Протидія дезінформації: посилення цифрової стійкості Альянсу

Альянсу потрібна дієва стратегія широкого спектра дії для протидії загрозі дезінформації, яка зростає. Засоби штучного розуму (ШР) можуть допомогти виявляти і уповільнювати поширення фальсифікованого і шкідливого контенту, підтримуючи при цьому цінності плюралістичних і відкритих суспільств.

Дезінформація: нічого особливо нового

Фальсифікована інформація і оманливі наративи використовуються як інструмент конфлікту і державної діяльності з часів падіння міфічної Трої перед давніми греками, а можливо і раніше. У далекому минулому це були дерев'яні коні, брехливі свідки і фальшиві плани. Сьогодні ми маємо фейкові новини, фальшиві профілі соціальних мереж і сфабриковані наративи, створені задля введення в оману – інколи в рамках скоординованої когнітивної війни, кампаній.

Пов'язана стаття : Протидія когнітивній війні: інформованість і стійкістьСоціальні мережі і Інтернет призвели до дезінформаційної революції. Ми живемо у світі дешевих цифрових інструментів і засобів інформації, які мають дуже широкий доступ, масштаб і вплив. Занепокоєння викликає те, що ці легкодоступні інструменти можуть отримати не лише державні дійові особи, але й недержавні, приватні особи і будь-хто загалом.

Соціальні мережі і Інтернет призвели до дезінформаційної революції, яка торкнулась державних дійових осіб, недержавних, приватних осіб і будь-кого загалом. © Centre for Research and Evidence on Security Threats

Фальшиві повідомлення і підбурюючі **Пов'язана стаття** : Дезінформація на західних Балканах наративи в екстремальних випадках отримували великий розголос, зокрема на західних Балканах і в деяких країнах Альянсу. Проте вони несуть більш приховану небезпеку через шкоду, яку вони можуть завдавати довірі громадян до інститутів демократичного врядування і ресурсів публічної інформації і обговорення. Останнім часом спостерігається посилення політичної поляризації, історично низькі рівні довіри до інституцій врядування, і зростання кількості випадків заворушень і насильства, частково спричинених і фальсифікованою інформацією.

Пов'язана стаття : "Справа Лізи": Німеччина як мішень російської дезінформаціїКраїни - члени НАТО підтримують відкриті цивільні системи комунікації, деякі з дуже високими рівнями використання соціальних мереж і месенджерів. Плюралістичний характер їхніх суспільств, будучи перевагою і джерелом сили, водночас може створювати можливості для поширення наративів, які підбурюють і розділяють людей. У багатьох з цих країн регуляторні структури, покликані забезпечити стійкість і захист, усе ще не розвинені. Комбінація цих умов викликає особливе занепокоєння Альянсу щодо загрози з боку дезінформації.

В центрі уваги фальшиві факти

Глобальні компанії соціальних мереж взялись за зменшення фальшивої інформації на своїх платформах. Більшість з них наймають власних фахівців з перевірки фактів, які здійснюють моніторинг поширення фальшивої інформації. Інші покладаються на перевірку фактів третьою стороною. Або на інструменти модерації. Декілька популярних платформ, серед яких Фейсбук, Ютуб і Твіттер, надають своїм користувачам можливість інформувати про інших користувачів, які підозрюються в свідомому або несвідомому поширенні фальшивої інформації. І намагаючись ретроспективно усунути шкоду, завдану фальшивою інформацією, переважна більшість платформ соціальних мереж застосовує масове видалення контенту, ідентифікованого за допомогою цих методів як шкідливий або оманливий.

У кращому разі це занадто мало, занадто пізно. В гіршому, це веде до звинувачень у цензурі або видаленні інформації чи думок, які пізніше виявляються достовірними або такими, що заслуговують на

публічне обговорення.

Проблема обсягів

Лише Фейсбуком щомісяця активно користуються майже три мільярда користувачів, кожен з яких здатен розмістити онлайн щось, що підбурює. У Твіттера понад 350 мільйонів активних користувачів, серед яких видатні особистості, популярні лідери громадської думки і розумні та винахідливі маніпулятори.

Нинішній підхід до протидії дезінформації в основному полягає у перевірці фактів вручну, видаленні змісту і контролі за завданою шкодою. Хоча втручання людини може бути корисним у випадках, коли потрібне розуміння нюансів, або культурних особливостей, воно погано вписується у великі обсяги інформації, яка генерується щоденно. Навряд чи збільшення штату є реалістичним варіантом проактивного виявлення фальшивого або шкідливого змісту до того як він отримає можливість широкого поширення. Перевірка фактів людиною сама собою вразлива до помилок, хибного тлумачення і упередженості.

Мільярди активних користувачів соціальних мереж щомісяця здатні розмістити онлайн щось, що підбурює. Це величезний обсяг інформації для перевірки вручну. © The Globe and Mail

Що стає вірусним?

«Брехня летить на крилах, а правда шкутильгає за нею», - написав у XVII стріччі сатирик Джонатан Свіфт. Нещодавнє, проведене MIT, дослідження виявило, що у Твіттері фальшиві новини мають набагато більше шансів стати вірусними, а їх поширенням займаються постійні користувачі, а не автоматичні «боти». Люди також роблять «ретвіт» цих фальшивих новин з почуттям здивування і обурення. На противагу цьому, правдиві історії викликають почуття смутку, сподівання і довіри (і ними діляться набагато рідше).

Тут з’являється потенційна можливість: можливо замість фактів нам варто зосереджуватись на емоціях? І чи можуть цьому навчитись комп’ютери, а не люди?

