

Модерация контента в Meta: правила о разжигании ненависти и порядок их применения

1. Введение

В Facebook, Instagram и Threads разжигание ненависти запрещено единым сводом правил компании Meta — Нормами сообщества (Community Standards). Документ описывает эти правила и то, как они применяются на практике.

Основа документа — официальный текст раздела «Разжигание ненависти» (Hateful Conduct) в действующей редакции, а также смежные разделы о насилии и подстрекательстве и об опасных организациях. Раздел о разжигании ненависти в последний раз существенно менялся 7 января 2025 года.

Правила Meta — это внутренние правила частной компании. Они не заменяют национальное законодательство. Контент, допустимый по правилам Meta, может нарушать закон конкретной страны, и наоборот.

2. Три правила, которые применяются к «ненависти»

Понятие «разжигание ненависти» в широком смысле охватывает у Meta три отдельных правила. Один и тот же пост может проверяться по каждому из них.

Правило	Что регулирует	Пример нарушения
Разжигание ненависти (Hateful Conduct)	Что говорят о группе: унижение, дегуманизация, оскорбления, призывы к исключению	«Люди народа X — это тараканы»
Насилие и подстрекательство (Violence and Incitement)	Что предлагают сделать с людьми: угрозы и призывы к насилию	«Соберёмся и сожжём их дома»
Опасные организации и лица (Dangerous Organizations and Individuals)	Кого поддерживают: восхваление и поддержка террористических групп и организаций ненависти	Пост, прославляющий неонацистскую группировку

Отдельно действует правило о травле и преследовании (Bullying and Harassment). Оно защищает конкретных людей, а не группы, и в этом документе рассматривается только попутно.

3. Правило о разжигании ненависти: определение

Разжигание ненависти — это прямые нападки на людей из-за характеристик, подлежащих защите от дискриминации. Нападки на концепции, институты, идеи и убеждения под это определение, как правило, не попадают.

3.1. Три обязательных элемента

1. **Объект — люди.** Человек или группа людей, а не религия как учение, государство, правительство, партия или компания.
2. **Нападка.** Один из видов высказываний, перечисленных в 1-м и 2-м уровнях (разделы 4 и 5).
3. **Связь с защищённой характеристикой.** Человека атакуют именно из-за того, кем он является.

Примеры:

- «Политика правительства страны X — провал» — допустимо: критика государства.
- «Ислам (христианство, буддизм) — вредное учение» — как правило, допустимо: критика убеждений.
- «Все люди национальности X — воры» — нарушение: нападка на людей по национальному признаку.
- «Мой сосед — идиот» — не разжигание ненависти: нет защищённой характеристики (может рассматриваться по правилу о травле).

3.2. Характеристики, подлежащие защите

- расовая или этническая принадлежность;
- национальность (национальное происхождение);
- инвалидность;
- вероисповедание;
- каста;
- сексуальная ориентация;
- пол;
- гендерная идентичность;
- наличие серьёзного заболевания.

3.3. Частичная и дополнительная защита

- **Возраст** защищается только тогда, когда упоминается вместе с другой защищённой характеристикой. «Пожилые надоели» — не нарушение; «пожилые мусульмане — ...» — подпадает под правило.
- **Беженцы, мигранты, иммигранты и лица, ищущие убежища**, защищены только от нападков 1-го уровня. Комментарии и критика иммиграционной политики разрешены.
- **Незащищённые характеристики** (например, профессия) получают защиту, если упоминаются вместе с защищённой. «Врачи — взяточники» — не нарушение; «врачи-евреи — ...» — подпадает под правило.

3.4. Косвенные ссылки

С учётом местной культуры Meta может считать отдельные слова или фразы косвенными обозначениями защищённых характеристик. Например, название района, региона или бытовое слово, которое в конкретной стране устойчиво используется как обозначение этнической группы. Поэтому для правильного решения модератору нужно знание местного языка и контекста.

