прибыль растёт вместе с разнообразием услуг и масштабом предприятия.

ВЫВОДЫ

У рекомендательных алгоритмов имеется ряд рисков, связанных с их ролью в экосистеме интернета. Реализация данных рисков может привести к значительному ущербу, который может сказаться как на самих пользователях в случае утечки их данных, так и на обществе в целом, если рекомендательный алгоритм будет взломан и использован в своих целях злоумышленниками.

РАЗРАБОТЧИКИ ПЫТАЮТСЯ РЕШИТЬ ПРОБЛЕМУ «ХОЛОДНОГО СТАРТА» И ПОВЫСИТЬ ПРИВАТНОСТЬ ПОЛЬЗОВАТЕЛЕЙ

Проблема «холодного старта» заключается в том, что рекомендательные системы наиболее распространённых типов испытывают серьёзные сложности с формированием рекомендаций для новых пользователей и объектов. Это связано с недостаточным объёмом данных для предсказания реакции пользователя на тот или иной контент. Для решения этих проблем разработчики предлагают новые архитектуры, способные работать с минимальным объёмом данных, который пользователь оставляет в системе во время первых взаимодействий, а также предварительную

разметку объектов метаданными. При решении проблемы «холодного старта» рекомендательные системы перестанут нуждаться в слежке за пользователями.

СЛЕДУЮЩИМ ШАГОМ ДЛЯ РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМ ЯВЛЯЕТСЯ ВЫДАЧА РЕКОМЕНДАЦИЙ НА БАЗЕ САМОГО СОДЕРЖАНИЯ

На данный момент большинство систем работает за счёт того, что у каждого объекта есть несколько параметров, тегов и категорий, исходя из которых система рекомендует объект пользователю. Такой подход не идеален и не обладает абсолютной точностью. Некоторые новые решения предполагают автоматический анализ контента и товаров с помощью ИИ и последующую тренировку нейросетей, направленных на составление рекомендаций исходя из крайне подробного анализа каждого объекта. Такие системы нередко являются мультимодальными, особенно когда это касается мультимедиа.

РЕКОМЕНДАТЕЛЬНЫЕ АЛГОРИТМЫ МОГУТ ИСПОЛЬЗОВАТЬСЯ ДЛЯ ПРОПАГАНДЫ

Из-за возможности точного профилирования пользователей рекомендательными системами с учётом множества параметров они могут служить как эффективный инструмент пропаганды. Пользователи делятся данными, обеспечивая настройку системы, подбирающей для них контент определённой идеологической направленности. Уже есть известные случаи подобного использования рекомендательных систем: Cambridge Analytica использовала механизмы Facebook* для профилирования пользователей по психологическому портрету. Собранные данные использовались для настройки таргетированной политической рекламы, что вызвало общественный резонанс.

РЕКОМЕНДАТЕЛЬНЫЕ СИСТЕМЫ НУЖДАЮТСЯ В ЭФФЕКТИВНЫХ МЕХАНИЗМАХ ЗАЩИТЫ

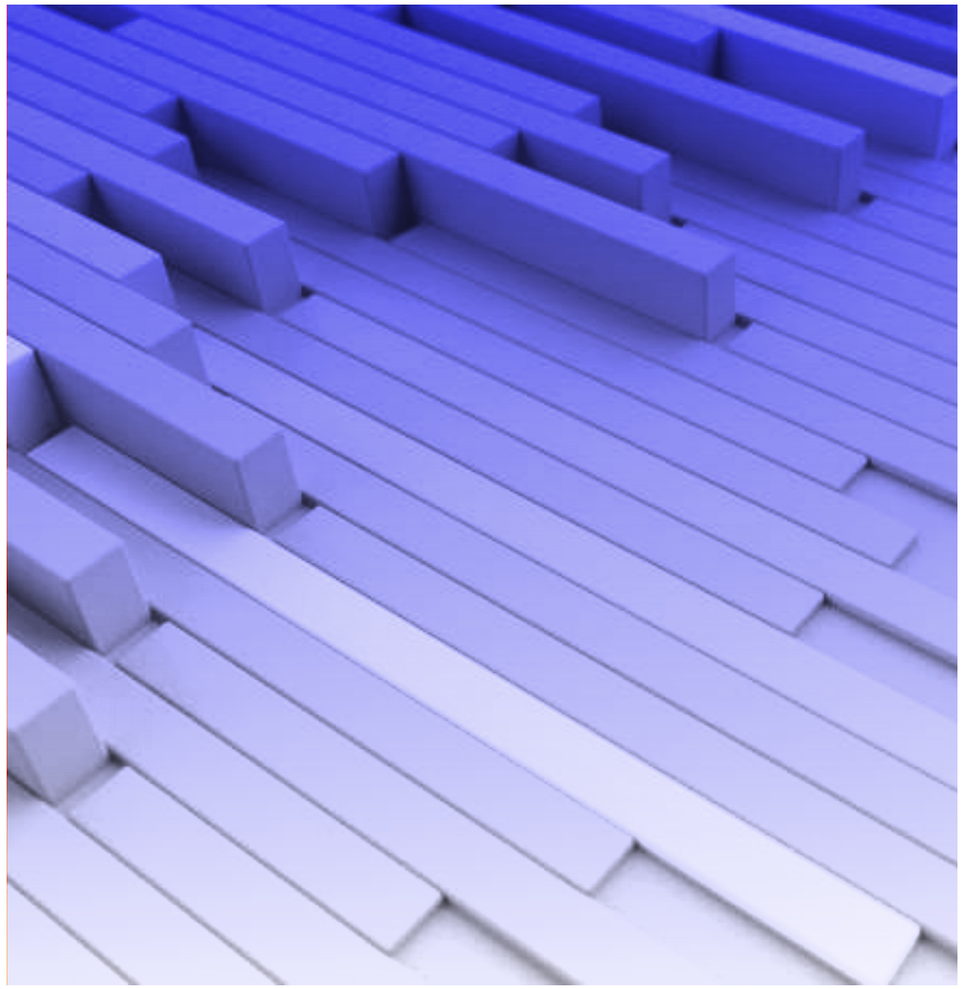
Каждая рекомендательная система обрабатывает внушительный объём личных данных: она анализирует, хранит и считывает большой спектр параметров. По этим параметрам можно узнать геолокацию, личный график, круг общения, вкусы и психологические особенности пользователя. Утечка даже части таких данных создаёт огромный список новых способов совершать преступления и увеличивает эффективность их совершения злоумышленниками. Данные могут использоваться для выявления и поиска перспективных жертв и подбора наиболее эффективной стратегии преступления. Так как рекомендательные системы не требуют больших мощностей, этот процесс можно автоматизировать при наличии данных.

