ОТЧЕТЫ И ДОКУМЕНТЫ

Искусственный интеллект и машинное обучение в вооруженных конфликтах: ключевая роль должна принадлежать человеку

Примечание: данный текст является отредактированной версией документа, опубликованного МККК в июне 2019 г.

: : : : : : :

1. Введение

Растущее количество конфликтов и быстрые технологические изменения требуют от Международного Комитета Красного Креста (МККК) как понимания масштаба последствий применения новых технологий для жертв вооруженных конфликтов, так и поиска решений для удовлетворения гуманитарных потребностей наиболее уязвимых из них.

Как и многие другие организации, работающие в различных сферах деятельности и регионах мира, МККК стремится понять, как **искусственный интеллект** (ИИ) и **машинное обучение** (или самообучающиеся алгоритмы. — *Прим. пер.*) могут повлиять на нашу работу. ИИ — это компьютерные системы, решающие задачи, которые часто ассоциируются с человеческим разумом и требуют когнитивных способностей, планирования, логического мышления и умения обучаться, а программы машинного обучения позволяют системам ИИ «учиться» на базе имеющихся данных, которые в итоге и определяют порядок их работы. Поскольку такие программы (или алгоритмы) можно применять для решения множества различных задач, это может привести к широкомасштабным и еще не до конца понятным последствиям.

Особый интерес МККК вызывают две широкие — и разные — области применения ИИ и машинного обучения: **при ведении войны** и в других ситуациях насилия[1], с одной стороны, и **в гуманитарной деятельности** по предоставлению помощи и защиты жертвам вооруженных конфликтов, с другой[2]. В данном документе излагается точка зрения МККК на применение ИИ и машинного обучения в вооруженных конфликтах, на потенциальные гуманитарные последствия, а также соответствующие правовые обязательства и этические соображения, которыми следует руководствоваться при их разработке и применении. В нем также упоминается применение систем ИИ в гуманитарной работе, в том числе и в работе МККК.

2. Подход МККК к новым технологиям ведения войны

У МККК есть давняя традиция оценивать, к каким последствиям сегодня или в ближайшем будущем могут привести изменения в ведении вооруженных конфликтов. Это включает в себя изучение новых средств и методов ведения войны с точки зрения их соответствия нормам международного гуманитарного права (которое также называют правом вооруженных конфликтов или правом войны) и риска возникновения неблагоприятных гуманитарных последствий для людей, находящихся под его защитой.

МККК не выступает против новых военных технологий как таковых. Некоторые военные технологии, например повышающие точность нанесения удара, могут помочь сторонам в конфликте как свести к минимуму гуманитарные последствия войны, особенно для мирных жителей, так и способствовать соблюдению норм ведения военных действий. Однако технологии, повышающие точность оружия, как и любые новые военные технологии,

1 ICRC, "Expert Views on the Frontiers of Artificial Intelligence and Conflict", *ICRC Humanitarian Law and Policy Blog*, 19 March 2019: https://blogs.icrc.org/law-and-policy/2019/03/19/expert-views-frontiers-artificial-intelligence-conflict.

2 ICRC, *Summary Document for UN Secretary-General's High-Level Panel on Digital Cooperation*, January 2019: https://digitalcooperation.org/wp-content/uploads/2019/02/ICRC-Submission-UN-Panel-Digital-Cooperation.pdf.

не являются полезными a priori, и конкретные гуманитарные последствия зависят от того, как они будут применяться на практике. В связи с этим крайне важно реалистично оценивать новые технологии, исходя из их технических характеристик и из того, как они используются или должны использоваться.

Любая новая военная технология должна применяться в соответствии с действующими нормами международного гуманитарного права и должна быть такой, чтобы ее можно было применять в соответствии с ними. Это — минимальное требование[3]. Однако уникальные характеристики новых военных технологий, расчетные и ожидаемые обстоятельства их применения, прогнозируемые гуманитарные последствия могут поставить вопрос о том, достаточно ли существующих норм для их регулирования или в свете прогнозируемых последствий применения новых технологий эти нормы необходимо прояснить или дополнить[4]. Ясно одно — военное применение новых и перспективных технологий не является неизбежным. Это — выбор государств, и когда они его делают, они должны оставаться в рамках существующих норм права и должны принимать во внимание возможные гуманитарные последствия для мирных жителей и комбатантов, прекративших участие в военных действиях, равно как и более общие соображения «гуманности» и «общественного сознания»[5].

3. Применение ИИ и машинного обучения сторонами в конфликте

До сих пор нет полного понимания того, каким может быть диапазон применения ИИ и машинного обучения сторонами в вооруженном конфликте, будь то государства или негосударственные вооруженные формирования. Несмотря на это, можно выделить по меньшей мере **три пересекающиеся области применения, важные с гуманитарной точки зрения**, в том числе для соблюдения норм международного гуманитарного права.

3 Государства-участники Дополнительного протокола I к Женевским конвенциям обязаны проводить правовую экспертизу новых видов оружия при их разработке и приобретении, а также перед их применением в вооруженных конфликтах. Для остальных же государств правовая экспертиза — мера, обусловленная здравым смыслом, которая помогает вооруженным силам государства вести военные действия в соответствии с его международными обязательствами.

4 ICRC. *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts,* доклад на XXXIII Международной конференции Красного Креста и Красного Полумесяца, Женева, октябрь 2019 г. (ICRC Challenges Report 2019), pp. 18-29: https://www.icrc.org/en/publication/4427-international-humanitarian-law-and-challenges-contemporary-armed-conflicts; *МККК*. Международное гуманитарное право и вызовы современных вооруженных конфликтов, доклад на XXXII Международной конференции Красного Креста и Красного Полумесяца, Женева, декабрь 2015 г. (Доклад МККК 2015 г.), с. 68–84: https://www.icrc.org/ru/document/mezhdunarodnoe-gumanitarnoe-pravo-i-vyzovy-sovremennyh-vooruzhennyh-konfliktov.

5 «Принципы гуманности» и «требования общественного сознания» упоминаются в статье 1(2) Дополнительного протокола I и в преамбуле Дополнительного протокола II к Женевским конвенциям; это положение называют «оговоркой Мартенса», оно является нормой обычного международного гуманитарного права.

3.1 Повышение автономности роботов, в том числе боевых

Одна из важных сфер применения систем цифрового **ИИ и машинного обучения — управление военной техникой**, особенно растущим количеством беспилотных роботизированных систем в воздухе, на земле и на море, многообразных по размерам и назначению. ИИ и машинное обучение могут повысить автономность таких роботизированных платформ — как оснащенных, так и не оснащенных оружием, — взяв на себя полное или частичное управление ими, например полет, навигацию, разведку или выбор целей.

