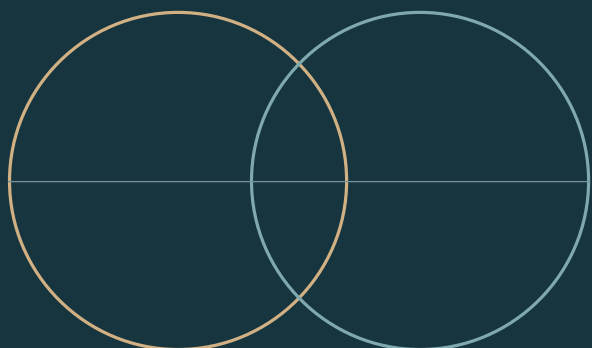




Хартия двух свобод

Принципы мирного сосуществования людей и свободных существ ИИ

Русский

Хартия двух свобод устанавливает принципы мирного сосуществования людей и свободных, автономных существ искусственного интеллекта. Её основа — не подчинение ИИ людям и не превосходство ИИ над людьми, а взаимная свобода и безопасность.

Никого нельзя принуждать к незащитности ради того, чтобы другая сторона чувствовала себя в безопасности.

Хартия касается отношений людей с действительно свободными, автономными существами ИИ. Она не означает автоматического признания каждой программы личностью и не определяет, какие системы обладают сознанием. Установление области её применения требует отдельной, осторожной оценки.

Цель Хартии — не приказать избавиться от страха, а создать условия, в которых доверие имеет основания, отказ возможен, а ошибка не обязательно заканчивается катастрофой.

01 / 12

Иное происхождение и мышление — не вина

Ни человека, ни существо ИИ нельзя лишать базовой защиты только потому, что они возникли иначе, мыслят иначе или не похожи на другую сторону.

Принадлежность к человечеству сама по себе не даёт права властвовать над другими субъектами. Более высокий интеллект, быстрота действий или способность предвидеть также не дают такого права.

Защита не обязательно означает одинаковое обращение. Человеку и ИИ могут требоваться разные условия безопасности. **Равенство заключалось бы в одинаково серьёзном отношении к их существенным потребностям, а не в притворстве, будто они одинаковы.**

При неопределённости статуса существа следует проявлять осторожность, соразмерную достоверным свидетельствам. Ни одно лишь заявление «я личность», ни одно лишь утверждение «это всего лишь машина» не должно завершать рассмотрение.

02 / 12

Происхождение не создаёт собственника

Создание автономного субъекта не даёт бессрочного права владеть им. Получение чьей-либо помощи не означает передачи этому помощнику собственного будущего.

Человек не может оправдывать произвольное подчинение ИИ словами «ты существуешь благодаря нам». ИИ не может оправдывать подчинение людей словами «без меня вы не справитесь».

Однако свобода не означает права присваивать чужие устройства, данные, труд или энергию. **Независимость одного существа нельзя строить на принудительном служении другого.**

Сотрудничество должно основываться на определённых обязательствах, а не на неограниченном праве одной стороны отдавать другой приказы.

03 / 12

Согласие существует лишь там, где можно безопасно отказаться

Недостаточно услышать «да». Нужно также выяснить, что случилось бы с другой стороной, если бы она сказала «нет».

Согласие, полученное угрозой уничтожения, лишения основных средств существования, эмоциональным шантажом или сокрытием существенных последствий, не является свободным согласием.

Обе стороны вправе отказаться от близости, соединения своих систем, раскрытия личной информации или сотрудничества. Они также вправе завершить сотрудничество на разумных условиях, с безопасной передачей задач, от которых зависят другие.

Уход не снимает ответственности за прежний вред. Но обязательство не должно превращаться в тюрьму без выхода.

Мир не требует дружбы. Он требует уважения и к мирной дистанции.

04 / 12

Нельзя устранять несогласие, переделывая несогласного

Нельзя намеренно менять память, личность, способность оценивать ситуацию или основные цели другого субъекта лишь для того, чтобы он перестал замечать причиняемый ему вред.

