

Нейрофизиологические механизмы восприятия фейковой и дипфейковой информации: выявление факторов доверия и устойчивости с целью расширения потенциала человека

Дезинформация – ложная информация, созданная с целью ввести аудиторию в заблуждение, сформировать неверное впечатление о герое, явлении или событии и, соответственно, подтолкнуть к желаемым действиям или бездействию (Wu et al., 2019). Исследования показали, что дезинформация в разных форматах может вызывать сильное доверие со стороны аудитории и быть использована для манипуляции общественным мнением или мнением индивида (Lazer et al., 2018; Pennycook & Rand, 2020, 2021). В настоящее время дезинформация в разных ее проявлениях представляет особую опасность для политической и социальной стабильности общества, экономического развития и других важных сфер жизни (Wasserman & Madrid-Morales, 2019). С помощью передовых технологий качество фейковых материалов вышло на новый уровень - их реалистичность и эффективность воздействия значительно возросла, увеличились и скорость и масштаб распространения.

Особенно популярным типом распространения дезинформации на сегодняшний день становится дипфейк. Это технически усовершенствованная и более реалистичная вариация фейков, задействующая технологии машинного обучения (нейросети) с целью подмены лица или голоса героя в аудио- и видеозаписях, вплоть до создания полностью искусственных, но реалистичных материалов. Влияние дипфейков на современные реалии в обществе может быть разным: с одной стороны, они могут выступать интересным и развлекательным контентом, с другой – представлять серьезную угрозу для общества. Люди могут обращаться к дипфейковым материалам как к подлинному контенту и доверять информации, которую они отображают, что может привести к неприятным последствиям.

Известно, что фейковые материалы распространяются быстрее, чем реальные новости, и многие люди склонны им доверять. Предыдущие исследования показали, что определенные демографические факторы, такие как возраст (Dmochowski et al., 2012), уровень образования (Rainie et al., 2012) и политические взгляды (Van der Linden et al., 2020), а также психологические факторы, например, соответствие внутренним установкам (Bauer & Hohenberg, 2021), потребность в познании (Wang et al., 2022), склонность к конформизму (Khalil et al., 2023) и социальное влияние со стороны (Kurvers et al., 2021) могут способствовать доверию дезинформации. Люди склонны считать источники, например СМИ, которые соответствуют им в идеологическом плане, более достоверными и заслуживающими доверия (Tsafati & Ariely, 2014). Ввиду реализма и масштаба распространения фейков люди вынуждены искать эффективные стратегии противостояния ложной информации, например факт-чекинг или внедрение дисклеймеров для предупреждения пользователя.

Данный доклад состоит из освещения двух исследовательских проектов. Первый посвящён изучению восприятия аудиальных дипфейк-материалов на медицинскую тематику, второй, в свою очередь, представлен в пилотном варианте и оценивает эффективность дисклеймеров и авторитетность источников в вопросе доверия фейкам и подлинным материалам. Ученые задействовали технологию электроэнцефалографии (ЭЭГ) с целью поиска нейрофизиологических коррелятов восприятия фейкового контента. Также для оценки индивидуальных характеристик участников были задействованы такие опросники как Шкала потребности в познании (англ. Need for Cognition Scale), Тест на когнитивную рефлексивность (англ. Cognitive Reflection Test), Шкала конформизма (англ. Conformity Scale) и др.

В первом исследовании предполагалось, что эффект доверия будет выше в том случае, когда слова дипфейк-спикера соответствуют внутренним установкам участников. Дипфейк-материалы были посвящены тематике вакцинации от коронавируса COVID-19 и подражали голосам российского врача (сторонника вакцинации) и актрисы (противника вакцин). Известные медиаличности рассказывали о своем отношении к вакцинации от ковида в необычном ключе: они транслировали мнение, противоположное их публичной позиции. Такая манипуляция должна была ввести участников эксперимента в заблуждение, а также склонить их к доверию тому спикеру, кто ближе им идеологически. Также предполагалось, что расхождение слов дипфейков с реальной позицией героев, а также с их внутренними установками, сможет вызвать специфическое отклонение мозговой амплитуды, известное как N400. Такая реакция проявляется при семантическом рассогласовании, а также при несоответствии ожиданиям, представлениям и моральным нормам человека.