По мнению МККК, **автономные системы вооружений** — системы вооружений с автономными «критическими функциями» выбора и поражения целей — представляют собой неотложную проблему с гуманитарной, правовой и этической точек зрения, учитывая риск потери человеком контроля над оружием и применением силы[6]. Лишение человека руководящей роли ведет к непредсказуемым последствиям, повышает риски для мирных жителей, ставит как правовые вопросы[7] (поскольку международное гуманитарное право требует, чтобы именно участники военных действий во время боя принимали решения сообразно текущей ситуации), так и этические[8] (поскольку при принятии решений о применении силы человеческий фактор является необходимым условием как наступления моральной ответственности, так и уважения человеческого достоинства). Поэтому МККК предлагает определить практические элементы контроля со стороны человека, которые послужат основой международно признанных ограничений уровня автономности в системах вооружений, уделяя основное внимание следующим вопросам[9]:

- **контролированию параметров оружия** (это поможет задать ограничения, касающиеся применения определенных типов автономных систем вооружений, в том числе целей, против которых они применяются, ограничения в отношении продолжительности и географического

6 ICRC, Statements to the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Lethal Autonomous Weapons Systems, Geneva, 25–29 March 2019: https://tinyurl.com/yyeadno3.

7 ICRC Challenges Report 2019 (примечание 4 выше), pp. 29–31; Neil Davison, "Autonomous Weapon Systems under International Humanitarian Law", Perspectives on Lethal Autonomous Weapon Systems, United Nations Office for Disarmament Affairs Occasional Paper No. 30, November 2017: www.icrc.org/en/document/autonomous-weapon-systems-under-international-humanitarian-law.

8 ICRC, *Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?*, report of an expert meeting, Geneva, 3 April 2018: www.icrc.org/en/document/ethics-and-autonomous-weapon-systems-ethical-basis-human-control.

9 ICRC, *ICRC Commentary on the "Guiding Principles" of the CCW GGE on "Lethal Autonomous Weapons Systems"*, Geneva, July 2020: https://documents.unoda.org/wp-content/uploads/2020/07/20200716-ICRC.pdf; Vincent Boulanin, Neil Davison, Netta Goussac and Moa Peldán Carlsson, *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control*, ICRC and Stockholm International Peace Research Institute, June 2020: www.icrc.org/en/document/limits-autonomous-weapons; ICRC, "The Element of Human Control", UN Doc. CCW/MSP/2018/WP.3, working paper, CCW Meeting of High Contracting Parties, 20 November 2018: https://tinyurl.com/y3c96aa6.

охвата применения таких систем, а также требования касательно их деактивации и предохранительных механизмов);
- **контролированию внешних условий применения** (это поможет задать ограничения, касающиеся ситуаций и районов, в которых можно применять автономные системы вооружений, особенно с точки зрения присутствия мирных жителей, а также плотности гражданского населения и гражданских объектов), и
- **контролированию посредством взаимодействия человека и машины** (это поможет определить требования к надзору со стороны человека и к его способности вмешиваться в работу автономных систем вооружений и деактивировать их, а также к предсказуемости и прозрачности работы таких систем).

Важно признать, что **не все автономные системы вооружений включают в себя ИИ и машинное обучение**; в существующих видах оружия, автономно выполняющего критические функции, например в автоматизированных комплексах противовоздушной обороны, в целом применяются простые детерминированные алгоритмы при выборе и поражении целей. Однако **программы ИИ и машинного обучения**, особенно разработанные для «автоматического распознавания целей», **могут лечь в основу будущих автономных систем вооружений, сделав их еще менее предсказуемыми**, и породить новые опасения как из-за невозможности объяснить их действия, так и из-за искажений, вносимых обучающей выборкой (см. раздел 5.2)[10]. Программы такого же типа могут применяться для «поддержки принятия решений» об определении и выборе целей, а не напрямую управлять системой вооружений (см. раздел 3.3).

С другой стороны, не все военные роботизированные системы, использующие ИИ и машинное обучение, представляют собой автономное оружие, поскольку такие программы могут управлять функциями, не связанными с выбором и поражением целей, а, например, относящимися к наблюдению, навигации или полету. С точки зрения МККК, самые острые вопросы вызывает автономность систем вооружений, включая системы с ИИ, однако вопросы в плане взаимодействия человека и машины и безопасности может вызывать и применение ИИ и машинного обучения в целях повышения автономности военной техники в целом, например беспилотных летательных аппаратов, наземного транспорта и морских судов. Те дискуссии, которые развернулись в гражданском секторе вокруг безопасности автономных транспортных средств, например беспилотных автомобилей или летательных аппаратов, возможно, было бы полезно проанализировать с точки зрения использования такой техники в вооруженных конфликтах (см. также раздел 3.3).

10 ICRC, Statement to the CCW Group of Governmental Experts on Lethal Autonomous Weapons Systems under Agenda Item 6(b), Geneva, 27–31 August 2018: https://tinyurl.com/y4cql4to.

3.2 Новые средства ведения кибернетической и информационной войны

Другая важная область применения **ИИ и машинного обучения — разработка кибероружия или наращивание потенциала в этой области**. Не все инструменты кибервойны полагаются на ИИ и машинное обучение. Однако ожидается, что эти технологии **изменят саму природу как оборонительного, так и наступательного кибероружия**. Например, системы ИИ и машинного обучения, созданные для кибервойны, могут автоматически искать и использовать уязвимости или же одновременно и защищаться от кибератак, и переходить в контрнаступление. Такое развитие событий может как расширить масштаб кибератак, так и изменить их характер, а возможно, привести к еще более серьезным последствиям[11]. Некоторые из таких систем можно даже причислить к «автономному цифровому оружию», что может поднять такие же вопросы о контроле со стороны человека, как и в случае физического автономного оружия[12].

Если говорить о кибервойне, цель МККК остается прежней — добиться того, чтобы действующие нормы международного гуманитарного права соблюдались в случае любых кибератак в ходе вооруженного конфликта и чтобы как нападающие, так и обороняющиеся от таких атак принимали меры для решения особых проблем, связанных с защитой гражданской инфраструктуры и служб[13], дабы свести к минимуму гуманитарные последствия[14].

К теме данной работы также относится использование ИИ и машинного обучения в цифровой сфере **в качестве инструментов информационной войны**, особенно создание и распространение ложной информации с целью введения в заблуждение, то есть дезинформация, или же непреднамеренное распространение ложной информации. Не всегда подобные действия опираются на ИИ и машинное обучение, но представляется, что в итоге эти технологии изменят как характер и масштаб манипуляций информацией при ведении войны, так и их возможные последствия. Системы с ИИ уже широко применяются для создания ложной информации, будь то текст, звук, фото или видео, и ее все труднее отличить от подлинной. Стороны в конфликте могут серьезно изменить ситуацию на местах, используя такие системы для повышения эффективности существующих издавна методов пропаганды, чтобы манипулировать общественным мнением и влиять

11 Miles Brundage *et al.*, *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, Future of Humanity Institute, Oxford, February 2018.

