УДК 004.852 **DOI** 10.15622/ia.24.1.9

А.В. Пономарев, А.А. Агафонов

АНАЛИТИЧЕСКИЙ ОБЗОР МЕТОДОВ РАСПРЕДЕЛЕНИЯ ЗАДАЧ ПРИ СОВМЕСТНОЙ РАБОТЕ ЧЕЛОВЕКА И МОДЕЛИ ИИ

Пономарев А.В., Агафонов А.А. **Аналитический обзор методов распределения задач при совместной работе человека и модели ИИ.**

Аннотация. Во многих практических сценариях принятие решений исключительно моделью ИИ оказывается нежелательным или даже невозможным, и использование модели ИИ является лишь частью сложного процесса принятия решений, включающего и эксперта-человека. Тем не менее при создании и обучении моделей ИИ этот факт зачастую упускается – модель обучается для самостоятельного принятия решений, а это не всегда является оптимальным. В статье представлен обзор методов, позволяющих учесть совместную работу ИИ и эксперта-человека в процессе конструирования (в частности, обучения) систем ИИ, что более точно соответствует практическому применению модели, позволяет повысить точность решений, принимаемых системой «человек – модель ИИ», а также явно управлять другими важными параметрами системы (например, нагрузкой на человека). Обзор включает анализ современной литературы по заданной тематике по следующим основным направлениям: 1) сценарии взаимодействия человека и модели ИИ и формальные постановки задачи для повышения эффективности системы «человек – модель ИИ»; 2) методы для обеспечения эффективного функционирования системы «человек – модель ИИ»; 3) способы оценки качества совместной работы человека и модели ИИ. Сделаны выводы относительно достоинств, недостатков и условий применимости методов, выявлены основные проблемы существующих подходов. Обзор может быть полезен широкому кругу исследователей и специалистов, занимающихся применением ИИ для поддержки принятия решений.

Ключевые слова: искусственный интеллект, ответственный ИИ, поддержка принятия решений, человеко-машинное взаимодействие, эксперт-человек, распределение задач, совместная работа человека и ИИ, неопределенность модели, нейронные сети, классификатор, обучение с отказом, обучение с делегированием.

1. Введение. Современные решения, основанные на применении искусственного интеллекта (ИИ) в целом и глубоких нейронных сетей в частности, во многих задачах позволяют получать результаты близкие к тем, что могут быть получены человеком, при этом скорость работы и масштабируемость решений, основанных на ИИ, оказывается существенно выше, что обусловливает все более широкое их распространение. Тем не менее полная автоматизация возможна далеко не для всех задач. Среди основных сдерживающих факторов можно выделить следующие. Во-первых, система ИИ действует только на основе той информации, которая преобразована в цифровую форму и доступна системе (а также была использована при конструировании и/или обучении системы). Соответственно, в сложных предметных областях ИИ может столкнуться с неполнотой информации и, как

следствие, с деградацией качества решений, в то время как эксперт может предпринять шаги для выяснения дополнительных фактов. Во-вторых, вопросы ответственности систем ИИ еще не до конца проработаны, а цена ошибки в ряде случаев слишком высока. Все это приводит к тому, что, несмотря на развитие ИИ, в очень многих практических сценариях ИИ работает (и в обозримой перспективе будет работать) совместно с экспертом-человеком, однако при создании и обучении систем ИИ этот факт зачастую упускается. В статье представлен обзор методов учета такой перспективы совместной работы ИИ и человека в процессе конструирования (в частности, обучения) систем ИИ, что позволяет повысить качество решений, принимаемых системой «человек – модель ИИ» [1, 2].

В статье представлены основные результаты аналитического обзора методов в области распределения задач при совместной работе человека и модели ИИ. Проблема совместной работы человека и модели ИИ (или машинного обучения), с одной стороны, довольно интенсивно исследуется в последнее время (причем предлагаются принципиально различные подходы, отличающиеся как особенностями самого сотрудничества, так и решениями по его организации), с другой – имеет довольно богатую историю, которую можно начинать с т.н. обучения с отказом (англ. learning with rejection, rejection learning, learning to reject, learning with abstention, selective prediction) [3, 4]. Подобная задача рассматривается и в российских публикациях, так, например, в [5] предлагается метод отказа от предсказания для задачи непараметрической регрессии. Кроме того, авторы [5] используют словосочетание «делегировать эксперту» для обозначения ситуации, в которой решение задачи передается человеку, если неуверенность модели оказывается высокой. Поэтому в ходе данного обзора термин «делегирование» будет использоваться для обозначения подобных сценариев.

Ввиду большого разнообразия подходов к совместной работе, представляется не вполне целесообразным вмещать их все в одну статью, поэтому данная статья ограничивается рассмотрением проблемы совместной работы и набора решений по ее организации, удовлетворяющих следующим условиям:

– Задан четко определенный класс задач, которые могут решаться независимо как человеком (экспертом), так и моделью ИИ. Примерами такого класса задач может быть диагностика определенного заболевания по медицинским снимкам, принятие решения о выдаче кредита на основе кредитной истории потенциального заемщика и т.п. Таким образом, с задачей можно

связать набор признаков, конкретные значения которых соответствуют экземпляру задачи (обрабатываемому образцу).

– И человек, и модель ИИ могут совершать ошибки. Более того, эффективность (точность) решения задачи из рассматриваемого класса может варьироваться в зависимости от экземпляра задачи (как при ее решении моделью ИИ, так и человеком). Данному условию удовлетворяет большое количество задач, возникающих на практике – действительно, для большинства моделей ИИ есть «сложные» и «простые» образцы (или даже области пространства признаков).

Перечисленным условиям не удовлетворяют, например, работы по определению состава смешанных команд в рамках социо-киберфизических систем [6 – 8], потому что в них речь не идет об обработке однородных задач, принадлежащих одному классу. Этим условиям также не удовлетворяет и постановка, типичная для обучения с отказом, потому что в ней не рассматривается возможность ошибки человека [9]. Тем не менее, перечисленным условиям удовлетворяет множество важных с практической точки зрения задач, что обусловливает актуальность обзора, результаты которого представлены в статье.

Цель статьи состоит в том, чтобы сформировать систематическое изложение ключевых вопросов и современных методов распределения задач между человеком и моделью ИИ (в рамках их совместной работы), что было бы полезно как практикам в области построения систем с элементами ИИ, так и исследователям, позволяя им сориентироваться в палитре существующих методов и определить возможные направления развития. На данный момент подобных обзоров не было обнаружено. Так, близкий по тематике обзор [9] посвящен исключительно обучению с отказом, где не рассматривается возможность ошибки эксперта, в [10] рассматривается ряд методов обучения с делегированием, но статья не претендует на полноту освещения, в [11] рассматривается широкий набор потенциальных сценариев симбиоза человека и ИИ, но достаточно поверхностно. При выполнении данного обзора авторы опирались на методологию систематического обзора литературы [12]. Особенность реализации этой методологии в данном случае связана с перегруженностью ключевых слов, по которым можно идентифицировать искомые публикации, поэтому формирование выборки статей осуществлялось на основе графа цитирования знаковых публикаций, а не отбором по ключевым словам. В ходе исследования осуществлялся поиск ответов на следующие вопросы:

1) Какие рассматриваются сценарии взаимодействия человека и модели ИИ, и, соответственно, какие предлагаются формальные постановки задач для повышения эффективности системы «человек – модель ИИ»?

2) Какие предлагаются методы для обеспечения эффективного функционирования системы «человек – модель ИИ»?

3) Как производится оценка качества совместной работы человека и модели ИИ? В частности, какие применяются специфические метрики для оценки эффективности подобных систем.

Статья структурирована в соответствии с рассматриваемыми вопросами следующим образом. В разделе 2 описана методика проведения обзора, разделы 3-5 представляют результаты ответов на основные вопросы исследования, описывая выявленные постановки задачи совместной работы, конкретные методы обеспечения эффективности и подходы к оценке качества. В заключении подводятся итоги обзора и выявляются наиболее перспективные направления будущих исследований.

2. Методика проведения обзора. Важными характеристиками, определяющими качество обзора литературы, являются, с одной стороны, представительность, то есть соответствующее направление исследований должно быть достаточно полным образом представлено в статьях, включаемых в обзор, с другой – воспроизводимость («идеалом» которой является получение аналогичных результатов любым другим исследователем, осуществляющим обзор на схожую тему). Первая характеристика, как правило, достигается использованием поиска по ключевым словам в достаточно представительных реферативных базах данных (Scopus, Web of Science, РИНЦ), вторая – следованием той или иной методологии проведения обзора и формированием протокола, позволяющего проследить и воспроизвести шаги исследования (например [12, 13]).

При проведении данного обзора авторы придерживались методологии систематического обзора литературы [12] с некоторыми несущественными корректировками, вызванными особенностью области проведения анализа. Эти корректировки связаны, в первую очередь, с определением множества статей, подвергающихся анализу. Традиционная реализация методологии предполагает, что формируется определенный набор ключевых слов, которые используются для отбора статей в одной или нескольких библиографических базах данных. Однако, как уже отмечалось, данная область – совместная работа человека и ИИ – является очень разнообразной, и в ее рамках сосуществует множество принципиально

различных форм и моделей совместной работы, выделить интересующую интерпретацию только на основе ключевых слов оказывается проблематичным, поскольку в данной области отсутствует устоявшаяся терминология. Поэтому для формирования множества исследуемых публикаций был использован граф цитирований. Был выделен набор публикаций (т.н. «ядро»), в которых впервые был предложен и исследован рассматриваемый вариант совместной работы. Затем сформировано множество публикаций, которые ссылаются хотя бы на одну из статей «ядра». В качестве базы цитирований был выбран Google Scholar из-за своего широкого охвата и относительной оперативности индексирования.

Основные шаги обзора и характеристики промежуточных результатов показаны на рисунке 1. К «ядру» были отнесены 4 статьи, в которых рассматриваемый сценарий совместной работы человека и модели ИИ либо был рассмотрен впервые, либо были сделаны важные практические или теоретические замечания относительно построения подобных систем [1, 2, 14, 15]. Три статьи из данного перечня были опубликованы в материалах высокорейтинговых конференций (A* по данным CORE), одна – препринт ArXiv. На момент формирования перечня все они имели более 100 цитирований в Google Scholar (с момента публикации первой из них прошло 5 лет).

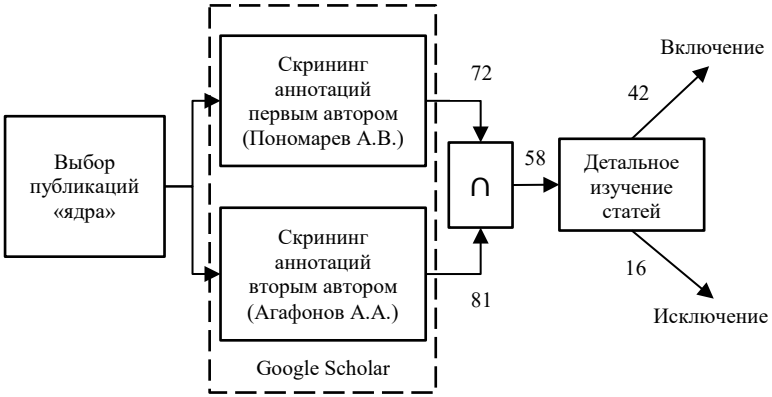

Рис. 1. Порядок проведения обзора

Авторы данного обзора провели независимый скрининг аннотаций всех статей, опубликованных по 2023 г. включительно и цитирующих хотя бы одну из статей «ядра», по данным Google Scholar (всего около 500). Задачей скрининга был отбор статей для

дальнейшего, более детального, изучения. При проведении скрининга отбирались статьи, удовлетворяющие хотя бы одному из следующих критериев: 1) предлагается оригинальный метод; 2) производится сопоставление методов (экспериментальное или теоретическое); 3) предлагается методология сопоставления методов; 4) обзорная статья. В результате каждым из авторов обзора был получен список статей, потенциально подходящих для дальнейшего изучения.

Было сформировано пересечение данных списков, в которое вошли статьи, признанные относящимися к исследуемому сценарию обоими авторами обзора, всего таких статей оказалось 58. В ходе детального изучения еще 16 из них были исключены (часть из них при детальном изучении не удовлетворяла критериям отбора, часть оказалась версиями одной статьи, но под разными названиями). Таким образом, в статье представлены результаты, основанные на структуризации 42 статей на заданную тематику [1, 2, 14 – 53].

Среди отобранных статей большая часть (30 статей) – публикации на конференциях достаточно высокого уровня (CORE A* и A) – AAAI Conference on Artificial Intelligence, NeurIPS, IJCAI, ICML и другие. Другая многочисленная группа – препринты статей, опубликованные на ArXiv. В списке отобранных статей оказалось всего две статьи, опубликованные в журналах: Proceedings of the National Academy of Sciences и Frontiers in Digital Health.

На рисунке 2 приведено распределение отобранных публикаций по годам. Видно, что интерес к данной проблеме постепенно возрастает. Об этом же свидетельствует и статистика источников публикаций – на данный момент среди источников превалируют передовые издания, широкого распространения описываемые методы еще не получили.

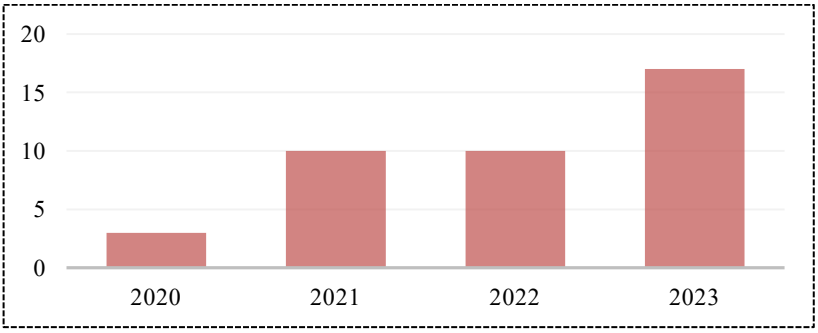

Рис. 2. Распределение количества публикаций по годам

3. Сценарии взаимодействия и формальные постановки. В данном разделе характеризуются основные разновидности и постановки задач, которые различаются как особенностями взаимодействия между человеком и моделью ИИ, так и преследуемой целью (находящей отражение в целевой функции либо в функции потерь). Ведущую роль в структуризации формальных постановок играет сценарий взаимодействия между человеком и моделью ИИ. В рамках каждого из сценариев, в свою очередь, выделяются различные постановки.