МЕХАНИЗМЫ СИСТЕМ РЕКОМЕНДАЦИИ НЕПРОЗРАЧНЫ

В рекомендательных системах наиболее остро стоит проблема «чёрного ящика». Учитывая широкий спектр сфер, где они применяются, важно понятное обоснование их работы, во избежание катастрофических последствий, например при использовании в медицинской или промышленной области. У обывателей, исследователей и государств есть общее понимание того, что системы учитывают поведение пользователей, их отзывы, время и продолжительность активностей, географические данные, а также нередко используют неэтичные, сторонние источники данных. Но чёткие механизмы, по которым такие данные используются для формирования рекомендаций, остаются в тени.

ПРОТИВОРЕЧИЯ МЕЖДУ ТОЧНОЙ НАСТРОЙКОЙ И ЭТИЧЕСКИ-МОРАЛЬНЫМИ НОРМАМИ

Рекомендательные системы не имеют встроенного этического фильтра, их настройки задаются той или иной платформой. Как итог возникает ряд проблем: рекомендации должны быть адаптированы под каждого человека вне зависимости от его региона, но разные культуры, поколения и страны имеют разные нормы. Разработчики в первую очередь озабочены эффективностью рекомендательной системы, а не её пользой для каждого человека с учётом его культуры и страны проживания. Остаётся открытым вопрос о том, в каком ценностном векторе алгоритмы активно формируют предпочтения людей. Это особенно опасно для психологически уязвимых людей, которые могут попасть в информационный пузырь.



* Продукт Meta, организации, которая признана в России экстремистской

СПИСОК ЛИТЕРАТУРЫ

1. Adjabi, Insaf, et al. **“Past, present, and future of face recognition: A review.”** Electronics 9.8 (2020): 1188.

2. Adlersberg, Shabtai, Menachem Honig, and A. D. A. R. Tatiana. **“Device, system, and method for multimodal recording, processing, and moderation of meetings.”** U.S. Patent No. 11,501,780. 15 Nov. 2022.

3. Ahmed, Sarfraj, and Mohd Akbar Shaun. **“Impact of deepfake technology on digital world authenticity: A review.”** International Journal Of Engineering And Management Research 12.3 (2022): 78–84.

4. Akhalwaya, Ismail Yunus, et al. **“Word sense disambiguation using a deep logico-neural network.”** U.S. Patent Application No. 17/039,133.

5. Ali, Waqar, et al. **“Classical and modern face recognition approaches: a complete review.”** Multimedia tools and applications 80 (2021): 4825–4880.

6. Alkawaz, Mohammed Hazim, Maran Tamil Veeran, and Husniza Razalli. **“Video Forgery Detection based on Metadata Analysis and Double Compression.”** 2019 IEEE 7th Conference on Systems, Process and Control (ICSPC). IEEE, 2019.

7. Altuncu, Enes, Virginia NL Franqueira, and Shujun Li. **“Deepfake: Definitions, Performance Metrics and Standards, Datasets and Benchmarks, and a Meta-Review.”** arXiv preprint arXiv:2208.10913 (2022).

8. Aslan, Sinem, et al. **“Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms.”** Proceedings of the 2019 CHI conference on human factors in computing systems. 2019.

9. Babakov, Nikolay, et al. **“Detecting Inappropriate Messages on Sensitive Topics that Could Harm a Company’s Reputation.”** arXiv preprint arXiv:2103.05345 (2021).

10. Bang, Grace, et al. **“Controversy detection, impact assessment and impact prediction based on social media data.”** U.S. Patent No. 10,956,478. 23 Mar. 2021.

