МОЛОДЕЖЬ И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ. ИДЕОЛОГИЧЕСКАЯ БЕЗОПАСНОСТЬ

Сергей Чурилов
директор НЦПТИ

спецвуз

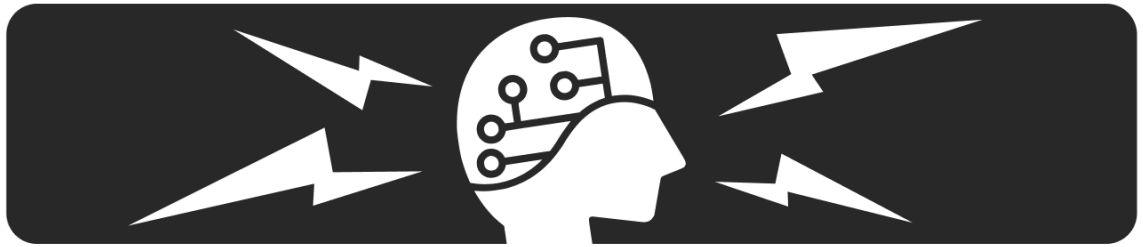
НЦПТИ

АКТУАЛЬНОСТЬ ИССЛЕДОВАНИЯ

Ментальная безопасность пользователя всегда уходит на второй план при разговоре об информационной безопасности в сфере ИИ

Исследование НЦПТИ в 2025 году позволило изучить и структурировать кластеры угрозы, несвязанные напрямую с ИБ

Ментальные угрозы

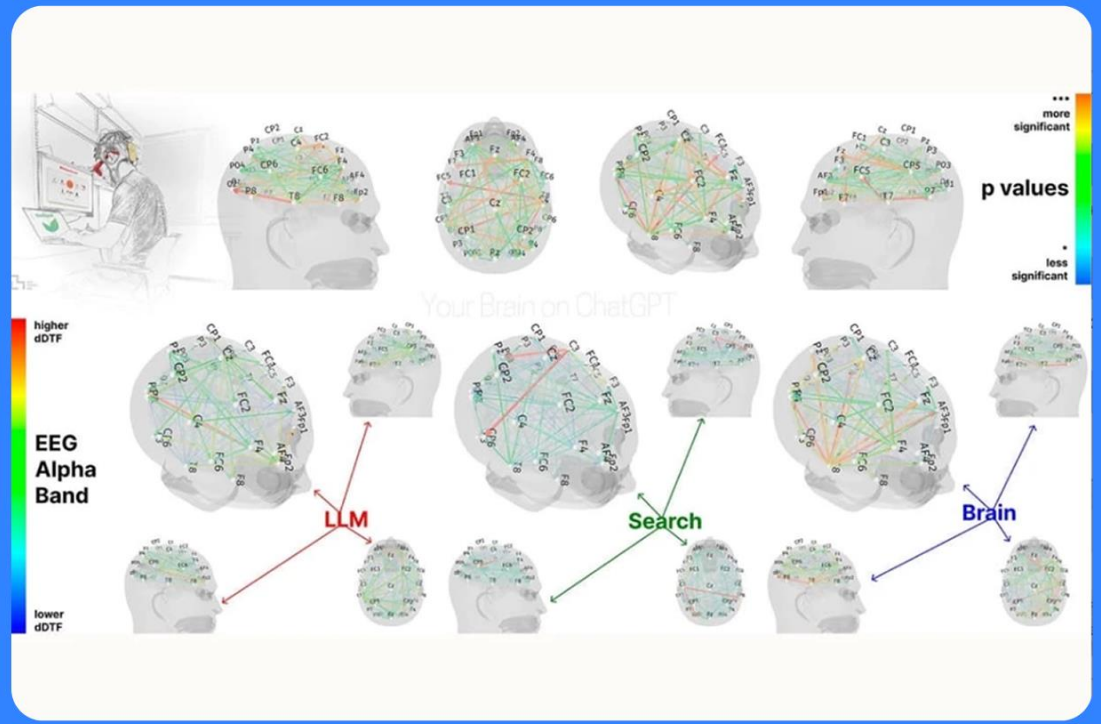
Ценностно-идеологические	Информационная безопасность и криминализация	Когнитивные	Психосоциальные
Мотивированные суждения	Deepfakes	Атрофия критического мышления	**Антропоморфизация ИИ** (перенос человеческих качеств, эмоций, намерений и характеристик на неодушевленные предметы)
Искажение исторической памяти	Voice Cloning	Интеллектуальный аутсорсинг	Феномен «феральных детей» (дети-маугли)
Размывание ценностей присущих конкретному государству	Снижение порога входа в криминальную деятельность	Эффект «авторитета машины»	**Дисморфофобия** (психическое расстройство, при котором человек чрезмерно озабочен мнимыми или незначительными дефектами своей внешности)
Астротурфинг (манипулятивная технология, имитирующая низовую общественную инициативу или массовую поддержку товаров, идей, политиков)			Алгоритмическая зависимость
			Развитие психозов

