SECURITY 2040

인공지능이 어떻게 핵전쟁의 위험에 영향을 미칠 수 있습니까?

EDWARD GEIST | ANDREW J. LOHN

전망

시의적절한 정책 문제에 관한 전문가의 통찰

RAND CORPORATION

오늘날의 핵 균형은 유지하기 힘든 몇 가지 조건에 의존합니다. 컴퓨팅 및 데이터 가용성의 발전으로 예전에는 인간의 노력이 필요했거나 완전히 불가능한 것으로 여겨졌던 수많은 작업을 기계가 수행할 수 있게 되었습니다. 이 인공 지능(AI)이 군비 경쟁에 박차를 가하거나 위기 시 국가들이 의도적이든 우발적이든 핵 사용으로 확전할 개연성을 높이는 새로운 성능이 될 수 있다는 조짐을 보입니다. [1] RAND Corporation은 AI 및 핵 안보 전문가들을 모아 2040년까지 AI가 이러한 위험을 안정화하는 힘이 될지, 아니면 불안정화하는 힘이 될지의 방식을 탐색하는 일련의 워크샵을 개최했습니다.

AI가 핵 전략에 미치는 영향은 핵이 실제로 저지를 수 있는 것만큼이나, 또는 그 이상으로 핵 성능에 대한 적의 인식에 따라 달라질 수 있습니다. 예를 들어, 한 국가가 적의 모든 핵무기 발사대를 찾아내 타격할 수 있는 능력을 개발하는 것은 기술적으로 극히 어려운 일이지만 그러한 능력이 있다면 이는 엄청난 전략적 이점이 될 것입니다. 따라서 국가들은 이러한 능력을 탐내며, 기술적 어려움과 경쟁국을 불안하게 거나 갈등의 가능성을 높이는 잠재력에도 불구하고 이러한 능력을 얻고자 합니다. 이러한 예시는 고도의 AI가 여전히 데이터의 한계와 정보 이론적 논쟁에서 기인하는 장애물을 극복하는데 어려움을 겪는 기술을 기반으로 만들어지겠지만, 추적 및 표적 시스템은 불안정화를 위해 가능하다고 오직 인식만 되면 됩니다. 거의 유효할 수 있는 성능은 이미 작동하는 성능보다 훨씬 더 위험할 수 있습니다.

AI의 개발 궤도는 상호보완적인 정보 기술 및 기타 발전과 더불어 향후 25년 동안 핵 안보 문제에 커다란 영향을 미칠 것입니다. AI 기술은 최근 몇 년 동안 그랬던 것처럼 계속해서 빠르게 발전하거나, 현재 기술의 성숙으로 정체기를 맞이할 수도 있습니다. 일부 이론가들은 기계가 일정 시점에서 자신의 지능을 향상시킬 수 있는 능력을 개발하여 인간이 이해할 수도 통제할 수도 없는 능력을 지닌 "슈퍼인텔리전스"를 탄생시킬 것이라 주장하지만, 슈퍼인텔리전스의 가능성을 포함하여 AI가 어떻게 발전할 것인가에 관해서는 거의 합의가 이루어지지 않고 있는 실정입니다. 일부 이론가들은 좌절이 뒤를 따르는 초기의 대성공을 생각하며, 다른 전문가들은 발전이 계속 될 것이라 추측합니다.

두 가지 극단적인 경우는 미래의 핵 전쟁과 제한적으로만 관련이 있습니다. 개발 지연(*인공지능의 겨울*이라고도 불림)은 현재의 핵 안보 환경에서 아주 사소한 변화를 이끌 것입니다. 슈퍼인텔리전스를 갖게 될 경우 AI는 세계를 혼란으로 몰고 갈 것이며, 그 과정에서 인류를 구하거나 파괴하거나 할 것입니다. AI가 상당히 발전하고 많은 새로운 능력을 선보이는 반면 여전히 오류가 있고 최소한 일부 측면에서는 인간보다 열등한 상태인 다른 두가지 경우는 -이런 역량이 국가 안보에 영향

을 미칠 것이라는 데에는 전문가들이 동의하지 않음에도 불구하고- 전문가 집단에서 더 많은 지지를 받는 것으로 보입니다. 일부 전문가들은 일종의 "자기 만족주의자들" 들로 구분되는데, 이들은 핵 균형을 깨뜨릴 수 있는 과제를 수행할 수 있는 AI의 생산은 실현이 불가능할 정도로 어렵다고 믿는 경향이 있습니다. "불안 조성자들"은 AI가 특정 과제를 수행할 수는 있지만 핵 전쟁의 어떠한 측면에도 포함되어서는 안 된다는 반대 견해를 갖고 있습니다. 제3의 집단인 "파괴주의자들" 은 AI를 변경하거나, 잘못 인도하거나, 전용하거나, 속이는 적의 능력에 초점을 맞추지만, 이러한 것은 안정화나 불안정화를 입증할 수 있습니다.

여러 워크샵에서 논의된 한 가지 예는 의사 결정 지원 시스템으로 행동하는 AI였습니다. 핵 발사 장치와 직접 연결되지 않아도 AI는 확전(擴戰) 문제에 대해 인간에게 여전히 조언을 제공할 수 있습니다. 최소한 일부 의사결정 과정에서 그러한 능력은 AI가 갈수록 복잡해지고 잘 명시되지 않은 과제에서 발전을 보일 경우 2040년 쯤 도달할 수 있습니다. 불안 조성자들은 충분히 강력해지기 전이나 그 것의 한계를 완전히 이해하지 못한 채 그러한 능력이 포함될까 우려할 수도 있습니다. 그러나 AI 조언자가 효과적이라고 입증된다면, 그 AI는 인간의 오류 가능성을 줄이고 계산 착오의 위험을 줄일 수 있는 철저한 투명성을 제공함으로써 안정성을 높일 수 있습니다. 하지만 많은 전문가들은 적이 해킹, 교육 데이터의 오염 또는 입력 조작을 통해 매우 유능한 AI조차도 파괴할 가능성이 있다며 우려를 표시했습니다.

향후 몇 십 년 간 전략적 안정성을 유지하려면 다극화된 세계 에서의 억제 이론의 토대들을 재검토해야 합니다. 효과적인 억제를 위해서는 AI의 발전으로 인해 빠르게 변화하는

핵 발사 장치와 직접 연결되지 않을 경우, AI는 단계적 확대 문제에 대해 인간에게 여전히 조언을 제공할 수 있습니다.

능력들과 씨름해야 합니다. 주요 고려사항으로는 실제 능력의 영향, 이런 능력(존재하든 않든)의 지각된 잠재력, 그리고 특히 적대적 행동의 결과로써 이런 능력들의 섣부른 사용 또는 오류 가능성이 있습니다. 주의를 기울여 조금만 더 생각해보면, 이런 위험들은 잠재적으로 인지되고 완화될 수 있습니다.