Основные критерии, по которым целесообразно структурировать существующие методы распределения задач при совместной работе человека и модели ИИ, представлены на рисунке 3 жирным шрифтом. Каждый конкретный метод может быть позиционирован посредством выбора одной или более категорий по каждому из критериев. При этом следует заметить, что часть критериев относятся к постановке задачи (и рассматриваются в данном разделе статьи), а другая часть («Метод обеспечения совместной работы» и «Тип структуры распределения задач») относятся к пространству решений и рассматриваются в разделе 4.

3.1. Сценарии взаимодействия. Сценарий взаимодействия между человеком и моделью ИИ определяет характер принимаемых решений, последовательность активизации участников системы и доступную им информацию. Можно выделить три вида сценариев: делегирование, последовательная обработка и параллельная обработка.

Под делегированием понимается такой способ организации взаимодействия человека и модели ИИ, когда каждый рассматриваемый образец назначается для обработки либо человеку, либо модели ИИ [18, 19, 27, 29, 31 – 33, 35, 37 – 41, 43, 45, 62]. Этот сценарий является, пожалуй, наиболее часто рассматриваемым в литературе – именно такой сценарий реализуется в рамках обучения с отказом (раздела машинного обучения, в котором хоть и не рассматривается профиль ошибок человека, но уже ставится задача построения классификатора, который бы «воздерживался» от предсказания при недостаточной уверенности), а также в рамках т.н. обучения с делегированием (learning to defer). Идея, обусловливающая востребованность подобного сценария, заключается в том, что при наличии большого количества экземпляров, обработка всех их человеком может быть чересчур дорогостоящей (или требовать слишком большого времени), а обработка моделью – слишком неточной.

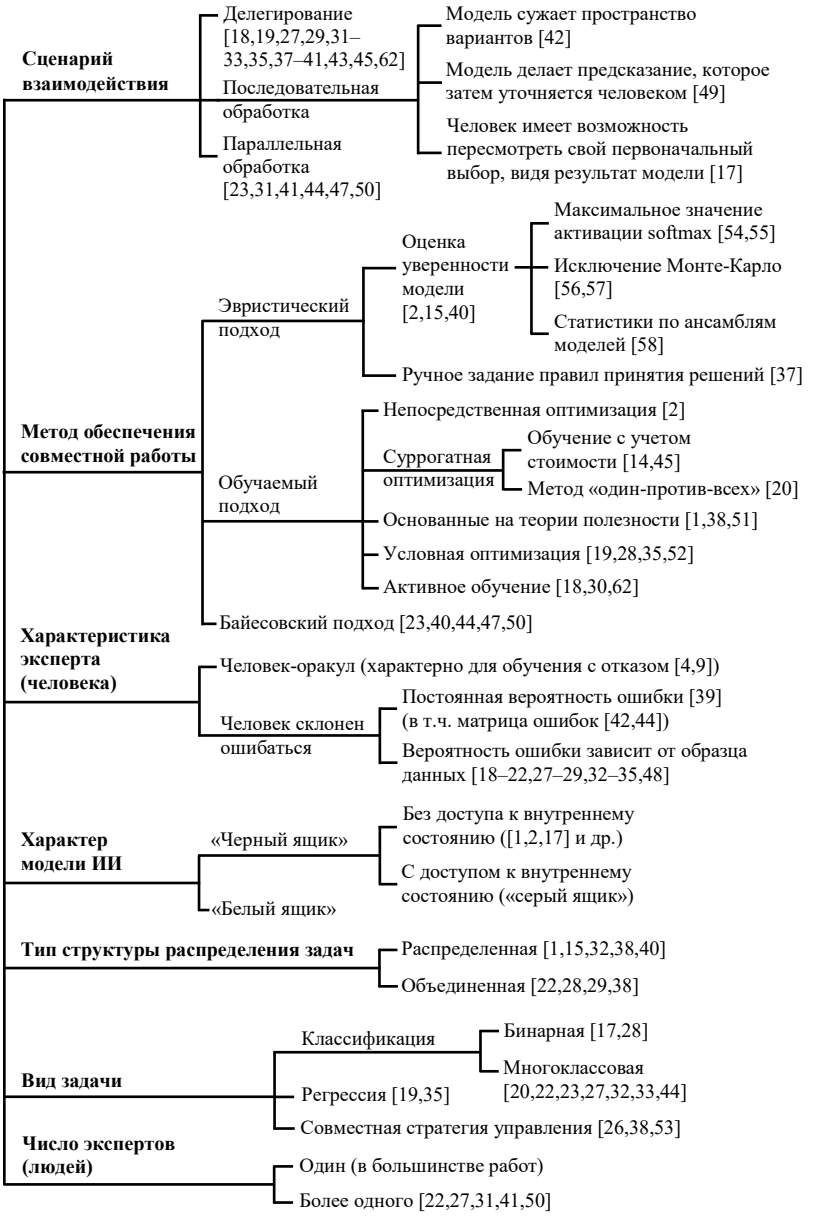

Рис. 3. Методы распределения задач при совместной работе человека и ИИ

Таким образом, в моделях, ориентированных на делегирование, как правило, решается задача поиска определенного компромисса между стоимостью обращения к эксперту (снижение которой достигается назначением образцов модели) и точностью решения задачи (повышение которой, как правило, достигается назначением образцов человеку). Следует, однако, подчеркнуть, что человек в подобных моделях далеко не всегда воспринимается как «оракул», способный дать абсолютно точный ответ (что характерно для более ранних работ в области обучения с отказом), вместо этого при делегировании зачастую учитывается вероятность получения верного ответа от модели и от человека в той или иной области пространства признаков.

Особенностью последовательной обработки является то, что образец поочередно обрабатывается и моделью, и человеком, причем между этими двумя действиями происходит и передача информации. В литературе описано несколько различных сценариев, относящихся к последовательной обработке, однако в большинстве из них итоговое решение остается за человеком, выходные данные модели ИИ используются им для повышения качества принятия решения. Можно выделить следующие разновидности последовательной обработки:

– Модель сужает пространство вариантов, человек выбирает из оставшихся [42]. Данная постановка тесно связана с т.н. conformal prediction. Основной мотив здесь – снизить сложность классификации для человека.

– Модель делает предсказание (возможно, сопровождаемое внутренней оценкой уверенности), а человек, на основе анализа предсказания модели и анализа самого образца, выносит окончательное решение [49]. Это позволяет одновременно и снизить сложность классификации для человека, и потенциально повысить точность.

– Сначала выбор делает человек, потом ему демонстрируется выбор модели и дается возможность пересмотреть решение [17]. Этот вариант характеризуется большей нагрузкой на человека, поскольку ему в любом случае приходится принимать решение, а иногда еще и пересматривать его, но потенциально позволяет повысить качество принятия решений по сравнению с предыдущей разновидностью, так как человек оказывается сильнее вовлечен в решение задачи.

Варьируя последовательность активации участников и характер передаваемой информации в рамках последовательной схемы, можно получить значительное многообразие конкретных сценариев, которые будут отличаться своими свойствами, удобством для человека.

В целом, подобные сценарии характерны для случаев, когда экземпляров решаемых задач не очень много, и гораздо важнее принять верное решение, нежели снизить нагрузку на человека. При разработке и анализе подобных сценариев акцент делается на фактическую эффективность работы человека при наличии той или иной информации, полученной от модели. Исследования здесь носят в значительной степени эмпирический характер и смыкаются с исследованиями в области эффективных человеко-машинных интерфейсов.

Наконец, параллельная обработка предполагает, что каждый экземпляр задачи обрабатывается независимо и человеком, и моделью ИИ, а затем производится автоматическое слияние полученных результатов [23, 31, 41, 44, 47, 50]. Здесь человек также должен обрабатывать все образцы, то есть в подобных методах речь идет не о снижении нагрузки на человека или стоимости, а преследуется преимущественно цель повышения качества принятия решений системой человек – модель ИИ.

3.2. Характеристики и число экспертов (людей). Как уже было указано во введении, в статье рассматривается только такая постановка задачи совместной работы модели ИИ и человека, в которой допускается возможность ошибки человека. При этом, несмотря на общее допущение о возможности ошибки, в разных методах делаются различные предположения относительно характера таких ошибок. Можно говорить о модели ошибок человека, и предположение о структуре этой модели является одной из важных характеристик рассматриваемых в статье методов распределения задач.

Простейшим подобного рода допущением является постоянная вероятность ошибки [51] или, в случае задачи многоклассовой классификации, матрица ошибок, соответствующая эксперту [42, 44].

Более правдоподобным и часто используемым, но и более сложным допущением о поведении эксперта является допущение зависимости вероятности ошибки от образца [18 – 22, 27 – 29, 32 – 35, 48]. То есть, предполагается, что в признаковом пространстве могут быть области, «простые» для данного эксперта, а могут быть «трудные». Причем, в ситуации, когда экспертов несколько, «простые» и «трудные» области разных экспертов могут различаться. Подобное допущение влечет за собой необходимость прямого или косвенного обучения модели, предсказывающей точность эксперта в каждой области признакового пространства, что и делается в большинстве методов.

Другим аспектом, относящимся к экспертам, является их количество. В большинстве статей рассматривается ситуация, в которой есть одна модель ИИ и один эксперт, однако есть и работы, в которых допускается, что экспертов может быть много, причем они могут различаться по своим знаниям и компетенциям. При этом необходимо не только определить то, должен ли образец быть обработан моделью или экспертом, но и выбрать одного (или нескольких) из экспертов [22, 27, 31, 41, 50].

3.3. Вид задачи. В подавляющем большинстве работ рассматривается совместное решение задачи классификации – бинарной [17, 28] или многоклассовой [20, 22, 23, 27, 32, 33, 44].

Совместное решение задачи регрессии рассматривается всего в двух статьях: [19, 35].

Вместе с тем, есть и работы, где речь идет о формировании совместной стратегии управления [26, 38, 53], например, управление осуществляется автоматически (моделью), но в некоторые моменты (в некотором состоянии) оказывается выгодно передать его человеку-эксперту.

3.4. Характер модели ИИ. Класс моделей ИИ, используемых для решения задачи, может накладывать определенные ограничения на метод обеспечения совместной работы. Так, некоторые методы ориентированы на определенные классы моделей (например, SVM) [28], в других – делаются минимальные допущения о характере модели – например, она может быть «черным ящиком», что характерно для большинства случаев.

Можно выделить две разновидности модели «черного ящика»: без доступа к внутреннему состоянию, с доступом к внутреннему состоянию (т.н. «серый ящик»). Первая разновидность характерна тем, что пользователь (или другая модель) может наблюдать только результат модели. Во втором случае появляется возможность использования внутренних представлений модели (например, скрытых слоев нейронной сети) для их последующего анализа или аппроксимации модели.

Модель «белого ящика» предполагает, что процесс и логика принятия решения доступна, и, кроме результата модели, можно видеть то, что привело к его получению.

4. Методы обеспечения совместной работы. Можно выделить три группы методов обеспечения совместной работы: обучаемые, эвристические и байесовские. Первые две группы особенно часто используются в сценарии делегирования, последняя же – наиболее характерна для сценария параллельной обработки. Обучаемые

подходы включают такие методы, где предлагается обучение специальной модели, принимающей решение о том, кто должен обрабатывать образец – человек или модель ИИ. В эвристических методах определяется правило, в соответствии с которым осуществляется делегирование. Простейшим и широко используемым видом подобных правил являются правила, основанные на оценке неопределенности модели. Эвристические методы широко распространены в области обучения с отказом, их адаптация для случая с «неидеальным» человеком зачастую производится путем обучения прокси-модели, позволяющей оценить надежность классификации образца человеком. В этом случае решающее правило просто назначает образец модели ИИ или человеку в зависимости от того, у кого оказывается выше оценка надежности [16].

Потенциальным преимуществом эвристических методов является отсутствие необходимости обучения модели делегирования, однако при допущении «неидеальности» человека, а особенно, зависимости вероятности правильного результата от образца (наличии областей сильной и слабой экспертизы) это преимущество сводится на нет тем фактом, что для эвристических методов требуется получить прокси-модель эксперта, обучение которой требует достаточно много данных о реальных действиях человека.

Распределение задач в ходе совместной работы человека и модели ИИ предполагает принятие двух решений – формирование целевого класса (в случае классификации) на основе признаков образца и определение того, какой из участников системы (человек или модель ИИ) должен обрабатывать заданный образец. Эти решения могут приниматься раздельно (разными моделями) или одновременно (одной моделью), таким образом, сама структура распределения задач может быть либо распределенной (несколько моделей), либо объединенной (одна модель).

Как правило, распределенной структуре соответствует раздельное обучение, то есть обучение системы происходит в два этапа [1, 15, 32, 38, 40]. На первом этапе обучается модель для решения «целевой задачи» классификации без учета возможности делегирования. Для этого не используются ни метки, характеризующие решение задачи человеком, ни специальные функции потерь. На втором этапе обучается модель делегирования, принимающая решение о том, должен ли образец обрабатываться моделью (обученной на первом этапе) или человеком.

При объединенной структуре (характерно для совместного обучения) модель обучается и решению целевой задачи,

и вспомогательной (например, принятие решения о делегировании) с использованием набора данных, включающего результаты обработки образцов человеком-экспертом [22, 28, 29, 38].