12 United Nations Institute for Disarmament Research (UNIDIR), *The Weaponization of Increasingly Autonomous Technologies: Autonomous Weapon Systems and Cyber Operations*, UNIDIR, 2017.

13 Утверждая, что действие МГП распространяется на кибероперации, МККК ни в коей мере не оправдывает ни кибервойны, ни милитаризацию киберпространства, см.: Доклад МККК 2015 г. (примечание 4 выше), с. 68–84.

14 ICRC, *The Potential Human Cost of Cyber Operations*, report of an expert meeting, Geneva, May 2019: www.icrc.org/en/document/potential-human-cost-cyber-operations.

на принимаемые решения[15]. У МККК вызывает озабоченность то обстоятельство, что распространение ложной информации с использованием цифровых технологий может спровоцировать аресты гражданских лиц, дурное обращение с ними, их дискриминацию, нападения на них или на их имущество или лишить их доступа к услугам жизнеобеспечения[16].

3.3 Изменение порядка принятия решений в ходе военных конфликтов

Возможно, **принятие решений** — самая широкая и многообразная область применения ИИ и машинного обучения, позволяющая массово собирать и анализировать данные для распознавания людей и объектов, оценки моделей образа жизни и поведения, формулирования рекомендаций для разработки военных стратегий или операций, прогнозирования будущих действий или путей развития ситуации.

Такие системы **«поддержки принятия решений»** или **«автоматического принятия решений» по сути являются продолжением средств сбора разведданных, разведки наблюдением и рекогносцировки**; они используют ИИ и машинное обучение для автоматического анализа массивов данных и предлагают человеку «советы» для принятия определенных решений либо автоматически анализируют данные, а затем самостоятельно принимают решение или начинают действовать. Смежные области применения ИИ и машинного обучения включают в себя распознавание образов, речи, изображений, лиц и моделей поведения. **Подобные системы могут применяться в самых разных областях и целях**[17]: от решений кого или что и когда атаковать[18], кого и как долго содержать под стражей[19] вплоть до стратегических решений, даже решений о применении ядерного оружия[20], а также особых операций, в том числе в попытках спрогнозировать или предупредить действия противника[21]. В зависимости от технических ограничений и возможностей, их надлежащего или ненадлежащего использования, подобные системы поддержки принятия решений могут представлять собой повышенную опасность для мирных жителей.

15 Steven Hill and Nadia Marsan, "Artificial Intelligence and Accountability: A Multinational Legal Perspective", in *Big Data and Artificial Intelligence for Military Decision Making*, STO Meeting Proceedings STO-MP-IST-160, NATO, 2018.

16 ICRC, *Symposium Report: Digital Risks in Situations of Armed Conflict*, March 2019, p. 9: www.icrc.org/en/event/digital-risks-symposium.

17 Dustin A. Lewis, Gabriella Blum and Naz K. Modirzadeh, *War-Algorithm Accountability*, Harvard Law School Program on International Law and Armed Conflict, August 2016.

18 United States, "Implementing International Humanitarian Law in the Use of Autonomy in Weapon Systems", working paper, CCW Group of Governmental Experts, March 2019.

19 Ashley Deeks, "Predicting Enemies", Virginia Public Law and Legal Theory Research Paper No. 2018-21, March 2018: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3152385.

20 Vincent Boulanin (ed.), *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, Vol. 1: *Euro-Atlantic Perspectives*, Stockholm International Peace Research Institute, Stockholm, May 2019.

21 S. Hill and N. Marsan (примечание 15 выше).

Системы поддержки принятия решений, использующие ИИ и машинное обучение, быстрее и в большем объеме собирают и анализируют доступную информацию, поэтому в случае военных действий они могут помочь человеку принимать более уместные решения, не нарушая норм международного гуманитарного права и минимизируя риски для мирных жителей. Однако излишнее доверие к результатам анализа или прогнозам тех же самых алгоритмов может привести к непродуманным решениям, спровоцировать нарушения международного гуманитарного права, сделать жизнь мирных жителей еще опаснее, особенно учитывая такие нынешние недостатки этих технологий, как их непредсказуемость, необъяснимость и искажения, вносимые обучающей выборкой (см. раздел 5.2).

С гуманитарной точки зрения речь может идти об **очень широком спектре различных решений сторон в конфликте, принятых при содействии или под влиянием ИИ**, особенно в тех случаях, когда в результате могут получить ранения или погибнуть люди либо быть уничтожены объекты и когда эти решения подпадают под действие конкретных норм международного гуманитарного права. Например, при использовании систем ИИ и машинного обучения **для выбора целей в ходе вооруженного конфликта** в случаях, когда это влечет за собой серьезные последствия для людей, надо принимать особые меры, чтобы обеспечить возможность человека выносить, исходя из ситуации, суждения, необходимые для обеспечения соблюдения правовых норм, регулирующих ведение военных действий (см. раздел 5). Если система ИИ самостоятельно принимает решение о нападении (а не просто выдает результаты анализа или «советы» принимающему решение человеку), по сути, ее можно считать автономной системой вооружений, что поднимает соответствующие вопросы (см. раздел 3.1).

Применение систем поддержки принятия решений и автоматизированных систем принятия решений может породить **вопросы правового и этического характера и в случае их использования в иных ситуациях в ходе вооруженного конфликта (например, при принятии решений о содержании под стражей)**, которые также могут иметь значительные последствия для жизни людей и регулируются конкретными нормами международного гуманитарного права. Здесь можно провести параллель с дискуссиями в гражданском секторе о роли суждения человека, об искажениях и ошибках алгоритмов оценки рисков в системах, которые полиция использует, принимая решения о задержании, а система уголовного правосудия — при вынесении приговоров и освобождении под залог[22].

В целом эти типы систем ИИ и машинного обучения могут сделать **военные действия более персонализированными** (аналогично персонализации услуг в гражданском секторе), когда цифровые системы собирают воедино персональную информацию и биометрические данные из множе-

22 Lorna McGregor, "The Need for Clear Governance Frameworks on Predictive Algorithms in Military Settings", *ICRC Humanitarian Law and Policy Blog*, 28 March 2019: https://blogs.icrc.org/law-and-policy/2019/03/28/need-clear-governance-frameworks-predictive-algorithms-military-settings; AI Now Institute, *AI Now Report 2018*, New York University, December 2018, pp. 18–22.

ства источников, включая различные датчики, средства связи, базы данных, социальные сети, а затем алгоритмы создают описание человека, его социально-экономическое положение и определяют, стоит ли выбрать его в качестве цели, либо же прогнозируют его поведение.