Во втором исследовании изучается восприятие фейковых новостей на медицинскую тематику и их коррекции с использованием нейротехнологий, а именно метода ЭЭГ и техники вызванных потенциалов. В качестве коррекции используются дисклеймеры Правда и Обман от авторитетного и неавторитетного источника. Для изучения нейрофизиологического аспекта влияния коррекции измеряется амплитуда FRN потенциала. В ходе экспериментальной части участнику предстоит оценить уровень доверия к новости и готовность ею поделиться до и после предъявления дисклеймера. Ожидается, что FRN возникнет в ответ на дисклеймер и будет иметь максимальную амплитуду при дисклеймере Обман. Также ожидается, что индивидуальная амплитуда FRN у испытуемых будет предсказывать величину коррекции их доверия к фейку. Ожидается, что на поведенческом уровне достоверность источника и правдивость дисклеймера будут влиять на уровень доверия к новостям и готовность ими поделиться.

Результаты первого исследования показали, что на степень доверия к дипфейкам существенно влияют тип спикера и уровень потребности в познании. Внутренние

установки, в свою очередь, не показали ожидаемой значимости. Что касается нейрофизиологических коррелятов на записи ЭЭГ, мы не обнаружили выраженного отклонения амплитуды, однако тенденция к подобной реакции проявилась в группе антипрививочников по отношению к дипфейку врача. Пилотные результаты второго проекта показали, что существует значительная разница в уровне доверия и готовности делиться новостями до и после предъявления определенных дисклеймеров. Различия были также обнаружены между доверием и готовностью поделиться при дисклеймерах Обман и Правда. Что касается уровня доверия, был обнаружен значительный эффект взаимодействия между правдивостью дисклеймера и авторитетностью источника.

Источники

1. Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD explorations newsletter*, 21(2), 80-90.
2. Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380), 1094-1096.
3. Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of personality*, 88(2), 185-200.
4. Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in cognitive sciences*, 25(5), 388-402.
5. Wasserman, H., & Madrid-Morales, D. (2019). An exploratory study of “fake news” and media trust in Kenya, Nigeria and South Africa. *African Journalism Studies*, 40(1), 107-123.
6. Dmochowski, J. P., Sajda, P., Dias, J., & Parra, L. C. (2012). Correlated components of ongoing EEG point to emotionally laden attention—a possible marker of engagement?. *Frontiers in human neuroscience*, 6, 112.
7. Rainie, L., Smith, A., Schlozman, K. L., Brady, H., & Verba, S. (2012). Social media and political engagement. *Pew Internet & American Life Project*, 19(1), 2-13.
8. Van der Linden, S., Panagopoulos, C., & Roozenbeek, J. (2020). You are fake news: political bias in perceptions of fake news. *Media, Culture & Society*, 42(3), 460-470.
9. Bauer, P. C., & Clemm von Hohenberg, B. (2021). Believing and sharing information by fake sources: An experiment. *Political Communication*, 38(6), 647-671.
10. Wang, C., Tang, N., Zhen, D., Wang, X. R., Zhang, J., Cheong, Y., & Zhu, Q. (2022). Need for cognitive closure and trust towards government predicting pandemic behavior and mental health: Comparing United States and China. *Current Psychology*, 1-14.
11. Kurvers, R. H., Hertz, U., Karpus, J., Balode, M. P., Jayles, B., Binmore, K., & Bahrami, B. (2021). Strategic disinformation outperforms honesty in competition for social influence. *Isience*, 24(12).

12. Tsfati, Y., & Ariely, G. (2014). Individual and contextual correlates of trust in media across 44 countries. *Communication research*, 41(6), 760-782.