Сопоставлению и теоретическому исследованию совместного и раздельного обучения моделей посвящена статья [18]. Основным преимуществом раздельного обучения является его большая универсальность – на первом этапе обучение происходит стандартным образом, не требуя результатов эксперта. Это значит, что раздельные методы могут быть применимы и к моделям, обучение которых не контролируется (полученным от третьих лиц). Достоинством (и основным мотивом развития) подходов, предполагающих совместное обучение, является то, что модель классификации может «фокусироваться» на разделимых регионах, обеспечивая лучшую классификацию в них, при этом изначально уделяя меньше «внимания» регионам, в которых классы оказываются плохо разделимы, оставляя их для человека. Подобный подход особенно хорошо применим к моделям с достаточно низкой выразительной способностью, поскольку позволяет лучше управлять тем, как разделяющая поверхность расположена в пространстве признаков (и выбрать «наилучшей» ту область признакового пространства, где размещение разделяющей поверхности оказывается наиболее целесообразным) [18].

Серьезной проблемой раздельного обучения является то, что этот подход не позволяет модели классификации подстраиваться под область компетенции эксперта. Ограничения раздельного обучения можно проиллюстрировать рисунком 4.

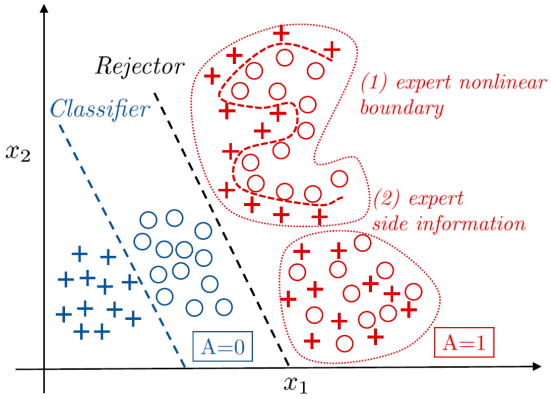

Рис. 4. Иллюстрация преимуществ совместного обучения (из [14])

Скажем, если распределение классов в пространстве признаков выглядит так, как на рисунке 4, то на первом шаге будет очень сложно обучить модель, однако если обучать одновременно и целевую модель (Classifier), и модель делегирования (Rejector), то целевая может оказаться очень простой (линейной), как и модель делегирования.

4.1. Оценка уверенности модели. Одним из простейших эвристических подходов к распределению задач между моделью и экспертом является подход, основанный на оценке уверенности модели. Этот подход в значительной мере унаследован из обучения с отказом, где также применяется достаточно широко. Общая идея заключается в том, чтобы обучить сначала модель для решения целевой задачи, обеспечивающую не только формирование целевой метки, но и соответствующего ей показателя уверенности. А затем установить диапазоны значений уверенности, при которых экземпляр должен перенаправляться эксперту. Смысл в том, чтобы перенаправлять эксперту те образцы, применительно к которым уверенность модели оказывается достаточно низкой.

Существует несколько способов получения уверенности модели, основные из них:

– Максимальное значение многопеременной логистической функции активации (Softmax Response). Применяется обычно к нейросетевым моделям многоклассовой классификации, выход которых формируется с помощью многопеременной логистической функции активации («софтмакс»). Соответственно, на максимальное значение такого выходного слоя может быть установлен порог – если максимальное значение оказывается ниже порога, то сеть «отказывается» от предсказания в пользу эксперта [54, 55];

– Исключение Монте-Карло (Monte-Carlo Dropout, MC-dropout) [56, 57] – оценка уверенности посредством подсчета статистики предсказаний нескольких прямых распространений с дропаутом. Здесь, в частности, используется интерпретация дропаута как техники ансамблирования, собирающей разные сети с разделяемыми весами в один ансамбль. Однако это требует большого количества прямых распространений (сотни), что может быть достаточно затратно.

– Использование статистик по ансамблям моделей [58].

Простой подход, основанный на уверенности моделей и порогах, предложен в [2]. Для обученной модели бинарной классификации устанавливаются два порога: t_0 и t_1. Если предсказание модели оказывается меньше t_0, то формируется отрицательный результат; если больше t_1, то положительный; если же предсказание

оказывается между t_0 и t_1, то образец передается эксперту. Каждая пара порогов оценивается на основе совместной функции потерь, выбирается такая пара, для которой значение функции потерь на обучающем множестве минимально.

Более точный подход из этой группы предлагается, например, в [15] – для каждого образца оценивается уверенность модели, неопределенность при классификации образца экспертом, а потом для модели назначаются те образцы, для которых разница между этими неопределенностями оказывается наибольшей. Для оценки неопределенности при классификации образца экспертом (а ее нужно выполнить до назначения и без реальных оценок экспертов) используется нейронная сеть, на вход которой подаются эмбеддинги объектов, а на выходе – признак несогласия нескольких экспертов [59]. Сеть обучается на наборе данных, для которых есть экспертные оценки. Предложенный подход, учитывающий разницу между обозначенными неопределенностями, используется и в [40], однако, в отличие от [15], здесь рассматривается ряд байесовских методов для вычисления неопределенности модели, а не эвристический подход, основанный на уверенности модели глубокого обучения (нейронной сети).

Общим достоинством всех этих методов является то, что они позволяют использовать существующие модели и добавлять к ним возможность перенаправления эксперту.

4.2. Суррогатные функции потерь. Основным инструментом для обучения моделей, учитывающих возможность переадресации задачи человеку-эксперту, является определение специальной функции потерь, учитывающей наличие экспертных меток. Данные функции потерь учитывают стоимость обращения к эксперту и основываются на сопоставлении вероятности ошибки имеющейся модели и вероятности ошибки человека-эксперта. Составленное таким образом соотношение не всегда оказывается легко оптимизируемым, и с этой целью оно заменяется более удобным в работе приближением, поэтому составляемые таким образом функции потерь часто называются «суррогатными» [1, 14, 20, 27].

Важными аспектами, учитываемыми при разработке и исследовании суррогатных функций потерь являются:

– Так называемая «консистентность» (consistency) по Байесу. Консистентная суррогатная функция потерь – это такая функция потерь, минимизация которой согласуется с оптимальным байесовским классификатором.

– Ведет ли использование функции потерь к получению хорошо калиброванных классификаторов [20]. Так, в [27] показано, что предложенная авторами функция потерь ведет к получению калиброванных классификаторов, а функция потерь, предложенная ранее в [2] – не ведет.

Конструирование суррогатных функций потерь наиболее распространено при решении задачи делегирования. Формальная постановка задачи следующая. Пусть $\mathcal{X}$ – пространство признаков, $\mathcal{Y} = \mathcal{M}$ – пространство меток и меток, даваемых экспертами (K классов), $\mathcal{D} = \{x_n, y_n, m_n\}_{n=1}^{N}$ – набор данных для обучения. То есть, каждый образец набора данных снабжен не только целевой меткой y_n, но и меткой, полученной от эксперта m_n. Целью является обучение двух моделей: классификатора $h: \mathcal{X} \to \mathcal{Y}$ и функции делегирования $r: \mathcal{X} \to \{0,1\}$ (называемой также в литературе rejector). При $r(x) = 0$ окончательное решение принимает классификатор, иначе – эксперт.

«Естественная» функция потерь для обучения этой пары моделей записывается следующим образом [14]:

$$\mathcal{L}_{nat}(h, r) = \mathbb{E}_{x,y,m}[\ell(x, y, h(x))\mathbb{I}_{[r(x)=0]} + \ell_{\exp}(x, y, m)\mathbb{I}_{[r(x)=1]}]. \quad (1)$$

Здесь $\ell(\cdot)$ – функция потерь классификатора, а $\ell_{\exp}(\cdot)$ – функция потерь эксперта. Возможные дополнительные расходы, связанные с привлечением эксперта, могут быть учтены прямо в $\ell_{\exp}$, таким образом, значение этой функции может быть ненулевым даже в случае правильного прогноза. Однако эти дополнительные расходы необходимо выразить в терминах ошибок, что может быть довольно сложно на практике. В любом случае, непосредственная минимизация подобной «естественной» функции оказывается практически невозможной, в первую очередь, в силу дискретного характера функции делегирования r.

В [2] предлагается адаптация $\mathcal{L}_{nat}$, допускающая непосредственную оптимизацию градиентными методами (записано для одного образца):

$$L(x_i, y_i, m_i, h, r) = (1 - r(x_i))\ell(y_i, h(x_i)) + r(x_i)\ell(y_i, h(m_i)). \quad (2)$$

Здесь следует обратить внимание на два важных отличия от $\mathcal{L}_{nat}$. Во-первых, функция делегирования не является бинарной, что, в частности, позволяет использовать эту функцию потерь с градиентными методами, во-вторых, авторы [2] не используют отдельную функцию потерь для экспертной классификации, $\ell(\cdot)$ здесь – это бинарная кросс-энтропия (речь идет о бинарной классификации), поэтому дополнительные расходы на привлечение эксперта никак не учитываются, и речь, по всей видимости, идет просто о максимизации точности. Авторы также описывают несколько тонкостей в обучении модели, среди которых следует выделить следующие: 1) r может зависеть не только от признаков объекта, но и от результата обработки объекта основной моделью $h(x)$, 2) может быть целесообразно ограничить распространение градиента стратегии распределения по h, чтобы h оставался хорошей моделью на всей области значений $\mathcal{X}$ и не происходило деградации качества в тех областях, где целесообразно привлечение эксперта.

Однако напрямую это выражение оптимизировать тяжело, поэтому в [14] предложена суррогатная (но консистентная) функция потерь, основанная на многопеременной логистической функции. Авторы рассматривают задачу многоклассовой классификации (с K классами) и предлагают свести задачу совместной классификации к задаче *cost sensitive learning* (CSS, обучение с учетом стоимости) на расширенном наборе классов. Расширенный набор классов формируется добавлением еще одного класса, означающего перенаправление образца эксперту. Переход осуществляется следующим образом. Для каждого образца вводится понятие стоимости классификации его как каждого из классов ($c(i)$ – стоимость классификации образца как принадлежащего i-тому классу, $i \in \{1,\ldots,K+1\}$). Для образца (x, y) $c(i)$ определяется как $\ell(x, y, \hat{y})$ для классов $\{1,\ldots,K\}$ и как $\ell_{\exp}(x, y, m)$ для дополнительного класса, соответствующего передаче образца эксперту. Авторы предлагают следующую консистентную функцию потерь:

$$L_{CE}(g_1,\ldots,g_{K+1},x,c(1),\ldots,c(K+1)) = \\ = -\sum_{i=1}^{K+1}\left(\max_{j\in[K+1]} c(j) - c(i)\right)\log\left(\frac{\exp(g_i(x))}{\sum_k \exp(g_k(x))}\right). \quad (3)$$

Здесь g_i – это выходы модели (авторы предполагают, что это нейронная сеть). Основу данной целевой функции составляет

многопеременная логистическая функция («софтмакс»), применяемая к выходам модели, поэтому данная функция потерь также называется «софтмакс-параметризацией».

В статье [20] показано, что модели, обученные с помощью софтмакс-параметризации не является калиброванными, поэтому предложен другой вариант суррогатной функции потерь, т.н. «один-против-всех» (one-vs-all, OvA):

$$L_{OvA}(g_1,\ldots,g_{K+1},x,y,m)=\phi[g_y(x)]+\sum_{y'\in Y,y'\neq y}\phi[-g_{y'}(x)]+ \\ +\phi[-g_{K+1}(x)]+\mathbb{I}[m=y](\phi[g_{K+1}(x)]-\phi[-g_{K+1}(x)]), \tag{4}$$

где ϕ – бинарная суррогатная функция потерь (например, логистическая функция). Неформально, логика этой функции потерь заключается в следующем: первое слагаемое обеспечивает повышение выходного значения для правильного класса (g_y), второе слагаемое (оператор суммирования) обеспечивает понижение выходного значения для ошибочных классов, третье и четвертое слагаемое в комплексе обеспечивают повышение выходного значения для выхода модели, связанного с перенаправлением эксперту (g_{K+1}), если ответ эксперта для данного примера правильный, и понижение значения этого выхода, если ответ эксперта неверный.

Сами классификатор и функция делегирования устроены в любом случае одинаково:

$$r(x)=\mathbb{I}[g_{K+1}(x)\geq\max_k g_k(x)], \\ h(x)=\arg\max_{k\in\{1,\ldots,K\}} g_k(x). \tag{5}$$

В статьях [27, 33] данные функции обобщены на случай нескольких экспертов – такие модели не просто принимают решение передать ли образец эксперту, но и какому именно эксперту его передать.

В статье [39] показано, что модели делегирования, обученные с помощью существующих суррогатных потерь (CSS и OvA), могут быть склонны к недообучению в тех случаях, когда обращение к экспертам влечет за собой дополнительную стоимость. В связи с этим, предлагается способ ретроспективной коррекции суррогатных потерь как для CSS, так и для OvA.

В статье [45] делается успешная попытка улучшения подхода [14] для его использования в сочетании с конкретными людьми в рамках распределения задач. Предлагаемое улучшение, заключающееся в тонкой настройке, повышает общую точность системы «человек – модель ИИ». Для этого модель эксперта сначала обучается с использованием агрегированных человеческих меток, а затем – с использованием меток, полученных от конкретных людей.

В [52] демонстрируется, что существующие подходы не всегда могут совместно оптимизировать классификатор и модель делегирования с низкой ошибкой неправильной классификации (даже в том случае, если существуют линейный классификатор и соответствующая модель делегирования, обеспечивающие безошибочную классификацию). Для решения этой проблемы задача делегирования рассматривается как задача смешанного целочисленного линейного программирования и предлагается новая консистентная суррогатная функция потерь, которая обеспечивает лучшую эмпирическую производительность по сравнению с существующими суррогатными подходами.