핵 균형 앞에 놓인 중요한 변화들의 암시

2015년 11월, 러시아는 궁극적인 "킬러 로봇", 즉 거대한 열핵탄두를 운반하기 위한 원자력 해저 무인 비행 물체를 개발 중이라고 밝혔습니다. 러시아 텔레비전은 대부분의 서방 관측통들이 의도적이라고 결론지은 "우연한" 유출을 통해 이 악몽 같은 무기의 존재를 드러내었습니다. TV 카메라들은 대외적으로는 기밀인 "해양 다목적 시스템 Status-6" 를 설명하는 블라디미르 푸틴 대통령을 위한 브리핑 슬라이드를 잠시 내보냈습니다. 거대한 어뢰처럼 생겼고 소형 원자로(3페이지 그림 참조)로 구동되는 Status-6는 바다 속의 거의 모든 것보다 더 빠를 수 있는 속도와 범위를 결합했기 때문에 적의 방어를 무력화할 것입니다 (Sutyagin, 2016). 이 무인 비행 물체는 아마도 크렘린을 파괴하려는 야비한 미국의 첫 번째 타격 공격이 시작된 후에 러시아 북해에 위치한 잠수함에서 발진하여, 대잠함 방어선을 자동으로 우회하여 약 100km/hr 속도로 바다를 가로질러 미국 해안선에 치

명적인 타격을 가할 것입니다. 수중 통신이 어렵기 때문에 무인 비행 물체는 최근에서야 AI의 발전으로 가능해진 자율적인 능력이 필요합니다.[2]

Status-6는 단순히 AI를 구체적으로 적용한 결과가 아니라, 적국이 한 국가나 그 동맹국들을 공격하지 못하도록 하기 위한 보복성 위협을 사용하는 것인, 핵 억제력에 대한 AI의 잠재적이고 무시무시한 영향력을 반영한 것입니다.[3] 핵 무인 비행 물체는 미국의 대군사 목표물 표적 능력과 미사일 방어 시스템에 맞서는 러시아의 보복 능력에 대한 러시아 지도자들의 걱정을 나타내는 가장 최근의 징후입니다. 이런 능력에 실제로 맞서지 못하기 때문에 현재의 러시아는 자국의 억제력에 대한 신뢰를 보장하기 위해 AI 개발을 희망하고 있습니다. 크렘린이 군사적 목적으로 AI를 사용할 새로운 방법들을 계속 개발하고 있기 때문에 이는 2040년 경에는 성공할 수 있습니다. 이런 노력은 미국의 우월한 능력에 대해 "비대칭적 대응"을 개발하는 수십 년 된 전략과 함께 유지되고 있습니다. 러시아의 해저 "둠스데이 드론(Doomsday Drone)"은 지금까지 드러난 이런 현상의 가장 극단적인 예에 불과합니다.[4]

핵 억제력이 2040년에는 가시화 될 것입니까? Status-6는 기술적 발전으로 핵 보유국들의 안도감이 약해질 경우 핵 보유국들이 전례 없는 신무기 시스템을 채택하고 강력한 태도를 보임으로써 자신들의 핵 억제력을 회복하려 시도할 수 있다는 냉혹한

Status-6의 구성요소

거대한 어뢰처럼 생겼고 소형 원자로로 구동되는 "해양 다목적 시스템 Status-6"는 바다 속의 거의 모든 것보다 더 빠를 수 있는 속도와 범위로 적의 방어를 무력화할 것입니다.

경고입니다. 이처럼 낮선 전략적 배치는 미국과 소련 연방이 불편한 평화를 유지했던 때보다 덜 안정적이라고 입증될 수 있으며, 이런 불안정성이 핵전쟁의 가능성을 높입니다. 위험이 증가하는 정도는 핵무기를 인도하고 핵 공격에 방어하는 새로운 방법을 모두 가능하게 하는 AI의 발전 속도와 정도에 상당 부분 의지합니다. 2017년 5월과 6월, RAND Corporation은 핵 안보에 대한 AI의 영향을 논의하기 위해 핵 안보 전문가들과 AI 연구자들을 초청해 3차례의 워크샵을 개최했습니다. 참가자들은 발전한 AI가 핵 전략 안정성을 심하게 훼손하고 그럼으로써 핵 전쟁의 위험을 증가시킨다는 데 동의한 것처럼 보였습니다. 그러나, 각각의 지지자들 내에서도 AI가 어떻게 그리고 왜 이런 영향력을 갖게 될 것인지에 대한 합의는 없었습니다.

워크샵의 방법론과 설명

25년 후 발전한 AI가 핵 안보에 미치는 잠재적인 영향을 조사하기 위해, RAND는 2017년 5월과 6월에 일련의 워크샵을 개최했습니다. 이 워크샵에는 핵 안보 전문가들과 AI 연구자들 뿐만 아니라 정부 및 산업 관계자들을 포함한 다양한 전문가 그룹들이 참가해 매우 다양한 전망을 내놓았습니다.

워크샵 1

첫 번째 워크샵은 2017년 5월 1일 RAND의 산타 모니카 사무소에서 열렸습니다. 16명의 참가자 중 다수가 핵 관련 또는 AI 관련 분야에서 일하는 RAND 연구자들이었습니다. 워크샵의 목표는 미래의 전략 지정학적 질서가 AI 기술의 발전보다 더 예측 가능하다는 전제에 입각하여 AI가 상호작용할 수 있는 전략적 환경을 구상하는 것이었습니다. 논의는 핵 보유국들 간의 갈등이

이처럼 낮선 전략적 배치는 미국과 소련 연방이 불편한 평화를 유지했던 때보다 덜 안정적이라고 입증될 수 있습니다.

더 첨예해지는 여러 개의 구체적인 시나리오들과 함께 시작되었습니다. 여기에 포함된 시나리오는 다음과 같습니다.

1. 뉴 스타트(New START) 조약이 붕괴되고 러시아가 2030년대 초반까지 전략적 핵무기에 대해 미국보다 상당한 우위를 갖는다는 "되살아나는 러시아" 시나리오
2. 중국이 전략적 핵 군비를 점차 확장하여 미국과 러시아와 동등한 수준으로 올라선다는 "떠오르는 중국" 시나리오
3. 파키스탄이 전술 핵무기를 성공적으로 사용하여 인도로 하여금 침략을 위한 군사력을 감축하게 만들고 핵 금기를 깨뜨린다는 "성공적인 제한적 사용" 시나리오
4. 붕괴 직전의 북한 정권이 한국과 일본, 중국을 공격하여 지역 황폐화를 이끈다는 "국지적 핵 전쟁" 시나리오

워크샵 참가자들은 핵 보유국이 관심을 가질 수 있는 AI 적용 사례를 찾아내기 위해 운반 시스템과 C4ISR(명령, 통제, 통신, 컴퓨터, 정보, 감시 및 정찰)의 개수와 능력을 포함한 각 핵 보유국들의 핵 무기력의 기술적 세부 사항을 제시하면서 이 시나리오들을 구체적으로 설명하라는 요청을 받았습니다. 미래의 전투 시스템은 군수 지원 시간은 느리지만 AI 발전이 방어 체계 도입보다 더 빠르게 이루어질 수 있기 때문에 현재 개발되고 있는 시스템과 비슷할 것으로 추정되었습니다. 참가자들은 이런 시스템

에 발전한 AI를 적용하면 미래의 균형 상태에 불안정한 영향을 미칠 것이라는 데에 동의하는 것처럼 보였습니다. 그러나 참가자들은 또한 모든 불안정화 기술을 대응하는 안정화 기술이 있다고 가정했습니다. 이 주제는 두 번째 워크샵에서 더 발전되었습니다.