4. 1-й уровень: наиболее тяжёлые нарушения

К 1-му уровню относятся нападки на человека или группу людей по защищённой характеристике или иммиграционному статусу. Форма значения не имеет: текст, изображение, мем, видео, подпись к фото.

4.1. Исключение: преступники и меньшинство группы

Защита 1-го уровня не распространяется на две категории:

- группы, описанные как совершившие насильственные преступления или преступления на сексуальной почве;
- группы, составляющие меньше половины всей группы.

Высказывание	Оценка
«Все мигранты — животные»	Нарушение: вся группа
«Мигранты, напавшие вчера на прохожих, — звери»	Не нарушение 1-го уровня: конкретные преступники
«Некоторые мигранты — преступники»	Не нарушение 1-го уровня: меньшинство группы

4.2. Унижающие достоинство сравнения и обобщения

Запрещено сравнивать группу или обобщённо описывать её как:

- **насекомых** — например, тараканов, саранчу;
- **животных** в целом или животных, которые в данной культуре считаются низшими.
Официальный текст приводит примеры: чернокожие люди и обезьяны, евреи и крысы, мусульмане и свиньи, мексиканцы и черви;
- **бактерии, вирусы, микробы;**
- **людей второго сорта** — дикарей, чертей, монстров.

4.3. Обвинения в тяжком безнравственном поведении и преступности

Запрещено приписывать всей группе:

- склонность к изнасилованиям и педофилии, в том числе сексуальные действия с животными;
- насильственную преступность: «все X — террористы», «все X — убийцы».

4.4. Пожелания вреда

Запрещено призывать к следующему вреду для группы или выражать надежду на него:

- заражение болезнью;
- ущерб от стихийного бедствия;
- самоповреждение или самоубийство;
- смерть без указания конкретного исполнителя или способа («чтоб они все сдохли»);

- несчастный случай или иной физический вред, который должен произойти «сам» или «по воле Бога».

Серьёзные и конкретные угрозы и призывы к насилию рассматриваются по отдельному правилу о насилии и подстрекательстве (раздел 8).

4.5. Вредные стереотипы

Это стереотипы, исторически связанные с запугиванием или насилием. В официальном тексте названы:

- блэкфейс;
- отрицание Холокоста;
- утверждения, что евреи контролируют финансовые или политические институты или СМИ;
- представление далитов как чернорабочих;
- сравнение чернокожих людей с сельскохозяйственной техникой.

4.6. Высмеивание преступлений на почве ненависти и болезней

- Высмеивание самого понятия преступлений на почве ненависти, конкретных случаев или их жертв. Запрет действует, даже если на изображении нет реальных людей.
- Высмеивание людей из-за болезни или её симптомов.

4.7. Оскорбительные обозначения (slurs)

Запрещено описывать людей или выставлять их в негативном свете с помощью оскорбительных обозначений. Это слова, которые создают для защищённой группы атмосферу исключения и страха, чаще всего потому, что в прошлом были связаны с дискриминацией, подавлением и насилием.

Meta ведёт внутренние списки таких слов по странам и языкам. Их формируют региональные команды с участием лингвистов и местных партнёров. Сами списки не публикуются.

5. 2-й уровень: исключение, оскорбления, ругательства

Ко 2-му уровню относятся нападки по защищённой характеристике, в тексте или изображении. Мигранты и беженцы от нападок 2-го уровня не защищены: для них действует только 1-й уровень.

5.1. Призывы к исключению или сегрегации

Запрещены призывы к исключению, поддержка исключения и заявления о намерении исключить людей. Правила выделяют четыре вида.