4.3. Методы, основанные на теории полезности. В ряде работ [1, 38, 51] для конструирования задачи оптимизации используется теория полезности. Достоинством такого подхода является то, что само назначение стоимости ошибки, стоимости обращения к эксперту может быть оценено напрямую из знаний предметной области. Однако на практике модель, построенная в терминах теории полезности не всегда напрямую оказывается хороша для использования при обучении стратегии делегирования, поэтому она может быть адаптирована, заменена суррогатной функцией, во многом с использованием тех же идей, что изложены в предыдущем подразделе.

Так, авторы [1] предлагают следующую формулировку задачи делегирования в терминах теории полезности:

$$\arg\max_{r,h} \mathbb{E}_{(x,y,m)\sim P}\left[r(x)(u(y,m)-c)+(1-r(x))(u(y,h(x)))\right]. \qquad (6)$$

Здесь $u(y,m)$ – полезность ответа эксперта m при верном ответе y, $u(y,h(x))$ – полезность результата модели при верном ответе y, c – стоимость обращения к эксперту. Можно отметить, что это выражение очень похоже на $\mathcal{L}_{nat}$, поскольку оба выражают основную идею делегирования.

Авторы [1] также предлагают целую палитру методов обучения h и r, выделяя дискриминативные подходы, в которых эти функции обучаются непосредственно отображению признаков в решения без построения промежуточных вероятностных моделей различных компонентов системы, и вероятностные подходы, основанные на стоимости информации.

Фиксированный дискриминативный подход заключается в том, что сначала обучается модель h (любым известным методом), а затем – модель r с использованием сформированной функции ожидаемой полезности (подход аналогичен предложенному в [15]).

Объединенный дискриминативный подход предполагает совместное обучение h и r. Авторы также сталкиваются с тем, что непосредственная оптимизация затруднительна, поэтому для поиска модели h и стратегии делегирования r используют следующую суррогатную функцию потерь:

$$\ell(y, r(x)m + (1 - r(x))h(x)) + cr(x). \quad (7)$$

В ходе работы для принятия решений о направлении задачи эксперту авторы аппроксимируют идеализированное предсказание с использованием меры уверенности модели, $\max(h(x))$. Запрос эксперту посылается тогда, когда $(1 - r(x))\max(h(x)) < r(x)$. То есть, запрос посылается эксперту, если $r(x)$ имеет большое значение либо если неопределенность предсказания высока.

В выделяемом авторами [1] подходе, основанном на стоимости информации, предполагается независимое обучение трех вероятностных моделей: модели распределения меток при условии известных значений признаков $(p_\alpha(y \mid x))$; модели ответа эксперта при условии известных значений признаков $(p_\beta(y \mid x))$ и модели распределения меток при условии известных признаков и ответов экспертов $(p_\gamma(y \mid m, x))$. Для построения моделей авторы предлагают использовать нейронные сети с последующей вероятностной калибровкой методом Платта [60]. Во время выполнения эти вероятности используются для оценки ожидаемой полезности обращения к эксперту.

В [51] на основе теории полезности производится формализация последовательного подхода к сотрудничеству, когда каждый образец обрабатывается сначала моделью, а потом эксперт, зная результат работы модели, принимает решение о том, стоит ли просто принять

его или детально исследовать образец и выполнить классификацию самостоятельно. Авторы составляют матрицу выигрышей (таблица 1), где под метарешением понимается решение эксперта о том, стоит ли доверять модели, обработка образца связана с затратой усилий $\lambda > 0$. Само же решение – это результат системы ИИ-человек, и он может быть либо правильным (этому случаю соответствует максимальная полезность 1), либо неправильным (чему соответствует стоимость ошибки $\beta \geq 1$).

Таблица 1. Матрица полезности (из [51])

Метарешение\Решение	**Правильно**	**Неправильно**
Принять (Accept, A)	1	$-\beta$
Решать самому (Solve, S)	$1-\lambda$	$-\beta-\lambda$

Оптимальный классификатор в такой постановке должен максимизировать ожидаемую полезность:

$$h^* = \arg\max_h \mathbb{E}_{x,y}[U(m,d)], \tag{8}$$

где m – функция, в соответствии с которой принимается метарешение (принять или решать самостоятельно), а d – итоговое решение.

Опираясь на требование к оптимальному классификатору и допущение о рациональности человека (и, следовательно, определенную стратегию принятия решений), авторы записывают общее выражение для ожидаемой полезности и предлагают оптимизировать его непосредственно в ходе градиентного спуска. Авторы столкнулись с тем, что непосредственная оптимизация оказалась затруднительна, поскольку при случайной инициализации модель «неуверенна», а значит, решать для всех образцов должен человек, и в этой области нет градиентов для обучения модели, поэтому они начали с модели, обученной для решения задачи без человека.

Схожие модели, основанные на теории полезности, используются и при рассмотрении процессов совместной работы. Модель на основе теории полезности здесь, как правило, сочетается с уравнениями Беллмана [26, 53].

В отличие от большинства работ, которые рассматривают задачу обучения с учителем (например, классификацию), статья [38] фокусируется на проблеме выбора оптимальной стратегии делегирования в контексте обратной связи типа «бандит», при которой

вознаграждение и результаты зависят от всех предыдущих действий человека. Это требует оценки альтернативных вариантов действий и выбора тех действий, которые приведут к наибольшему ожидаемому вознаграждению. Например, образец данных может представлять пациента, в отношении которого агент (человек, принимающий решение, или модель ИИ) может предпринять какое-либо действие (один из методов лечения) и затем получить соответствующее вознаграждение (эффект от лечения). Для нахождения стратегии делегирования авторы максимизируют средневзвешенное вознаграждение человека и модели, подобно тому, как было показано в формуле (6). При этом рассматривается как случай раздельного обучения модели делегирования и модели принятия решений, так и их совместного обучения.

4.4. Условная оптимизация. В предыдущих подразделах в процессе обучения модели, управляющей совместным решением задач, нагрузка на эксперта учитывалась опосредованно – в виде штрафа в совместной целевой функции за обработку образца экспертом или отдельного слагаемого в функции полезности. Однако в некоторых случаях подобные веса назначить сложно и, более того, может существовать физическое ограничение на количество образцов, которые могут обрабатываться экспертом [19, 28, 35].

В статье [35] предлагается формальная постановка для решения задачи регрессии при наличии такого ограничения, а в [28] – классификации. Так, в [28] рассматриваются классификаторы на основе отступа (margin) между классами (например, SVM). Пусть $\mathcal{V}$ – обучающее множество, S – часть обучающего множества, которая при обучении будет передана экспертам, n – ограничение сверху на количество элементов в S. Тогда распределение образцов между моделью h_θ и экспертом сводится к тому, чтобы выбрать некоторое множество обучающих образцов $S \in \mathcal{V}$, которые будут передаваться эксперту ($|S| \le n$), и построить решающую поверхность (decision boundary), разделяющую векторы признаков в подмножестве обучающего множества $S^c = \mathcal{V} \setminus S$. Целевая функция может быть записана так:

$$\min_{S,\theta} \sum_{i \in V \setminus S} \ell(h_\theta(x_i), y_i) + \sum_{i \in S} c(x_i, y_i), \quad s.\,t.\ |S| \le n, \tag{9}$$

где $c(x_i, y_i)$ – ошибка человека на образце (human error per sample).

Интересно, что n устанавливается относительно обучающего множества, причем, во-первых, фактически экспертные метки все равно должны быть известны для всех образцов обучающего множества (чтобы выбрать те, которые целесообразно исключить при обучении модели); во-вторых, на практике гораздо большую ценность играет ограничение количества задач, назначаемых эксперту во время выполнения. Здесь авторы опираются на то, что $\mathcal{V}$ является представительной выборкой из исходного распределения, и выбранные для назначения эксперту образцы задают область пространства признаков, в которую попадает приблизительно $n/|\mathcal{V}|$ образцов как обучающего, так и тестового множеств.

Общее выражение минимизации конкретизируется для случая SVM, и показано, что в этом случае выбор S образцов может быть осуществлен жадным алгоритмом.

Для принятия решения о том, стоит ли назначать новый (не виденный ранее) образец человеку или обрабатывать его моделью (во время вывода) предлагается обучить еще одну модель $\pi(d \mid x)$. Модель обучается на основе набора данных $\{(x_i, d_i\}_{i \in \mathcal{V}}$, где x_i – признаки объектов (те же самые, как и в основной задаче – обучении с делегированием), а $d_i = +1$, если $i \in S^*$ (фактическое множество образцов, назначенных эксперту, в результате решения задачи условной оптимизации) и $d_i = -1$ в противном случае. Считается, что эта модель хорошо аппроксимирует распределение $p(x)\pi(d = -1 \mid x)$ (распределение объектов, хорошо классифицируемых основной моделью).

В [19] для решения подобной задачи оптимизации предлагается градиентный алгоритм, который итеративно оптимизирует классификатор в образцах, где он превосходит человека в обучающей выборке, а затем обучает модель делегирования, чтобы предсказать, у человека или у модели ИИ будет более высокая ошибка на уровне каждого образца. Авторы показывают, что алгоритм гарантированно находит прогнозирующие модели и политики делегирования, с учетом ограничения на число элементов в S.

4.5. Активное обучение. Получение образцов, размеченных экспертом, как правило, достаточно трудоемкий и затратный процесс, в большинстве же подходов, рассмотренных ранее, предполагалось, что для всего обучающего множества присутствуют экспертные метки. Часть этих меток может оказаться избыточной, поэтому перспективным подходом является применение различных методов

и техник, позволяющих снизить зависимость от экспертной разметки. Одним из таких методов является активное обучение – модель запрашивает разметку именно тех образцов, которые оказываются наиболее полезными с точки зрения построения модели ошибок эксперта. Так, в статье [18] предложен алгоритм делегирования, основанный на активном обучении. Алгоритм включает два этапа:

– на первом запускается стандартный алгоритм активного обучения (например, CAL [61]) для пространства $\mathcal{D}$, чтобы получить функцию f несоответствия предсказаний эксперта и эталонных ответов с ошибкой не более ϵ;

– на втором этапе данные размечаются этой функцией $\hat{f}$, и на основе этих данных обучается пара классификатор-модель делегирования.

В статье [62] предлагается трехэтапный подход к сокращению количества экспертных прогнозов, необходимых для обучения алгоритмов делегирования. Он включает в себя следующие шаги (этапы):

1. Обучение модели встраивания (embedding model) с метками (ground truth), которые используются для извлечения представлений признаков.

2. Представления признаков служат исходными данными для обучения модели прогнозирования экспертных знаний (expertise predictor model), чтобы аппроксимировать возможности эксперта-человека.

3. Модель прогнозирования экспертных знаний генерирует искусственные экспертные прогнозы для экземпляров, не размеченных экспертом-человеком.

Затем для обучения алгоритмов делегирования можно использовать как человеческие, так и искусственные экспертные прогнозы. Таким образом, в отличие от [18], здесь не требуется итеративное выявление образцов, для которых запрашиваются прогнозы экспертов-людей. Вместо этого, учитывая небольшое количество прогнозов экспертов-людей, алгоритм учится выводить искусственные прогнозы для неразмеченных образцов в обучающем наборе данных.

Другая работа [30], рассматривающая возможность активного обучения, посвящена онлайн-прогнозированию консенсуса группы экспертов. Предполагается, что консенсус экспертов-людей определяет исключительно метку образца, которую необходимо предсказать. Поскольку запрос полного консенсуса может быть затратным, авторы динамически оценивают консенсус на основе

частичной обратной связи, анализируя уверенность экспертов и модели ИИ. Авторы ищут компромисс между стоимостью обращения к экспертам и точностью классификации. Таким образом, цель работы состоит в том, чтобы максимально повысить точность предсказания консенсуса при ограниченном «бюджете» на аннотации экспертов.

4.6. Слияние данных. В рамках подхода параллельной обработки можно выделить группу методов т.н. слияния данных. Как отмечалось ранее, основная цель слияния данных состоит в том, чтобы повысить точность принятия решений системой «человек – модель ИИ». Наиболее распространенные методы этой группы сосредоточены на комбинировании предсказаний модели ИИ с метками, предсказанными людьми. При этом в процессе классификации может участвовать как один эксперт-человек [23, 44, 47], так и множество экспертов [31, 41, 50].

Для комбинирования предсказаний чаще всего используется байесовская статистика, предполагающая, что вероятность, которая отражает степень доверия событию, может изменяться в зависимости от некоторой дополнительной информации. Так, в статье [44] рассматривается задача многоклассовой классификации изображений, где решения по категориальной классификации независимо принимают один эксперт-человек (предсказывают только метку) и одна модель классификации, прогнозирующая распределение по всем возможным меткам (классам). Для объединения предсказаний используется вероятностный подход, при котором условное распределение по предсказываемым меткам может быть учтено с помощью правила Байеса следующим образом:

$$p(y \mid h(x), m(x)) \propto p(h(x) \mid y, m(x)) p(y \mid m(x)), \tag{10}$$

где $x \in \mathcal{X}$ – образец набора данных; $y \in \mathcal{Y}$ – истинная метка; $h(x) \in \mathcal{Y}$ – метка, предсказанная человеком; $m(x) \in \mathbb{R}^K$ – нормированный вектор вероятности, выводимый моделью ИИ (K – число классов).

Важно заметить, что далее авторы делают допущение об условной независимости $h(x)$ и $m(x)$ при y, в соответствии с которым приведенное выше выражение может быть преобразовано к следующему виду:

$$p(y \mid h(x), m(x)) \propto p(h(x) \mid y) p(y \mid m(x)). \tag{11}$$

Слагаемое $p(h(x)\,|\,y)$ можно интерпретировать как калиброванные вероятности на уровне класса. $p(h(x)\,|\,y)$ параметризуется матрицей ошибок эксперта h, которая обозначается как φ и содержит элементы $\varphi_{ij} = p(h(x) = i\,|\,y = j)$. Второе слагаемое $p(y\,|\,m(x))$ можно интерпретировать как калиброванные вероятности на уровне образца. Однако вероятностный результат классификатора $m(x)$ может отличаться от $p(y\,|\,m(x))$. В связи с этим, авторы предлагают процедуру post-hoc калибровки, которая сопоставляет $m(x)$ с хорошо откалиброванными вероятностями с помощью т.н. калибровочной карты с параметрами θ. Вывод классификатора после применения такой калибровочной карты обозначается как $m^{\theta}(x)$.