워크샵 2

2017년 5월 25일 샌프란시스코에서 열린 두 번째 워크샵에는 19명이 참가했습니다. 이 참가들 중 7명이 AI 분야에서, 5명은 국가 안보 분야에서, 3명은 AI 및 국가 안보 분야에서, 4명은 다른 분야에서 일하고 있다고 소개했습니다. AI 관련 분야 참가자들 중에는 기업, 대학 및 비정부 AI 연구기관 뿐만 아니라 AI 정책 커뮤니티에서 활동하는 저명한 인물들이 있었으며, 국가 안보 관련 참가자들 중에는 국가 연구소의 핵무기 전문가들이 있었습니다. 참가자들은 세부 그룹으로 나뉘어 3가지 문제에 대해 논의했습니다.

첫 번째 문제는 AI를 통해 국가들이 적국의 보복 능력을 추적하고 표적으로 삼을 수 있는지, 그렇게 함으로써 대부분의 핵 전략 이론의 기초를 형성하는 확실한 보복이라는 전제를 약화시킬 수 있는가(뒷 장에서 심도 있게 논의됨) 하는 것이었습니다. 핵 안보 전문가들이 다수 참가한 한 그룹은 AI가 이를 달성할 수 있다는 결론을 내렸지만 생성적 대립쌍 네트워크(Generative Adversarial Networks, 발생기 신경망이 분류기 신경망과 상호 작용하여 점점 더 현실적인 가짜 사례를 만드는 방법을 배우는 기술)의 저명한 전문가가 포함된 2차 그룹과는 견해를 달리 했습니다. 이 그룹의 견해에서는 적대적인 조작 공격에 대한 취약성이 대부분 기계의 학습 기술에 내재되어 있어, 각 국가는 이런 방법을 적이 발사 장치를 추적하지 못하게 하는 수단으로 사용할 수 있습니다.

이 워크샵에서 거론된 두 번째 문제는 위기나 갈등 시 불거지는 전략적 핵 문제에 대해 의사 결정자들에게 조언하는 의사 결정 지원 시스템에서 AI를 사용하는 것이었습니다. 각 그룹은 이런 과제에 대한 AI의 사용에 대해 대부분 동의하지 않았으며, 일부는 AI를 엄격하게 인간의 통제 하에 유지해야 한다고 주장한 반면 다른 일부는 그것이 비현실적이라 주장했습니다. 이 주제는 나중에 상세하게 논의됩니다.

마지막으로, 이 워크샵은 미래 AI 적용을 위해 핵무기 통제로부터 배울 수 있는 가능한 교훈들을 묘사하였습니다. 참가자들은 원자력 기술과 AI는 너무나 달라, AI의 군사적 적용을 방지하기 위해 핵 확산을 미연에 방지하는 데 사용되는 법률 구조와 규정을 복제할 수 없다는 데 동의하는 것처럼 보였습니다. 핵 전쟁 관련 과제와 연결된 AI의 구체적인 사례에서, 참가자들은 AI의 통제는 어렵겠지만 이런 적용에 필요한 다른 구성요소들(즉, 센서 플랫폼)은 모니터링과 통제를 받을 수 있다고 지적했습니다. 참가자들은 데이터 통제, 인간의 재능 또는 자원을 처리함으로써 AI를 통제할 수 있는지 여부를 논의했습니다. AI 연구자 참가자들 중 몇몇은 현재 인간 재능의 부족은 일시적이고 시뮬레이션이 개선됨에 따라 교육 데이터가 덜 중요해지겠지만 하드웨어가 그 다음에 제한 요인이 될 수 있다고 주장했습니다. 그들의 견해는 그래픽 처리 장치(GPU)와 같은 구성요소를 생산하는 공장의 수가 제한되어 있어 어느 정도의 통제 체제를 구성하는 것은 가능할 수 있다는 것이었습니다. 하지만 다른 많은 참가자들은 이 견해에 회의적이었습니다.

어떤 하위 그룹은 미래의 AI 시스템이 반드시 무기 통제 체제가 될 수 있고, 인간의 투입 없이도 준수를 모니터링하고 위반을 판정할 수 있다고 도발적인 주장을 했습니다.

워크샵 3

세 번째이자 마지막 워크샵은 2017년 6월 9일 버지니아 앨링턴의 RAND 사무소에서 개최되었습니다. 이 워크샵에는 핵 문제와 관련된 일을 하고 있는 8명, AI 관련 5명을 포함한 총 15명이 참가했습니다. AI와 관련된 일을 하는 참가자들은 AI 연구 전문가라기보다는 정책 분야에서 일하고 있었습니다. 나머지 2명의 참가자는 포착 정책과 관련하여 귀중한 전문지식을 제공했습니다. 이 그룹에는 RAND 연구자와 미 육군, 국방부 및 국무부 무기 통제, 검증 및 준법 감시국의 대표단이 모두 포함되었습니다. 이 워크샵은 이전 두 번의 워크샵에서 발견된 사항을 토대로 구성되었으며, 참석한 정책 전문가들에게 이전 두 번의 워크샵에서 확인된 문제점을 어떻게 해결할 것인지 질문했습니다.

첫 번째 논의에서는 추적 및 표적의 문제에 초점을 맞춰, 참가자들에게 AI를 사용해 전략적 군사력을 취약하게 만들고자 시도하는 적의 계획을 좌절시키는 방법을 생각해보라고 요청했습니다. 참가자들은 AI 자체보다는 관련 센서 및 통신 네트워크를 공격함으로써 이런 능력을 무력화시킬 것을 제안했습니다. 후속 논의에서, 참가자들은 어떤 합의도 도출하지 않았지만(아마도 이런 방법의 기술적 세부 사항에 대해 참가자 대부분이 익숙하지 않았기 때문에) 두 번째 워크샵에서 AI 연구자들이 강조한 생성적 대립쌍 기술이 제기한 문제에 대해 고민했습니다.

두 번째 논의는 AI가 전략적 배경을 크게 바꿀 수 있다는 가능성에 비추어 미국이 자체 핵 무기력 현대화 프로그램의 궤도를 재고할 필요가 있는지 여부를 다뤘습니다. 참가자들은 현재 프로그램은 많은 취약점이 있지만, 분명한 대안들이 몇몇 분석에서는 더 나을 수 있어도 명백히 더 좋지는 않다고 지적했습니다. 국내 및 해외 기관들은 미국이 대륙간 탄도 미사일(ICBM), 잠수함 및 유인 폭격기의 현재 "3축 체제" 에서 크게 벗어나지 못하도록 압력을 가합니다.

세 번째 논의는 AI가 핵무기 통제에 어떻게 기여할 것인지를 다뤘습니다. AI는 투명성과 신뢰를 증대시킬 수 있기 때문에 조약의 검증과 같은 업무에 사용될 수 있습니다. 어떤 하위 그룹은 미래의 AI 시스템이 반드시 무기 통제 체제가 *될* 수 있고, 인간의 투입 없이도 준수를 모니터링하고 위반을 판정할 수 있다고 도발적인 주장을 했습니다. 마지막으로, 참가자들은 AI 자체에 무기 통제력을 부여하는 것이 가능하거나 바람직한지 여부를 고민했습니다. 대부분의 참가자들은 이런 목표의 실행 가능성과 바람직함에 대해 회의적이었습니다. 많은 참가자들은 이 목표를 실질적으로 불가능한 것이거나 AI 연구자들을 억류하는것과 같은 극단적이고 용인할 수 없는 개입을 요구하는 것이라고 간주했습니다.