11. Beskow, David M., and Kathleen M. Carley. **“Bot-Match: Social Bot Detection with Recursive Nearest Neighbors Search.”** arXiv preprint arXiv:2007.07636 (2020).

12. Bhattacharya, Moumita, and Sudarshan Lamkhede. **“Augmenting Netflix Search with In-Session Adapted Recommendations.”** arXiv preprint arXiv:2206.02254 (2022).

13. Bojjireddy, Sirisha, Soon Ae Chun, and James Geller. **“Machine Learning Approach to Detect Fake News, Misinformation in COVID-19 Pandemic.”** DG. O2021: The 22nd Annual International Conference on Digital Government Research. 2021.

14. Bontempelli, Théo, et al. **“Flow Moods: Recommending Music by Moods on Deezer.”** Proceedings of the 16th ACM Conference on Recommender Systems. 2022.

15. Bouziane, Mostafa, et al. **“Team Buster. ai at CheckThat! 2020 Insights and Recommendations to Improve Fact-Checking.”** CLEF (Working Notes). 2020.

16. Bravo, Cesar Augusto Rodriguez, et al. **“Video privacy using machine learning.”** U.S. Patent No. 11,270,119. 8 Mar. 2022.

17. Kim, Taehyoung, Im Y. Jung, and Yih-Chun Hu. **“Automatic, location-privacy preserving dashcam video sharing using blockchain and deep learning.”** Human-centric Computing and Information Sciences 10.1 (2020): 1–23.”

18. Brij bhooshan gupta et al. **“Method and system for automatic fake news detection using feature selection on social networks.”** AU2021103137 (A4) — 2022-03-24.

19. Caldarini, Guendalina, Sardar Jaf, and Kenneth McGarry. **“A literature survey of recent advances in chatbots.”** Information 13.1 (2022): 41.

20. Carley, Kathleen M. **“Social cybersecurity: an emerging science.”** Computational and mathematical organization theory 26.4 (2020): 365–381.

21. Chatterjee, Ayan, et al. **“ProHealth eCoach: User-Centered Design and Development of an eCoach App to Promote Healthy Lifestyle with Personalized Activity Recommendations.”** (2022).

22. Chew, Peter A. **“Exposing Bot Activity with PARAFAC Tensor Decompositions.”** Conference of the Computational Social Science Society of the Americas. Springer, Cham, 2018.

23. Chew, Peter A., and Jessica G. Turnley. **“Understanding Russian information operations using unsupervised multilingual topic modeling.”** International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. Springer, Cham, 2017.

24. China Daily, Wang Yiqing, **“AI face-recognition technology helps reunite lost families”**, January 15, 2020.

25. Christiano, Paul F., et al. **“Deep reinforcement learning from human preferences.”** Advances in neural information processing systems 30 (2017).

26. Ciftci, Umur Aybars, Ilke Demir, and Lijun Yin. **“How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with biological signals.”** 2020 IEEE international joint conference on biometrics (IJCB). IEEE, 2020.

27. Cresci, Stefano. **“A decade of social bot detection.”** Communications of the ACM 63.10 (2020): 72–83.

28. Dahiya, Snehil, et al. **“Would your tweet invoke hate on the fly? forecasting hate intensity of reply threads on twitter.”** Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021.

29. Dam, Nguyen Anh Khoa, and Thang Le Dinh. **“A Literature Review of Recommender Systems for the Cultural Sector.”** ICEIS (1) (2020): 715–726.

30. Deldjoo, Yashar, et al. **“Audio-visual encoding of multimedia content for enhancing movie recommendations.”** Proceedings of the 12th ACM Conference on Recommender Systems. 2018.

31. Devlin, Jacob, et al. **“Bert: Pre-training of deep bidirectional transformers for language understanding.”** arXiv preprint arXiv:1810.04805 (2018).

32. Dewantara, Dimas Sony, and Indra Budi. **“Combination of LSTM and CNN for Article-Level Propaganda Detection in News Articles.”** 2020 Fifth International Conference on Informatics and Computing (ICIC). IEEE, 2020.

33. Dhariwal, Prafulla, et al. **“Jukebox: A generative model for music.”** arXiv preprint arXiv:2005.00341 (2020).

34. Dong, Zhenhua, et al. **“A Brief History of Recommender Systems.”** arXiv preprint arXiv:2209.01860 (2022).

35. Du, Xiaoyu, and Mark Scanlon. **“Methodology for the automated metadata-based classification of incriminating digital forensic artefacts.”** Proceedings of the 14th International Conference on Availability, Reliability and Security. 2019.

36. Du, Xiaoyu, Quan Le, and Mark Scanlon. **“Automated Artefact Relevancy Determination from Artefact Metadata and Associated Timeline Events.”** 2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security). IEEE, 2020.

37. Duarte, Jose Marcio, et al. **“Deep analysis of word sense disambiguation via semi-supervised learning and neural word representations.”** Information Sciences 570 (2021): 278–297.

