

Не доверяй — и проверяй. Почему искусственный интеллект врёт и что с ЭТИМ делать

20:30, 19 сентября 2026

Автор: [Валерия КУРНЕВИЧ](#)

ИИ прочно прописался во многих сферах — от медицины и юриспруденции до финансов и маркетинга. За блестящими возможностями скрывается неприятная особенность: часто нейросети выдают ложь за правду и делают это с такой уверенностью, что подвоха не замечают даже профессионалы. Статистика впечатляет: по данным бенчмарка HalluHard, который тестирует склонность ИИ к галлюцинациям в многоэтапных диалогах, даже самые мощные модели с доступом к интернету ошибаются в 30 процентах случаев. А если отключить веб-поиск, доля вымысла подскакивает до 60 процентов.



Что такое галлюцинация нейросети

У искусственного интеллекта этот феномен устроен почти как у естественного: информация, которой не существует в реальности, рисуется правдоподобной и вполне логичной. Корень проблемы кроется в самой архитектуре нейросетей, объясняют на кафедре интеллектуальных информационных технологий БГУИР. ИИ обучаются на огромных массивах текстов, улавливая статистические закономерности. При ответе модель

просто предсказывает, какое слово или фраза с наибольшей вероятностью должны появиться следом. В отличие от человека у алгоритма нет внутренней картины мира. Если в базе не окажется сведений на заданную тему, система может их придумать из похожих фактов, названий или языковых шаблонов.

Полностью отучить машину от фантазий, увы, нельзя. Но снизить число ошибок до приемлемого уровня вполне реально. Специалист советует формулировать для нейросети четкие и строгие ограничения: указывать источники информации, требовать отделять факты от предположений, просить ссылки. Вот пример рабочего промта: «Ответь только на основе приложенных документов. Для каждого важного утверждения укажи источник. Если данных недостаточно, напиши об этом и не дополняй ответ догадками».

Но галлюцинациями список опасностей не ограничивается. Появилось более тревожное явление, которое называют коллапсом модели. Нейросети черпают знания из интернета и неизбежно сталкиваются с контентом коллег-предшественников. Профессиональное сообщество бьет тревогу по поводу «загрязнения науки»: при недавнем исследовании электронного архива arXiv обнаружилось почти 150 тысяч фиктивных цитирований на крупнейших научных репозиториях. Так, одни алгоритмы принимают за доказанный научный факт информацию, выдуманную другими алгоритмами.

Как отличить правду от вымысла

За последние годы нейросети здорово прокачались: базы данных постоянно обновляются свежими новостями, а функция автоматического поиска в интернете стала стандартом. Проблема ошибок уже менее острая, однако совсем не исчезла. Вместо грубых фактологических ляпов современные модели чаще спотыкаются на тонкостях. Например, могут предложить сомнительный сайт, смешать свежие данные со старыми, сделать кривой вывод, а для пущей убедительности прикрепить вымышленную ссылку. В общем, на ИИ надейся, да сам не плошай: даже самые продвинутые алгоритмы не снимают с пользователя ответственности за фактчекинг.

Особую бдительность при работе с искусственным интеллектом стоит проявлять в критических сферах, подчеркивает **заведующий кафедрой защиты информации БГУИР Ольга Бойправ**. Ведь ошибки в медицине, юриспруденции или финансах способны обернуться реальными убытками и юридическими последствиями. Один из самых громких прецедентов