AI의 잠재적 영향을 평가하기 위한 이론적, 역사적 배경

냉전 시기, 미국과 소련 연방은 둘 다 *상호 확증 파괴* (MAD)의 조건을, 즉 전면적인 공격은 두 사회가 파괴될 수 있는 종말론적 보복 공격에 맞닥뜨리게 될 거라는 전제를 마지못해 받아들였습니다. MAD는 전략이라기보다는 두 초강대국

이 가능하면 피하기를 희망하는 조건이었습니다(Buchan et al., 2003). 상호 취약성이 일반적인 핵 전쟁 가능성을 낮추긴 했지만, 우연이나 계산 착오로 인해 여전히 어디서나 전쟁이 일어날 가능성이 초강대국 지도자들의 마음을 무겁게 짓눌렀습니다. 예를 들어, Ronald Reagan은 과학자들에게 핵무기를 "무력하고 쓸모 없는" 것으로 만드는 미사일 방어 체계를 고안하라 했고, 소련 연방은 정교한 민방위 프로그램을 개발했습니다(Garthoff, 1987; Geist, 2012). 마찬가지로 MAD도 미국이나 소련 핵 전략의 충분한 근거는 아니었습니다. MAD는 미국에 대한 소련의 선제 타격을 확실히 억제했지만, 핵 전쟁의 위험에서도 NATO의 유럽 동맹국들을 보호하겠다고 약속한 미국의 실행 가능성을 약화시켰습니다. 미국이 MAD만 믿는다면 소련은 서유럽을 침공하기 위해 우월한 재래식 무기를 개발할 수 있을 테고, 미국은 항복 또는 전면적인 핵 전쟁 사이에서 냉혹한 선택에 직면할 것입니다. 결과적으로, 미국 전략가들과 정부 관리들은 확증 보복이라는 더 포괄적인 정책, 즉 적절하고 효과적인 대응으로 적의 군사적 도발에 대응한다는 비전을 개발했습니다(Long, 2008). 적의 도발에 비례하는 보복 위협을 가함으로써, "확증 보복"은 국지적인 공격과 전면적인 공격을 모두 확실히 억제하려 했습니다. 냉전 종식 수십 년 후, 미국은 상쇄 전략이라 불리는 변형된 방법을 통해 선제 대항 공격을 포함한 모든 공격을 억제하려 했고, 미국의 보복으로 인해 그와 같은 공격이 목표를 달성하지 못할 것이라 확신했습니다(Slocombe, 1981).

핵 전략은 단순한 억제력 이상입니다(맞은 편 페이지의 표 참조). *억제력*은 보복 위협을 활용해 적이 자국 또는 그 동맹국을 공격하지 못하게 하는 것입니다. 억제력은 *중심 억제력*(한 국가의 본토에 대한 공격 억제력)과 *확장 억제력*(한 국가의 전력적 파트너 국가들에 대한 공격 억제력)으로 분류할 수 있습니다(Cimbala, 2002). 핵무기는 또한 *강압*, 즉 적으로 하여금 원하지 않는 일을 하게 강압하는 것으로도 사용될 수 있습니다(Long, 2008, p. 9). 강요적인 억제력과 강압 위협 외에도, 핵무기는 2차 세계대전이 끝날 때 사용된 것처럼 전투에 사용될 수 있습니다. 핵 전략의 실질적인 복잡성은 *확장* 억제력에 신뢰를 제공하는 확신의 어려움으로부터 비롯됩니다. 냉전 기간 동안 미국은 전략적 전술적 핵무기를 대량 비축함으로써 소련이 재래식 무기로 유럽을 공격할 경우 기꺼이 핵무기로 대응 보복할 것임을 동맹국들에게 확신시켰습니다. Denis Healey 영국 국방장관이 언급했듯이, "러시아를 억제하는 데에는 미국의 보복에 대한 신뢰도가 5%만 있어도 충분하지만, 유럽을 안심시키기 위해서는 95%의 신뢰도"가 필요했습니다(Healey, 1989). 그러나 미국의 핵 군비 규모는 미국이 소련을 선제 타격할 가능성을 모

핵 전략 목표 범주

측면	정의
강요	
억제력	적국이 하려는 것을 단념시킴
강압	적국이 하려 하지 않는 것을 하도록 강요
확신	안보 보장이 신뢰할 만하다고 동맹국을 설득
안심	도발 행위를 삼가면 공격받지 않을 거라고 적국을 설득

색하고 있을 지도 모른다고 생각했던 소련 지도자들을 놀라게 했습니다. 이런 불신으로 인해 억제되고 있는 행동을 삼간다면 공격 받지 않을 것이라고 적국을 확신시키는 *안심*의 필요성이 강조되었습니다(Schelling, 1966).

전략적 안정성은 적이 도발적인 행동에 나설만한 커다란 동기가 결여되어 있을 때 존재합니다.[5] 다양한 시간적 척도로 구별되는 여러 종류의 전략적 안정성이 있습니다. *선제 타격 안정성*은 황폐화시키는 보복에 대해 커다란 두려움 없이 적국을 무모하게 공격할 수 있는 국가가 없을 때 존재합니다. 그런 가능성은 확실한 2차 타격 군사력의 압도적이고 자동적인 보복 위협으로 가장 잘 억제됩니다(Cimbala, 2002, p. 66). 이와는 대조적으로, *위기 안정성*은 1960년대 초 베를린과 쿠바에서 일어난 것처럼 위기가 닥쳤을 때 확전을 방지하거나 관리하는 것을 목표로 합니다(Cimbala, 2002, p. 98). 이런 환경에서 국가 지도자들은 후퇴함으로써 약점을 보이지 말라는 커다란 압력을 받지만, 국가들이 신호를 보내기 위해 핵 군사력을 기동할 경우 예기치 못한 확전 기회가 크게 증가합니다. 이럴 경우, 선제 타격 안정성을 극대화하기에 이상적인 거대한 자동 보복은 재난의 비결이 됩니다.

마지막으로, 군비 경쟁 안정성은 적국들의 군사능력간에 유리하게 이용할 수 있는 불평등이 존재하지 않을 때 달성됩니다. (Cimbala, 2002, p. 110). 각 국가는 장기간 경쟁의 위험과 비용을 관리하기 위해, 그리고 미래의 선제 타격 안정성과 위기 안정성을 위태롭게 하는 것을 피하기 위해 이런 불평등에서 벗어나려 합니다. 핵 전략은 이런 목적들이 서로 긴장 관계에 있기 때문에 어렵습니다.

극단적인 경우, AI는 MAD의 조건을 약화시켜 핵 전쟁에서 승리할 수 있게 할 수 있지만 전략적 안정성을 급속히 약화시킵니다. AI가 발전함에 따라 어떤 수준의 충돌에서는 보복의 신뢰성을 의심하게 됩니다. 미국, 러시아, 중국과 같은 주요 핵 보유국은 중심 억제력의 신뢰성 유지에 비슷한 관심을 갖고 있지만, 그들의 핵심 전략 지역이라 간주하는 곳을 차지하기 위해 지역적 이점을 추구합니다. 특정한 확장 억제 확약과 같은 신뢰성이 이미 한계에 다다른 지역은 불안정성에 특히 취약합니다. 점점 더 다극화되는 전략적 환경 또한 안정성을 위협하는 경쟁을 부추기고 있습니다. 예를 들어, 미국은 비주류 핵 보유국의 이동식 미사일 발사 장치를 추적하고 표적화하는 능력을 개발하는 데 관심이 있지만, 러시아와 중국은 똑같은 기술이 그들의 더 정교한 보복 군사력에 위협이 될 수 있음을 두려워합니다. 위기 상황에서, AI로 가능한 정보, 감시 및 정찰(ISR) 또는 무기 시스템의 사용 또는 가용도는 긴장을 부추겨서 예기치 못한 확전 기회를 증가시킬 수 있습니다. 마지막으로, 고도의 군사 능력의 추구는 미사일 방어의 역사적 사례에서 볼 수 있듯이 이런 기술을 실행할 수 없다 해도 군비 경쟁 불안정성을 유발하는 경향이 있습니다.