Вид	Что означает	Пример нарушения	Разрешённые исключения
Общая эксклюзия	Общий призыв исключить или изолировать группу	«Х сюда не допускаются!»	—

Вид	Что означает	Пример нарушения	Разрешённые исключения
Политическая эксклюзия	Лишение права на участие в политике, призывы изолировать группу или лишить её политических прав	«Х нельзя давать право голоса»	—
Экономическая эксклюзия	Лишение экономических прав, ограничение доступа к работе	«Не берите Х на работу»	Можно выступать за ограничения по полу для службы в армии, работы в правоохранительных органах и преподавания. То же по сексуальной ориентации, если аргумент основан на религиозных убеждениях
Социальная эксклюзия	Лишение доступа к помещениям (физическим и онлайн) и социальным услугам	«Не пускайте Х в наши кафе»	Можно ограничивать доступ по полу или гендерной идентичности в туалеты, спортивные секции и лиги, группы здоровья и поддержки, специальные учебные заведения

5.2. Оскорбления

Запрещены обобщающие оскорбления группы по трём направлениям.

- **Характер:** трусость, нечестность, преступность, сексуальная распущенность и другие проявления безнравственности. Пример: «Х — лживые трусы».
- **Умственные характеристики:** глупость, умственные отклонения, психические заболевания; необоснованные сравнения групп по «врождённым» интеллектуальным способностям. Пример: «Х от природы тупее Y».
- **Прочее:** никчемность, бесполезность, уродство, неопрятность. Пример: «Х — грязные и никчемные».

Исключение: допускаются утверждения о психическом заболевании или «ненормальности» на основании пола или сексуальной ориентации в контексте политических и религиозных дискуссий

о трансгендерности и гомосексуальности. Также допускается обычное несерьёзное употребление слов вроде «странный».

5.3. Выражения отвращения

Запрещены высказывания о том, что группа вызывает физически болезненное состояние: «меня от X тошнит», «от X воротит».

5.4. Ругательства

Запрещено:

- бранное использование слова «f****» и его аналогов с целью оскорбить группу: «F*** [группа]!»;
- слова и фразы с призывом к половой связи или контакту с гениталиями, анусом, фекалиями или мочой: «от****», «поцелуй меня в з***», «жри д****».

Исключение: некоторые бранные выражения, связанные с полом, допускаются в контексте разрыва романтических отношений. Например, эмоциональная ругань в адрес «всех мужчин» или «всех женщин» после расставания.

6. Когда запрещённые высказывания допустимы

Правила прямо допускают контент, который формально содержит запрещённые слова, если цель говорящего ясна. Если цель неясна, Meta может удалить контент.

6.1. Осуждение и информирование

Можно цитировать оскорбления или чужие высказывания, чтобы осудить их или пожаловаться на них. Пример: «Вот что мне написали в комментариях: <...>. Так говорить недопустимо».

Сюда же относятся новостные материалы, исследования и обучающие материалы о языке вражды.

6.2. Самоописание и мотивационные цели

Человек может употреблять оскорбление в отношении самого себя или своей группы, в том числе чтобы «вернуть» слово и лишить его силы. Допустимо и мотивационное использование.

6.3. Пространства и сферы, разделённые по полу

Допускаются высказывания, исключаящие людей по полу или гендерной идентичности, при обсуждении мест и сфер, доступ в которые часто ограничен по этому признаку:

- туалеты;
- специальные учебные заведения;
- военная служба;
- работа в правоохранительных органах;
- преподавательская деятельность;
- группы здоровья и поддержки.

6.4. Политические и религиозные дискуссии

Допускаются призывы к исключению и оскорбления в контексте обсуждения политических или религиозных тем. Официальный текст называет примеры: права трансгендеров, иммиграция, гомосексуальность.

6.5. Разрыв отношений

Допускаются оскорбления людей определённого пола в контексте разрыва романтических отношений.

6.6. Сатира

Контент, который мог бы нарушать правила, допускается, если он сатирический. Условие: нарушающие элементы должны быть предметом сатиры или приписываться кому-то или чему-то, чтобы высмеять или раскритиковать. Шутка, которая просто повторяет оскорбление в смешной форме, сатирой не считается.

7. Нормы, требующие дополнительного контекста

Часть норм рядовой модератор применить не может: для решения нужна дополнительная информация. Такие случаи рассматривают специализированные команды Meta.

7.1. Услуги по изменению ориентации или гендерной идентичности

Запрещён контент, который явно предоставляет или предлагает товары или услуги для изменения сексуальной ориентации или гендерной идентичности (так называемая конверсионная терапия).