Не перевіряйте факти, перевіряйте емоції

Аналіз почуттів на основі штучного розуму представляє собою абсолютно інший підхід до зменшення впливу дезінформації через навчання комп’ютерів виявляти мессиджі і пости, які містять елементи здивування, обурення та інших емоційних провокацій. Вони більш вірогідно зв’язані з фальшивою інформацією і мають розпалити пристрасті користувачів соціальних мереж.

Алгоритми обробки природних мов дозволяють виявляти лінгвістичні індикатори відповідних емоцій. Вони дозволяють повністю уникнути втручання людини в перевірку фактів, зменшуючи вірогідність упередженості, знижуючи вартість та збільшуючи швидкість обробки. Група студентів з Університету Джонса Гопкінса створила перспективний робочий прототип, а їхні колеги по команді з Технічного інституту Джорджії і Лондонського імперського коледжу розробили техніко-економічні оцінки і потенційні регуляторні підходи.

Не зупиняйте дезінформацію, уповільнюйте її

Проте що робити після того, як виявляється вірусний (і вірогідно фальшивий) сигнал чи пост? Аналогія з фінансовими ринками пропонує таке рішення – автоматичний «запобіжник», який тимчасово призупиняє, або уповільнює поширення зарядженого емоціями змісту.

Біржі цінних паперів уникають панічних розпродаж тимчасово призупиняючи торгівлю акціями, які впали нижче певного відсоткового порогу. На Нью-Йоркській фондовій біржі акції, які втратили в ціні більше семи відсотків, спочатку призупиняються на 15 хвилин. Ідея полягає в тому, щоб уповільнити процеси і надати можливість більш холодним головам взяти гору. Подальші зниження ціни можуть спровокувати додаткові призупинення торгів.

Ефект охолодження від уповільнення процесів може бути істотним. В сфері соціальних мессиджів, мессидж, який подвоюється кожні 15 хвилин, може гіпотетично досягти мільйона користувачів за п’ять годин, 16 мільйонів – за шість годин. Проте якщо його уповільнити до рівня подвоювання кожні 30 хвилин, він досягне лише однієї тисячі користувачів за п’ять годин, і чотирьох тисяч за шість. Невелика різниця у вірулентності призводить до величезної різниці у враженні.

Такий механізм міг би діяти не через запобігання поширенню, а через уповільнення взаємодії; наприклад, накладаючи певний період «охолодження» між коментарями або рекомендуючи користувачам зважити на можливі наслідки перед тим, як переслати мессидж. Він ґрунтується на головній ідеї книги Нобелівського лауреата Даніеля Канемана «Швидке і повільне мислення». Повільне мислення раціональне і уникає емоційності швидких реакцій на неочікувані і шокуючі новини чи події.

Це може зменшити занепокоєння щодо цензури або свавільного обмеження свободи обміну ідеями. Мессиджі і пости не відміняються чи видаляються. Вони залишаються доступними для перегляду і обговорення, лише у повільнішому темпі. Це пом’якшує проблему «хто вирішує, що дозволено говорити», і захищає важливі свободи висловлювання і публічного дискурсу. Такий підхід можна запровадити за допомогою стимулів або регулювання в різних шарах комунікаційної інфраструктури: самих компаній-джерел, посередницьких шлюзів (або платформ «посередників», на рівні транспортування мессиджу («комунікаційні «труби»), або навіть на рівні самого пристрою (смартфону чи планшета).

На що треба зважати Альянсу

Дезінформація – одна із цифрових загроз, що стоять перед Альянсом. Нещодавні інформаційні кампанії і кібератаки продемонстрували, що навіть технічно просунуті країни-члени повинні більше робити для підготовки до нинішніх і нових цифрових викликів. Треба робити більше для становлення дієвих механізмів і регуляторних структур, що забезпечують стійкість.

Кібератаки загрожують навіть найбільш технічно просунутим країнам - членам НАТО. Треба робити більше для становлення дієвих механізмів і регуляторних структур, що забезпечують стійкість.

А ці загрози, схоже, зростають щодня, залишаючи Альянсу дуже мало часу. Використовуючи наявні технології (такі, про які йшлося раніше) і застосовуючи їх в інноваційні способи можна зекономити як час, так і ресурси. Концепції мінімально інвазійного пом'якшення, такі як уповільнення – але не цілковитого видалення – потенційно шкідливих меседжів і постів у соціальних мережах, може бути найбільш перспективним першим кроком з протидії загрозі дезінформації. Тоді Альянс зможе присвятити більше часу на додатковий технічний розвиток і більш комплексні підходи до регулювання у майбутньому.

Історично, стійкість і сила відкритих і плюралістичних суспільств полягає в їхній здатності до інноваційної адаптації до нових викликів і обставин. Основоположний механізм цього складається з вільного потоку ідей і інформації, а також відкритого і публічного обговорення і вивчення варіантів, стратегій і планів. Будь-яке рішення з протидії дезінформації повинно захищати цей механізм, якщо ми прагнемо зберегти цю перевагу. Більше того, прийняття країнами-членами того чи іншого рішення буде залежати від сприйняття його суспільством загалом і навряд чи буде успішним, якщо певні групи суспільства відчуватимуть себе маргіналізованими або виключеними з публічного діалогу.

НАТО може взяти на себе підтримку підходів на основі цієї технології і принципів, залишаючи за країнами-членами рішення щодо власної національної стратегії цифрової безпеки. Це забезпечить уряди країн-членів гнучкістю, необхідною для запровадження механізмів, які на їхню думку відповідають і вписуються в місцеві рівні впливу соціальних мереж, очікування населення щодо свободи слова і реалії цивільної комунікаційної інфраструктури.