В ближайшие годы может быть разработан искусственный интеллект общего назначения (AGI), способный выполнять большинство задач не хуже или лучше человека. Google DeepMind утверждает, что AGI может принести огромную пользу обществу, например, помочь открыть новые лекарства, улучшить образование и решить проблему изменения климата.

Но вместе с этой мощной технологией приходит и серьезная ответственность. DeepMind опубликовала новый документ, в котором описала, как она работает над обеспечением безопасности и надежности разработки AGI.

Они сосредоточены на предотвращении четырех ключевых рисков:

- **Неправильное использование** — когда люди используют ИИ для нанесения вреда, например, для распространения ложной информации или организации кибератак.
- **Несоответствие** — когда ИИ не следует целям человека должным образом, например, использует неэтичные короткие пути для решения задач.
- **Аварии** — неожиданные проблемы, возникающие из-за поведения ИИ.
- **Структурные риски** — более широкие опасности системного уровня.

Чтобы снизить эти риски, DeepMind тестирует свои самые передовые модели ИИ на предмет опасного поведения, разрабатывает более надежные инструменты безопасности и новые методы, позволяющие сделать решения ИИ более понятными. Для руководства своей работой компания также создала советы и партнерства с исследователями, правительствами и некоммерческими организациями.