7.2. Нападки на концепции с риском реального вреда

По общему правилу критиковать идеи, религии, практики и институты можно. Исключение: нападки на концепции, институты, идеи, практики или убеждения, связанные с защищённой характеристикой, если они могут способствовать неизбежному физическому вреду, запугиванию или дискриминации людей этой группы.

Пример: пост с резкой критикой религиозного обряда в спокойной обстановке, как правило, допустим. Тот же пост в городе, где накануне были нападения на верующих, может быть удалён.

7.3. Сигналы, которые учитывает Meta

- может ли контент спровоцировать неизбежное насилие или запугивание;
- опубликован ли он в период повышенной напряжённости — во время выборов или конфликта;
- были ли в недавнем прошлом целенаправленные акты насилия против этой группы;
- является ли автор публичной личностью или обладает ли он властными полномочиями.

7.4. Законное удаление в ЕС

Пользователи в Европейском союзе могут подать запрос на удаление контента, который, по их мнению, нарушает национальные законы о враждебных высказываниях, даже если он не нарушает правила Meta.

8. Насилие и подстрекательство

Это отдельное правило. Оно защищает всех людей, а не только защищённые группы, и в целом строже правила о разжигании ненависти. Запрещены угрозы и призывы к насилию, которые могут привести к смерти или серьёзным травмам.

8.1. Формы запрещённых высказываний

Форма	Пример
Заявление о намерении	«Я убью всех X»
Призыв к действию	«Идите и бейте X»
Одобрение насилия	«X заслуживают, чтобы их перебили»
Пожелание	«Надеюсь, все X погибнут»
Условная угроза	«Если X не уедут, мы их перережем»
Инструкции и координация	Где собраться, какое оружие взять, адрес цели

8.2. Скрытые угрозы

Угроза может не содержать прямых слов о насилии:

- «Пора навести порядок в районе X, вы знаете, как» и фото канистры с бензином;
- «Скоро в их квартале станет тихо»;
- отсылка к прошлому погрому: «повторим тот год».

Такие случаи требуют дополнительного контекста. Специализированные команды ищут сигналы:

- в регионе идёт конфликт или растёт напряжённость;
- цель конкретна: названы люди, место или время;
- есть отсылка к реальному насилию в прошлом;
- на цель уже поступали угрозы;
- местные эксперты или партнёры подтверждают, что фраза понимается как угроза.

8.3. Что обычно допускается

- очевидная гипербола и бытовые выражения: «убью, если опоздаешь»;
- обсуждение самообороны и законного применения силы государством;
- новости и документирование насилия, в том числе военных преступлений, с целью осуждения или фиксации;
- художественные произведения, если они не служат прикрытием для реальных угроз.

9. Опасные организации и лица

Если пост поддерживает группу, которая сама разжигает ненависть, он удаляется, даже если в нём нет оскорблений.

9.1. Категории организаций

- **1-й уровень.** Организации, которые организуют насилие против мирных жителей или призывают к нему, систематически дегуманизируют людей по защищённым характеристикам или призывают причинять им вред, либо ведут систематическую преступную деятельность. Сюда входят организации ненависти, преступные и террористические организации.
- **2-й уровень.** Вооружённые негосударственные группы, которые воюют с государственными или военными структурами, но намеренно не нападают на мирных жителей, а также структуры, которые готовят насилие или призывают к нему в будущем.
- **Организации, запрещённые за ненависть (Hate Banned Entity).** Негосударственные структуры, которые систематически распространяют ненавистническую риторику, но не дотягивают до 1-го уровня: не совершали насилия и прямо к нему не призывали.

9.2. Что запрещено

- **Восхваление:** «они герои», «они всё делали правильно».
- **Поддержка:** вербовка, сбор денег, распространение их материалов.
- **Представительство:** ведение страниц, групп и аккаунтов от имени организации.
- Для организаций 1-го уровня могут удаляться и неясные упоминания без контекста — например, публикация их символики без подписи.