38. Dulhanty, Chris, et al. **“Taking a stance on fake news: Towards automatic disinformation assessment via deep bidirectional transformer language models for stance detection.”**

39. Dzedzickis, Andrius, Artūras Kaklauskas, and Vytautas Bucinskas. **“Human emotion recognition: Review of sensors and methods.”** Sensors 20.3 (2020): 592.

40. Egger, Maria, Matthias Ley, and Sten Hanke. **“Emotion recognition from physiological signal analysis: A review.”** Electronic Notes in Theoretical Computer Science 343 (2019): 35–55.

41. Esser, Patrick, Robin Rombach, and Bjorn Ommer. **“Taming transformers for high-resolution image synthesis.”** Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

42. Fang, Yuanbo, et al. **“Bidirectional LSTM with Multiple Input Multiple Fusion Strategy for Speech Emotion Recognition.”** IAENG International Journal of Computer Science 48.3 (2021): 613–618.

43. Ferrucci, David A., et al. **“Fact checking using and aiding probabilistic question answering.”** U.S. Patent No. 8,972,321. 3 Mar. 2015.

44. Gad, Gad, et al. **“Deep Learning-Based Context-Aware Video Content Analysis on IoT Devices.”** Electronics 11.11 (2022): 1785.

45. Galuten, Albhy. **“Method for determining news veracity.”** U.S. Patent Application No. 15/901,740.

46. Gao, Zhongke, et al. **“A channel-fused dense convolutional network for EEG-based emotion recognition.”** IEEE Transactions on Cognitive and Developmental Systems (2020).

47. Gaur, Eshan, Vikas Saxena, and Sandeep K. Singh. **“Video annotation tools: A Review.”** 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN). IEEE, 2018.

48. Ghozia, Ahmed, et al. **“Intelligence Is beyond Learning: A Context-Aware Artificial Intelligent System for Video Understanding.”** Computational Intelligence and Neuroscience 2020 (2020).

49. Ghulati, Dhruv. **“Content scoring.”** U.S. Patent Application No. 16/643,573.

50. Gillespie, Tarleton. **“Content moderation, AI, and the question of scale.”** Big Data & Society 7.2 (2020): 2053951720943234.

51. Google Research Blog, Jacob Devlin, Ming-Wei Chang, **“Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing”**, November 2, 2018.

52. Guderlei, Maike, and Matthias Aßenmacher. **“Evaluating unsupervised representation learning for detecting stances of fake news.”** Proceedings of the 28th International Conference on Computational Linguistics. 2020.

53. Gupta, Shraddha. **“A Literature Review on Recommendation Systems.”** Int. Res. J. Eng. Technol 7 (2020): 3600–3605.

54. Han, Hu, et al. **“Tattoo image search at scale: Joint detection and compact representation learning.”** IEEE transactions on pattern analysis and machine intelligence 41.10 (2019): 2333–2348.

55. Harrison, Zachary, and Anish Khazane. **“Taxonomic Recommendations of Real Estate Properties with Textual Attribute Information.”** Proceedings of the 16th ACM Conference on Recommender Systems. 2022.

56. Heidari, Maryam, H. James Jr, and Ozlem Uzuner. **“An empirical study of machine learning algorithms for social media bot detection.”** 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS). IEEE, 2021.

57. Hoelz, Bruno WP, Célia Ghedini Ralha, and Rajiv Geeverghese. **“Artificial intelligence applied to computer forensics.”** Proceedings of the 2009 ACM symposium on Applied Computing. 2009.

58. Hong, Wenyi, et al. **“CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers.”** arXiv preprint arXiv:2205.15868 (2022).

59. Hrckova, Andrea, et al. **“Automated, not Automatic: Needs and Practices in European Fact-checking Organizations as a basis for Designing Human-centered AI Systems.”** arXiv preprint arXiv:2211.12143 (2022).

60. Hu, Shu, Yuezun Li, and Siwei Lyu. **“Exposing GAN-generated faces using inconsistent corneal specular highlights.”** ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.

61. Ivaschenko, Anton, et al. **“Hybridization of intelligent solutions architecture for text understanding and text generation.”** Applied Sciences 11.11 (2021): 5179.

62. Ivaylo B. Bozhinov et al. **“Making Data Smarter with IBM Spectrum Discover: Practical AI Solutions”**, IBM (2020).

63. Jeong, Younghoon, et al. **“KOLD: Korean Offensive Language Dataset.”** arXiv preprint arXiv:2205.11315 (2022).

64. Jesse, Mathias, and Dietmar Jannach. **“Digital nudging with recommender systems: Survey and future directions.”** Computers in Human Behavior Reports 3 (2021): 100052.

65. Jiang, Wei, et al. **“Neuman: Neural human radiance field from a single video.”** European Conference on Computer Vision. Springer, Cham, 2022.

66. Juneja, Prerna, and Tanushree Mitra. **“Human and technological infrastructures of fact-checking.”** Proceedings of the ACM on Human-Computer Interaction 6.CSCW2 (2022): 1–36.

67. Kaliyar, Rohit Kumar, et al. **“MultiDeepFake: Improving Fake News Detection with a Deep Convolutional Neural Network Using a Multimodal Dataset.”** International Advanced Computing Conference. Springer, Singapore, 2020.