AI가 전략적 안정성에 제기하는 도전은 이 특별한 기술에만 존재하는 것이 아닙니다. 하지만 이런 도전은 AI 때문에, 그리고 AI와 핵 전략의 많은 잠재적 교차점에서 기술적 발전이 빠르게 일어나기 때문에 더 첨예해집니다. 자율적인 센서 플랫폼을 통제하는 ISR 데이터 분석과 같이 AI가 사용될 가능성이 있는 대부분의 특정한 애플리케이션과 목표물 자동 인식(ATR) 기술은 수십 년 동안 열렬히 추구되고 있지만 사용 가능한 기술 능력 범위를 벗어났습니다. 더 큰 돌파구 없이도, 기존 AI 기술을 사용한

전략적 안정성은 적이 도발적인 행동에 나설만한 커다란 동기가 결여되어 있을 때 존재합니다.

점진적인 발전 덕분에 이 장기적인 목표들은 예측 가능한 미래에 실제로 구현될 수 있습니다.

러시아와 중국 모두 미국이 AI를 지렛대로 활용하여 그들의 전략적 핵 군사력의 생존 가능성을 위협하고, 위기가 닥쳐오면 파국으로 치달을 수 있는 상호 불신을 부추기고 있다고 믿는 것처럼 보입니다. Paul Bracken의 언급과 같이, AI와 같은 기술에서 진행되고 있는 개선은 "최소 억제 전략을 약화시키"고 "재래식 전쟁과 핵 전쟁의 경계를 불투명하게 만들" 위험이 있습니다(Bracken, 2017).

냉전 시대의 AI

선구자 Marvin Minsky는 AI를 "인간에 의해 수행될 경우 정보를 요구하는 일을 기계가 하게 만드는 과학"이라고 정의했습니다(Minsky, 1968, p. v). AI 연구는 1950년대에 시작되었기 때문에, 인간이 "정보"를 이해하는 방법을 컴퓨터가 개조하면서 이 분야의 경계가 바뀌었습니다. AI 역시 이론적 패러다임이 유행에 따라 진화합니다. 1950년대부터 1980년대까지 지배적이었던 높은 수준의 인간 추론을 복제하는 것이 목표였던 "상징적인" 패러다임은, 인공 신경망을 사용하여 인간의 인지 능력의 생물학적 토대를 모방하고자 했던 "연결주의" 패러다임에 의해 대체되었을 뿐입니다. 20세기에는 어떤 패러다임도 실험실 증명 밖에서는 잘 작동하지 않았습니다. 이로 인해 AI 연구를 위한 자금 조달이 잘 안 되던 시기(때로 인공지능의 겨울이라 불림)가 있었습니다. 수십 년에 걸친 컴퓨터 공학의 발전, 컴퓨팅과 통신 하드웨어 및 소프트웨어의 발전, 클라우드 컴퓨팅과 빅데이터의 부상 덕택에, AI는 지난 몇 년 동안 빠르게 발전해왔고, 특히 "심층 신경망(DNNs)" 또는 다층 신경망 분야에서 가장 두드러지게 발전했습니다(Goodfellow, Bengio, and Courville, 2016). DNNs는 성능이 너무나 급속히 개선되어 AI와 거의 동의어가 되었지만, 실제로는 이전 패러다임 역시 계속 발전하고 있고 상업적, 군사적으로 널리 사용되고 있습니다. 세계 바둑 챔피언을 물리친 Alphabet DeepMind의 AlphaGo 프로그램과 같은 인상적인 최근의 일부 AI 시스템은 가능한 동작 나무를 검색하는 것과 같은 이전 기술과 결합한 DNNs를 사용합니다. AI의 60년 역사에서 한결 같이 남아 있는 한 가지는 지지자들의 커다란 희망입니다. 충분한 정보가 있다면, 겉으로 보기에 불가능할 것 같은 빈곤과 질병의 정복, 핵 전쟁에서의 승리도 가능할까요?

AI와 핵 전쟁의 교차 지점은 50년 이상 전에 나온 상투적인 공상과학 소설이지만, 그 교차지점과 실제 세계가 연결된 것은 훨씬 오래 전입니다. 최초의 AI 연구자들은 그들의 이론적 연구가 곧 실질적인 군사적 목적으로 변환될 것이라고 주장함으로써 국가 안보 업무에 연루되었고, 정부의 보장된 지원을 받았습니다. Claude Shannon은 1950년에 쓴 기초적인 논문 "체스를 두는 컴퓨터 프로그래밍(Programming a Computer for Playing Chess)"에서 컴퓨터로 그토록 유서 깊은 게임을 하게 하면 "단순한 군사 작전에서 전략적 결정을 하는 기계들"이 "가까운 미래에" 사용될 수 있다는 이론적 통찰력을 갖게 될 것이라고 주장했습니다(Shannon, 1950, p. 256). 1950년대 중반에 연구자들은 미국 공군의 지원을 받아 최초로 작동하는 AI 프로그램을 만들었습니다 (Simon and Newell, 1958; Newell, Shaw, and Simon, 1959). 그런 기계들의 적용 가능성이 전략적 이론가들의 글에서 모습을 드러내기 시작했습니다. 1950년대 후반, Herman Kahn은 받아들일 수 없는 적의 도발과 보복을 인지하도록 프로그래밍된 컴퓨터를 채용하게 될 "인류 파멸의 흉기(doomsday machines)"라는 개념을 가정했습니다(Kahn, 1960, pp. 145–154). Kahn은 이 개념을 핵 전략을 수행하지 않는 방법을 보여주는 사

고 실험으로 의도했지만, 공상과학 소설 작가들은 핵무기를 통제하는 지능형 컴퓨터라는 개념에 착안해 많은 소설 및 콜로서스(*Colossus,* 1970), *위험한 게임(WarGames,* 1983), *터미네이터(Terminator,* 1984)와 같은 영화에 영감을 주었습니다.