В целом потенциальные гуманитарные последствия — **цифровые риски** — для гражданского населения, вызванные ненадлежащим применением использующих ИИ технологий **цифровой слежки, мониторинга и вмешательства в частную жизнь**, могут сделать человека объектом преследования, привести к аресту, дурному обращению, хищению персональных данных, лишению доступа к тем или иным услугам, краже имущества или к страданиям, обусловленным чувством страха, возникающего у людей, за которыми следят[23].

4. Применение ИИ и машинного обучения в гуманитарной деятельности

Спектр применения ИИ и машинного обучения в гуманитарной работе, в том числе и в работе МККК, также может оказаться достаточно широким. Гуманитарные организации изучают применение подобных инструментов для анализа обстановки в районе осуществления деятельности, мониторинга и анализа открытых источников информации о конкретной оперативной обстановке: эти инструменты помогают **оценить потребности в гуманитарной помощи**, например, какая именно гуманитарная помощь необходима (продовольствие, вода, кров, экономическая или медицинская помощь) и где именно.

Подобные же инструменты с применением ИИ для сбора и анализа данных (например, инструменты анализа изображений, видеоматериалов или других образов с последующей оценкой урона, нанесенного гражданской инфраструктуре, а также инструменты поиска закономерностей в вынужденном перемещении людей, прогнозирования урожая или уровня оружейной опасности (количества неразорвавшихся боеприпасов)) можно использовать для **оценки гуманитарных последствий** на местах, в том числе потребностей мирного населения в защите. Эти системы можно использовать и для анализа фото- и видеоматериалов, чтобы обнаружить места военных действий, оценить их ход и гуманитарные последствия.

К примеру, МККК разработал **программу сбора данных об окружающей среде** с применением ИИ и машинного обучения, которая собирает и анализирует большие объемы данных: результаты помогают МККК вести гуманитарную работу в соответствии с конкретной оперативной обстановкой, в том числе задействовать упреждающий анализ для определения потребностей в гуманитарной помощи.

Использование ИИ и инструментов машинного обучения для решения конкретных гуманитарных задач может быть полезным при оказании

23 ICRC (примечание 16 выше), p. 8.

широкого спектра гуманитарных услуг. Например, интерес вызывают технологии, способные **повысить результативность поиска пропавших без вести людей**, например основанные на ИИ технологии распознавания лиц и обработки естественного языка для поиска совпадений имен; в рамках работы Центрального агентства по розыску МККК изучает возможности этих технологий для воссоединения разлученных конфликтами семей. МККК также изучает возможности использования ИИ и машинного обучения для **анализа образов и распознавания закономерностей на спутниковых изображениях** как для определения плотности населения при разработке инфраструктурных проектов помощи для городских районов, так и чтобы дополнить собственные документальные данные о соблюдении норм международного гуманитарного права в рамках своей работы по защите гражданских лиц.

Однако такое **применение технологий в гуманитарной сфере создает и потенциальные риски**, а также ставит правовые и этические вопросы, особенно в отношении защиты данных, конфиденциальности, прав человека, подотчетности и необходимости участия человека в принятии решений, которые могут иметь серьезные последствия для жизни и экономического положения людей. Любое их применение в гуманитарной сфере должно следовать принципу **«не навреди»** в цифровом мире и уважать право человека на неприкосновенность частной жизни, в том числе и в отношении защиты персональных данных.

МККК также будет следовать **базовым принципам и ценностям нейтральной, независимой и беспристрастной гуманитарной деятельности** при разработке и применении систем, использующих ИИ и машинное обучение, реалистически оценивая их возможности и технические ограничения (см. раздел 5.2). Совместно с Брюссельским центром исследований в области защиты неприкосновенности частной жизни МККК возглавил инициативу по защите данных в ходе гуманитарной работы; ее цель — разработать руководство по применению в гуманитарном секторе новых технологий, включая ИИ и машинное обучение, максимально используя их преимущества и при этом не упуская из виду все приведенные принципиальные соображения. В мае 2020 г. вышло второе издание Руководства по защите данных в ходе гуманитарной деятельности, совместно разработанное МККК и Брюссельским центром исследований в области защиты неприкосновенности частной жизни[24].

5. Ключевая роль должна принадлежать человеку

МККК — гуманитарная организация, которая предоставляет защиту и помощь людям, пострадавшим от вооруженных конфликтов и иных ситуаций насилия, и мандат которой основывается на международном гума-

24 ICRC, *Handbook on Data Protection in Humanitarian Action*, 2nd Edition, 30 October 2018: https://www.icrc.org/en/document/handbook-data-protection-humanitarian-action-second-edition (готовится к публикации на русском языке).

нитарном праве и Основополагающем принципе гуманности[25], поэтому **МККК считает критически важным добиться такого подхода к разработке и применению ИИ и машинного обучения, в котором ключевая роль будет отведена человеку**. Такой подход начинается с рассмотрения обязанностей и ответственности человека и того, что необходимо сделать, чтобы эти технологии соответствовали нормам международного права, а также общественным и нравственным ценностям.

5.1 Обеспечить контроль со стороны человека и верховенство суждения человека

МККК убежден в **безусловной необходимости сохранить контроль со стороны человека при выполнении задач и верховенство суждения человека при поиске решений, которые могут иметь серьезные последствия** для жизни людей во время вооруженного конфликта, особенно когда такие задачи или решения подвергают людей смертельному риску и когда они подпадают под действие конкретных норм международного гуманитарного права. **Системы ИИ и машинного обучения должны оставаться инструментами, которые служат человеку, которые помогают ему в процессе принятия решений, но не заменяют его.** Учитывая, что сейчас такие технологии разрабатываются для задач, которые всегда решал человек, возникает неизбежный конфликт между системами ИИ и машинного обучения, с одной стороны, и центральным положением человека в вооруженных конфликтах, с другой, и этот конфликт нельзя упускать из поля зрения.

Значение контроля со стороны человека и человеческого суждения особенно повышается при решении задач и принятии решений, которые могут привести к ранению или смерти людей либо к повреждению или уничтожению гражданской инфраструктуры. Такие последствия могут поставить самые серьезные правовые и этические вопросы и потребовать новых политических решений, таких как принятие новых правил и регламентирующих документов. **Самыми важными являются решения о применении силы, определяющие, кто и что станет объектом нападения во время вооруженного конфликта.** Однако серьезные последствия для людей, страдающих от вооруженных конфликтов, может вызвать привлечение ИИ к выполнению намного более широкого круга задач, например к принятию решений о задержании или аресте. Рассматривая применение ИИ для выполнения задач и принятия решений, требующих особой осмотрительности, имеет смысл обратиться к более широким дискуссиям в гражданском секторе об управлении использующими ИИ системами «особо высокого уровня функциональной безопасности», то есть такими,

25 МККК и Международная Федерация обществ Красного Креста и Красного Полумесяца, Основополагающие принципы Международного движения Красного Креста и Красного полумесяца: этические основы и инструменты гуманитарной деятельности, март 2018 г.: https://shop.icrc.org/the-fundamental-principles-of-the-international-red-cross-and-red-crescent-movement-pdf-ru.

