УДК 004.8

ЭМЕРДЖЕНТНЫЕ СПОСОБНОСТИ: ВОЗНИКНОВЕНИЕ И РИСКИ

Е. А. Рожкова[1)], К. Р. Соколюк[2)]

[1)] *студент, Белорусский государственный университет, г. Минск, Беларусь, elen.lenss@yandex.by*
[2)] *студент, Белорусский государственный университет, г. Минск, Беларусь, kristinasokoluk2@gmail.com*

Научный руководитель: **Н. И. Шандора**

старший преподаватель, Белорусский государственный университет, г. Минск, Беларусь, shandoranatasha@tut.by

В статье рассматривается феномен эмерджентности одновременно является источником научного интереса и беспокойства по поводу непредвиденных последствий, влеча за собой новые, не очевидные на сегодняшний день, риски. Описываются связанные с эмерджентностью риски, возникающие в процессе формирования сложных систем. Акцентируется внимание на том, что понимание этих свойств имеет ключевое значение для предсказания и контроля сложных адаптивных систем в науке и технологиях. Также подчеркивается важность междисциплинарного подхода к изучению эмерджентности.

Ключевые слова: эмерджентность; большие языковые модели; производительность систем; фазовый переход; искусственный интеллект.

EMERGENT CAPABILITIES: OCCURRENCE AND RISKS

E. A. Rozhkova[1)], K. R. Sokolyuk[2)]

[1)] *student, Belarusian State University, Minsk, Belarus, elen.lenss@yandex.by*
[2)] *student, Belarusian State University, Minsk, Belarus, kristinasokoluk2@gmail.com*

Supervisor: **N. I. Shandora**

senior lecturer, Belarusian State University, Minsk, Belarus, shandoranatasha@tut.by

The article discusses the phenomenon of emergent nature, which is both a source of scientific interest and concern about unintended consequences, entailing new risks that are not obvious to date. It describes the risks associated with emergent nature that arise in the process of formation of complex systems. It is emphasized that understanding these properties is crucial for predicting and controlling complex adaptive systems in science and

technology. The importance of an interdisciplinary approach to the study of emergence is also emphasized.

Keywords: emergent; large language models; system performance; phase transition; artificial intelligence.

Эмерджентность означает новые свойства и возможности больших языковых моделей (LLM), которые появляются внезапно и непредсказуемо по мере увеличения размера модели, вычислительной мощности и объема обучающих данных. Даже простые объекты при правильном взаимодействии могут порождать высокоорганизованные и адаптивные системы. Качества LLM, которые отсутствуют в более мелких моделях, но при этом существуют в крупномасштабных системах, можно считать эмерджентными способностями. Данные качества являются неожиданным и непредсказуемым результатом взаимодействия определённого количества компонентов системы без первоначального программирования [1].

В некоторых исследованиях утверждается, что возникновение эмерджентных способностей подчиняется закону масштабирования (последовательное увеличение производительности). Считается, что системы масштабируются по таким основным факторам как: объем вычислительных операций, число параметров модели, размер обучающего набора данных. При проявлении эмерджентных качеств отмечается следующая закономерность: с увеличением количества параметров производительность системы носит случайный характер до достижения определённого уровня, на котором начинает проявляться новое свойство. Данное качественное преобразование также называется фазовым переходом – резкое изменение в общем поведении системы, которое не может быть предсказано при анализе систем меньшего масштаба. Его также называют эмерджентным, поскольку такой переход невозможно предсказать, рассматривая мелкомасштабную модель [3].

Однако Р. Шеффер, учёный стэндфордского университета, в своей статье утверждает, что эмерджентные способности – иллюзия, вызванная выбором определённой метрики, которая нелинейно или скачкообразно деформирует коэффициент ошибок, а не непредсказуемыми изменениями в поведении модели с увеличением масштаба. Выбор такого нелинейного измерения может привести к чему-то непредвиденному, что потом ошибочно принимают за эмерджентные свойства. Заменив нелинейное измерение результатов на линейные, развитие модели становится предсказуемым и плавным, исключая внезапное возникновение новых способностей. Поэтому на самом деле кривая производительности всё это время растёт плавно и не совершает резких скачков [1].

Примером эмерджентности является алгоритм ИИ AlphaGo, который был разработан для игры в го. Он продемонстрировал эмерджентные свойства, научившись играть в го на сверхчеловеческом уровне. Благодаря самостоятельной игре и обучению с подкреплением AlphaGo разработал стратегии и тактики, которые не были явно запрограммированы, а возникли в процессе обучения системы [2].

Также важно учитывать риски, которые возрастают с увеличением масштаба модели. К таким проблемам относится непреднамеренный обман, синтез вредоносного контента и автономные транспортные средства. Если данный вид транспорта столкнется с ситуацией, которая не встречалась в его обучающих данных, то это может привести к непредсказуемым и потенциально опасным последствиям, так как автомобиль начнёт демонстрировать нестандартное поведение, не соответствующее запрограммированным алгоритмам. Для обнаружения и снижения возникающих рисков используются подходы, включающие фильтрацию данных, прогнозирование, управление и автоматическое обнаружение вредоносного поведения [2].

Эмерджентность представляет собой как возможность, так и риск, проявляясь в сложных системах, требующих вдумчивого подхода к управлению и прогнозированию. Понимание данных свойств и связанных с ними рисков может помочь в создании более устойчивых систем, минимизируя нежелательные последствия. Адаптация к изменяющимся условиям и принятие эффективных стратегий управления рисками станут основой будущего не только в области научных исследований, но и в практическом применении на уровне общества и экономики.

Библиографические ссылки

1. Emergent Abilities of Large Language Models [Electronic resource]. URL: https://arxiv.org/abs/2206.07682 (date of access: 10.09.2024).

2. Emergent properties of AI [Electronic resource]. URL: https://digitaldaze.io/emergent-properties-ai/ (date of access: 10.19.2024).

3. Exploring the Emergent Abilities of Large Language Models [Electronic resource]. URL: https://www.deepchecks.com/exploring-the-emergent-abilities-of-large-language-models/ (date of access: 15.09.2024).