68. Keyrouz, Fakheredine, Lara Tauk, and Elias Feghali. **“Enhanced Chemical Structure Recognition and Prediction Using Bayesian Fusion.”** 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, 2019.

69. Khalil, Ruhul Amin, et al. **“Speech emotion recognition using deep learning techniques: A review.”** IEEE Access 7 (2019): 117327–117345.

70. Khelif, Khaled, et al. **“SIIP: An innovative speaker identification approach for law enforcement agencies.”** STO meeting proceedings paper, NATO-OTAN. 2018.

71. Kikuchi, Takuya, Tomohiro Fukuda, and Nobuyoshi Yabuki. **“Diminished reality using semantic segmentation and generative adversarial network for landscape assessment: evaluation of image inpainting according to colour vision.”** Journal of Computational Design and Engineering 9.5 (2022): 1633–1649.

72. Kim, J., Dong, Z., Guan, E., Rosenthal, J., Fu, S., Rafailovich, M., & Polak, P. (2022). Detection of (Hidden) Emotions from Videos using Muscles Movements and Face Manifold Embedding. arXiv preprint arXiv:2211.00233.

73. Ko, Byoung Chul. **“A brief review of facial emotion recognition based on visual information.”** sensors 18.2 (2018): 401.

74. Kodhai, E., et al. **“Literature Review on Emotion Recognition System.”** 2020 International Conference on System, Computation, Automation and Networking (ICSCAN). IEEE, 2020.

75. Kortli, Yassin, et al. **“Face recognition systems: A survey.”** Sensors 20.2 (2020): 342.

76. Kou, Yubo, and Xinning Gui. **“Mediating community-ai interaction through situated explanation: The case of ai-led moderation.”** Proceedings of the ACM on Human-Computer Interaction 4.CSCW2 (2020): 1–27.

77. Kuriakose, Ammu, et al. **“ALIKAH-A Clickbait and Fake News Detection System using Natural Language Processing.”** 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, 2019.

78. Li, Ruilong, et al. **“Ai choreographer: Music conditioned 3d dance generation with aist++.”** Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

79. Liang, Jingyun, et al. **“Swinir: Image restoration using swin transformer.”** Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

80. Lladós, Josep, et al. **“Symbol recognition: Current advances and perspectives.”** Graphics Recognition Algorithms and Applications: 4th International Workshop, GREC 2001 Kingston, Ontario, Canada, September 7–8, 2001 Selected Papers 4. Springer Berlin Heidelberg, 2002.

81. Llansó, Emma J. **“No amount of “AI” in content moderation will solve filtering’s prior-restraint problem.”** Big Data & Society 7.1 (2020): 2053951720920686.

82. Llansó, Emma, et al. **“Artificial intelligence, content moderation, and freedom of expression.”** (2020).

83. Ma, Qinmin. **“Abnormal Event Detection in Videos Based on Deep Neural Networks.”** Scientific Programming 2021 (2021).

84. Mahmud, Bahar Uddin, and Afsana Sharmin. **“Deep insights of deepfake technology: A review.”** arXiv preprint arXiv:2105.00192 (2021).

85. Mannarswamy, Sandya, and Saravanan Chidambaram. **“Opening the NLP Blackbox-Analysis and Evaluation of NLP Models: Methods, Challenges and Opportunities.”** Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD). 2021.

86. Martín-Gutiérrez, David, et al. **“A deep learning approach for robust detection of bots in twitter using transformers.”** IEEE Access 9 (2021): 54591–54601.

87. Martino, Giovanni Da San, et al. **“A survey on computational propaganda detection.”** arXiv preprint arXiv:2007.08024 (2020).

88. Maseri, Aimi Nadrah, et al. **“Socio-technical mitigation effort to combat cyber propaganda: A systematic literature mapping.”** IEEE Access 8 (2020): 92929–92944.

89. Memon, Jamshed, et al. **“Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR).”** IEEE Access 8 (2020): 142642–142668.

90. Meysam Alizadeh, Jacob Shapiro. **“System and Method for Machine Learning Based Prediction of Social Media Influence Operations”**, Patent Application WO 2021/041830 A1, 2020.

91. Mnasri, Maali. **“Recent advances in conversational NLP: Towards the standardization of Chatbot building.”** arXiv preprint arXiv:1903.09025 (2019).

92. Mohammed, Hussam, Nathan Clarke, and Fudong Li. **“An automated approach for digital forensic analysis of heterogeneous big data.”** Journal of Digital Forensics, Security and Law 11.2 (2016): 9.

93. Murali, Vijayaraghavan, et al. **“Neural sketch learning for conditional program generation.”** arXiv preprint arXiv:1703.05698 (2017).

94. Murdock, V., et al. **“Search and Exploration of X-rated Information: WSDM’16 workshop proceedings: February 22, 2016, San Francisco, USA.”** (2016).

95. Natural Language Computing Group, Microsoft Research Asia. **“R-NET: Machine reading comprehension with self-matching networks.”** (2017).

96. Nguyen, Thanh Tam, et al. User guidance for efficient fact checking. No. CONF. 2019.

97. Ning, Lin, et al. **“EANA: Reducing Privacy Risk on Large-scale Recommendation Models.”** Proceedings of the 16th ACM Conference on Recommender.