허구적인 스릴러물에서 핵무장한 컴퓨터들이 미친 듯이 날뛰는 이야기를 장황하게 늘어놓는 동안, 실제 세계에서 AI를 핵 전략 문제에 적용하려는 시도들은 훨씬 더 일상적이었습니다. 미국 관리들도 소련 관리들도 발사 결정을 컴퓨터에 맡기려 하지 않았습니다. 그들이 이런 특권을 교묘하게 손아귀에 넣었기 때문임과 동시에, 자동 보복이 강압이나 위기 안정성과 같은 어려운 전략적 문제에 대한 논리적인 대응이 아니었기 때문입니다. 주목할 만한 유일한 예외는 냉전 말기 소련에서 발생했습니다. 미국이 선제 타격 능력을 갖추기를 희망한다는 사실을 인지하고 자신들이 참수 작전의 목표물이 될 지도 모른다는 사실에 불안해하던 소련 지도자들은 자본주의 침략자들을 응징할 방법을 모색했습니다.[6] 소문에 따르면 소련 연방은 소련 정치 지도자와 접촉할 수 없는 경우 선제 타격을 받은 직후에 미국을 향해 생존을 위한 ICBM을 자동으로 발사하게 될 시스템 개발을 고려했습니다. "죽음의 손(Dead Hand)"이라 불리는 완전 자동화 버전은 야전 사령관들에게 발사 권한을 자동으로 위임하지만 항상 인간이 핵심적인 역할을 하도록 요구하는 "페리메트르(Perimetr)"라 불리는 버전으로 인해 거부되었던 것처럼 보입니다(Hoffman, 2009). 러시아 언론에 따르면, 페리메트르 시스템은 여전히 존재하여, AI의 일종을 사용합니다.[7] 그러는 사이에 미국은 자체 대항 능력을 강화하기 위해 AI의 사용 가능성을 연구했습니다. 1980년대 후반에 수행된 연구 프로젝트인 생존 가능 적응 계획 실험(SAPE)은 당시의 AI 기술을 사용하여 미국이 소련 연방의 이동식 ICBM 발사 장치를 표적화할 수 있는지 연구했습니다. SAPE는 핵무기를 직접 통제할 수는 없지만 전문가 시스템을 채용하여 정찰 데이터를 유인 B-2 폭격기로 운반되는 핵 표적화 계획으로 전환합니다. SAPE는 활성화될 경우 소련 연방의 핵 군비 생존 가능성을 심각하게 위협할 수 있었던 계획된 시스템 및 기능의 일부분에 불과했습니다(Roland and Shiman, 2002, p. 305; Long and Green, 2012).

AI와 새로운 지정학적 질서

20세기의 AI는 이와 같은 적용을 실현하려 노력했지만, 최근의 컴퓨팅 기술의 발전이 그들의 잠재력을 해방시킬 수 있었습니다. 딥 러닝과 같은 현대적 기술은 머신 비전과 다른 신호 처리 애플리케이션을 크게 발전시키고 있으며, 이로써 자율성과 센서 융합을 향상시킬 수 있습니다. 자율성과 센서 융합은 ISR, ATR 및 종말 유도 능력을 크게 향상시킬 수 있기 때문에 전략적 관련성에서 가장 중요할 수 있습니다. 이 모든 것은 핵 보유국이 자국 핵 전력의 생존 가능성을 확신하는 수단을 심각하게 침해할 수 있습니다. 무기 정확도의 증가로 인해 오래 전부터 사일로 기반 ICBM의 생존 가능성이 약화되었기 때문에, 미국, 러시아, 중국은 선제 타격에도 살아 남을 수 있는 것으로 간주되는 잠수함과 이동식 ICBM에 핵무기를 탑재했습니다. 생존 가능한 전력(잠수함과 이동식 미사일)이 표적화되어 파괴될 가능성을 더 높여 주는 기술로 인해 한 국가가 선제 타격을 위협할 수 있는 가능성이 더 커집니다. 이는 전략적 안정성을 약화시킵니다. 왜냐하면 이런 능력을 보유한 국가가 실제로 그 능력을 사용할 의도가 없다 해도 적이 그 사실을 확신하지 못하기 때문입니다. 따라서, 이런 능력은 여전히 잠재적인 적국에 압력을 가하고 위기 시 양보를 이끌어내기 위해 사용될 수 있습니다. 이런 능력은 위

기 시에 정치적으로 유용하게 사용하기 위해 개발할 필요는 없습니다. Alfred T. Mahan이 언급했듯이, "힘은 존재하는 것으로 알려져 있지만 행사되지 않을 때가 가장 잘 가동되는 때입니다." (Mahan, 1912, p. 105). 적국이 이런 능력이 존재할 지도 모른다고 두려워하는 한, 노골적인 대립 없이도 굴복을 이끌어낼 수 있습니다. 더 강력한 국가는 실제 위기 상황에서 선제적으로 "승리" 하는 국가입니다. 결과적으로, 대항 표적화 능력은 전략적 안정성의 타협 가능성에도 불구하고 많은 국가가 갖고 싶어하는 매력적인 능력입니다.

AI 기술은 추적 및 표적화에서 그리고 대잠수함전에서 새로운 돌파구를 찾을 수 있게 도와주거나 고정밀 재래식 무기로 단단한 ICBM 사일로를 더 쉽게 파괴하도록 도와줄 수 있습니다 (Holmes, 2016). 그런 기능은 정책 결정자들이 다른 핵 공격보다 훨씬 더 실행 가능한 재래식 무기를 사용하겠다고 위협할 수 있기 때문에 특히 불안정화를 조장할 것입니다. 재래식 무기 위협은 위기 시 적국에 커다란 압력을 행사할 수 있습니다. 이런 압력은 적국에 굴복을 강요할 수 있지만 핵 전쟁으로 이어질 수도 있습니다. 이와 같은 확전은 적국이 무장 해제되기 전에 자체 무기를 사용할 필요가 있다고 느끼거나 단지 위기가 우발적인 사용을 유발하기 때문에 발생할 수 있습니다.

러시아 같은 잠재적인 미국의 적국들은 미국이 AI 같은 기술에서 선제 핵 공격을 급격히 향상시키기 위해 그 이점을 활용할 가능성을 심각하게 받아들입니다. 지난 몇 년 동안, 러시아 군사 전문가들은 러시아의 전략적 취약성 정도에 대해 군사 언론에서 치열한 토론을 벌인 적이 있습니다.[8] 현재 및 미래 미국의 능력이 러시아의 안보에 대단히 심각한 위협을 제기한다는 그들의 가정이 이런 불안감을 부추깁니다.

핵 전략의 중요한 난제는 적국이 한 국가의 안보 보복 전력을 선제 타격 위협이나 인류 파멸의 흉기로 해석하고, 그에 따라 반응할 수 있다는 것입니다. 예를 들어, 러시아는 Status-6를 AI를 이용해 미국 방어선을 자율적으로 우회하는 최후의 2차 공격 옵션 수단이라 설명했지만, 서방 관측통들은 전면 핵추진론자의 "코발트 폭탄" 이라 해석했습니다. AI 발전은 러시아가 바람직하지 않은 전략적 특성을 갖고 있는 구형 시스템을 두 배로 줄이는 데에도 기여하고 있습니다. 예를 들어, RS-28 "사르마트 (Sarmat)" 미사일을 보유한 채로 러시아는 한 때 전략 무기 감축 협상 2조약에 따라 폐기하기로 계획했던 무기의 한 종류인 다탄두 각개 재돌입 발사체(MIRV) 탄두를 가진 거대한 사일로 기반 ICBM에 다시 투자하고 있습니다. 서방 전략 이론은 거대한 MIRV 장착 ICBM은 선제 타격에는 이상적이나 선제 방어에는 취약하기 때문에 일반적으로 이들을 불안정화 요인으로 간주합니다.

21세기 초에 러시아는 이동식 ICBM을 강조하고 소련 연방으로부터 물려 받은 거대한 사일로 기반 미사일을 폐기함으로써 자체 전력의 생존 가능성을 보장할 수 있다고 믿었습니다. 그러나 이동식 ICBM의 생존 가능성에 미치는 미국의 잠재적인 위협에 대한 소련 지도자들의 불안은 이런 계산을 변경하게 하고 미국의 공격 시 참고 견디는 대신 발사함으로써 확실히 보복하게 이끈 것처럼 보입니다. 이것은 위기 시 먼저 발사하도록 러시아 지도자들에게 커다란 압력을 가하여 우발적인 확전의 기회를 증

핵 전략의 중요한 난제는 적국이 한 국가의 안보 보복 전력을 선제 타격 위협이나 인류 파멸의 흉기로 해석하고, 그에 따라 반응할 수 있다는 것입니다.