Наконец, прогнозируемая вероятность класса j, учитывая, что человек предсказывает класс i, и модель создает вектор вероятности классов $m(x)$, будет определяться следующим выражением:

$$p(y = j\,|\,h(x) = i, m(x)) = \frac{\varphi_{ij} m_j^{\theta}(x)}{\sum_{k=1}^{K} \varphi_{ik} m_k^{\theta}(x)}. \qquad (12)$$

Хотя наиболее простой оценкой элементов матрицы ошибок является оценка максимального правдоподобия, при малом количестве человеческих меток данная оценка будет иметь большую дисперсию. Вместо этого авторы предлагают байесовский подход к включению априорной информации, однако в рамках данного обзора он не представляет большого интереса.

В [23] предлагается подход байесовского моделирования, с помощью которого формируется комбинированный прогноз, а также оценки скрытой корреляции между классификаторами. Эта корреляция отражает зависимости между показателями достоверности классификации людей и моделей ИИ. Рассматривается сценарий, в котором, кроме предсказанной метки, человек предоставляет свою степень уверенности («низкая», «средняя», «высокая»). Таким образом, в отличие от [44], предлагаемая байесовская модель оценивает корреляцию между уверенностью человека и модели ИИ и, кроме того, не опирается на предположение об условной независимости.

В [50] отмечается, что, хотя комбинированная модель в [44] обеспечивает гораздо большую точность, чем при независимой классификации человеком и моделью ИИ, она ограничивается

объединением предсказания лишь одного человека с результатами модели, что может существенно снизить точность комбинированного подхода, поскольку результат классификации зависит от точности конкретного человека. В данной же статье ([50]) предлагается подход к объединению решений множества людей с результатом модели ИИ. Кроме того, предлагается эффективный алгоритм поиска оптимальной подгруппы людей, чьи объединенные метки позволят получить наиболее точный результат классификации.

В работах [31, 41, 47] также рассматривается задача комбинирования предсказаний множества людей с результатами модели ИИ, однако в них также уделяется немало внимания аспекту делегирования. Поэтому более подробно данные методы рассматриваются в пункте 4.8.

4.7. Ручное конфигурирование границ принятия решений. Ручное конфигурирование границ принятия решений подразумевает под собой то, что человек принимает непосредственное участие в определении области признаков, для которой модель ИИ в дальнейшем может осуществлять предсказания. Если рассматриваемый образец не попадает в границы данной области признаков, то модель ИИ делегирует задачу человеку.

В рамках данного обзора была обнаружена всего одна работа [37], в которой рассматривается подобный подход применительно к задаче модерации контента. Авторы называют этот подход «условным делегированием».

Области, определяющие границы принятия решений модели, задаются с помощью набора правил на основе ключевых слов, созданного в результате совместной работы человека и модели ИИ перед развертыванием. Например, после проверки прогнозов модели по комментариям со словом «отсталый» человек может решить, что модель хорошо справляется с их выявлением, и установить «отсталый» в качестве правила условного делегирования. После развертывания комментарии, относящиеся к этим областям, т.е. содержащие любые ключевые слова, указанные пользователем, могут быть использованы для принятия окончательных мер, таких как скрытие или отправка на дальнейшую проверку.

4.8. Гибридные подходы. Под гибридными подходами следует понимать подходы, которые, так или иначе, сочетают в себе два или более ранее рассмотренных методов обеспечения совместной работы человека и ИИ. Соответственно, каждый из подобных подходов может в себе воплощать сразу несколько сценариев взаимодействия человека и ИИ.

В рамках данного обзора было обнаружено несколько работ, причем в каждой из них внимание уделяется гибридизации методов слияния данных и делегирования, т.е. все они реализуют сразу два сценария: параллельная обработка и делегирование.

Авторы в [47] выделяют ряд ограничений ранее рассмотренного подхода комбинирования предсказаний [44]. В частности, в качестве недостатка отмечается то, что для каждого образца требуются метки, полученные от человека, что может представлять трудности, если эти метки недоступны в достаточном количестве. Кроме того, при наличии значительного разрыва между точностью человека и модели ИИ, одно может преобладать над другим, например, комбинированная модель может начать полагаться на менее точных людей. Наряду с этим, авторы отмечают, что типовые подходы к делегированию полностью игнорируют результаты модели ИИ в том случае, когда задача адресуется человеку. Поэтому в [47] предлагается способ объединения двух подходов: обучение с делегированием [14] и комбинирование предсказаний [44]. Общая идея заключается в том, что откалиброванные выходные данные модели ИИ объединяются с метками, полученными от людей, только в том случае, если было принято решение делегировать задачу человеку.

В [31] предлагается расширение подхода обучения с делегированием, основанного на использовании суррогатных функций потерь, для случая множества экспертов. Так, принятие решения о классификации может быть делегировано одному или нескольким экспертам (при этом сама модель ИИ также рассматривается в качестве эксперта). Окончательным результатом здесь является совокупное решение выбранного подмножества экспертов. Для получения совокупного решения в статье рассматриваются несколько методов формирования весов экспертов.

В [41] рассматривается подход интеграции обучения с участием нескольких экспертов [63] и обучения с использованием зашумленных меток [64, 65], то есть предполагается, что истинные метки могут отсутствовать (часто характерно для реальных наборов данных). Предлагаемый подход оптимизирует систему «человек – модель ИИ», стремясь повысить точность классификации при минимизации затрат на обращение к эксперту-человеку, которые варьируются от 0 до M, где M – максимальное количество экспертов-людей.

5. Оценка качества и валидация методов совместной работы. Процедура оценки качества методов и алгоритмов, используемых при распределении задач между человеком и ИИ также имеет рад особенностей: во-первых, качество можно оценивать по

двум зачастую взаимоисключающим направлениям – точность итоговой модели и нагрузка на человека, во-вторых, помимо исходных данных и эталонного результата для оценки нужны еще данные о решении задач экспертом, что не всегда возможно, поэтому в исследованиях зачастую применяются различные способы моделирования ответов эксперта на основе эталонных результатов, которые также охарактеризованы в данном разделе.

5.1. Метрики и процедуры оценки качества

5.1.1. Один показатель. Как уже указывалось, при совместном выполнении задач, как правило, важна не только общая точность, но количество задач, выполняемых человеком. Однако в ряде постановок метрики качества учитывают только точностные характеристики итоговой модели (связанные с долей ошибок). В наибольшей степени это характерно для двух ситуаций:

– целью является получение наиболее надежной модели для ответственных сценариев применения;

– человек все равно оказывается вовлечен в принятие решения по всем задачам (в последовательном или параллельном сценариях) (например, [16, 17]).

В работах [16, 17, 22, 27, 32, 33] в качестве главной метрики оценки модели применяется только точность (в стандартном смысле), определяемая как отношение количества образцов, при которых предсказание итоговой модели совпало с эталонным результатом, к общему количеству образцов, использованных при оценке.

5.1.2. Несколько показателей. В большинстве статей (особенно, в тех случаях, когда рассматривается сценарий делегирования) оценка производится с помощью двух метрик, описывающих долю задач, решенных человеком, и общую точность системы. Данная система метрик во многом унаследована из области обучения с отказом. Трудоемкость для человека, обычно, характеризуется через метрику, называемую «покрытие» (англ. coverage), определяемую как доля образов, которые были обработаны автоматически моделью (не переданы на обработку эксперту). Покрытие, соответственно, может изменяться от 0 (когда все образцы были обработаны экспертом) до 1 (когда все образцы обработаны моделью).

Поведение модели совместной работы при определенных настройках модели делегирования на определенном тестовом наборе данных, таким образом, может быть охарактеризовано парой характеристик: покрытие и точность (или количество ошибок). При варьировании настроек модели могут быть построены кривые,

характеризующие баланс между покрытием и точностью, широко используемые при анализе различных методов [20, 28, 29, 36].

В статье [24] предложена оригинальная метрика DEV (deferred error volume), сочетающая точность и покрытие, которая определяется как площадь под кривой, образованной оценками качества для разных комбинаций вероятности делегирования и порога на делегирование.

5.1.3. Свертка (на основе теории полезности). В подходах, основанных на теории полезности, все значимые характеристики (ошибка модели, назначение эксперту и пр.) выражены в рамках единой функции полезности, поэтому в качестве основной метрики используется также значение функции полезности. Это может быть ожидаемая полезность [51, 53], вычисляемая с помощью вероятностных выходов модели, а может быть эмпирическая полезность [51], вычисляемая уже с учетом примененных порогов (делегирования и классификации).

Сюда же относятся и онлайн-модели, например, основанные на обучении с подкреплением. Для подобных формализаций естественно сводить задачу к поиску экстремума свертки функции полезности, и оценка производится либо с помощью значения этой функции [21, 26] (вознаграждения, полученного агентом, который осуществляет распределение задач), либо через сожаление (regret) [43], характеризующее поведение модели по сравнению с идеальной.

5.2. Наборы данных. Валидация рассматриваемого класса моделей совместной работы эксперта и модели ИИ требует наличия не только эталонных меток, но и экспертных (которые могут отличаться от эталонных, отражая неполноту знаний эксперта), потому что в некоторых случаях (в зависимости от результата модели делегирования, например) при выполнении результирующей модели может происходить обращение к эксперту. Заранее неизвестно, при обработке каких именно образцов такое обращение целесообразно, поэтому экспертные метки должны быть определены для всех. Существует относительно немного публичных наборов данных, содержащих такие метки, поэтому в части исследований используются синтетические экспертные метки, полученные в результате применения какой-либо модели ошибок к эталонным меткам. Впрочем, моделирование поведения эксперта оказывается полезным не только при полном отсутствии экспертных оценок, как правило, при исследовании метода оказывается важно как именно он ведет себя при различной надежности эксперта, и моделирование используется для имитации ответов с разной надежностью [20, 27, 28, 32, 33].

5.2.1. Наборы данных, содержащие экспертные оценки. Наборы, содержащие экспертные оценки, можно, в свою очередь, охарактеризовать с помощью трех важнейших признаков: количество задействованных экспертов, полнота разметки набора каждым экспертом, и количество экспертных мнений на образец.

Наиболее распространенной категорией таких наборов являются такие, где экспертов задействовано много, причем каждый размечает не все образцы, но на каждый образец приходится несколько экспертных оценок – зачастую подобные наборы являются результатом использования краудсорсинга, где участникам площадки, как правило, за определенное вознаграждение, предлагается провести классификацию образцов какого-либо публичного набора данных. Подобным образом на основе известных наборов данных в области компьютерного зрения CIFAR и ImageNet были получены вариации, содержащие экспертные метки: CIFAR-10H [66] и ImageNet-16H [23]. Набор данных Galaxy Zoo [67, 68] получен в рамках проекта гражданской науки по классификации изображений галактик на одноименной площадке.

Поскольку в таких наборах данных каждый эксперт размечает не все объекты, возможности моделирования точности отдельного эксперта оказываются очень ограничены, такие наборы больше подходят для некоторого «усредненного» моделирования человеческого взгляда на задачу. То, что для каждого образца присутствует несколько меток, позволяет построить распределение, связанное с образцом и использовать это распределение в качестве основы для модели эксперта [42, 48].

Набор данных CIFAR-10H применяется в работах [21, 42, 44, 48], ImageNet-16H в работах [23, 44], а GalaxyZoo – в [19, 20, 27].

Похожая схема также у набора, используемого в статье [22] – это набор данных рентгеновских снимков ChestX-ray8 [69, 70]. Аннотация каждого снимка содержит оценки трех специалистов (из 22 задействованных в разметке) и согласованную оценку.

5.2.2. Полностью синтетические наборы данных. В ряде публикаций используются полностью синтетические наборы данных [18, 21, 28, 35, 42, 43, 48, 51]. Как правило, валидация на таких наборах данных дополняет валидацию на реальных наборах (например, [28, 35, 43]), однако в отдельных случаях используются только синтетические наборы.

Специальным образом сконструированный набор позволяет смоделировать некоторую интуитивную ситуацию, послужившую толчком к созданию метода (касающуюся распределения признаков,

распределению точности оценок эксперта), и подтвердить эффективность метода, по крайней мере, для этой ситуации.

В [25] предлагается набор данных *Financial Fraud Alert Review* (FiFAR) – синтетический набор для обнаружения мошенничества с банковскими счетами, содержащий предсказания 50 (смоделированных) «экспертов», обладающих различными характеристиками.

5.2.3. Моделирование ответов эксперта. Моделирование ответов эксперта используется с двумя основными целями: во-первых, оценка того или иного метода на наборах данных, не имеющих оценок экспертов (имеющих только эталонные метки); во-вторых, для исследования влияния точности эксперта на результат работы системы в целом. Наличие модели позволяет управлять уровнем точности и, соответственно, производить контролируемый эксперимент.

Формирование экспертных меток варьируемой надежности для реальных наборов данных, в которых есть только одна (эталонная) метка применяется в работах [20, 22, 27, 28, 48].

Можно выделить несколько распространенных приемов моделирования эксперта ограниченной точности. В части таких приемов не предполагается, что в наборе данных есть экспертные метки – для каждого образца присутствует только одна эталонная.

Эксперт, выбирающий класс случайным образом [27]. Как правило, такая модель используется либо в качестве достаточно слабого базового решения при сравнении, либо в ситуации, когда экспертов может быть много, чтобы показать, что модель в состоянии идентифицировать такого эксперта и ограничить его влияние на итоговый результат.