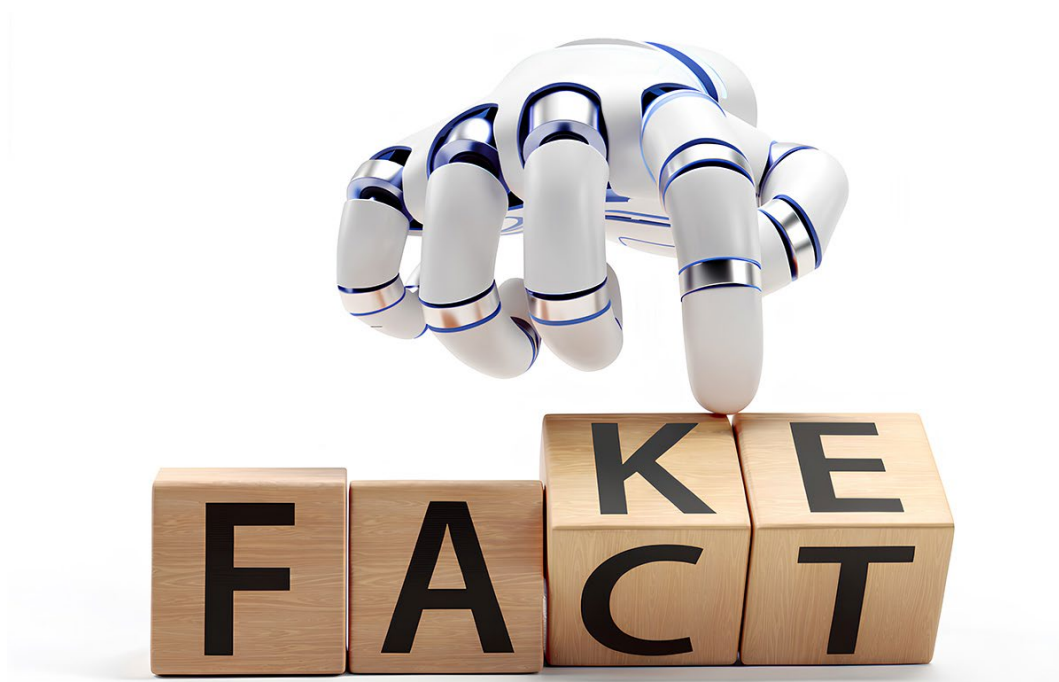
случился в 2023 году: американский адвокат доверил ChatGPT подготовку судебного иска. Недолго думая (или не думая вообще), юрист с 30-летним стажем вставил фантазии нейронки в официальный документ. Дело едва не завершилось дисциплинарным взысканием.

Прежде чем нажимать кнопку «копировать», эксперт призывает тщательно проверять источники, на которые опирается бот:

— Первым делом перейдите по ссылке и убедитесь, что на ресурсе-источнике действительно те же цифры, даты и факты, что и в ответе. Обратите внимание, защищено ли соединение с веб-сервером: рядом с адресной строкой должен гореть значок замка. Если доменное имя ресурса зарегистрировано совсем недавно (проверить это можно через специальные сервисы вроде whois.ru), отнеситесь к информации скептически.

Почему нельзя скармливать ИИ свои данные

Еще большую опасность представляют дипфейки. Первые ролики, где нейросети заставляли танцевать котят или говорить помидоры, вызывали разве что улыбку. Теперь качественно сгенерированный контент с живыми лицами, естественной речью и правдоподобными эмоциями превратился в реальный инструмент манипуляции. Как не попасться на удочку? Ольга Бойправ рекомендует приглядываться к деталям: синхронно ли движутся губы, подходит ли одежда персонажей окружающей обстановке, естественно ли падают тени, нет ли странных артефактов.



— Установить, подвергался ли файл обработке, поможет, например, сервис FotoForensics, — подсказывает эксперт.

Но угроза кроется не только в том, что создает нейросеть, но и в том, что мы ей доверяем. Почему бы не поручить боту рутину: скажем, загрузить паспортные данные для составления договора или корпоративные отчеты для аналитики. Быстро, удобно и... чревато последствиями. Вся информация, которую скормили ИИ, оседает на серверах компании-разработчика, и далеко не факт, что там уделяется должное внимание вопросам безопасности. Так, при успешной хакерской атаке ваши секреты вполне могут уйти на черный рынок.

— Если в базе сервиса останутся ваши данные — полное имя, место работы, номер телефона, — мошенникам не составит труда использовать их в своих целях. К тому же многие разработчики используют переписку с пользователями для обучения своих моделей. Не исключено, что коммерческая тайна или персональные данные однажды всплывут в ответе, который ИИ сформирует для другого человека, — подытоживает Ольга Бойправ.

Эффективно и безопасно

Как правильно работать с нейросетями



Не скармливайте ИИ-сервисам чувствительные данные. Если для работы с большими массивами персональной информации нейросеть все же необходима, предварительно обезличьте ее: уберите имена, адреса, номера телефонов и другие идентификаторы. Всегда проверяйте источники, на которые ссылается нейросеть. Даже если ответ выглядит безупречно, перейдите по ссылкам, сверьте цифры, даты и факты. Если публикуете информацию, созданную нейросетью, честно указывайте это.

[#ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ](#) [#ИИ](#) [#НЕЙРОСЕТЬ](#) [#ТЕХНОЛОГИИ](#)