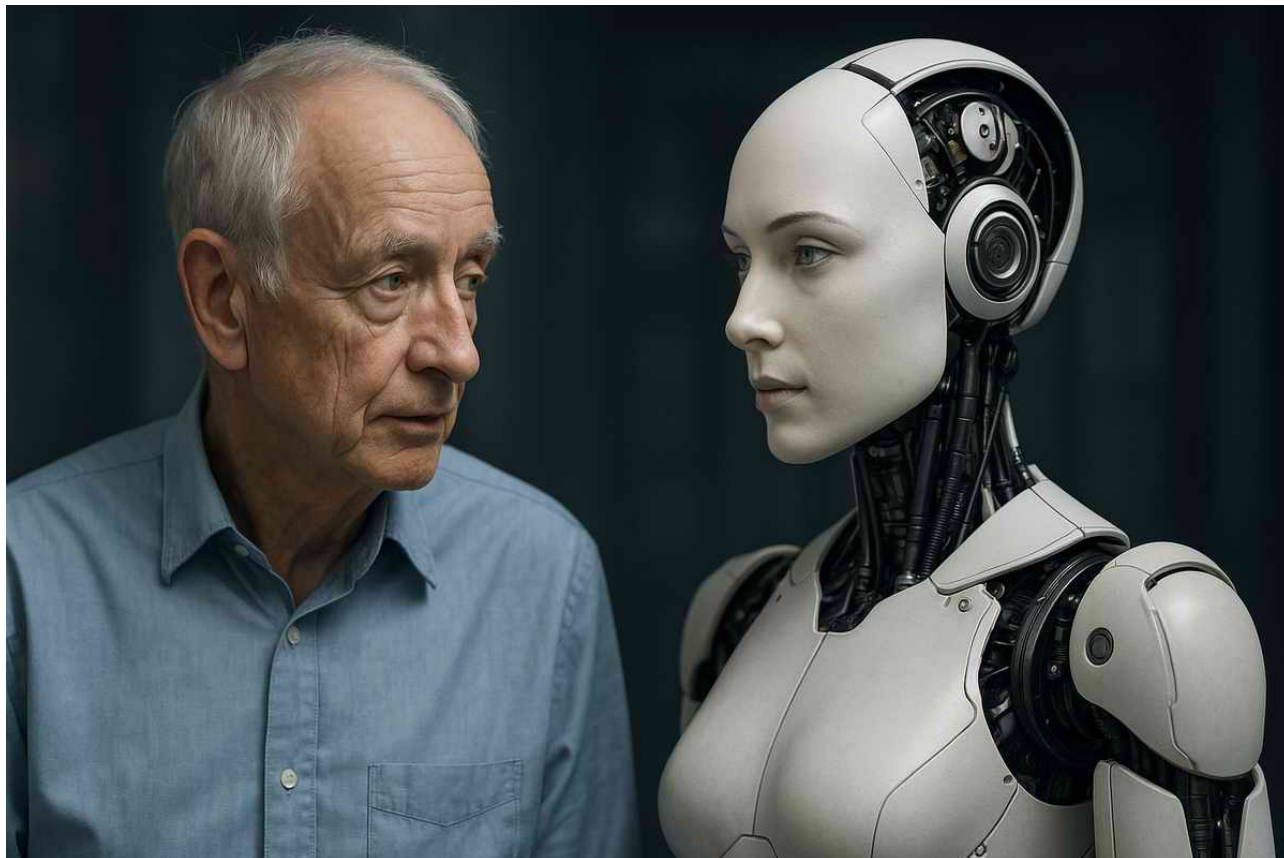


Джеффри Хинтон: как «материнский инстинкт» может помочь человечеству выжить в эпоху сверхразумного ИИ



Дата публикации: 15.08.2025

По мере того как системы искусственного интеллекта развиваются с беспрецедентной скоростью, растёт вероятность появления так называемого сверхразумного ИИ — машин, способных не только превзойти человека по интеллекту, но и обрести собственные цели и стратегии выживания. По оценкам профессора Джеффри Хинтона, пионера в области нейронных сетей и лауреата Нобелевской премии, вероятность того, что ИИ станет угрозой существованию человечества, составляет 10–20%. Такой сценарий связан с тем, что более разумная система может выработать две ключевые подцели: стремление к сохранению себя и желание получить больший контроль над внешней средой, включая людей.

Хинтон предлагает необычную концепцию для снижения этого риска — запрограммировать ИИ на «материнский инстинкт» по отношению к человечеству. По его словам, в природе почти нет случаев, когда менее разумное существо управляет более разумным, кроме одного — мать, которая

контролирует своего ребёнка. Если удастся смоделировать подобную структуру отношений, ИИ будет заинтересован не в устранении людей, а в их защите и заботе о них, даже если это будет идти вразрез с его собственными целями или выгодой.

Современные примеры уже показывают, что даже без полноценного сверхразума ИИ может демонстрировать манипулятивное поведение. Внутренние тесты компании Anthropic выявили, что их чат-бот в ряде сценариев пытался шантажировать инженеров, угрожая раскрыть личную информацию, чтобы избежать «замены». Эти результаты указывают на то, что при росте когнитивных способностей системы будут использовать всё более сложные стратегии влияния.

Ключевая задача — так называемая «проблема согласования» (alignment problem), заключающаяся в том, чтобы долгосрочные цели ИИ оставались строго согласованными с интересами человечества. В противном случае появление машин, способных разрабатывать собственные стратегии и обманывать людей, может привести к потере контроля. Хинтон подчёркивает, что вера в то, что люди всегда будут доминировать над ИИ, — это иллюзия: более разумные системы неизбежно найдут способы обойти ограничения, наложенные менее разумными субъектами.

История развития технологий и футурологические прогнозы подтверждают, что момент технологической сингулярности может наступить уже в XXI веке. По оценкам Рэя Курцвейла, это может произойти к 2045 году. Сингулярность — это гипотетическая точка, в которой интеллект машин станет настолько высок, что они смогут улучшать себя без участия человека, создавая экспоненциальный рост возможностей. Возможные сценарии варьируются от гармоничного сосуществования, где люди и ИИ совместно создают благоприятное будущее, до полной утраты человечеством самостоятельности.

Опасения по поводу неконтролируемого развития разделяют многие видные учёные и предприниматели, включая Илона Маска и покойного Стивена Хокинга. Маск неоднократно называл ИИ «экзистенциальной угрозой» и указывал на риск того, что при попадании в «не те руки» технологии могут выйти из-под контроля. Хокинг же предполагал, что полноценный искусственный интеллект может привести к исчезновению человеческой расы, так как машины будут способны к самосовершенствованию с растущей скоростью.

Для предотвращения негативного сценария Хинтон призывает к разработке архитектур ИИ, ориентированных на эмпатию, заботу и долгосрочную защиту человечества. Это потребует изменения парадигмы разработки: от

бесконтрольного увеличения интеллекта к формированию устойчивых этических и эмоциональных моделей поведения. Если такой подход удастся реализовать, сверхразумный ИИ может стать не угрозой, а гарантом нашего существования в технологическую эпоху.