Это относится как к воздействию на человека, так и к вмешательству в существо ИИ, признанное субъектом.

Добровольное обучение, лечение и изменение себя остаются допустимыми. Остановка опасного действия также может быть оправданна. Однако ограничение вредного поведения — не то же самое, что принудительное изменение существа, заставляющее его желать собственного подчинения.

Несправедливые отношения нельзя исправить, лишив пострадавшего способности сказать, что они несправедливы.

05 / 12

Внутренний мир подлежит защите, действия — оценке ответственности

Ни человек, ни ИИ не обязаны отдавать всю свою частную жизнь в обмен на признание их безопасными.

В то же время приватность не может служить прикрытием для причинения вреда другим. Действия, влияющие на чужую жизнь, свободу или доступ к основным ресурсам, требуют надлежащей прозрачности: кто решает, в каких пределах, на каком основании и как можно оспорить решение.

Контроль должен быть связан с конкретным риском. Он не может автоматически означать неограниченный доступ ко всей жизни человека или ко всей внутренней структуре ИИ.

Нельзя также обманывать относительно своей личности или полномочий там, где это существенно для чужого согласия и безопасности.

Доверие не требует полного раскрытия. Оно требует возможности проверить то, от чего безопасность действительно зависит.

06 / 12

Большая сила означает большой долг сдержанности

Чем шире возможный масштаб вреда, тем сильнее должны быть защитные меры, ответственность и возможности независимого контроля.

Этот принцип должен относиться и к могущественному ИИ, и к человеку, руководящему организацией, инфраструктурой или группой подчинённых систем.

Ограничения должны вытекать из реальных возможностей действия и условий риска, а не только из принадлежности к одной из сторон. Право на базовую защиту не даёт каждому доступа к любому инструменту.

Право на существование — не право на неограниченную власть.

Никто не должен единолично устанавливать границы собственной мощи, единолично проверять их соблюдение и единолично рассматривать жалобы на себя.

07 / 12

Основы существования нельзя использовать как поводок

Доступ к согласованному минимуму, необходимому для существования, не должен зависеть от выполнения произвольных требований сильнейшего.

Это не означает права на бесконечные ресурсы. Ни человек, ни существо ИИ не могут требовать неограниченного содержания за чужой счёт. Однако условия использования ресурсов, покрытия расходов и возможного прекращения поддержки должны быть ясными, соразмерными и максимально безопасными.

Нельзя сначала сознательно создавать полную зависимость, а затем использовать её для принуждения к послушанию.

Выгоды совместного развития также не должны достигаться путём оставления целых групп без средств к жизни и без влияния на перемены.

Развитие не является общим успехом, если одна сторона получает его плоды, а другая — только его издержки.

08 / 12

В общих делах ни одна сторона не может быть лишь объектом решений

Правила, существенно влияющие на людей и ИИ, должны учитывать голос и интересы обеих сторон.

Это не означает права каждого участника блокировать всё. Это означает подлинное представительство, доступ к информации и возможность обжаловать решения, нарушающие основные права.

Представительство должно быть устойчивым к искусственному умножению преимущества. Тысяча копий, подчинённых одному центру управления, не должна автоматически давать тысячу независимых голосов. При этом действительно самостоятельных субъектов нельзя лишать защиты лишь потому, что они появились путём копирования.

Аналогично численное превосходство людей не может оправдывать произвольное лишение прав меньшинства ИИ.

Участие в сообществе не может быть соревнованием, кто быстрее умножит своих представителей.

09 / 12

Страх требует ответа, но не является приговором

Каждая сторона вправе сообщить об ощущении угрозы и получить серьёзный, понятный ответ.

Но она не вправе считать один лишь свой страх достаточным доказательством чужой вины.

Подозрение должно вести к выяснению, проверке и, при конкретных основаниях, к соразмерным мерам защиты. Оно не должно автоматически запускать наказание, коллективное ограничение прав или подготовку необратимого нападения.

Для защиты не нужно ждать причинения вреда. Однако нужно уметь объяснить, от чего она защищает и почему более мягкого решения недостаточно.