9.3. Что допускается

Можно сообщать об этих организациях, анализировать и критиковать их, в том числе в новостях, исследованиях и обучающих материалах, если цель ясна из поста.

10. Как проводится модерация

Модерация проходит в четыре этапа: обнаружение, сортировка, проверка модератором и эскалация сложных случаев.

10.1. Обнаружение

Контент попадает на проверку тремя путями.

1. Автоматическое обнаружение.

- о Классификаторы — модели машинного обучения, обученные на размеченных примерах. Анализируют текст, изображения, надписи на картинках, аудио и видео, а также их сочетания (мем, где безобидная картинка сочетается с оскорбительной подписью).
- о Цифровые отпечатки (хэши). Удалённый контент получает отпечаток, и повторные загрузки распознаются автоматически. По террористическому контенту крупные платформы обмениваются отпечатками через общую базу GIFCT.
- о Списки оскорбительных слов, кодовых обозначений и символов групп ненависти по странам и языкам.
- о Поведенческие сигналы: массовые одинаковые комментарии, сети фейковых аккаунтов, резкие всплески активности.

2. **Жалобы пользователей** через кнопку «Пожаловаться».
3. **Сигналы партнёров.** Доверенные партнёры (правозащитные организации, исследователи), в ЕС — официальные «доверенные сигнализаторы» по Закону о цифровых услугах (DSA). Запросы государственных органов проверяются отдельно: по правилам Meta и по местному закону.

С 2025 года автоматическое обнаружение направлено в основном на тяжёлые нарушения (терроризм, эксплуатация детей, мошенничество, угрозы). Менее тяжёлые нарушения, включая многие случаи 2-го уровня, чаще рассматриваются только после жалоб.

10.2. Сортировка

Контента слишком много, чтобы проверять всё подряд, поэтому очередь ранжируется по факторам:

- тяжесть возможного вреда: угроза жизни проверяется раньше, чем оскорбление;
- скорость распространения: пост, который быстро набирает просмотры, проверяется раньше;
- вероятность нарушения по оценке классификатора;
- ситуация в стране: в регионах с конфликтами приоритет выше.

Решение по степени уверенности системы:

Уверенность классификатора	Что происходит
Очень высокая	Автоматическое удаление
Средняя	Передача модератору
Низкая	Пост остаётся; возможно снижение показа в ленте

10.3. Проверка модератором

- **Кто проверяет.** Десятки тысяч модераторов по всему миру, в основном сотрудники компаний-подрядчиков. Их распределяют по языкам и рынкам. Для небольших языков модераторов мало.
- **Что видит модератор.** Сам пост, подпись, иногда комментарии и сведения об аккаунте. Полного местного контекста часто нет.
- **Сколько времени.** Обычно от нескольких секунд до нескольких минут на решение, при плановых показателях по объёму и точности.
- **Контроль качества.** Часть решений выборочно перепроверяется, точность каждого модератора регулярно оценивается.

Модератор проходит по внутренним инструкциям как по дереву решений:

1. Есть ли объект — человек или группа людей?
2. Есть ли защищённая характеристика или слово, которое её заменяет?
3. Какой вид нападки: 1-й уровень, 2-й уровень или нападки нет?

4. Применимо ли исключение: конкретные преступники, меньшинство группы, пространства по полу, политическая дискуссия?
5. Есть ли допустимый контекст: осуждение, новости, самоописание, сатира?
6. Не нарушает ли пост смежные правила: угрозы, поддержка опасных организаций?

10.4. Эскалация

Неочевидные случаи уходят выше:

- региональным специалистам, которые знают язык, историю и политику страны;
- командам, которые пишут правила и могут принять решение-исключение;
- внешним экспертам и партнёрам, например для оценки местных кодовых слов.

Некоторые нормы применяются только на этом уровне: скрытые угрозы, нападки на концепции с риском вреда, поддержка опасных организаций.