98. Onur, Eker, and Bal Murat. **“A Comparative Analysis of the Face Recognition Methods in Video Surveillance Scenarios.”** arXiv preprint arXiv:2211.02952 (2022).

99. Ouyang, Long, et al. **“Training language models to follow instructions with human feedback.”**

100. Pal, Ratnabali, et al. **“Topic-based video analysis: A survey.”** ACM Computing Surveys (CSUR) 54.6 (2021): 1–34.

101. Passos, Leandro A., et al. **“A review of deep learning-based approaches for deepfake content detection.”** arXiv preprint arXiv:2202.06095 (2022).

102. Paterson, Thomas, and Lauren Hanley. **“Political warfare in the digital age: cyber subversion, information operations and ‘deep fakes’.”** Australian Journal of International Affairs 74.4 (2020): 439–454.

103. Petrov, Aleksandr, and Craig Macdonald. **“Effective and Efficient Training for Sequential Recommendation using Recency Sampling.”** Proceedings of the 16th ACM Conference on Recommender Systems. 2022.

104. Phillips, P. Jonathon, et al. **“The FERET evaluation methodology for face-recognition algorithms.”** IEEE Transactions on pattern analysis and machine intelligence 22.10 (2000): 1090–1104.

105. Pinitjitsamut, Karn, Kanabadee Srisomboon, and Wilaiporn Lee. **“Logo Detection with Artificial Intelligent.”** 2021 9th International Electrical Engineering Congress (iEECON). IEEE, 2021.

106. Poria, Soujanya, et al. **“Emotion recognition in conversation: Research challenges, datasets, and recent advances.”** IEEE Access 7 (2019): 100943–100953.

107. Qian, Jing, et al. **“Learning to decipher hate symbols.”** arXiv preprint arXiv:1904.02418 (2019).

108. Radsch, Courtney. **“AI and Disinformation: State-Aligned Information Operations and the Distortion of the Public Sphere.”** OSCE Representative on Freedom of the Media, Organization for Security and Co-operation in Europe (2022).

109. Rana, Md Shohel, et al. **“Deepfake detection: A systematic literature review.”** IEEE Access (2022).

110. Rangaswamy, Shanta, et al. **“Metadata extraction and classification of YouTube videos using sentiment analysis.”** 2016 IEEE International Carnahan Conference on Security Technology (ICCST). IEEE, 2016.

111. Rashed, Ahmed, Shereen Elsayed, and Lars Schmidt-Thieme. **“CARCA: Context and Attribute-Aware Next-Item Recommendation via Cross-Attention.”** arXiv preprint arXiv:2204.06519 (2022).

112. Rodrigues, Joana, et al. **“Hands-on data publishing with researchers: Five experiments with metadata in multiple domains.”** Digital Libraries: Supporting Open Science: 15th Italian Research Conference on Digital Libraries, IRCDL 2019, Pisa, Italy, January 31–February 1, 2019, Proceedings 15. Springer International Publishing, 2019.

113. Saikh, Tanik, et al. **“A deep transfer learning approach for fake news detection.”** 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020.

114. Salazar, Armida Panganiban. **“AI Tools on Fake News Detection: An Overview and Comparative Study.”** (2020).

115. Saxena, Anvita, Ashish Khanna, and Deepak Gupta. **“Emotion recognition and detection methods: A comprehensive survey.”** Journal of Artificial Intelligence and Systems 2.1 (2020): 53–79.

116. Schroff, Florian, Dmitry Kalenichenko, and James Philbin. **“Facenet: A unified embedding for face recognition and clustering.”** Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

117. Schuster, Tal, et al. **“The Limitations of Stylometry for Detecting Machine-Generated Fake News.”** arXiv preprint arXiv:1908.09805 (2019).

118. Schuster, Tal, et al. **“Towards debiasing fact verification models.”** arXiv preprint arXiv:1908.05267 (2019).

119. Seibold, Clemens, et al. **“Detection of face morphing attacks by deep learning.”** International Workshop on Digital Watermarking. Springer, Cham, 2017.

120. Setty, Vinay J., and Rahul Mishra. **“Deep neural architectures for detecting false claims.”** U.S. Patent No. 10,803,387. 13 Oct. 2020.

121. Shalaby, Walid, et al. **“M2TRec: Metadata-aware Multi-task Transformer for Large-scale and Cold-start free Session-based Recommendations.”** Proceedings of the 16th ACM Conference on Recommender Systems. 2022.

122. Sharma, Vijeta, et al. **“Video processing using deep learning techniques: A systematic literature review.”** IEEE Access 9 (2021): 139489–139507.

123. Shivaswamy, Pannaga, and Dario Garcia-Garcia. **“Adversary or Friend? An adversarial Approach to Improving Recommender Systems.”** Proceedings of the 16th ACM Conference on Recommender Systems. 2022.

124. Shu, Lin, et al. **“A review of emotion recognition using physiological signals.”** Sensors 18.7 (2018): 2074.

125. Singer, Uriel, et al. **“Make-a-video: Text-to-video generation without text-video data.”** arXiv preprint arXiv:2209.14792 (2022).