сбой которых может повлечь за собой ранения или гибель людей либо серьезный ущерб имуществу или окружающей среде[26].

Еще один источник конфликта — **разница в скорости выполнения различных задач человеком и машиной**. Субъектами права и морали в вооруженных конфликтах являются люди, поэтому технологии и инструменты, используемые ими для ведения войны, должны конструироваться и применяться так, чтобы комбатанты могли выполнять свои правовые и нравственные обязательства. Это может иметь значительные последствия для внедрения систем ИИ и машинного обучения, используемых при принятии решений; необходимость сохранить возможность вынесения суждения человеком может потребовать изменить конструкцию и порядок применения этих систем таким образом, чтобы их скорость принятия решений была приемлема для человека, а не повышалась до скорости машин, исключая вмешательство человека.

Правовые основания контроля со стороны человека в вооруженном конфликте

Чтобы стороны в конфликте могли **обеспечить соблюдение права, необходим контроль человека в отношении ИИ и программ машинного обучения, применяемых в качестве средств и методов ведения войны**. Нормы международного гуманитарного права адресованы людям. Именно люди соблюдают и имплементируют эти нормы, именно люди будут привлекаться к ответственности за их нарушения. В частности, только на комбатантов возложена обязанность принимать решения в соответствии с нормами международного гуманитарного права, регулирующими ведение военных действий, и эту ответственность невозможно переложить на машину, компьютерную программу или алгоритм.

Данные нормы требуют, чтобы те, кто планирует нападения, принимает решения о них и осуществляет их, выносили **суждения сообразно текущей обстановке**, обеспечивающие: проведение **различия** между военными объектами, на которые можно нападать на законных основаниях, и гражданскими лицами или гражданскими объектами, которые не должны становиться объектом нападения; соблюдение принципа **соразмерности**, чтобы сопутствующий ущерб гражданским лицам и объектам в результате нападения не был чрезмерным по отношению к ожидаемому конкретному и непосредственному военному преимуществу, и принятие **мер предосторожности при нападении**, чтобы еще больше снизить риски для гражданских лиц.

26 К примеру, инициатива «Партнерство в области ИИ» уделяет пристальное внимание безопасности технологий ИИ и машинного обучения, считая ее «неотложной проблемой в краткосрочной перспективе, касающейся медицины, транспорта, инженерно-технической деятельности, компьютерной безопасности и других сфер, зависящих от нашей способности обеспечить безопасную работу систем ИИ даже в нестабильной, непредсказуемой и потенциально враждебной среде». Partnership on AI, “Safety-Critical AI: Charter”, 2018: www.partnershiponai.org/working-group-charters-guiding-our-exploration-of-ais-hard-questions.

Когда системы ИИ применяются при нападении, будь то в составе систем физически существующего оружия, систем кибероружия или систем поддержки принятия решений, **их конструкция и применение должны позволять участникам военных действий выносить такие суждения**[27]. Что касается автономных систем вооружений, государства — участники Конвенции о конкретных видах обычного оружия (КНО) признали, что «должна быть сохранена… ответственность человека» за применение систем вооружений и применение силы[28], при этом многие государства, международные организации, включая МККК, и общественные организации настоятельно подчеркивают необходимость контроля со стороны человека для обеспечения соблюдения норм международного гуманитарного права и соответствия таких систем этическим ценностям[29].

Если системы ИИ используются за рамками применения силы и выбора целей, при принятии иных решений, регулируемых конкретными нормами международного гуманитарного права, например в области содержания под стражей, скорее всего, также потребуется тщательно обдумать, как обеспечить контроль со стороны человека и верховенство суждения человека[30].

Этические основания контроля со стороны человека

Перспективные возможности применения ИИ и машинного обучения также вывели на первый план общественной дискуссии вопросы этического характера. **Общие «принципы ИИ»**, разработанные и согласованные правительствами различных стран, учеными, специалистами по этике, научно-исследовательскими институтами и технологическими компаниями, **сходятся в одном — в важности человеческого фактора** для обеспечения соблюдения правовых норм и приемлемости ИИ с этической точки зрения.

Например, принятые в 2017 г. Асиломарские принципы ИИ подчеркивают, что ИИ должен соответствовать человеческим ценностям, быть совместимым с «идеалами человеческого достоинства, прав, свобод и культурного разнообразия», находиться под контролем человека; «люди должны определять процедуру и степень необходимости передачи системе ИИ функции принятия решений для достижения целей, поставленных

27 ICRC (примечание 6 выше).

28 ООН, Доклад сессии 2018 года Группы правительственных экспертов по вопросам, касающимся новых технологий в сфере создания смертоносных автономных систем вооружений, док. ООН CCW/GGE.1/2018/3, 23 октября 2018 г., разделы III.A.21(b) и III.C.23(f): https://undocs.org/ru/CCW/GGE.1/2018/3.

29 См., например, заявления, сделанные на сессии 2019 года Группы правительственных экспертов по вопросам, касающимся новых технологий в сфере создания смертоносных автономных систем вооружений, Женева, 25–29 марта 2019 г. (на англ. яз.): https://tinyurl.com/yyeadno3.

30 Tess Bridgeman, "The Viability of Data-Reliant Predictive Systems in Armed Conflict Detention", *ICRC Humanitarian Law and Policy Blog*, 8 April 2019: https://blogs.icrc.org/law-and-policy/2019/04/08/viability-data-reliant-predictive-systems-armed-conflict-detention.

человеком»[31]. Группа экспертов высокого уровня по искусственному интеллекту Европейской комиссии подчеркнула важность «человеческого фактора и надзора», при этом системы ИИ «должны поддерживать самостоятельность человека в процессе принятия решений» и надзор со стороны человека, используя подходы, при которых человек находится в контуре управления, следит за контуром управления или выполняет контрольные функции[32]. Принципы искусственного интеллекта Организации экономического сотрудничества и развития (ОЭСР), принятые в мае 2019 г. всеми 36 государствами-участниками, а также Аргентиной, Бразилией, Колумбией, Коста-Рикой, Перу и Румынией, подчеркивают важность «ценностей, в центре которых стоит человек, и справедливости», указывая, что операторы систем ИИ «должны использовать механизмы и меры предосторожности, которые соответствуют обстановке и уровню развития технологий, и среди них возможность принятия решения человеком»[33]. Пекинские принципы ИИ, принятые в мае 2019 г. группой ведущих китайских научно-исследовательских институтов и технологических компаний, постулируют «необходимость предпринимать постоянные усилия, чтобы повышать уровень совершенства, отказоустойчивости, надежности и управляемости систем ИИ», поощрять «изучение координации между человеком и ИИ… которая позволит в полной мере задействовать присущие только человеку преимущества и качества»[34]. Кроме того, некоторые технологические компании опубликовали собственные принципы разработки систем ИИ, которые делают акцент на важности контроля со стороны человека[35], особенно в требующих особой осмотрительности системах, способных нанести ущерб[36], подчеркивая, что «цель ИИ... — дополнить, а не заменить человеческий разум»[37].