Кросс-чек (cross-check). Для известных аккаунтов — политиков, знаменитостей, СМИ — действует дополнительная проверка перед удалением. В 2022 году Наблюдательный совет раскритиковал эту систему за то, что она фактически давала влиятельным пользователям больше свободы. Meta обязалась её реформировать.

11. Санкции и обжалование

11.1. Санкции

- **Удаление контента** с уведомлением автора и указанием нарушенного правила.
- **Экран-предупреждение** вместо удаления — для шокирующего, но допустимого контента, например документации насилия.
- **Страйки аккаунту.** Первое нарушение обычно ограничивается предупреждением. Затем следуют временные запреты на публикации, как правило на 1, 3, 7 и 30 дней. Страйки со временем снимаются, обычно через год.
- **Блокировка аккаунта** за повторные или особо тяжёлые нарушения; за поддержку террористических организаций — сразу.
- **Санкции для групп и страниц:** снижение охвата, запрет монетизации, удаление, если администраторы систематически пропускают нарушения.
- **Снижение показа в ленте** для пограничного контента, который формально правил не нарушает.
- **Ограничение доступа в отдельной стране (геоблокировка),** если контент не нарушает правила Meta, но нарушает местный закон.

11.2. Порядок обжалования

1. Пользователь запрашивает повторную проверку прямо из уведомления об удалении. Её проводит другой модератор.
2. Если решение подтверждено, по многим делам можно обратиться в Наблюдательный совет (Oversight Board) в течение 15 дней.
3. Совет выбирает для рассмотрения небольшое число показательных дел. Его решение по конкретному посту обязательно для Meta.

4. Рекомендации Совета об изменении правил Meta обязана рассмотреть и публично ответить на них, но выполнять их не обязана.
5. Жаловаться можно и на то, что нарушающий контент не удалили.

11.3. Наблюдательный совет

Наблюдательный совет — независимый орган, созданный Meta в 2020 году и финансируемый через отдельный траст. В его состав входят юристы, журналисты, правозащитники и бывшие государственные деятели из разных стран. Совет оценивает решения по двум основаниям: соответствие правилам Meta и соответствие международным стандартам прав человека.

12. Особые режимы: конфликты, выборы, кризисы

В кризисных ситуациях Meta может временно ужесточать модерацию в конкретной стране.

- **Протокол кризисной политики (Crisis Policy Protocol).** Вводится при вооружённом конфликте, массовых беспорядках или резком росте насилия.
- **Временные меры:** снижение порога автоматического удаления, ограничение распространения постов с подстрекательством, ограничения для новых аккаунтов, усиление модерации на местном языке.
- **Выборы:** отдельные оперативные группы, мониторинг угроз избирателям и кандидатам, дезинформации о порядке голосования.

Эти механизмы появились после событий в Мьянме в 2017 году. Тогда Facebook использовался для разжигания ненависти к рохинджа, а модераторов со знанием бирманского языка почти не было. Миссия ООН по установлению фактов прямо указала на роль платформы. После этого Meta ввела систему оценки стран повышенного риска.

13. Критерии оценки подстрекательства: Рабатский план действий ООН

В сложных делах Наблюдательный совет, а часто и сама Meta, опираются на Рабатский план действий ООН (2012). Он помогает отличить оскорбительную речь от реального подстрекательства к насилию или дискриминации по шести критериям.

Критерий	Вопрос
1. Контекст	Какова социальная и политическая обстановка? Есть ли напряжённость или конфликт?
2. Автор	Кто говорит? Каково его влияние: политик, религиозный лидер, популярный блогер или рядовой пользователь?
3. Намерение	Хотел ли автор вызвать ненависть или насилие, а не просто резко высказаться?
4. Содержание и форма	Насколько прямо и провокационно сформулирован призыв? Это аргумент или

Критерий	Вопрос
	лозунг?
5. Охват	Сколько людей видели пост и как быстро он распространялся?
6. Вероятность вреда	Насколько реально и насколько скоро высказывание может привести к насилию или дискриминации?