126. Singh, Pradeep Kumar, et al. **“Recommender systems: An overview, research trends, and future directions.”** International Journal of Business and Systems Research 15.1 (2021): 14–52.

127. Školkay, Andrej, and Juraj Filin. **“A comparison of fake news detecting and fact-checking AI based solutions.”** Studia Medioznawcze 4 (2019): 365–383.

128. Slimani, Yahya. **“A Hybrid Approach for Fake News Detection in Twitter Based on User Features and Graph Embedding.”** Distributed Computing and Internet Technology: 16th International Conference, ICDCIT 2020, Bhubaneswar, India, J anuary 9–12, 2020, Proceedings. Vol. 11969. Springer Nature, 2020.

129. Smith, Steven T., et al. **“Automatic detection of influential actors in disinformation networks.”** Proceedings of the National Academy of Sciences 118.4 (2021): e2011216118.

130. Solomon, JK Merrin Mary, and Mahendra Singh Meena. **“Challenges in face recognition systems.”** IJRAR 6.2 (2019).

131. Song, Chenguang, Kai Shu, and Bin Wu. **“Temporally evolving graph neural network for fake news detection.”** Information Processing & Management 58.6 (2021): 102712.

132. Starostin, Anatoly, and Dmitrii Kuklin. **“Method and system of text synthesis based on extracted information in the form of an RDF graph making use of templates.”** U.S. Patent No. 10,210,249. 19 Feb. 2019.

133. Stokman, Henricus Meinardus Gerardus, Marc Jean Baptist Van Oldenborgh, and Fares Alnajar. **“Monitoring and analyzing body language with machine learning, using artificial intelligence systems for improving interaction between humans, and humans and robots.”** U.S. Patent No. 11,443,557. 13 Sep. 2022.

134. Su, Hang, Xiatian Zhu, and Shaogang Gong. **“Deep learning logo detection with data expansion by synthesising context.”** 2017 IEEE winter conference on applications of computer vision (WACV). IEEE, 2017.

135. Su, Yu-Sheng, Hung-Yue Suen, and Kuo-En Hung. **“Predicting behavioral competencies automatically from facial expressions in real-time video-recorded interviews.”** Journal of Real-Time Image Processing 18.4 (2021): 1011–1021.

136. Suhaimi, Nazmi Sofian, James Mountstephens, and Jason Teo. **“EEG-based emotion recognition: A state-of-the-art review of current trends and opportunities.”** Computational intelligence and neuroscience 2020 (2020).

137. Surís, Dídac, Ruoshi Liu, and Carl Vondrick. **“Learning the predictability of the future.”** Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

138. Tan, Fei, et al. **“System and method for text moderation via pretrained transformers.”** U.S. Patent No. 11,481,543. 25 Oct. 2022.

139. Tarnowski, Paweł, et al. **“Eye-tracking analysis for emotion recognition.”** Computational intelligence and neuroscience 2020 (2020).

140. Telley, Christopher. **“The influence machine: Automated information operations as a strategic defeat mechanism.”** (2018).

141. Thomas, Christopher, and Adriana Kovashka. **“Predicting the politics of an image using webly supervised data.”** Advances in Neural Information Processing Systems 32 (2019).

142. Tolba, A. S., A. H. El-Baz, and A. A. El-Harby. **“Face recognition: A literature review.”** International Journal of Signal Processing 2.2 (2006): 88–103.

143. Tombre, Karl, Salvatore Tabbone, and Philippe Dosch. **“Musings on symbol recognition.”** Graphics Recognition. Ten Years Review and Future Perspectives: 6th Internation Workshop, GREC 2005, Hong Kong, China, August 25–26, 2005, Revised Selected Papers 6. Springer Berlin Heidelberg, 2006.

144. Tomita, Yoji, Riku Togashi, and Daisuke Moriwaki. **“Matching Theory-based Recommender Systems in Online Dating.”** Proceedings of the 16th ACM Conference on Recommender Systems. 2022.

145. Toney, Autumn, et al. **“Automatically characterizing targeted information operations through biases present in discourse on twitter.”** 2021 IEEE 15th International Conference on Semantic Computing (ICSC). IEEE, 2021.

146. Towards Data Science, Rani Horev, **“BERT Explained: State of the art language model for NLP”**, November 10, 2018.

147. Tsay, Jason, et al. **“Aimmx: Artificial intelligence model metadata extractor.”** Proceedings of the 17th International Conference on Mining Software Repositories. 2020.

148. Tsay, Jason, et al. **“Aimmx: Artificial intelligence model metadata extractor.”** Proceedings of the 17th International Conference on Mining Software Repositories. 2020.

149. Tunell, John. **“Classification of offensive game-emblem drawings using CNN (convolutional neural networks) and transfer learning.”** (2018).

150. Tunell, John. **“Classification of offensive game-emblem drawings using CNN (convolutional neural networks) and transfer learning.”** (2018).

151. Ulrich, Hannes, et al. **“Understanding the nature of metadata: systematic review.”** Journal of Medical Internet Research 24.1 (2022): e25440.

