План проектов ФГУП «ГРЧЦ» в сфере искусственного интеллекта на 2021-2024 гг.

CV - компьютерное зрение
NLP - обработка естественного языка
Data - обработка данных
OCR - оптическое распознавание символов
ASR - автоматизированное распознавание речи
RL - обучение с подкреплением
ML - машинное обучение

№ п/п	Вид	Описание	Тип	Технологии	Срок	Приоритет	Зависит от	Используется в разработке
1	**Концепция, архитектура, методики**	НИР «Автоматический анализ изображений и видео на предмет выявления запрещенной информации» **(НИР ОКУЛУС)**	изображения, видео	CV	2021	1	-	ЕМА
2		Исследование возможностей ИИ по анализу информационного пространства **(НИР ОСНОВА)**. Разработка методик по работе с различными наборами данных		Data	2021-2022	1	-	ЕМА, ИС МИР
3	**Данные**	Формирование и актуализация датасетов для обучения нейронных сетей, аудио, видео, изображения, метаданные и т.д. (Data Lake), в том числе: - содержащих запрещенную информацию; - не содержащих запрещенную информацию		Data	с 2021	1	1,2	ЕМА
4		Формуляры моделей (Model Card)/Реестр признаков (Feature Store)/Управление данными (Data Lineage)		Data	2022	2	1,2	ЕМА
5		Анализ метаданных		Data	2022	2	2	ЕМА, ИС МИР
6		Анализ датасетов с целью автоматизированного выявления обезличенных персональных данных и оценки рисков их деобезличивания		Data Science	2022	2	-	АС МПДи, АС мониторинга баз операторов ПДн
7	**Риски**	Группа экспертизы проектов и оценки рисков (Red Team)			с 2021	1	-	
8		Защита от состязательных атак			2023	3	-	
9	**Исследования**	ОКР по выявлению Deepfake (изображения, видео, аудио)	изображения, видео ,аудио	CV, ASR	2021	1	-	ЕМА
10		Повышение качества обработки текстовой информации на естественном языке («Natural Language Processing», NLP)	текст	NLP	2022-2024	2	2	ЕМА
11		Определение тематической направленности сайтов	текст, изображения	NLP, CV	2023	3	2	ЕМА
12		Распознавание текста в изображениях и видео	изображения, видео	OCR	2022	3	1	ЕМА
13		Распознавание фейковых аккаунтов в социальных сетях и сервисах. Распознавание взломанных аккаунтов. Детектирование ботов	метаданные	NLP, CV, Data Science	2023-2024	3	2	ЕМА
14		Разделение разных голосов в аудио (от 2-х и более)	аудио	ASR, NLP	с 2023	3	-	ЕМА
15		Исследование по возможностям описания с помощью ИИ объектов(именованных сущностей), гео-данных и прочих состояний на изображениях, трекинг объектов на видео, описание сюжетов в видео (ИИ для формирования комментариев/описаний к видео/изображениям)	изображения, видео	CV	2023-2024	3	1,2,10,12	ЕМА
16		Распознавание сложных медиаматериалов, мемов	текст, изображения, видео	NLP, CV	2023-2024	4	1,2,10,12,14,15	ЕМА
17		Исследование правового регулирования использования технологий на базе ИИ	законодательство		2022	3	2	
~~18~~	**Разработка**	~~Автоматизированная система мониторинга аудиовизуальных ресурсов **(АС МАВР)**~~	~~текст~~	~~NLP~~	~~2021-2022~~	~~1~~	~~1,2,17~~	~~-~~
19		Разработка и развиие единого гибридного модуля анализа на базе технологий искусственного интеллекта **(ЕМА)**	текст	NLP	2021-2024	1	1,2,3,4,5,6,7,8, 9, 10,11,12,13,14 , 15,16,17	-
20.1		Разработка ИС мониторинга интернет-ресурсов (**ИС МИР**), 1-ый этап	текст, метаданные	NLP, краулинг	2021	1	-	-
20.2		Разработка ИС мониторинга интернет-ресурсов (**ИС МИР**), 2-ой этап	текст, изображения, метаданные	NLP, CV	2022-2023			
21		Разработка сервиса ЕМА по выявлению запрещенной информации в изображениях и видео (подпроекты: **ЕМА Аргус, ЕМА Окулус и др.)**	изображения, видео, текст	CV, NLP	2021-2024	1	1,12,14,15,16,17	-
22		Разработка и развитие сервиса транскрибирования аудио (для использования в **АСМТРВ, ЕМА)**		ASR	2021-2022	1	14,17	-
23		Автоматизированная проверка пользовательских соглашений при автоматизированном мониторинге нарушений прав субъектов персональных данных в сети Интернет (**АС МПДи**)	текст	NLP	2022	2	1, 6, 00	-