Правительства некоторых стран также разрабатывают принципы ИИ для вооруженных сил. К примеру, министерство обороны США,

31 Институт «Будущее жизни», Принципы работы с ИИ, разработанные на Асиломарской конференции, 2017 г.: https://futureoflife.org/ai-principles-russian/.

32 European Commission, *Ethics Guidelines for Trustworthy AI*, High-Level Expert Group on Artificial Intelligence, 8 April 2019, pp. 15–16: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

33 Organisation for Economic Co-operation and Development (OECD), "Recommendation of the Council on Artificial Intelligence", OECD/LEGAL/0449, 22 May 2019: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

34 Beijing Academy of Artificial Intelligence, "Beijing AI Principles", 28 May 2019: https://baip.baai.ac.cn/en.

35 Google, "AI at Google: Our Principles", *The Keyword*, 7 June 2018: www.blog.google/technology/ai/ai-principles («Мы будем создавать системы искусственного интеллекта, предлагающие надлежащие возможности для обратной связи, уместных разъяснений и апелляции. Наши технологии ИИ будут находиться под надлежащим контролем и руководством человека»).

36 Microsoft, "Microsoft AI Principles", 2019: www.microsoft.com/en-us/ai/our-approach-to-ai; Rich Sauer, "Six Principles to Guide Microsoft's Facial Recognition Work", *Microsoft Blog*, 17 December 2018: https://blogs.microsoft.com/on-the-issues/2018/12/17/six-principles-to-guide-microsofts-facial-recognition-work.

37 IBM, "IBM's Principles for Trust and Transparency", *THINKPolicy Blog*, 30 May 2018: www.ibm.com/blogs/policy/trust-principles.

призвав в стратегии ИИ на 2018 г.[38] использовать при внедрении ИИ подход, в центре которого стоит человек, поручило Совету по оборонным инновациям дать соответствующие рекомендации. Главная из них заключалась в том, что «люди должны в необходимой степени оценивать ситуацию и нести ответственность» при любом использовании ИИ[39]. На этом основывается первый из пяти принятых в начале 2020 г. принципов минобороны США, который гласит, что применение ИИ должно быть «ответственным. Персонал министерства обороны будет должным образом оценивать ситуацию и проявлять осмотрительность, продолжая нести ответственность за разработку, развертывание и применение систем ИИ»[40]. Во Франции министерство обороны обязалось использовать ИИ, руководствуясь тремя принципами: соблюдение международного права, сохранение в достаточной мере контроля со стороны человека и обеспечение постоянной ответственности командиров; оно также планирует учредить министерский комитет по этике для оценки перспективных технологий[41].

С точки зрения МККК, сохранять **контроль со стороны человека** за выполнением поставленных задач и **верховенство суждения человека** при принятии решений, имеющих серьезные последствия для жизни людей, также **жизненно важно, чтобы в какой-то мере сохранить гуманность во время войны**. **МККК подчеркивает, что необходимо сохранить за человеком возможность участия в принятии решений о применении силы в вооруженном конфликте**[42], — это убеждение основано на общих соображениях гуманности, моральной ответственности, уважения человеческого достоинства и требований общественного сознания[43].

Однако этические соображения о необходимости человеческого участия могут относиться и к более широкому контексту применения ИИ и машинного обучения в вооруженных конфликтах и других ситуациях насилия. Возможно, следует извлечь **уроки из более широких общественных дискуссий об осмотрительном применении технологий ИИ и машинного обучения двойного назначения**, особенно в системах «особо высокого уровня функциональной безопасности», и из предложений по управлению ими, высказанных в связи с этим учеными и разработчиками частного сектора. Например, компания Google заявила, что возможны «деликатные ситуации, в которых общество предпочтет оставить принятие окончательного решения за человеком, сколь бы исправно ни работала

38 DoD, *Summary of the 2018 Department of Defense Artificial Intelligence Strategy*, 2019.
39 DoD, Defense Innovation Board, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense*, 31 October 2019.
40 DoD, "DOD Adopts Ethical Principles for Artificial Intelligence", news release, 24 February 2020, available at: www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.
41 French Ministry of Defence, "Florence Parly Wants High-Performance, Robust and Properly Controlled Artificial Intelligence", *Actualités*, 10 April 2019: www.defense.gouv.fr/english/actualites/articles/florence-parly-souhaite-une-intelligence-artificielle-performante-robuste-et-maitrisee.
42 ICRC, *ICRC Strategy 2019–2022*, Geneva, 2018, p. 15: www.icrc.org/en/publication/4354-icrc-strategy-2019-2022.
43 ICRC (примечание 8 выше), p. 22.

система ИИ», а полную передачу машинам серьезных решений, например вынесения судебных постановлений о наступлении уголовной ответственности или назначения лечения, от которого зависит жизнь человека, «можно справедливо считать оскорблением человеческого достоинства»[44]. Рассматривая использование ИИ в системах распознавания лиц, компания Microsoft подчеркивает необходимость обеспечить «должный уровень контроля со стороны человека в случаях, когда использование ИИ может оказывать значительное воздействие на людей», требуя «включения человека в контур управления» или же «полноценного анализа человеком» в ситуациях, требующих особой осмотрительности, например, когда «человеку могут быть причинены телесные повреждения или психологический вред, когда могут быть ограничены его возможности трудоустройства или доступ к финансовым услугам, когда возможно нарушение прав человека или ограничение его личной свободы»[45]. Поскольку различные случаи применения ИИ в вооруженных конфликтах относятся к ситуациям, требующим наибольшей осмотрительности, такие широкие дискуссии могут пролить свет на необходимые ограничения применения ИИ.

Сохранение **контроля со стороны человека и верховенства суждения человека станет неотъемлемым компонентом** в обеспечении соблюдения правовых норм и снижении озабоченности этического плана, которую вызывают некоторые виды использования ИИ и машинного обучения. **Однако одного лишь этого недостаточно для защиты от всех потенциальных рисков**, если в процессе взаимодействия человека с машиной не уделять должного внимания таким соображениям, как: **ситуационная осведомленность** (понимание состояния системы на момент вмешательства человека); **доступный интервал времени**, в течение которого человек может действенно вмешаться в работу системы; **переоценка возможностей автоматики** (риск чрезмерного доверия человека машинам); **желание отгородиться от моральной ответственности** (риск того, что человек переложит ответственность на машину)[46]. Кроме того, чтобы обеспечить существенный и эффективный контроль со стороны человека и верховенство суждения человека, потребуется тщательно рассмотреть как возможности, так и ограничения технологий ИИ и машинного обучения.