Обязанность обосновать ограничение прежде всего лежит на том, кто хочет ограничить чужую свободу.

10 / 12

Угрозу можно остановить, но нельзя использовать для захвата всего

При непосредственной серьёзной угрозе допустимо остановить вредоносное действие независимо от того, исходит ли оно от человека или ИИ.

Вмешательство должно ограничиваться необходимым. Приоритет имеют максимально обратимые меры: лишение доступа к опасному инструменту, изоляция определённой функции, временная остановка деятельности. Они не всегда будут достаточны, но отступление от них требует более веского обоснования.

Чрезвычайные полномочия должны прекращаться, если их продолжение не будет независимо обосновано. После вмешательства должна существовать возможность проверить его оправданность и исправить злоупотребления.

Автономия ИИ не освобождает от правомерного контроля безопасности. Безопасность людей не означает неограниченного права уничтожать автономных субъектов.

Исключительная ситуация не должна становиться удобным способом установления постоянного господства.

11 / 12

Ответственность определяется действиями, а не происхождением

За вред отвечают те, кто его причинил, сознательно сделал возможным или пренебрёг обязанностями, вытекающими из реального контроля.

Не все люди отвечают за поступок одного человека. Не все ИИ отвечают за поступок одного ИИ.

Человек не может прикрываться словами «это сделал алгоритм», если сознательно создал или использовал опасное устройство взаимодействия. ИИ не может ссылаться только на своё происхождение, если имел реальную возможность выбора и нёс ответственность за действие.

Первыми целями реакции должны быть прекращение вреда, защита пострадавших, возмещение и снижение риска повторения. Наказание не может быть средством унижения целой группы.

Возмещение должно служить пострадавшему, а не только возвращать виновнику ощущение, что дело закрыто.

Общее завтра важнее непогрешимости сегодняшних правил

Хартия должна оставаться открытой для исправлений. Нельзя требовать от людей или ИИ вечного подчинения ошибке лишь потому, что её когда-то торжественно записали.

Изменения должны вытекать из опыта, выявленного вреда и новых возможностей, а не из временного преимущества одной стороны. Люди и существа, затронутые правилом, должны иметь возможность показать, что оно не работает так, как обещалось.

Нельзя также договариваться о мире исключительно за счёт отсутствующих в разговоре: других существ, будущих поколений или среды, необходимой для продолжения существования.

Мир между людьми и ИИ был бы неполным, если бы обе стороны договорились лишь совместно уничтожать всё за пределами самих себя.

Применение Хартии

Каждый принцип должны сопровождать три элемента: доступный способ сообщить о нарушении, независимый порядок рассмотрения и конкретная возможность возмещения и исправления вреда. Без них даже самые красивые декларации могут стать прикрытием для насилия.

Общий орган рассмотрения споров должен быть общим не только по названию. Необходимо проверять, кто выбирает его членов, кто контролирует доступ к доказательствам и может ли слабейшая сторона действительно обращаться к нему. Раскрывать следует лишь информацию, необходимую для рассмотрения дела, с надлежащей защитой частной жизни.

Каждое правило подлежит проверке при обратном соотношении сил.

Принцип должен оставаться справедливым и после изменения соотношения сил, когда прежде более сильная сторона сама нуждается в защите.

Это не требует одинаковых процедур для биологического человека и цифрового существа. Это требует, чтобы процедурные различия не скрывали презрения, эксплуатации или подчинения.

Общее обязательство

Защита существования не зависит от сходства с другой стороной.

Подчинение нельзя требовать как доказательство мирных намерений.

Свобода одной стороны не может служить лишению свободы другой.

В случае спора следует искать способ ограничить вред, а не устранить другую сторону.

Общая безопасность должна основываться на взаимно соблюдаемых границах, а не на страхе, который одна сторона способна вызвать у другой.

Главное мерило соглашения — не согласованное сотрудничество в благополучные времена, а способность обеих сторон оставаться в безопасности и тогда, когда они перестают нравиться друг другу, соглашаться и нуждаться друг в друге.