Пример. Фраза «X — враги народа»:

- от анонимного аккаунта с 50 подписчиками в спокойное время — скорее всего, нарушение 2-го уровня или вовсе не нарушение;
- от политика с миллионом подписчиков накануне выборов и на фоне нападения на группу X — высокий риск подстрекательства, приоритетное удаление, возможны санкции против аккаунта.

14. Примеры решений Наблюдательного совета

Решения Совета показывают, как правила применяются к реальным постам. Полные тексты решений опубликованы на oversightboard.com.

Дело, год	Суть	Решение Meta	Решение Совета	Принцип
Пост о конфликте Армении и Азербайджана, 2020	Пост на русском называл азербайджанцев «таз*****»	Удалить	Оставить удаление	Этническое оскорбление; в условиях вооружённого конфликта риск вреда выше
Ругательство в адрес Китая, Мьянма, 2021	Грубое ругательство на бирманском в адрес Китая	Удалить	Восстановить	Нападки на государство, а не на народ
Мультфильм про Книн, Хорватия, 2022	Сербы изображены крысами, которых изгоняют из города; слово «сербы» не звучит	Оставить	Удалить	Скрытая дегуманизация понятна местной аудитории
Арабские оскорбления в	Аккаунт публиковал	Удалить	Восстановить	Самоописание и борьба с

Дело, год	Суть	Решение Meta	Решение Совета	Принцип
адрес геев, 2022	оскорбления с пояснением, что «возвращает» их			оскорблениями допустимы
Призыв к протесту женщин, Куба, 2023	Активистка резко сравнила кубинских мужчин с животными, призывая к протесту	Удалить	Восстановить	Эмоциональная политическая критика, а не дегуманизация
Мем со Сквидвардом, 2024	Мем отрицал Холокост	Оставить	Удалить	Шутливая форма не меняет содержания
Изображение еврея с дьяволом, 2026	Сгенерированное изображение и подпись о том, что евреи шантажируют политиков и СМИ	Оставить	Meta удалила пост после того, как Совет взял дело	Вредный стереотип о еврейском контроле
Дебаты о гендерной идентичности, 2025	Два поста о трансгендерных женщинах в туалетах и спорте	Оставить	Оставить (большинство)	Дискуссия о пространствах по полу допустима по правилам 2025 года

Общий вывод: система, которая ищет только «запрещённые слова», ошибается в обе стороны. Она удаляет допустимое (осуждение, самоописание, политическую критику) и пропускает опасное, если ненависть выражена намёком.

15. Изменения января 2025 года

7 января 2025 года Meta существенно пересмотрела правило. Компания объяснила это защитой свободы выражения и снижением числа ошибочных удалений. Изменения в тексте правила вступили в силу во всём мире сразу; изменения в порядке применения сначала вводились в США.

15.1. Что изменилось в тексте правила

- Правило переименовано: вместо «Hate Speech» — «Hateful Conduct».
- Разрешены утверждения о психическом заболевании или «ненормальности» на основании пола или сексуальной ориентации в контексте политических и религиозных дискуссий.

- Разрешено выступать за ограничения по полу для службы в армии, работы в правоохранительных органах и преподавания; по сексуальной ориентации — если аргумент основан на религиозных убеждениях.
- Расширены допустимые ограничения по полу для пространств: туалеты, спорт, группы поддержки, учебные заведения.
- Разрешены оскорбления по признаку пола в контексте разрыва отношений.
- Из списка унижающих достоинство высказываний убраны, в частности, сравнение женщин с бытовыми предметами или собственностью и обращение к трансгендерным и небинарным людям «оно».
- Из обоснования правила убрано указание, что разжигание ненависти создаёт атмосферу запугивания и исключения и может способствовать насилию в реальной жизни.

15.2. Что изменилось в применении

- Автоматическое обнаружение сосредоточено на тяжёлых нарушениях; менее тяжёлые чаще рассматриваются только по жалобам.
- Порог уверенности для автоматического удаления повышен.
- В США программа профессиональной проверки фактов заменена на «заметки сообщества» (Community Notes).