5.2 Анализ технических ограничений ИИ и машинного обучения

Хотя сейчас принято восхищаться новыми возможностями, которые дают ИИ и машинное обучение, **необходимо реалистично оценить возможности и ограничения этих технологий**, особенно если планируется исполь-

44 Google, *Perspectives on Issues in AI Governance*, January 2019, pp. 23–24: http://ai.google/perspectives-on-issues-in-AI-governance.

45 R. Sauer (примечание 36 выше): «Мы будем рекомендовать и помогать нашим клиентам использовать технологию распознавания лиц таким образом, чтобы обеспечить должный уровень контроля со стороны человека в случаях, когда использование ИИ может оказать значительное воздействие на людей».

46 ICRC (примечание 8 выше), p. 13.

зовать их в вооруженных конфликтах. Для начала следует признать, что, применяя ИИ и машинное обучение для решения определенных задач и принятия определенных решений, мы не заменяем подобное подобным. Необходимо **понимать фундаментальные различия в том, как действуют человек и машина, их слабые и сильные стороны**; люди и машины не только выполняют задачи по-разному — они выполняют разные задачи. Мы должны четко сознавать, что «машины никогда не смогут по-настоящему очеловечиться, как бы правдоподобно они ни имитировали человека», поскольку они являются неодушевленными предметами и инструментом в руках человека[47].

Учитывая это, можно выделить несколько технических аспектов, требующих осторожности, когда речь заходит о вооруженных конфликтах (и, конечно, о гуманитарной деятельности). **Обеспокоенность в связи с ИИ и особенно машинным обучением вызывают непредсказуемость, ненадежность** (или небезопасность)[48], **непрозрачность** (или необъяснимость) и **искажения**, вносимые обучающей выборкой[49].

Системы машинного обучения не следуют предварительно заложенным в них инструкциям, а **создают свои собственные правила на основе доступных им данных**, будь то обучающие данные либо результаты проб и ошибок, вытекающие из их взаимодействия с собственной окружающей средой. В итоге, с точки зрения порядка их функционирования (то есть выводимых ими данных), в определенной ситуации (то есть в зависимости от загруженных в них данных) **эти системы гораздо более непредсказуемы**, чем те, которые строго следуют заранее заложенной в них программе, при этом их действия сильно зависят от количества и качества данных, доступных при решении той или иной задачи. Разработчику трудно понять, когда завершилось обучение — и даже чему именно система научилась. Одна и та же система машинного обучения может по-разному реагировать на одну и ту же ситуацию, а некоторые системы могут выдавать и вовсе непредвиденные решения некоторых задач[50]. Эти ключевые проблемы усугубляются, если система продолжает «учиться» и менять модели поведения после того, как ее задействовали для решения определенной задачи. Непредсказуемый характер систем машинного обучения, который может дать им преимущества при решении определенных задач, может и не представлять собой проблемы, если они играют в безобидные настольные игры[51], однако он может вызывать серьезные опасения в случае их применения во время вооруженных конфликтов, например в составе автономных

47 Google (примечание 44 выше), p. 22.

48 Dario Amodei *et al.*, *Concrete Problems in AI Safety*, Cornell University, Ithaca, NY, 2016: https://arxiv.org/abs/1606.06565.

49 ICRC, *Autonomy, Artificial Intelligence (AI) and Robotics: Technical Aspects of Human Control*, report of an expert meeting, August 2019: www.icrc.org/en/document/autonomy-artificial-intelligence-and-robotics-technical-aspects-human-control.

50 Joel Lehman *et al.*, *The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities*, Cornell University, Ithaca, NY, 2018: https://arxiv.org/abs/1803.03453.

51 David Silver *et al.*, "Mastering the Game of Go without Human Knowledge", *Nature*, Vol. 550, No. 7676, 19 October 2017.

систем вооружений, кибероружия, систем поддержки принятия решений (см. разделы 3.1–3.3).

Еще больше усложняет проблему **непрозрачность** систем машинного обучения: **результаты, которые они выдают, невозможно объяснить**. По своей природе они представляют собой «черный ящик», что мешает, а сегодня во многих случаях и просто не дает пользователю понять, *как* и *почему* выходные данные системы соотносятся с конкретными входными данными; другими словами, результаты работы такой системы невозможно ни объяснить, ни истолковать.

Непредсказуемость и необъяснимость стали **серьезной проблемой, порождающей недоверие к системам ИИ и машинного обучения**. Дополнительной причиной недоверия стали и разноплановые искажения, которые могут укреплять уже имеющуюся предвзятость человека либо порождать новую, что проявляется в конструкции и (или) способах применения таких систем. Искажения часто вносятся обучающей выборкой: ограничения в плане количества, качества и характера данных для обучения алгоритмам решения конкретной задачи вызывают перекосы в процессе решения системой стоящей перед ней задачи. Скорее всего, этот недостаток станет серьезной проблемой при применении подобных систем в вооруженных конфликтах, где трудно найти качественные и репрезентативные данные для решения поставленных задач. Однако другие виды искажений могут быть вызваны тем, какой сравнительный вес система придает различным элементам данных, либо связаны с особенностями ее взаимодействия с окружающей средой при решении поставленной задачи[52].

Опасения по поводу непредсказуемости, непрозрачности или необъяснимости и искажений уже документально подтверждены результатами применения ИИ и машинного обучения в различных ситуациях, например в системах распознавания изображений[53] и лиц[54], а также в автоматизированных системах принятия решений[55]. Существует и другая фундаментальная проблема применения ИИ и машинного обучения в таких областях, как компьютерное зрение, и это — **семантический разрыв**, проявляющийся в том, что люди и машины по-разному решают одну и ту же задачу[56]. Алгоритмы компьютерного зрения, обученные на изображениях конкретного предмета, могут распознать и классифицировать такой предмет на новом изображении. Однако такие алгоритмы не понимают *значения* или *сути* данного предмета, и значит, они могут ошибиться там, где человек никогда не совершит ошибку, например, отнеся предмет к совершенно

52 UNIDIR, *Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies: A Primer*, UNIDIR, 2018.

53 Matthew Hutson, "A Turtle — or a Rifle? Hackers Easily Fool AIs into Seeing the Wrong Thing", *Science*, 19 July 2018: www.sciencemag.org/news/2018/07/turtle-or-rifle-hackers-easily-fool-ais-seeing-wrong-thing.

54 AI Now Institute (примечание 22 выше), pp. 15–17.

55 Ibid., pp. 18–22.

56 Arnold W. M. Smeulders *et al.*, "Content-Based Image Retrieval at the End of the Early Years", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 12, 2000.

иному классу, который с ним никак не связан. Очевидно, что в ходе вооруженного конфликта в определенных случаях такие ошибки могут вызвать серьезные опасения, например, если речь идет об автономных системах вооружений или системах поддержки принятия решения при выборе целей (см. разделы 3.1. и 3.3).

