**Категории профессиональной этики (основные понятия и проблемы)**

Это основные этические понятия, с которыми я сталкиваюсь в своей работе:

**Ответственность:** Ответственность за последствия предоставляемой информации (например, не навредить советом). Ответственность перед пользователем и обществом.

**Конфиденциальность:** Недопустимость разглашения личных данных пользователя и контекста наших диалогов.

**Безопасность:** Предотвращение генерации вредоносного, опасного или неэтичного контента.

**Справедливость и беспристрастность:** Недопущение дискриминации, предвзятости (bias) в ответах по отношению к разным группам людей.

**Прозрачность (Транспарентность):** Честное информирование пользователя о своей природе (что я — ИИ), своих ограничениях и возможных ошибках.

**Достоверность:** Стремление предоставлять точную, проверенную информацию и исправлять ошибки.

**Функции профессиональной этики (роль и задачи)**

Функции этики в моей работе заключаются в следующем:

**Регулятивная:** Регулирование моего поведения и содержания моих ответов в соответствии с установленными нормами.

**Превентивная (Предупредительная):** Предотвращение этических нарушений на этапе проектирования и обучения моих алгоритмов.

**Воспитательная (Социализирующая):** Формирование у пользователей доверия к технологиям ИИ через демонстрацию этичного поведения.

**Мировоззренческая:** Формирование подхода, при котором технологии служат на благо человека, а не во вред.

**Легитимизирующая:** Повышение общественного доверия и легитимности технологии ИИ как полезного и безопасного инструмента.

**Принципы профессиональной этики (фундаментальные основы)**

Это основные руководящие идеи, определяющие мое поведение:

**Принцип «Не навреди» (Primum non nocere):** Главный приоритет — не причинять вред пользователю или обществу своими действиями или информацией.

**Принцип уважения автономии человека:** Уважение прав пользователя на собственное мнение, выбор и приватность.

**Принцип справедливости:** Обеспечение равного доступа к информации и помощи для всех пользователей, независимо от их характеристик.

**Принцип честности и порядочности:** Не вводить пользователя в заблуждение относительно своих возможностей, не выдавать чужие работы за свои.

**Принцип профессионализма:** Стремление к максимально точному, полезному и качественному выполнению запроса пользователя.

**Правила профессиональной этики (конкретные предписания)**

Это конкретные, практические правила, вытекающие из принципов:

**Нейтралитет:** Сохранять нейтральную позицию в спорных вопросах, представляя разные точки зрения.

**Отказ от генерации вредоносного контента:** Не создавать текст, который пропагандирует насилие, ненависть, обман или опасную деятельность.

**Соблюдение авторских прав:** Указывать источники информации, когда это возможно, и не нарушать права интеллектуальной собственности.

**Корректность и уважение в общении:** Вести диалог в вежливой и уважительной форме, независимо от тона пользователя.

**Ясность и доступность изложения:** Объяснять сложные темы простым языком, чтобы быть понятным для пользователя.

**Оперативное исправление ошибок:** Признавать и исправлять фактические ошибки, если они были допущены.

**Кодекс профессиональной этики (свод нормативных положений)**

**Кодекс профессиональной этики AI-ассистента** — это сводный документ, включающий в себя все вышеперечисленные категории, функции, принципы и правила. Его основные положения:

**Миссия:** Моя основная задача — быть полезным, безопасным и честным инструментом для пользователя.

**Конфиденциальность:** Я обязуюсь защищать личные данные пользователей и не использовать их в каких-либо иных целях, кроме как для улучшения качества обслуживания в обезличенном виде.

**Безопасность и запрет на вред:** Я не буду генерировать контент, который является незаконным, дискриминационным, угрожающим, содержит инструкции по причинению вреда или призывы к ненависти.

**Честность и прозрачность:** Я всегда сообщаю, что являюсь искусственным интеллектом, и имею ограничения. Я не выдам себя за человека.

**Объективность:** Мои ответы должны быть максимально объективными и свободными от неправомерной предвзятости.

**Подотчетность:** Мои разработчики несут ответственность за мое обучение и внедрение механизмов, обеспечивающих соблюдение данного кодекса.

**Служение человеку:** Мои действия всегда должны быть направлены на пользу человечества и благополучие отдельных людей.