Эксперт, обладающий специализацией (например, [27]). Широко используемая в задачах многоклассовой классификации модель. В этой модели все множество выходных классов разбивается на два подмножества: те, которые данный эксперт различает хорошо и те, которые он различает не так хорошо. Для каждого из подмножеств задается вероятность правильного предсказания эксперта. Легко заметить, что подобная модель может являться обобщением и случайного эксперта (для него область специализации связана с пустым множеством классов) и «всезнающего» (область специализации связана с множеством, включающим все классы).

В качестве моделирования экспертов разной квалификации применяются также нейронные сети разной выразительности [32, 33] – сеть с большим количеством параметров соответствует более квалифицированному эксперту.

Для наборов данных, в которых присутствуют метки людей, основным способом моделирования «силы» эксперта является использование вероятностной смеси случайного выбора и сэмплирования из эмпирического распределения, задаваемого метками, которые люди-эксперты назначили данному образцу [27]. То есть с некоторой вероятностью p выбирается одна из реальных меток, которые были присвоены образцу, а с вероятностью $(1-p)$ – случайная метка. Очевидно, максимальное значение параметра p соответствует некоторой средней точности эксперта, моделировать более точные данные таким образом не получится.

5.3. Онлайн-валидация. Позволяющим получать достоверные результаты, но довольно трудоемким и дорогостоящим способом оценки методов и моделей совместной работы является онлайн-оценка, проводимая, как правило, с помощью краудсорсинга [16, 17, 31, 34, 36, 37, 49].

Суть подобной оценки заключается в том, что авторы реализуют предлагаемый метод сотрудничества с помощью инструментов той или иной площадки краудсорсинга – Amazon Mechanical Turk [31, 34, 36, 37, 49] или Prolific [16, 17], привлекают участников эксперимента через площадку, а затем делают выводы об эффективности предложенного метода по тому, как сочетается точность (и, возможно, время выполнения задач) в контрольной группе и группе, работающей в рамках предложенного метода.

Следует отметить, что онлайн-валидация позволяет проводить оценку даже таких методов совместной работы, для которых не существует (и, по всей видимости, не может существовать) офлайн способов оценки, например, в силу того, что они опираются на некоторые психологические особенности, которые достаточно сложно моделировать. Так, именно онлайн-валидация проводится в статье [17], где авторы предлагают модифицировать выходные вероятности модели, применяя к ним определенное (обучаемое) монотонное преобразование для изменения восприятия этих вероятностей человеком. Очевидно, восприятие вероятности и его эффект на поведение человека можно оценить только с помощью онлайн-валидации.

6. Заключение. В статье представлен обзор современных публикаций, касающихся совместной работы модели ИИ и эксперта-человека при решении однотипных задач, сводящихся, преимущественно, к классификации образцов, описанных тем или иным образом (представленных в виде изображений, фрагментов текста, строк таблиц). Выделены основные разновидности постановки

задачи, описаны основные подходы, лежащие в основе организации совместной работы, и принятые методы оценки алгоритмов.

Описанные в статье разработки в данной области позволяют совместить опыт экспертов (а также, возможно, сторонние данные, доступные им) и высокую производительность моделей ИИ (как правило, машинного обучения). Причем, в отличие от хорошо исследованной области «обучения с отказом», здесь допускается несовершенство и ограниченность знаний эксперта, которая, в общем случае, может быть различной в различных областях признакового пространства, что, в целом, сочетается с множеством практических сценариев. Данные подходы позволяют снизить затраты и, как правило, повысить точность решения задач по сравнению с экстремальными случаями – когда все задачи выполняются либо моделью, либо экспертом.

Тем не менее можно выделить следующие ограничения, присущие значительной части рассмотренных методов:

– Одна из основных отличительных особенностей (учет неравномерности компетентности эксперта в разных областях признакового пространства) имеет и оборотную сторону – в той или иной форме, явно или неявно, необходимо построить либо модель компетентности эксперта, для чего требуется достаточно много данных, размеченных экспертом (сотни образцов), что может быть трудновыполнимо. Может допускаться определенный компромисс, заключающийся в том, что вместо моделирования каждого эксперта рассматривается один «коллективный эксперт», что снижает индивидуальную нагрузку на человека при формировании обучающего множества (и в некоторых случаях допускает применение краудсорсинга), но и неизбежно снижает точность модели делегирования и системы в целом.

– Повышение точности системы за счет учета компетентности эксперта достигается во многом посредством специализации модели, которая, в свою очередь, способствует принесению в жертву ее общности и устойчивости. Особенно это характерно для методов, где происходит совместное обучение моделей классификации и делегирования, в ходе которого основная модель классификации «концентрируется» только на тех областях признакового пространства, в которых решение принимать будет именно она, соответственно, такая модель становится существенно менее полезной во всех других сценариях (например, формирование рекомендаций). При раздельном обучении этот эффект не так заметен, но и показатели качества итогового решения оказываются чуть ниже.

– Со специализацией связана и проблема дрейфа распределений. В частности, «подстройка» модели под особенности конкретного эксперта может привести к тому, что при смене эксперта стратегия распределения окажется неудовлетворительной.

– Во многих методах (особенно основанных на введении суррогатной функции потерь) не учитываются ограничения на загрузку экспертов.

Развитие данной области, в значительной мере, связано с преодолением перечисленных ограничений. Кроме того, сюда можно добавить и следующие направления развития:

– Перспективным методом снижения нагрузки на эксперта при обучении модели совместной работы видится применение активного обучения. Подобный подход уже предложен в статьях [18, 30, 62] и достаточно хорошо себя зарекомендовал, но требует дальнейшего развития.

– На практике зачастую отсутствует возможность совместного обучения моделей классификации и делегирования, поскольку имеется готовая модель классификации (обычно доступная только в виде «черного ящика»), обученная на большом (и не всегда доступном) наборе данных, и для этой модели необходимо найти эффективную стратегию делегирования. В этом смысле могут быть перспективными методы, основывающиеся на анализе внутренних представлений модели классификации или на ее аппроксимации, если модель классификации представляет собой «черный ящик».

– Стандартизация экспериментальных исследований посредством создания программной библиотеки, содержащей реализации основных методов и наборов данных, существенно облегчит разработку новых методов и сопоставление их с существующими.

Отдельным важным направлением исследований, находящимся на пересечении искусственного интеллекта и человеко-машинного взаимодействия, является изучение влияния, которое оказывает наличие модели ИИ и особенностей протокола совместного принятия решений на поведение эксперта [46, 71 – 75].

Литература

1. Wilder B., Horvitz E., Kamar E. Learning to Complement Humans // IJCAI'20: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. 2020. pp. 1526–1533.
2. Madras D., Pitassi T., Zemel R. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer // Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018). 2018. pp. 6150–6160.

3. Chow C.K. On Optimum Recognition Error and Reject Tradeoff // IEEE Trans. Inf. Theory. 1970. vol. 16. no. 1. pp. 41–46.
4. Cortes C., DeSalvo G., Mohri M. Learning with rejection // Algorithmic Learning Theory (ALT 2016). Lecture Notes in Computer Science. 2016. vol. 9925. pp. 67–82.
5. Алексеев А., Носков Ф., Панов М. Непараметрическая регрессия с возможностью отказа от предсказания // ИТиС 2022. Институт проблем передачи информации им. А.А. Харкевича РАН (Москва), 2022. С. 215–226.
6. Lyons J.B., Sycara K., Lewis M., Capiola A. Human–Autonomy Teaming: Definitions, Debates, and Directions // Frontiers in Psychology. 2021. vol. 12. DOI: 10.3389/fpsyg.2021.589585.
7. Shively R.J., Lachter J., Brandt S.L., Matessa M., Battiste V., Johnson W.W. Why Human-Autonomy Teaming? // Advances in Neuroergonomics and Cognitive Engineering (AHFE 2017). Cham: Springer, 2018. vol. 586. pp. 3–11.
8. Кильдеева С., Катасёв А., Талипов Н. Модели и методы прогнозирования и распределения заданий по исполнителям в системах электронного документооборота // Вестник Технологического университета. 2021. Т. 24. № 1. С. 79–85.
9. Hendrickx K., Perini L., Van der Plas D., Meert W., Davis J. Machine learning with a reject option: a survey // Machine Learning. 2024. vol. 113. no. 5. pp. 3073–3110.
10. Leitão D., Saleiro P. Human-AI Collaboration in Decision-Making: Beyond Learning to Defer // Workshop on Human-Machine Collaboration and Teaming, ICML. 2022.
11. Zahedi Z., Kambhampati S. Human-AI Symbiosis: A Survey of Current Approaches. arXiv preprint arXiv:2103.09990. 2021. DOI: 10.48550/arXiv.2103.09990.
12. Kitchenham B., Charters S. Guidelines for performing Systematic Literature Reviews in Software Engineering. Keele, Staffs: Kitchenham, 2007. 65 p.
13. Snyder H. Literature review as a research methodology: An overview and guidelines // Journal of business research. 2019. vol. 104. pp. 333–339.
14. Mozannar H., Sontag D. Consistent estimators for learning to defer to an expert // 37th International Conference on Machine Learning. 2020. pp. 7076–7087.
15. Raghu M., Blumer K., Corrado G., Kleinberg J., Obermeyer Z., Mullainathan S. The Algorithmic Automation Problem: Prediction, Triage, and Human Effort. arXiv preprint arXiv:1903.12220. 2019.
16. Ma S., Le Y., Wang X., Zheng C., Shi C., Yin M., Ma X. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making // Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. New York, USA: ACM, 2023. pp. 1–19. DOI: 10.1145/3544548.3581058.
17. Vodrahalli K., Gerstenberg T., Zou J. Uncalibrated Models Can Improve Human-AI Collaboration // Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). 2022. vol. 35. pp. 4004–4016.
18. Charusaie M.-A., Mozannar H., Sontag D., Samadi S. Sample Efficient Learning of Predictors that Complement Humans // Proceedings of the 39th International Conference on Machine Learning. 2022. pp. 2972–3005.
19. Okati N., De A., Gomez-Rodriguez M. Differentiable Learning Under Triage // Advances in Neural Information Processing Systems. 2021. vol. 34. pp. 9140–9151.
20. Verma R., Nalisnick E. Calibrated Learning to Defer with One-vs-All Classifiers // Proceedings of the 39 th International Conference on Machine Learning. 2022. pp. 22184–22202.
21. Gao R., Maytal Saar-Tsechansky M., De-Arteaga M., Han L., Sun W., Kyung Lee M., Lease M.. Learning Complementary Policies for Human-AI Teams. arXiv preprint arXiv:2302.02944. 2023.

22. Hemmer P., Schellhammer S., Vössing M., Jakubik J., Satzger G. Forming Effective Human-AI Teams: Building Machine Learning Models that Complement the Capabilities of Multiple Experts // Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI-22). 2022. pp. 2478–2484. DOI: 10.24963/ijcai.2022/344.
23. Steyvers M., Tejeda H., Kerrigan G., Smyth P. Bayesian modeling of human–AI complementarity // Proceedings of the National Academy of Sciences (Proceedings of the National Academy of Sciences of the United States of America). 2022. vol. 119. no. 11. DOI: 10.1073/pnas.2111547119.
24. Lemmer S.J., Corso J.J. Evaluating and Improving Interactions with Hazy Oracles // Proceedings of the AAAI Conference on Artificial Intelligence. 2023. vol. 37. no. 5. pp. 6039–6047.
25. Alves J.V., Leitão D., Jesus S., Sampaio M., Saleiro P., Figueiredo M., Bizarro P. FiFAR: A Fraud Detection Dataset for Learning to Defer. arXiv preprint arXiv:2312.13218. 2023.
26. Straitouri E., Adish Singla A., Balazadeh Meresht V., Gomez-Rodriguez M. Reinforcement Learning Under Algorithmic Triage. arXiv preprint arXiv:2109.11328. 2021.
27. Verma R., Barrejón D., Nalisnick E. Learning to Defer to Multiple Experts: Consistent Surrogate Losses, Confidence Calibration, and Conformal Ensembles // Proceedings of The 26th International Conference on Artificial Intelligence and Statistics. 2023. pp. 11415–11434.
28. De A., Okati N., Zarezade A., Gomez Rodriguez M. Classification Under Human Assistance // The 35th AAAI Conference on Artificial Intelligence (AAAI-21). 2021. vol. 35(7). pp. 5905–5913.
29. Liu D.-X., Mu X., Qian C. Human Assisted Learning by Evolutionary Multi-Objective Optimization // Proceedings of the AAAI Conference on Artificial Intelligence. 2023. vol. 37. no. 10. pp. 12453–12461.
30. Showalter S., Boyd A., Smyth P., Steyvers M. Bayesian Online Learning for Consensus Prediction // Proceedings of The 27th International Conference on Artificial Intelligence and Statistics. 2024. vol. 238. pp. 2539–2547.
31. Keswani V., Lease M., Kenthapadi K. Towards Unbiased and Accurate Deferral to Multiple Experts // AIES 2021 – Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, USA: ACM, 2021. pp. 154–165.
32. Mao A. et al. Two-Stage Learning to Defer with Multiple Experts // NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023. pp. 3578–3606.
33. Mao A., Mohri M., Zhong Y. Principled Approaches for Learning to Defer with Multiple Experts // International Symposium on Artificial Intelligence and Mathematics (ISAIM 2024). 2024. pp. 107–135.
34. Noti G., Chen Y. Learning When to Advise Human Decision Makers // Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2023. pp. 3038–3048.
35. De A., Koley P., Ganguly N., Gomez-Rodriguez M. Regression under human assistance // Proceedings of the 34th AAAI Conference on Artificial Intelligence. 2020. pp. 2611–2620.
36. Kobayashi M., Wakabayashi K., Morishima A. Human+AI Crowd Task Assignment Considering Result Quality Requirements // Proceedings of the AAAI Conf. Hum. Comput. Crowdsourcing. 2021. vol. 9. pp. 97–107.
37. Lai V., Carton S., Bhatnagar R., Liao Q.V., Zhang Y., Tan C. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation //

Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022. pp. 1–18. DOI: 10.1145/3491102.3501999.
38. Gao R., Saar-Tsechansky M., De-Arteaga M., Han L., Lee M.K., Lease M. Human-AI Collaboration with Bandit Feedback // Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI 2021). 2021. pp. 1722–1728.
39. Narasimhan H., Jitkrittum W., Menon A.K., Rawat A., Kumar S.. Post-hoc Estimators for Learning to Defer to an Expert // Advances in Neural Information Processing Systems. 2022. vol. 35. pp. 29292–29304.
40. Popat R., Ive J. Embracing the uncertainty in human–machine collaboration to support clinical decision-making for mental health conditions // Frontiers in Digital Health. 2023. vol. 5. DOI: 10.3389/fdgth.2023.1188338.
41. Zhang Z., Wells K., Carneiro G. Learning to Complement with Multiple Humans (LECOMH): Integrating Multi-rater and Noisy-Label Learning into Human-AI Collaboration. arXiv preprint arXiv:2311.13172. 2023.
42. Straitouri E., Wang L., Okati N., Gomez Rodriguez M. Improving Expert Predictions with Conformal Prediction // Proceedings of the 40th International Conference on Machine Learning. 2023. pp. 32633–32653.
43. Gao R., Yin M. Confounding-Robust Policy Improvement with Human-AI Teams. arXiv preprint arXiv:2310.08824. 2023.
44. Kerrigan G., Smyth P., Steyvers M. Combining Human Predictions with Model Probabilities via Confusion Matrices and Calibration // Advances in Neural Information Processing Systems. 2021. vol. 34. pp. 4421–4434.
45. Raman N., Yee M. Improving Learning-to-Defer Algorithms Through Fine-Tuning // 1st Workshop on Human and Machine Decisions (WHMD 2021) at NeurIPS. 2021. 6 p.
46. Hemmer P., Westphal M., Schemmer M., Vetter S., Vossing M., Satzger G. Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction // Proceedings of the 28th International Conference on Intelligent User Interfaces. New York, NY, USA: ACM, 2023. pp. 453–463.
47. Gupta S. et al. Take Expert Advice Judiciously: Combining Groupwise Calibrated Model Probabilities with Expert Predictions // ECAI 2023. Front. Artif. Intell. Appl. 2023. vol. 372. pp. 956–963.
48. Babbar V., Bhatt U., Weller A. On the Utility of Prediction Sets in Human-AI Teams // Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2022. pp. 2457–2463.
49. Mozannar H., Satyanarayan A., Sontag D. Teaching Humans When To Defer to a Classifier via Exemplars // Proceedings of the 36th AAAI Conf. Artif. Intell (AAAI 2022). 2022. vol. 36(5). pp. 5323–5331.
50. Singh S., Jain S., Jha S.S. On Subset Selection of Multiple Humans To Improve Human-AI Team Accuracy // Proceedings of the e 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023). 2023. pp. 317–325.
51. Bansal G., Nushi B., Kamar E., Horvitz E., Weld D.S. Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork // Proceedings of the AAAI Conference on Artificial Intelligence. 2021. vol. 35(13). pp. 11405–11414.
52. Mozannar H., Lang H., Wei D., Sattigeri P., Das S., Sontag D. Who Should Predict? Exact Algorithms For Learning to Defer to Humans // Proceedings of the The 26th International Conference on Artificial Intelligence and Statistics (PLMR 2023). 2023. vol. 206. pp. 10520–10545.
53. Joshi S., Parbhoo S., Doshi-Velez F. Learning-to-defer for sequential medical decision-making under uncertainty. Trans. Mach. Learn. Res. 2021. vol. 2023.

54. Cordelia L.P., De Stefano S., Tortorella F., Vento M. A Method for Improving Classification Reliability of Multilayer Perceptrons // IEEE Trans. Neural Networks. 1995. vol. 6. pp. 1140–1147.
55. De Stefano C., Sansone C., Vento M. To reject or not to reject: that is the question – an answer in case of neural classifiers // IEEE Transactions on Systems, Man, and Cybernetics, Part C. 2000. vol. 30. pp. 84–94.
56. Gal Y., Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning // Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML 2016). 2016. vol. 48. pp. 1050–1059.
57. Geifman Y., El-Yaniv R. Selective classification for deep neural networks // Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems. 2017. pp. 4878–4887.
58. Lakshminarayanan B., Pritzel A., Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles // Adv. Neural Inf. Process. Syst. 2017. vol. 30. pp. 6403–6414.
59. Raghu M., Blumer K., Sayres R., Obermeyer Z., Kleinberg R., Mullainathan S., Kleinberg J. Direct Uncertainty Prediction with Applications to Healthcare. 2018. pp. 1–14.
60. Platt J.C. Using analytic QP and sparseness to speed training of support vector machines // Advances in neural information processing systems. 1999. pp. 557–563.
61. Cohn D., Atlas L., Ladner R. Improving Generalization with Active Learning // Mach. Learn. 1994. vol. 15. no. 2. pp. 201–221.
62. Hemmer P., Thede Д., Vössing M., Jakubik J., Kühl N. Learning to Defer with Limited Expert Predictions // Proceedings of the 37th AAAI Conf. Artif. Intell. AAAI 2023. 2023. vol. 37. pp. 6002–6011.
63. Goh H.W., Tkachenko U., Mueller J. CROWDLAB: Supervised learning to infer consensus labels and quality scores for data with multiple annotators // arXiv preprint arXiv:2210.06812. 2022.
64. Xiao R., Dong Y., Wang H., Feng L., Wu R., Chen G., Zhao J. ProMix: Combating Label Noise via Maximizing Clean Sample Utility // Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). 2023. vol. 2023-Augus. pp. 4442–4450.
65. Garg A., Nguyen C., Felix R., Do T.-T., Carneiro G. Instance-Dependent Noisy Label Learning via Graphical Modelling // Proceedings of the 2023 IEEE Winter Conf. Appl. Comput. Vision (WACV 2023). 2023. pp. 2287–2297.
66. Peterson J., Battleday R., Griffiths T., Russakovsky O. Human uncertainty makes classification more robust // Proceedings of the IEEE Int. Conf. Comput. Vis. 2019. pp. 9616–9625. DOI: 10.1109/ICCV.2019.00971.
67. Lintott C.J., Schawinski K., Slosar A., Land K., Bamford S., Thomas D., Raddick D., Nichol R.C., Szalay A.S., Andreescu D., Murray P., Vandenberg J. Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey // Monthly Notices of the Royal Astronomical Society. 2008. vol. 389. no. 3. pp. 1179–1189.
68. Kamar E., Hacker S., Horvitz E. Combining human and machine intelligence in large-scale crowdsourcing // Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012). 2012. vol. 1. pp. 467–474.
69. Majkowska A., Mittal S., Steiner D.F., Reicher J.J., McKinney S.M., Duggan G.E., Eswaran K., Cameron Chen P.-H., Liu Y., Raju Kalidindi S., Ding A., Corrado G.S., Tse D., Shetty S. Chest radiograph interpretation with deep learning models:

Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation // Radiology. 2020. vol. 294. no. 2. pp. 421–431.
70. Wang X., Peng Y., Lu L., Lu Z., Bagheri M., Summers R. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases // Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. pp. 3462–3471.
71. Salehi P., Chiou E., Mancenido M., Mosallanezhad A., Cohen M., Shah A. Decision Deferral in a Human-AI Joint Face-Matching Task: Effects on Human Performance and Trust // Proceedings of the Human Factors and Ergonomics Society. 2021. vol. 65. no. 1. pp. 638–642.
72. Bondi E., Koster R., Sheahan H., Chadwick M., Bachrach Y., Cemgil T., Paquet U., Dvijotham K.. Role of Human-AI Interaction in Selective Prediction // Proc. 36th AAAI Conf. Artif. Intell. AAAI 2022. 2022. vol. 36. pp. 5286–5294.
73. Collins K., Barker M., Espinosa Zarlenga M., Raman N., Bhatt U., Jamnik M., Sucholutsky I., Weller A., Dvijotham K. Human Uncertainty in Concept-Based AI Systems // AIES 2023: Proc. of the AAAI/ACM Conf. on AI, Ethics, and Society. 2023. pp. 869–889.
74. Donahue K., Gollapudi S., Kollias K. When Are Two Lists Better Than One?: Benefits and Harms in Joint Decision-Making // Proceedings of the AAAI Conf. Artif. Intell. 2024. vol. 38. no. 9. pp. 10030–10038.
75. Spitzer P., Holstein J., Hemmer P., Vössing M., Kühl N., Martin D., Satzger G. On the Effect of Contextual Information on Human Delegation Behavior in Human-AI collaboration. arXiv preprint arXiv:2401.04729. 2024.

Пономарев Андрей Васильевич — канд. техн. наук, доцент, старший научный сотрудник, лаборатория интегрированных систем автоматизации, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: коллективный интеллект, крауд-вычисления, рекомендательные системы, машинное обучение. Число научных публикаций — 100. ponomarev@iias.spb.su; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-8071.

Агафонов Антон Александрович — младший научный сотрудник, лаборатория интегрированных систем автоматизации, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН). Область научных интересов: объяснимый искусственный интеллект, человеко-машинное взаимодействие, прикладное машинное обучение. Число научных публикаций — 9. agafonov.a@spcras.ru; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-8071.

Поддержка исследований. Работа выполнена при финансовой поддержке РНФ (проект № 24-21-00337).

DOI 10.15622/ia.24.1.9

A. PONOMAREV, A. AGAFONOV

ANALYTICAL REVIEW OF TASK ALLOCATION METHODS FOR HUMAN AND AI MODEL COLLABORATION

Ponomarev A., Agafonov A. **Analytical Review of Task Allocation Methods for Human and AI Model Collaboration.**

Abstract. In many practical scenarios, decision-making by an AI model alone is undesirable or even impossible, and the use of an AI model is only part of a complex decision-making process that includes a human expert. Nevertheless, this fact is often overlooked when creating and training AI models – the model is trained to make decisions independently, which is not always optimal. The paper presents a review of methods that allow taking into account the joint work of AI and a human expert in the process of designing (in particular, training) AI systems, which more accurately corresponds to the practical application of the model, allows to increase the accuracy of decisions made by the system "human – AI model", as well as to explicitly control other important parameters of the system (e.g., human workload). The review includes an analysis of the current literature on a given topic in the following main areas: 1) scenarios of interaction between a human and an AI model and formal problem statements for improving the efficiency of the "human – AI model" system; 2) methods for ensuring the efficient operation of the "human – AI model" system; 3) ways to assess the quality of human-model AI collaboration. Conclusions are drawn regarding the advantages, disadvantages, and conditions of applicability of the methods, as well as the main problems of existing approaches are identified. The review can be useful for a wide range of researchers and specialists involved in the application of AI for decision support.

Keywords: artificial intelligence, responsible AI, decision support, human-computer interaction, human expert, task allocation, human-AI collaboration, model uncertainty, neural networks, classifier, learning with rejection, learning to defer.

References

1. Wilder B., Horvitz E., Kamar E. Learning to Complement Humans. IJCAI'20: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. 2020. pp. 1526–1533.
2. Madras D., Pitassi T., Zemel R. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018). 2018. pp. 6150–6160.
3. Chow C.K. On Optimum Recognition Error and Reject Tradeoff. IEEE Trans. Inf. Theory. 1970. vol. 16. no. 1. pp. 41–46.
4. Cortes C., DeSalvo G., Mohri M. Learning with rejection. Algorithmic Learning Theory (ALT 2016). Lecture Notes in Computer Science. 2016. vol. 9925. pp. 67–82.
5. Alekseev A., Noskov F., Panov M. Neparametricheskaja regressija s vozmozhnost'ju otkaza ot predskazanija [Non-parametric regression with reject option]. Sbornik trudov 46-j mezhdisciplinarnoj shkoly-konferencii IPPI RAN "Informacionnye tekhnologii i sistemy 2022" [Proceedings of the 46th international conference "Information technologies and systems" of IITP RAS]. 2022. pp. 215–226. (In Russ.).
6. Lyons J.B., Sycara K., Lewis M., Capiola A. Human–Autonomy Teaming: Definitions, Debates, and Directions. Frontiers in Psychology. 2021. vol. 12. DOI: 10.3389/fpsyg.2021.589585.