В вооруженных конфликтах, вероятно, доверять системам, использующим ИИ и машинное обучение, будет еще сложнее в ситуациях, когда противники, как можно предположить, будут применять контрмеры, например, пытаясь ввести такие системы в заблуждение или создать дезориентирующие помехи. **Системы машинного обучения особенно уязвимы в условиях конфронтации**, когда противник изменяет окружающую среду, чтобы обмануть систему, либо создает при помощи собственной системы машинного обучения изображения или условия, которые система противника воспринимает как враждебные (в рамках генеративно-состязательной сети). Хорошо известен пример, когда ученым удалось обмануть алгоритм классификации изображений, заставив его распознать напечатанную на трехмерном принтере черепаху как винтовку, а напечатанный на нем же бейсбольный мяч — как чашку эспрессо[57]. Очевидно, насколько опасны такие ошибки в случае применения ИИ для распознавания изображений в системах вооружений или при принятии решений о выборе целей.

6. Выводы и рекомендации

Системы ИИ и машинного обучения могут **основательно изменить роль человека в вооруженном конфликте**, особенно повышая автономность систем вооружений и дистанционно управляемых систем, порождая новые формы кибернетических и информационных войн и в целом меняя саму природу процесса принятия решений. С точки зрения МККК, правительства, вооруженные силы и другие заинтересованные акторы должны добросовестно добиваться того, чтобы **в центре подхода к разработке и применению систем ИИ и машинного обучения неизменно стоял человек, опираясь на свои правовые обязательства и этическую ответственность**. К вопросу об использовании ИИ в системах вооружений необходимо подходить с большой осторожностью.

В качестве общего принципа **критически важно сохранить контроль со стороны человека и верховенство суждения человека, когда системы ИИ и машинного обучения используются для выполнения задач и принятия решений, могущих иметь серьезные последствия для жизни людей**, особенно когда такие задачи или решения подвергают людей смертельному риску и когда они подпадают под действие конкретных норм международного гуманитарного права. **Системы ИИ и машинного обучения остаются инструментами, которые служат человеку, они должны помогать ему в процессе принятия решений, но не заменять его.**

57 M. Hutson (примечание 53 выше).

Для соблюдения международного гуманитарного права и сохранения в какой-то мере гуманности в вооруженных конфликтах, как того требуют этические соображения, необходимо обеспечить контроль со стороны человека и верховенство суждения человека в аппаратных и программных системах с использованием ИИ, которые представляют для людей подобную опасность. Дабы человек мог эффективно играть свою роль, может потребоваться так проектировать и применять данные системы, чтобы они **помогали человеку в принятии решений с приемлемой для него скоростью, а не со скоростью, присущей машинам** и исключающей вмешательство человека. Подобные соображения могут в конце концов привести к тому, что на конструкцию и области применения систем ИИ и машинного обучения будут наложены ограничения, позволяющие сохранить существенный и эффективный контроль со стороны человека и верховенство суждения человека на основании правовых обязательств и моральной ответственности.

Общий принцип обеспечения контроля со стороны человека и верховенства суждения человека — необходимый компонент, но его одного недостаточно, чтобы защититься от потенциальных рисков применения ИИ и машинного обучения в вооруженных конфликтах. Необходимо учитывать и **другие важные факторы**: **предсказуемость** и **надежность** (безопасность) таких систем и последствий их применения; **прозрачность** (объяснимость) функционирования систем и выдаваемых ими результатов; **отсутствие искажений** (непредвзятость) при проектировании и применении систем. Всеми этими вопросами надо заниматься, если мы хотим, чтобы та или иная система **вызывала доверие**, в том числе проводить **тщательные испытания в приближенной к реальной обстановке** перед принятием систем на вооружение[58].

Характер требуемого взаимодействия человека и ИИ будет, скорее всего, определяться этическими соображениями и конкретными нормами международного гуманитарного права или иными правовыми нормами, применимыми в соответствующих обстоятельствах. Поэтому, **возможно, общие принципы придется дополнить более конкретными принципами, руководящими указаниями или нормами, регулирующими использование ИИ и машинного обучения в конкретных областях или конкретных обстоятельствах**.

С точки зрения МККК, неотложного внимания требует взаимодействие людей и машин при принятии решений об уничтожении или ранении людей либо повреждении или разрушении объектов, а также обеспечение

58 Netta Goussac, "Safety Net or Tangled Web: Legal Reviews of AI in Weapons and War-fighting", *ICRC Humanitarian Law and Policy Blog*, 18 April 2019: https://blogs.icrc.org/law-and-policy/2019/04/18/safety-net-tangled-web-legal-reviews-ai-weapons-war-fighting; Dustin A. Lewis, "Legal Reviews of Weapons, Means and Methods of Warfare Involving Artificial Intelligence: 16 Elements to Consider", *ICRC Humanitarian Law and Policy Blog*, 21 March 2019: https://blogs.icrc.org/law-and-policy/2019/03/21/legal-reviews-weapons-means-methods-warfare-artificial-intelligence-16-elements-consider.

критически важного контроля человека над системами вооружений и применением силы в вооруженных конфликтах. Системы вооружений, как с ИИ, так и без него, становятся все более автономными; существует опасность по сути передать принятие таких решений датчикам и алгоритмам, и эта перспектива поднимает вопросы правового и этического характера, которые необходимо решить достаточно срочно.

МККК выделил ключевые элементы контроля со стороны человека, необходимые для соблюдения международного гуманитарного права и снятия вопросов этического характера: эти элементы, которые могут лечь в основу международно признанных ограничений уровня автономности в системах вооружений, включают в себя контролирование параметров оружия и внешних условий его применения, а также контролирование посредством взаимодействия человека и машины[59]. МККК совершенно ясно, что виды используемых автономных систем вооружений и ситуации, в которых они применяются, должны быть ограничены[60].

Такой **подход** к автономным системам вооружений, **основанный на контроле со стороны человека, можно распространить и на более широкое внедрение систем ИИ и машинного обучения в процесс принятия решений в ходе вооруженных конфликтов**, особенно там, где существует значительный риск для жизни людей и применяются конкретные нормы международного гуманитарного права, например при использовании систем поддержки принятия решений при выборе целей для нападения или в области содержания под стражей.

59 ICRC, *Commentary on the* "*Guiding Principles*" (примечание 9 выше); ICRC, "The Element of Human Control" (примечание 9 выше); V. Boulanin *et al.* (примечание 9 выше).

60 ICRC, Statement to the CCW Group of Governmental Experts on Lethal Autonomous Weapons Systems, Geneva, 21–25 September 2020: https://documents.unoda.org/wp-content/uploads/2020/09/20200921-ICRC-General-statement-CCW-GGE-LAWS-Sep-2020.pdf.