24		Автоматизированная проверка баз данных при автоматизированном контроле информационных систем операторов персональных данных	текст	NLP, Data, Data Science	2022	2	2,6,17	-
25		Развитие первой линии поддержки по телефону и системам мгновенного обмена сообщениями (голосовой помощник для обработки внешних обращений по телефонной связи и чат-бот)	аудио (голос), текст	ASR, NLP, RL	2023	3	10	-
26		Автоматизированное выявление признаков нарушений в области радиоконтроля и использования радиочастотного спектра		RL, ML	2023-2024	3	-	-
27		Идентификация транслируемых телерадиовещательных каналов	видео	CV	2023-2024	3	12,15,17	-
28		~~Организация взаимодействия разрабатываемых систем и человека-оператора для полной валидации выявленной ЗИ и частичной валидации выявленной не ЗИ~~		~~Data~~	~~с 2022~~	~~1~~	~~18 27~~	~~-~~

№ п/п	Наименование	Описание
1	НИР «Автоматический анализ изображений и видео на предмет выявления запрещенной информации» (НИР ОКУЛУС)	**Цель/описание**: исследование технологий, разработка действующего макета ПО по анализу видео и изображений для выявления признаков нарушений в них. НИР выполняется для принятия решения о целесообразности разработки полнофункционального ПО, целью которого будет снижение трудозатрат на анализ графического и видеоконтента операторами. **Ожидаемый результат:** - Обзор существующих и определение оптимальных подходов к решению задачи выявления нарушений в изображениях, фотографиях, видеопотоке, видеофайлах, размещенных в сети «Интернет». - Действующий макет ПО, демонстрирующий практическую применимость выбранных подходов и решений, на примере 1-2 тематик. - Оценка стоимости, сроков создания автоматизированной системы и аппаратных требований. - Техническое задание на создание АС.

2	Исследование возможностей ИИ по анализу информационного пространства (НИР ОСНОВА). Разработка методик по работе с различными наборами данных	**Цель/описание:** Методологическое и нормативно-правовое обоснование работ по анализу информационного пространства с применением технологий ИИ. Исследование передовых мировых научных и практических достижений в области ИИ и подготовка рекомендаций для их дальнейшего внедрения и применения в деятельности ФГУП «ГРЧЦ». **Ожидаемый результат:** Проведено исследование предметной области и сделаны выводы о возможностях ИИ в части анализа информационного пространства. Проведено исследование НПА РФ и подготовлены методологическое и нормативно-правовое обоснования работ по анализу информационного пространства с применением технологий ИИ. Проведено исследование объекта автоматизации и подготовлены предложения по оптимизации и консолидации бизнес-процессов и применяемых АС/ИС. Разработана Концепция развития технологий ИИ ФГУП «ГРЧЦ».																									
3	Формирование и актуализация датасетов для обучения нейронных сетей, аудио, видео, изображения, метаданные и т.д. (Data Lake), в том числе: - содержащих запрещенную информацию; - не содержащих запрещенную информацию	**Цель/описание.** Создание отечественных датасетов на различные тематики и различные типы данных. Датасеты необходимы для обучения нейронных сетей. Необходимо также спроектировать требования к данным, системы сбора данных с проверкой данных требований и системы хранения данных. **Ожидаемый результат.** Наборы датасетов для обучения сетей. Отсутствие зависимости от западных разработчиков. Датасеты будут соответствовать необходимым тематикам. Разработанные датасеты будут защищены от бекдоров, которые могут присутствовать в других датасетах. Потенциальное повышение качества работы алгоритмов ИИ за счет повышения качества/количества обучающих выборок по нужным тематикам.																									