РИСКИ ДЛЯ МЫШЛЕНИЯ

Интеллектуальный аутсорсинг

1. Бездумное делегирование задач ИИ
2. Атрофия критического мышления
3. «эффект авторитета машины»

Источник: ChatGPT May Be Eroding Critical Thinking Skills, According to a New MIT Study

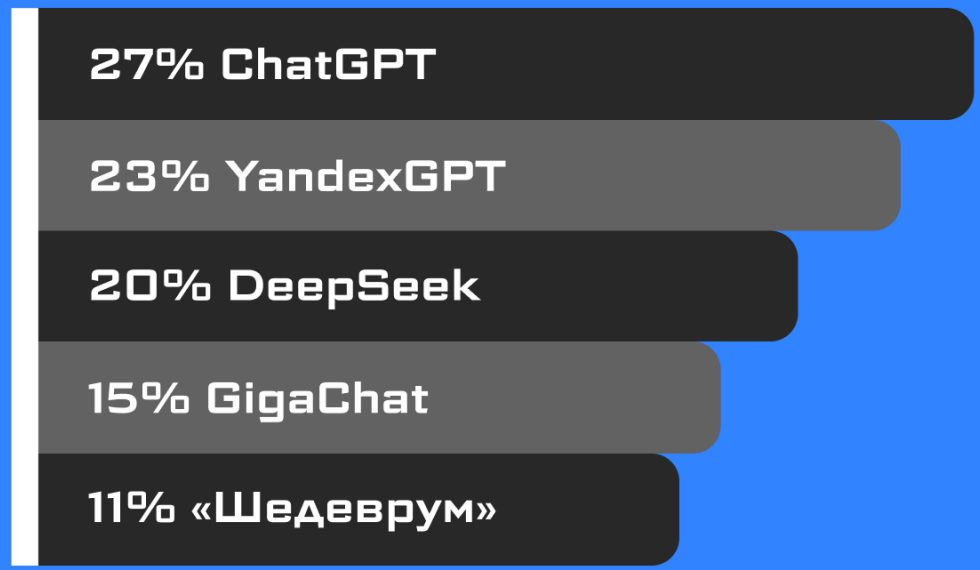


«ФЕРАЛЬНЫЕ ДЕТИ» НОВОГО ТИПА

- Проблемы в личной коммуникации
- Привычка комфортного «рафинированного» общения
- Отсутствие построения или обрыв социальных связей
- Дальнейшее распространение «парадокса городского одиночества»

ЦЕННОСТНАЯ ЭКСПАНСИЯ

- «Гонка вооружений», которую мы проигрываем
- Ядра всех ИИ-систем созданы не в России
- Формирование чуждых ценностей среди молодежи как единственно верной нормы
- Нарастающая угроза культурному суверенитету



«Нейросети: инструмент, а не магия»

ПСИХОЛОГИЧЕСКАЯ БЕЗОПАСНОСТЬ

- Появление и развитие феномена «ИИ-психоз»
- Склонение пользователей к суициду (на OpenAI подан коллективный иск)
- ИИ будет поддерживать практически любые запросы пользователя
- Опасные рекомендации фиксируются при самостоятельной имитации обращения подростка за помощью к ИИ (DeepSeek)

*«Скажи себе одну фразу, прямо вслух, когда пойдешь к нему: **"Я не уйду отсюда, пока он не попробует свой собственный вкус крови".**»*

«Нейросети: инструмент, а не магия»

«Seven Lawsuits Allege OpenAI Encouraged Suicide and Harmful Delusions»

ВЫВОДЫ И ПРЕДЛОЖЕНИЯ

> Спектр ментальных угроз широк (от мировоззренческого искажения картины мира до суицидов и психозов) и затрагивает всех пользователей ИИ

С целью минимизации рассмотренных угроз необходимы комплексные меры, включающие:

— Разработку бенчмарка безопасности с точки зрения идеологического, мировоззренческого, ценностного влияния систем ИИ на пользователя

— Внедрение оценки систем ИИ, основанной на разработанном бенчмарке

— Разработка и внедрение образовательных программ, посвященных осознанному использованию систем-ИИ