점점 다극화되어가는 핵 환경은 AI의 잠재적 전략적 영향도 악화시키고 있습니다.

가시킬 수 있는 전략 핵대응 발사 자세를 채택하는 것과 마찬가지입니다. 러시아는 사르마트 사일로가 선제 공격을 견뎌낼 수 없을 것임을 알고 있으며, 이것의 생존 가능성은 적 탄두를 사일로에서 약간 떨어진 곳에서 폭파시켜 핵 폭발에서 살아남을 수 있게 하려는 코드명 "모지르(Mozyr')"라는 능동 방어 시스템에 달려 있습니다.

점점 다극화되어가는 핵 환경 역시 AI의 잠재적 전략적 영향을 악화시키고 있습니다. 냉전 시기에는 6개 국가가 핵 폭탄을 보유했지만, 그 중 5개 국가가 소련 연방을 주적으로 간주함으로써 전략적 질서를 필연적으로 양극화로 몰아갔습니다. 이 양극화는 위기 및 무기 경쟁 안정성을 모두 장려합니다. 오늘날에는 9개 핵무기 보유 국가와 다수의 전략적 경쟁 국가들이 서로에게 간접적인 영향을 주고 있습니다. 미국은 러시아와 중국에 대해 우려하고 있습니다. 러시아는 미국과 중국 모두와의 대결 국면을 계획하고 있고, 중국은 미국, 러시아 그리고 인도를 잠재적 적국으로 간주합니다. 인도는 중국과 파키스탄과 전략적 경쟁으로 얽혀 있고, 북한은 거의 모든 국가의 두통거리입니다.

이처럼 복잡한 다극적 갈등에 적용 가능한 전략적 안정성 이론을 개발하기 위해 아직 할 일이 많습니다. 이처럼 어려운 문제를 해결할 방법을 개발할 때, 분석가들은 의도적이거나 우발적인 핵 전쟁의 위험을 AI가 높이거나 낮출 수 있는 다양한 수단을 고려해야 합니다. 현재의 불완전한 이해에도 불구하고, 일부 새로운 능력과 그 상호 작용의 영향을 고려하기 시작하는 것은 가능합니다.

가능한 AI 미래에 대한 공개적인 전문가 의견 수렴

앞서 논의한 것처럼, 몇 가지 전망이 워크샵의 논의를 지배했습니다.

AI의 발전 예측

AI의 발전과 관련하여 4개의 주요 관점이 있습니다. 각각의 관점에 대해 어느 것이 다른 것보다 낫다는 엄격한 판단을 제공하기는 어렵습니다. 그렇지만 지지자들은 해당 관점을 대표하여 종종 격론을 벌입니다. 대부분의 경우 워크샵에 참가한 전문가들은 다양한 관점에 대해 잘 알고 있었으며, 모든 관점에서 논의할 수 있었습니다.

AI의 가능한 각각의 미래 상태에 대해서는, 핵 안보에 대한 각 범주의 전문가들의 견해가 13페이지의 표에 요약되어 있습니다. 이런 견해와 이를 지지하거나 반박하는 논쟁들은 다음 섹션에서 심도 깊게 논의됩니다.

슈퍼인텔리전스

일부 사람들은 기계가 도리 없이 인간을 지적으로 능가하게 되는 필연적인 상태가 슈퍼인텔리전스라고 예측합니다. 옥스포드의 철학자 Nick Bostrom(2014) 같은 이론가들은 슈퍼인텔리전스가 호의적으로 인류의 모든 문제를 해결하거나, 악의적이든 우발적이든 인류를 파괴한다는 두 가지 결과를 주장합니다. Bostrom은 반복적으로 자기 개선을 실행하는 AI는 실수를 거의 또는 전혀 하지 않으면서 지극히 신속하게 초인간의 지성으로 진화할 거라고 생각합니다. 이 "지성 폭발"은 몇 시간 또는 몇 분이 걸릴 수 있습니다.

이 경우, 핵 안보는 사소한 역할이 됩니다. 호의적일 경우 슈퍼인텔리전스는 핵 전쟁으로부터 인류를 구할 것이고, 악의적일 경우 핵 타격이 가능한 많은 전멸 방법 중 단 하나가 될 것입니다.

대부분의 AI 전문가들은 슈퍼인텔리전스를 임박하거나 필연적인 것으로 생각하지 않는 것 같지만, 많은 지지자들은 슈퍼인텔리전스의 비용과 이익이 아주 극단적이기 때문에 그 발생 가능성이 낮다 해도 주목할 만한 가치는 있다고 생각합니다.

제한적인 발생

진정한 슈퍼인텔리전스가 창조되지는 않아도 거대하고 불연속적인 지능의 커다란 향상이 가능할 수 있으며, 최소한 일부 측면에서는 여전히 인간보다 낮지만 그래도 크게 향상된 지능으로 이어질 수 있습니다. 예를 들어, 반복적으로 재프로그래밍 가능한 소프트웨어 시스템이 하드웨어가 도달 가능한 최고 수준에 도달하여 한층 더 발전할 수 없을 때까지 지능을 급속히 증가시킨다면 이런 일이 발생할 수 있습니다.

이 경우, 인간과 비교해 드러난 정확한 능력에 따라 AI는 자신의 비교 우위를 이용하는 데 사용될 가능성이 가장 크고, 인간은 그 비교 우위를 최대화하기 위해 사용될 수 있습니다. AI는 인간과 마찬가지로 여전히 오류를 범하며, 가능한 결과의 범위와 영향은 그와 같이 오류를 범하는 성격에 크게 의존합니다.

지속적으로 증가하는 발전

세 번째 가능성은 위에서 대략 설명한 상태와 비슷한 상태가 불연속적인 지능 향상을 통해서가 아니라 지속적으로 증가하는 발전의 결과로 AI의 발전에 도달한다는 것입니다. 이런 발전은 컴퓨팅 속도, 하드웨어 아키텍처, 알고리즘 발전, 데이터 가용도의 증가 및 비용 감소에 의존합니다. 이것은 거의 틀림 없이 최신 AI 추세에 대한 가장 적절한 해석이라 할 수 있습니다. 현재

전문가 의견으로 본 대체 AI 미래

	가능한 AI의 미래 상태		
전문가 의견의 범주	인공지능의 겨울	제한적인 발생 또는 지속적으로 증가하는 발전	슈퍼인텔리전스
자기 만족주의자들	가능함	AI의 발전에도 불충분한 데이터와 너무 복잡한 문제들	존재할 가능성은 낮지만 인간보다는 더 안전할 듯함
불안 조성자들	가능성 낮음	거의 작동하지 않는 알고리즘은 적을 경보할 수 있고, 사용할 경우 실패할 수 있음	궁극적으로 피할 수 없고, 의도적이든 우발적이든 인류를 파괴할 수 있음
파괴주의자들	중립	AI가 비극적으로 실패하도록 만들어질 수 있음, 또는 AI의 실패 유도가 안정화 확신을 제공할 수 있음	파괴와 인간 통제에 모두 저항하는 슈퍼인텔리전스

AI 능력의 증가는 어느 한 순간을 기준으로 볼 때는 일상적이지만 몇 년을 놓고 볼 때는 놀라운 것입니다. 2040년까지 20년 이상을 바라볼 때, AI 발전은 오늘날 인터넷이 1990년대 중반에 기술 문외한들에게 거의 알려지지 않았던 것처럼 놀라운 것일 수 있습니다.