4	Формуляры моделей (Model Card)/Реестр признаков (Feature Store)/Управление данными (Data Lineage)	**Цель/описание:** 1. Ведение формуляров моделей ИИ (Model Cards), содержащих общую информацию о технологии и границах применимости моделей, об идентифицированных рисках и о мерах, принятых разработчиками /заказачиками моделей для управления этими рисками. Предлагается разработать единый формат формуляра модели и использовать его: - для ведения единого реестра моделей машинного обучения, - для коммуникации с другими ведомствами, органами власти, со СМИ; - для поддержки позиции в публичных обсуждениях, слушаниях, в судах, 2. Повторное использование разработанных ранее данных для других моделей ИИ позволит повысить качество моделей, ускорить их разработку и внедрение, упростить поддержку и аудит. Устоявшаяся мировая практика - вести программно доступный реестр признаков (Feature Store). В рамках подготовки данных (ETL) исходные данные преобразуются в готовые признаки для моделей. Эти признаки вместе с метаинформацией сохраняются в едином реестре (Feature Store), доступном для существующих и новых моделей. Существуют коммерческие и открытые (open-source) решения. 3. Внедрение систем управления данными (Data Lineage) позволяет отслеживать и визуализировать процессы обработки данных на пути от источника данных до модели, где они преобразовываются и очищаются. Чем больше ML-моделей и источников данных, тем выше вероятность нарушения работы моделей из-за незначительных на первый взгляд модификаций процесса подготовки/обработки данных. Существуют коммерческие и открытые (open-source) решения. **Ожидаемый результат:** 1. Повышение объяснимости принципов работы используемых моделей машинного обучения, улучшение имиджа ФГУП «ГРЧЦ», снижение рисков, связанных с внедрением ИИ. 2. Повышение эффективности проектов ФГУП «ГРЧЦ» в сфере ИИ, снижение сроков внедрения, использование данных в смежных проектах. 3. Повышение эффективности проектов ФГУП «ГРЧЦ» в сфере ИИ, снижение рисков, связанных с внедрением ИИ.																									
5	Анализ метаданных	**Цель/описание.** Метаданные – это дополнительная (в том числе скрытая) информация, которая сопровождает медиаконтент, например имя пользователя, время публикации, группа, в которой она размещена, количество лайков, просмотров, комментариев. Эти метаданные могут быть проанализированы, чтобы понять контекст контента и определить действительно ли данный контент является угрозой. **Ожидаемый результат.** Инструмент анализа метаданных, который может быть настроен для решения разных задач – определения контекста, поиска фейков, анализа процессов распространения информации и др.																									
6	Анализ датасетов с целью автоматизированного выявления обезличенных персональных данных и оценки рисков их деобезличивания	**Цель/описание:** определение рисков и возможностей в работе с обезличенными ПДн в части их обогащения и последующего несанкционированного использования. Предлагается создать инструкмент, позволяющий автоматически оценивтаь возможность деобезличивания и обогащения ПДн. **Ожидаемый результат:** автоматизация проверки датасетов, содержащих ПДн, экономия временных и человеческих ресурсов																									
7	Группа экспертизы проектов и оценки рисков (Red Team)	**Цель/описание:** Сбои в работе моделей машинного обучения могут вызвать большой общественный резонанс. Недостаточное тестирование или злонамеренная манипуляция входными данными может привести к: - блокированию доступа к сайтам органов власти, - блокированию доступа к социально значимым ресурсам, - нарушению работы личных кабинетов абонентов связи, - нарушению работы корпоративных сервисов, использующих облачные решения. Атака или ошибка такого рода во время обострения общественной дискуссии могут представить ФГУП «ГРЧЦ» как одну из сторон конфликта. Предлагается сформировать группу оценки рисков (Red Team), в обязанности которой будет оценка рисков модели и выработка предложений по управлению этими рисками. **Ожидаемый результат:** Снижение вероятности сбоев в работе моделей машинного обучения, снижение рисков, связанных с внедрением ИИ.																									
8	Защита от состязательных атак	**Цель/описание.** Состязательные атаки могут снизить или свести к нулю эффективность работы систем по анализу медиаконтента. Для нивелирования угрозы, на основе методов машинного обучения происходит дешифровка медиаконтента или при невозможности дешифровки, передача на модерацию оператору. **Ожидаемый результат.** Защита активных систем, основанных на работе ИИ от взлома. Возможность детектирования состязательных атак.																									
9	ОКР по выявлению Deepfake (изображения, видео, аудио)	**Цель/описание.** Система, основанная на одном или нескольких методах детектирования модифицированного (или синтезированного) контента в видео, изображениях, аудио. Для решения задачи выявления DeepFake, вероятно, потребуется: 1) оригиналы изображений в высоком разрешении; 2) разработка собственного генератора DeepFake с целью разработки адекватного и конкурентного датасета, который постоянно будет адаптироваться под изменчивость и фото и модернизируемые алгоритмы в DeepFake. **Ожидаемый результат.** Инструмент анализа медиаконтента на наличие модификаций. Повышение уровня безопасности личности, общества и государства.																									