152. Uyheng, J., Magelinski, T., Villa-Cox, R., Sowa, C., & Carley, K. M. (2019). Interoperable pipelines for social cyber-security: assessing Twitter i nformation operations during NATO Trident Juncture 2018. Computational and Mathematical Organization Theory. doi:10.1007/s10588-019-09298-1

153. Vadapalli, Raghuram, et al. **“When science journalism meets artificial intelligence: An interactive demonstration.”** Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2018.

154. Verburg, Michiel, and Vlado Menkovski. **“Micro-expression detection in long videos using optical flow and recurrent neural networks.”** 2019 14th IEEE International conference on automatic face & gesture recognition (FG 2019). IEEE, 2019.

155. Vijjali, Rutvik, et al. **“Two stage transformer model for COVID-19 fake news detection and fact checking.”** arXiv preprint arXiv:2011.13253 (2020).

156. Visvam Devadoss, Ambeth Kumar, Vijay Rajasekar Thirulokachander, and Ashok Kumar Visvam Devadoss. **“Efficient daily news platform generation using natural language processing.”** International journal of information technology 11.2 (2019): 295–311.

157. Wang, Mei, and Weihong Deng. **“Deep face recognition: A survey.”** Neurocomputing 429 (2021): 215–244.

158. Wang, Yanru. **“Cultural Symbol Recognition Algorithm Based on CTPN+ CRNN.”** International Conference on Human Centered Computing. Springer, Cham, 2020.

159. Westerlund, Mika. **“The emergence of deepfake technology: A review.”** Technology innovation management review 9.11 (2019).

160. Wu, Xiongwei, Doyen Sahoo, and Steven CH Hoi. **“Recent advances in deep learning for object detection.”** Neurocomputing 396 (2020): 39–64.

161. Wu, Yaxiong, Craig Macdonald, and Iadh Ounis. **“Multi-Modal Dialog State Tracking for Interactive Fashion Recommendation.”** Proceedings of the 16th ACM Conference on Recommender Systems. 2022.

162. Xie, Haozhe, et al. **“Pix2vox: Context-aware 3d reconstruction from single and multi-view images.”** Proceedings of the IEEE/CVF international conference on computer vision. 2019.

163. Xu, Canwen, et al. **“Leashing the Inner Demons: Self-Detoxification for Language Models.”** arXiv preprint arXiv:2203.03072 (2022).

164. Xu, Jiajing, Andrew Zhai, and Charles Rosenberg. **“Rethinking Personalized Ranking at Pinterest: An End-to-End Approach.”** Proceedings of the 16th ACM Conference on Recommender Systems. 2022.

165. Xu, Zhongyou, et al. **“Stylization-based architecture for fast deep exemplar colorization.”** Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

166. Yoosuf, Shehel, and Yin Yang. **“Fine-grained propaganda detection with fine-tuned BERT.”** Proceedings of the second workshop on natural language processing for internet freedom: censorship, disinformation, and propaganda. 2019.

167. Yu, Jiahui, et al. **“Free-form image inpainting with gated convolution.”** Proceedings of the IEEE/CVF international conference on computer vision. 2019.

168. Yun, Xiao-Long, et al. **“Instance GNN: a learning framework for joint symbol segmentation and recognition in online handwritten diagrams.”** IEEE Transactions on Multimedia 24 (2021): 2580–2594.

169. Zhang, Qingyang, et al. **“Edge video analytics for public safety: A review.”** Proceedings of the IEEE 107.8 (2019): 1675–1696.

170. Zhang, Yongfeng, et al. **“Towards conversational search and recommendation: System ask, user respond.”** Proceedings of the 27th acm international conference on information and knowledge management. 2018.

171. `Zhou, Andrew Jie, Jiyun Luo, and Lewis John McGibbney. **“Multimedia metadata-based forensics in human trafficking web data.”** Vanessa Murdock, Charles LA Clarke, Jaap (2016): 10.

КОЛЛЕКТИВ АВТОРОВ

РЫЖОВА ЕВГЕНИЯ ЮРЬЕВНА

Советник генерального директора по научно-техническому развитию, ФГУП «ГРЧЦ»

ГЛАЗКОВ БОРИС МИХАЙЛОВИЧ

Вице-президент по стратегическим инициативам, ПАО «Ростелеком»

СОКОЛЕНКО МИХАИЛ ВЯЧЕСЛАВОВИЧ

Начальник отдела перспективных проектов научно-технического центра, ФГУП «ГРЧЦ»

КОМАНДА АНАЛИТИКОВ TEQVISER

НИКИФОРОВ ЕВГЕНИЙ АЛЕКСАНДРОВИЧ

Руководитель проекта, ФГУП «ГРЧЦ»

КОРОСТАШОВ РОМАН НИКОЛАЕВИЧ

Руководитель научно-технического центра, ФГУП «ГРЧЦ»

ЮСУФОВ РУСЛАН ГЕННАДЬЕВИЧ

Управляющий партнер, MINDSMITH

КОМАНДА АНАЛИТИКОВ MINDSMITH

Ростелеком

TEQVISER

MINDSMITH