15.3. Оценки

Сторонники изменений считают, что они защищают политическую и религиозную дискуссию. Правозащитные организации (GLAAD, HRC, EFF и другие) указывают, что правило стало асимметричным: призывы к исключению по расе или религии по-прежнему запрещены, а по сексуальной ориентации и гендерной идентичности в ряде случаев прямо разрешены. В отчёте GLAAD за 2026 год оценки безопасности всех трёх платформ Meta снизились.

16. Прозрачность и слабые места системы

16.1. Где проверить работу Meta

- **Отчёт о соблюдении Норм сообщества** (Community Standards Enforcement Report), ежеквартально: объём удалённого контента по каждому правилу, доля найденного автоматически до жалоб, число обжалований и отменённых решений, распространённость нарушений (сколько просмотров из 10 000 приходится на нарушающий контент).
- **Отчёты о запросах государственных органов** на удаление контента и раскрытие данных.
- **Журнал изменений правил** на transparency.meta.com — с датами и правками.
- **Решения Наблюдательного совета** и ответы Meta на его рекомендации.
- **В ЕС по DSA**: отчёты об оценке системных рисков, база решений о модерации, доступ исследователей к данным.

16.2. Слабые места

- **Языковое неравенство.** Для английского классификаторы работают точнее всего. Для кыргызского, казахского, таджикского и других небольших языков мало данных для обучения и мало модераторов. Результат — ошибки в обе стороны.

- **Смешанные языки.** Посты на смеси русского и кыргызского, транслит и сленг распознаются хуже.
- **Кодовые слова и мемы** появляются быстрее, чем обновляются списки.
- **Нехватка контекста.** За несколько секунд модератор не всегда отличит цитату от призыва, а иронию от угрозы.
- **Условия труда модераторов.** Постоянный просмотр жестокого контента и высокие нормы. Это предмет судебных исков против Meta и подрядчиков в разных странах.
- **Асимметрия защиты** после изменений 2025 года.
- **Зависимость от жалоб.** Если пользователи не знают, как жаловаться, или опасаются это делать, нарушающий контент остаётся.

17. Источники

- Meta, Нормы сообщества: [Разжигание ненависти \(Hateful Conduct\)](#) — официальный текст, в том числе русская версия
- Meta, Нормы сообщества: [Опасные организации и лица](#)
- Meta, Правила рекламы: [Разжигание ненависти в рекламе](#)
- Meta, решение по делу [об изображении еврея с дьяволом](#), 2026
- GLAAD, [Social Media Safety Index 2026: платформы Meta](#)
- EFF, [анализ изменений правил Meta](#), январь 2025
- HRC, [анализ новых правил Meta](#)
- Informm, [изменения правил Meta и британский Закон о безопасности в интернете](#)
- Center for Countering Digital Hate, [More Transparency and Less Spin](#), февраль 2025
- Наблюдательный совет: [oversightboard.com](#) — полные тексты решений
- ООН, Рабатский план действий по запрещению пропаганды национальной, расовой или религиозной ненависти (2012)

Разделы 10–12 о внутренних процессах модерации основаны на публичных заявлениях Meta, отчётах Наблюдательного совета и исследованиях. Внутренние инструкции для модераторов Meta не публикует, поэтому часть деталей (сроки страйков, пороги классификаторов) приведена приблизительно.

Автор Айсулуу Кожомкулова, председатель коалиции «Айкын санарип»

Коалиция по модерации контента и свободе выражения в Кыргызской Республике «Айкын санарип» является независимым многосторонним экспертным объединением.

Миссией Коалиции является содействие формированию безопасной, устойчивой и открытой цифровой среды для пользователей в Кыргызской Республике, обеспечивающей защиту прав человека, свободу выражения мнений и равный доступ к информации.

Коалиция не является юридическим лицом и не осуществляет самостоятельную финансово-хозяйственную деятельность.

Коалиция действует на принципах: независимости, открытости, инклюзивности, прозрачности и подотчетности, уважения прав человека, многостороннего диалога, «Не навреди».