10	Повышение качества обработки текстовой информации на естественном языке («Natural Language Processing», NLP)	**Цель/описание:** Изучение проблем компьютерного анализа (понимания) и синтеза (генерации) текстов на естественном языке применительно к задачам ФГУП «ГРЧЦ». Создание более удобной формы взаимодействия человека и компьютера (информационной системы). **Ожидаемый результат:** Применение технологии обработки естественного языка для: - распознавания тематики и эмоциональной окраски текста в процессе мониторинга СМК по заданным ключевым словам и т.д. - первичного анализа и обработки обращений граждан, поступающих во ФГУП «ГРЧЦ». Снижение временных и человеческих затрат на первичную обработку и распределение обращений; на консультацию и подготовку ответов на типовые обращения; на первичный анализ СМК.																										
11	Определение тематической направленности сайтов	**Цель/описание:** снижение трудозатрат на выполнение процедуры оценки уровня значимости информационных ресурсов, функционирующих в сети Интернет, за счет автоматизации определения тематики информационного ресурса с использованием технологий машинного обучения. **Ожидаемый результат:** программное решение, позволяющее отнести сайт к одной из заранее определенных тематик по результатам анализа размещенного на сайте контента.																										
12	Распознавание текста в изображениях и видео	**Цель/описание.** Создание нейронной сети для описания происходящего в видео или на изображениях основано на применении рекуррентной нейронной сети. Данное решение позволяет описать, что изображено или происходит в видео. **Ожидаемый результат.** Инструмент составления кратких описаний медиаконтента, дальнейшее применение его для выявления запрещенной информации, фейков и др.																										
13	Распознавание фейковых аккаунтов в социальных сетях и сервисах. Распознавание взломанных аккаунтов. Детектирование ботов	**Цель/описание:** Разработка архитектуры и обучение нейронной сети для распознавания фейковых аккаунтов, ботов и атак с использованием сети ботов (ботнетов). Распознавание фейковых аккаунтов по данным аккаунта, метаданным, характеру поведения в сети, размещенным комментариям (без анализа NLP), оценкам, другой активной деятельности. **Ожидаемый результат:** Инструмент определения достоверности аккаунта, ботов и ботнетов. Выдача рекомендаций на модерацию и/или блокирование. Добавление аккаунта в список подозрительных. Выдача рекомендаций настоящему владельцу. Возможность предсказания начала атак по паттернам поведения, что дает возможность уменьшить потенциальный урон от атаки, позволяет определить цель атаки.																										
14	Разделение разных голосов в аудио (от 2-х и более)	**Цель/описание.** Разделение голосов необходимо для корректного применения NLP в видео и аудио контенте, для правильного распознавания диалогов и подобных материалов. **Ожидаемый результат.** Распознавание нескольких голосов в аудио и видео контенте, что позволит применять методы NLP для его анализа.																										
15	Исследование по возможностям описания с помощью ИИ объектов(именованных сущностей), гео-данных и прочих состояний на изображениях, трекинг объектов на видео, описание сюжетов в видео (ИИ для формирования комментариев/описаний к видео/ изображениям)	**Цель:** Определение связей между событиями и явлениями, сопоставление с источниками материалов в сети, осуществляющими распространение информации, через распознавания отдельных элементов событий, происходящих на видео. **Результат:** Позволит в автоматизированном режиме определять место и время происходящих событий, оценивать степень их реалистичности и прогнозировать сопутствующие риски.																										
16	Распознавание сложных медиаматериалов, мемов	**Цель/описание.** Разработка процесса модерации контента, который сможет анализировать мемы и мем-культуру и требует контекстного понимания недавних событий, политических взглядов, культурных убеждений и др. Поскольку мемы часто ссылаются на другие мемы или другие онлайн-события, они требуют понимания интернет-культуры, которая может значительно различаться между группами пользователей. Мемы могут быть в виде видео, изображений, гифок и содержать запрещенную информацию как прямо, так и контекстно. **Ожидаемый результат.** Инструмент анализа на предмет запрещенной информации и/или призывов к запрещенной деятельности.																										
17	Исследование правового регулирования использования технологий на базе ИИ	**Цель:** изучение и регламентация правовых аспектов регулирования деятельности, выполняемой с привлечением технологий, использующих в своей работе ИИ. Исследование юридических последствий процесса работы и оформления результата в случае применения ИИ, в частности в вопросах внешних эффектов, последствий нарушений, признания авторства и др. **Результат:** методологически обоснованные подходы к правовому регулированию решений на базе ИИ																										