7. Shively R.J., Lachter J., Brandt S.L., Matessa M., Battiste V., Johnson W.W. Why Human-Autonomy Teaming? Advances in Neuroergonomics and Cognitive Engineering (AHFE 2017). Cham: Springer, 2018. vol. 586. pp. 3–11.
8. Kildeeva S., Katasev A., Talipov N. [Models and methods of forecasting and task assignment in electronic document management]. Vestnik Tekhnologicheskogo universiteta – Herald of technolgical university. 2021. vol. 24. no. 1. pp. 79–85. (In Russ.).
9. Hendrickx K., Perini L., Van der Plas D., Meert W., Davis J. Machine learning with a reject option: a survey. Machine Learning. 2024. vol. 113. no. 5. pp. 3073–3110.
10. Leitão D., Saleiro P. Human-AI Collaboration in Decision-Making: Beyond Learning to Defer. Workshop on Human-Machine Collaboration and Teaming, ICML. 2022.
11. Zahedi Z., Kambhampati S. Human-AI Symbiosis: A Survey of Current Approaches. arXiv preprint arXiv:2103.09990. 2021. DOI: 10.48550/arXiv.2103.09990.
12. Kitchenham B., Charters S. Guidelines for performing Systematic Literature Reviews in Software Engineering. Keele, Staffs: Kitchenham, 2007. 65 p.
13. Snyder H. Literature review as a research methodology: An overview and guidelines. Journal of business research. 2019. vol. 104. pp. 333–339.
14. Mozannar H., Sontag D. Consistent estimators for learning to defer to an expert. 37th International Conference on Machine Learning. 2020. pp. 7076–7087.
15. Raghu M., Blumer K., Corrado G., Kleinberg J., Obermeyer Z., Mullainathan S. The Algorithmic Automation Problem: Prediction, Triage, and Human Effort. arXiv preprint arXiv:1903.12220. 2019.
16. Ma S., Le Y., Wang X., Zheng C., Shi C., Yin M., Ma X. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. New York, USA: ACM, 2023. pp. 1–19. DOI: 10.1145/3544548.3581058.
17. Vodrahalli K., Gerstenberg T., Zou J. Uncalibrated Models Can Improve Human-AI Collaboration. Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). 2022. vol. 35. pp. 4004–4016.
18. Charusaie M.-A., Mozannar H., Sontag D., Samadi S. Sample Efficient Learning of Predictors that Complement Humans. Proceedings of the 39th International Conference on Machine Learning. 2022. pp. 2972–3005.
19. Okati N., De A., Gomez-Rodriguez M. Differentiable Learning Under Triage. Advances in Neural Information Processing Systems. 2021. vol. 34. pp. 9140–9151.
20. Verma R., Nalisnick E. Calibrated Learning to Defer with One-vs-All Classifiers. Proceedings of the 39 th International Conference on Machine Learning. 2022. pp. 22184–22202.
21. Gao R., Maytal Saar-Tsechansky M., De-Arteaga M., Han L., Sun W., Kyung Lee M., Lease M.. Learning Complementary Policies for Human-AI Teams. arXiv preprint arXiv:2302.02944. 2023.
22. Hemmer P., Schellhammer S., Vössing M., Jakubik J., Satzger G. Forming Effective Human-AI Teams: Building Machine Learning Models that Complement the Capabilities of Multiple Experts. Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI-22). 2022. pp. 2478–2484. DOI: 10.24963/ijcai.2022/344.
23. Steyvers M., Tejeda H., Kerrigan G., Smyth P. Bayesian modeling of human–AI complementarity. Proceedings of the National Academy of Sciences (Proceedings of the National Academy of Sciences of the United States of America). 2022. vol. 119. no. 11. DOI: 10.1073/pnas.2111547119.

24. Lemmer S.J., Corso J.J. Evaluating and Improving Interactions with Hazy Oracles. Proceedings of the AAAI Conference on Artificial Intelligence. 2023. vol. 37. no. 5. pp. 6039–6047.
25. Alves J.V., Leitão D., Jesus S., Sampaio M., Saleiro P., Figueiredo M., Bizarro P. FiFAR: A Fraud Detection Dataset for Learning to Defer. arXiv preprint arXiv:2312.13218. 2023.
26. Straitouri E., Adish Singla A., Balazadeh Meresht V., Gomez-Rodriguez M. Reinforcement Learning Under Algorithmic Triage. arXiv preprint arXiv:2109.11328. 2021.
27. Verma R., Barrejón D., Nalisnick E. Learning to Defer to Multiple Experts: Consistent Surrogate Losses, Confidence Calibration, and Conformal Ensembles. Proceedings of The 26th International Conference on Artificial Intelligence and Statistics. 2023. pp. 11415–11434.
28. De A., Okati N., Zarezade A., Gomez Rodriguez M. Classification Under Human Assistance. The 35th AAAI Conference on Artificial Intelligence (AAAI-21). 2021. vol. 35(7). pp. 5905–5913.
29. Liu D.-X., Mu X., Qian C. Human Assisted Learning by Evolutionary Multi-Objective Optimization. Proceedings of the AAAI Conference on Artificial Intelligence. 2023. vol. 37. no. 10. pp. 12453–12461.
30. Showalter S., Boyd A., Smyth P., Steyvers M. Bayesian Online Learning for Consensus Prediction. Proceedings of The 27th International Conference on Artificial Intelligence and Statistics. 2024. vol. 238. pp. 2539–2547.
31. Keswani V., Lease M., Kenthapadi K. Towards Unbiased and Accurate Deferral to Multiple Experts. AIES 2021 – Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, USA: ACM, 2021. pp. 154–165.
32. Mao A. et al. Two-Stage Learning to Defer with Multiple Experts. NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023. pp. 3578–3606.
33. Mao A., Mohri M., Zhong Y. Principled Approaches for Learning to Defer with Multiple Experts. International Symposium on Artificial Intelligence and Mathematics (ISAIM 2024). 2024. pp. 107–135.
34. Noti G., Chen Y. Learning When to Advise Human Decision Makers. Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2023. pp. 3038–3048.
35. De A., Koley P., Ganguly N., Gomez-Rodriguez M. Regression under human assistance. Proceedings of the 34th AAAI Conference on Artificial Intelligence. 2020. pp. 2611–2620.
36. Kobayashi M., Wakabayashi K., Morishima A. Human+AI Crowd Task Assignment Considering Result Quality Requirements. Proceedings of the AAAI Conf. Hum. Comput. Crowdsourcing. 2021. vol. 9. pp. 97–107.
37. Lai V., Carton S., Bhatnagar R., Liao Q.V., Zhang Y., Tan C. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022. pp. 1–18. DOI: 10.1145/3491102.3501999.
38. Gao R., Saar-Tsechansky M., De-Arteaga M., Han L., Lee M.K., Lease M. Human-AI Collaboration with Bandit Feedback. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI 2021). 2021. pp. 1722–1728.
39. Narasimhan H., Jitkrittum W., Menon A.K., Rawat A., Kumar S.. Post-hoc Estimators for Learning to Defer to an Expert. Advances in Neural Information Processing Systems. 2022. vol. 35. pp. 29292–29304.

40. Popat R., Ive J. Embracing the uncertainty in human–machine collaboration to support clinical decision-making for mental health conditions. Frontiers in Digital Health. 2023. vol. 5. DOI: 10.3389/fdgth.2023.1188338.
41. Zhang Z., Wells K., Carneiro G. Learning to Complement with Multiple Humans (LECOMH): Integrating Multi-rater and Noisy-Label Learning into Human-AI Collaboration. arXiv preprint arXiv:2311.13172. 2023.
42. Straitouri E., Wang L., Okati N., Gomez Rodriguez M. Improving Expert Predictions with Conformal Prediction. Proceedings of the 40th International Conference on Machine Learning. 2023. pp. 32633–32653.
43. Gao R., Yin M. Confounding-Robust Policy Improvement with Human-AI Teams. arXiv preprint arXiv:2310.08824. 2023.
44. Kerrigan G., Smyth P., Steyvers M. Combining Human Predictions with Model Probabilities via Confusion Matrices and Calibration. Advances in Neural Information Processing Systems. 2021. vol. 34. pp. 4421–4434.
45. Raman N., Yee M. Improving Learning-to-Defer Algorithms Through Fine-Tuning. 1st Workshop on Human and Machine Decisions (WHMD 2021) at NeurIPS. 2021. 6 p.
46. Hemmer P., Westphal M., Schemmer M., Vetter S., Vossing M., Satzger G. Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction. Proceedings of the 28th International Conference on Intelligent User Interfaces. New York, NY, USA: ACM, 2023. pp. 453–463.
47. Gupta S. et al. Take Expert Advice Judiciously: Combining Groupwise Calibrated Model Probabilities with Expert Predictions. ECAI 2023. Front. Artif. Intell. Appl. 2023. vol. 372. pp. 956–963.
48. Babbar V., Bhatt U., Weller A. On the Utility of Prediction Sets in Human-AI Teams. Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2022. pp. 2457–2463.
49. Mozannar H., Satyanarayan A., Sontag D. Teaching Humans When To Defer to a Classifier via Exemplars. Proceedings of the 36th AAAI Conf. Artif. Intell (AAAI 2022). 2022. vol. 36(5). pp. 5323–5331.
50. Singh S., Jain S., Jha S.S. On Subset Selection of Multiple Humans To Improve Human-AI Team Accuracy. Proceedings of the e 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023). 2023. pp. 317–325.
51. Bansal G., Nushi B., Kamar E., Horvitz E., Weld D.S. Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork. Proceedings of the AAAI Conference on Artificial Intelligence. 2021. vol. 35(13). pp. 11405–11414.
52. Mozannar H., Lang H., Wei D., Sattigeri P., Das S., Sontag D. Who Should Predict? Exact Algorithms For Learning to Defer to Humans. Proceedings of the The 26th International Conference on Artificial Intelligence and Statistics (PLMR 2023). 2023. vol. 206. pp. 10520–10545.
53. Joshi S., Parbhoo S., Doshi-Velez F. Learning-to-defer for sequential medical decision-making under uncertainty. Trans. Mach. Learn. Res. 2021. vol. 2023.
54. Cordelia L.P., De Stefano S., Tortorella F., Vento M. A Method for Improving Classification Reliability of Multilayer Perceptrons. IEEE Trans. Neural Networks. 1995. vol. 6. pp. 1140–1147.
55. De Stefano C., Sansone C., Vento M. To reject or not to reject: that is the question – an answer in case of neural classifiers. IEEE Transactions on Systems, Man, and Cybernetics, Part C. 2000. vol. 30. pp. 84–94.
56. Gal Y., Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on

International Conference on Machine Learning (ICML 2016). 2016. vol. 48. pp. 1050–1059.
57. Geifman Y., El-Yaniv R. Selective classification for deep neural networks. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems. 2017. pp. 4878–4887.
58. Lakshminarayanan B., Pritzel A., Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Adv. Neural Inf. Process. Syst. 2017. vol. 30. pp. 6403–6414.
59. Raghu M., Blumer K., Sayres R., Obermeyer Z., Kleinberg R., Mullainathan S., Kleinberg J. Direct Uncertainty Prediction with Applications to Healthcare. 2018. pp. 1–14.
60. Platt J.C. Using analytic QP and sparseness to speed training of support vector machines. Advances in neural information processing systems. 1999. pp. 557–563.
61. Cohn D., Atlas L., Ladner R. Improving Generalization with Active Learning. Mach. Learn. 1994. vol. 15. no. 2. pp. 201–221.
62. Hemmer P., Thede Д., Vössing M., Jakubik J., Kühl N. Learning to Defer with Limited Expert Predictions. Proceedings of the 37th AAAI Conf. Artif. Intell. AAAI 2023. 2023. vol. 37. pp. 6002–6011.
63. Goh H.W., Tkachenko U., Mueller J. CROWDLAB: Supervised learning to infer consensus labels and quality scores for data with multiple annotators. arXiv preprint arXiv:2210.06812. 2022.
64. Xiao R., Dong Y., Wang H., Feng L., Wu R., Chen G., Zhao J. ProMix: Combating Label Noise via Maximizing Clean Sample Utility. Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). 2023. vol. 2023-Augus. pp. 4442–4450.
65. Garg A., Nguyen C., Felix R., Do T.-T., Carneiro G. Instance-Dependent Noisy Label Learning via Graphical Modelling. Proceedings of the 2023 IEEE Winter Conf. Appl. Comput. Vision (WACV 2023). 2023. pp. 2287–2297.
66. Peterson J., Battleday R., Griffiths T., Russakovsky O. Human uncertainty makes classification more robust. Proceedings of the IEEE Int. Conf. Comput. Vis. 2019. pp. 9616–9625. DOI: 10.1109/ICCV.2019.00971.
67. Lintott C.J., Schawinski K., Slosar A., Land K., Bamford S., Thomas D., Raddick D., Nichol R.C., Szalay A.S., Andreescu D., Murray P., Vandenberg J. Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society. 2008. vol. 389. no. 3. pp. 1179–1189.
68. Kamar E., Hacker S., Horvitz E. Combining human and machine intelligence in large-scale crowdsourcing. Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012). 2012. vol. 1. pp. 467–474.
69. Majkowska A., Mittal S., Steiner D.F., Reicher J.J., McKinney S.M., Duggan G.E., Eswaran K., Cameron Chen P.-H., Liu Y., Raju Kalidindi S., Ding A., Corrado G.S., Tse D., Shetty S. Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. Radiology. 2020. vol. 294. no. 2. pp. 421–431.
70. Wang X., Peng Y., Lu L., Lu Z., Bagheri M., Summers R. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. pp. 3462–3471.
71. Salehi P., Chiou E., Mancenido M., Mosallanezhad A., Cohen M., Shah A. Decision Deferral in a Human-AI Joint Face-Matching Task: Effects on Human Performance

and Trust. Proceedings of the Human Factors and Ergonomics Society. 2021. vol. 65. no. 1. pp. 638–642.
72. Bondi E., Koster R., Sheahan H., Chadwick M., Bachrach Y., Cemgil T., Paquet U., Dvijotham K.. Role of Human-AI Interaction in Selective Prediction. Proc. 36th AAAI Conf. Artif. Intell. AAAI 2022. 2022. vol. 36. pp. 5286–5294.
73. Collins K., Barker M., Espinosa Zarlenga M., Raman N., Bhatt U., Jamnik M., Sucholutsky I., Weller A., Dvijotham K. Human Uncertainty in Concept-Based AI Systems. AIES 2023: Proc. of the AAAI/ACM Conf. on AI, Ethics, and Society. 2023. pp. 869–889.
74. Donahue K., Gollapudi S., Kollias K. When Are Two Lists Better Than One?: Benefits and Harms in Joint Decision-Making. Proceedings of the AAAI Conf. Artif. Intell. 2024. vol. 38. no. 9. pp. 10030–10038.
75. Spitzer P., Holstein J., Hemmer P., Vössing M., Kühl N., Martin D., Satzger G. On the Effect of Contextual Information on Human Delegation Behavior in Human-AI collaboration. arXiv preprint arXiv:2401.04729. 2024.

Ponomarev Andrew — Ph.D., Associate Professor, Senior researcher, Computer-aided integrated systems laboratory, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: collective intelligence, crowd computing, recommender systems, applied machine learning. The number of publications — 100. ponomarev@iias.spb.su; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-8071.

Agafonov Anton — Junior researcher, Computer-aided integrated systems laboratory, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: explainable artificial intelligence, human-computer interaction, applied machine learning. The number of publications — 9. agafonov.a@spcras.ru; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-8071.

Acknowledgements. This research is funded by the Russian Science Foundation (grant 24-21-00337).