철학적 또는 예방적 정책 관점에서는 다를 수 있지만, 20년 이상 지속적으로 증가하는 발전은 핵 안보 세계에 대한 결과와 영향 측면에서 제한된 발생이라는 관점과 대동소이합니다. 제한된 발생의 경우와 마찬가지로, 인간과 기계 둘 다 우월한 측면들이 있을 수 있지만 둘 다 오류를 범하며 그런 오류가 위험을 유발합니다.

AI 정체기

몇몇 워크샵 참가자들이 내놓은 마지막 전망은 현재 기술이 기술적인 성숙에 도달하면 AI 발전이 정체되리라는 것이었습니다. 그런 결과는 그 시기 동안 AI 연구는 자금 조달과 대중적인 관심이 현저히 떨어진 상태에서도 계속 발전했었던 과거의 인공지능의 겨울과는 다를 수 있습니다. 예를 들어, 컴퓨터 하드웨어가 침체기를 맞이하여 AI에게서 이론적 가능성에 도달하는 데 필요한 전산 자원을 박탈할 수 있습니다.

현재 활발한 AI 연구자들은 상상할 수 있는 일이라 인정하면서도 이런 생각에 대해 굉장히 회의적입니다. AI 발전을 위한 현재의 전 세계적 직접 투자 수준은 유례 없는 것이고 지속 가능하지 않을 수도 있지만, 사기업들과 중국과 같은 정부들이 공공연히 약속한 수준을 보면 예측 가능한 미래에 견고한 자금 조달이 지속될 것임을 알 수 있습니다. 그토록 많은 투자가 지속되면, 무어(Moore)의 법칙이 수십 년 동안 반도체의 발전을 이끌었듯 발전 전망은 저절로 달성될 수 있습니다(Mack, 2011). 또한 "AI 정체기"는 현재의 능력으로부터 상당한 발전이 일어난 후에 발생할 수 있고, 이전 두 가지 시나리오와 유사한 많은 핵 안보 난제들을 만들어낼 수 있습니다.

핵 안보에 예상되는 영향

다른 두 가지 관점이 핵 전망과의 관련성이 제한되어 있기 때문에, 이런 관점의 초점은 주로 제한된 발생과 지속적으로 증가하는 발전에 맞춰져 있습니다. 제한된 발생과 지속적으로 증가하는 발전 계획의 특징은 AI가 계속 복잡해지고 데이터로 제한된 업무에서 인간의 능력을 훨씬 뛰어 넘는다는 것입니다. 전문가들은 이런 능력이 핵 안보에 의미하는 바에 대해 동의하지 않습니다.

자기 만족주의자들

한 가지 공통된 견해는 AI가 효율성 및 투명성 제고를 제외하고는 현재 상태를 크게 변화시키지 않으리란 것입니다. 이 "자기 만족주의자들"의 견해에 대한 지지자들은 전략이나 정책보다는 기술 문제에 초점이 맞춰지는 경향이 있습니다. 세계에서 가장 유능한 AI 엔지니어들 중 일부가 포함되었지만, 참가자들은 자기 분야에서보다는 AI 안전성에 아마 더 관심을 보였고, 따라서 이런 견해가 실제보다 적게 나타났을 수 있다는 점을 유념해야 합니다. 자기 만족주의자들은 핵 전쟁이 너무 복잡해서 AI가 크게 기여할 수 없으며, 따라서 기존 균형에 대한 AI의 영향이 미미할 수 있다고 생각할 수 있습니다. 예를 들어, 자기 만족주의자들은 데이터 수집의 어려움과 실제 시스템이나 행동과 미끼 사이의 구별의 어려움은 2040년까지도 AI가 극복하기 불가능할 거라고 주장합니다. 이들은 핵 확산과 관련된 결정을 하기 위한 투입들을 식별하고 해석하는 문제를 AI에 전적으로 맡기기에는 충

또한 "AI 정체기"는 현재의 능력으로부터 상당한 발전이 일어난 후에 발생할 수 있고, 이전 두 가지 시나리오와 유사한 많은 핵 안보 난제들을 만들어낼 수 있습니다.

분히 광범위하다고 볼 것입니다. 다시 말하면, 이와 같은 과제에서 인간을 능가할 수 있는 컴퓨터라면 일반적으로 인간을 능가할 수 있을 정도로 반드시 향상되었을 것입니다.

불안 조성자들

그 반대 쪽에 AI가 기존 문제를 취약하게 만들거나 현재의 전략적 균형이 심각한 문제가 되도록 충분히 흔들 거라고 믿는 불안 조성자들이 있습니다. 이 진영에는 핵 의사 결정의 모든 측면을 결코 알고리즘에 맡기지 않으려 하는 사람들이 포함됩니다. 참가자들 중 일부는 알고리즘을 만들어 관련 목표를 달성하려는 역사적인 시도를 개인적으로 경험했습니다. 일부 경우, 이런 알고리즘 기술의 부족과 결정의 정서적 윤리적 측면을 고려하지 못한 것 때문에 참가자들은 AI와 핵 문제의 교차 지점에 대해 불편해 했습니다. 불안 조성자들은 또한 적국 발사 장치의 추적 및 표적화와 관련해 AI가 매우 효과적인 불안정화 요인으로 *인식되어야만* 한다고 주장합니다. 2차 타격 능력의 잠재적 손실로 위협을 느낄 경우, 적국은 선제적인 타격을 가하라거나 자국의 군비를 확장하라는 압력을 받게 될것이며 둘 다 바람직한 결과는 낳지 못할 것입니다.

파괴주의자들

자기 만족주의자 진영과 불안 조성자 진영 사이에 위치하는 세 번째 관점은 적대 행위에 대한 AI의 민감성에 관한 우려에 뿌리를 두고 있습니다. 이 관점은 이런 적대적인 공격이 매우 효과적일 수 있다는 이론적 고찰과 증거에서 파생되었습니다. 이 견해는 AI가 제한된 발전을 보여주거나 효과적이라고 인정받지만, 그렇지 않은 경우들과 겹치는 면이 있을 수 있지만 항상 같은 결론을 이끄는 것은 아니라고 보고 있습니다.

많은 설득력 있는 증거를 통해 머신 러닝 알고리즘을 파괴하려는 적대적인 소수의 노력은 커다란 효과를 보여주었습니다. 일부 연구자들은 이게 머신 러닝의 보편적인 특징이며, 앞으로도 몇 년 동안 지속될 거라고 기대된다고 주장합니다. 추적과 표적화에 효과적인 AI가 세태를 불안정하게 만들고 확산 또는 악화를 이끌 경우, 적국이 이런 적대적인 방법을 사용하여 탐지를 미연에 방지하는 자체 능력을 신뢰할 경우, 적국은 2차 타격 전력의 생존 가능성을 다시 신뢰할 수 있고, 그럼으로써 전략적 안정성을 재설정할 수 있습니다. 반면 어떤 이는 그렇게 하면 AI의 선제 타격을 식별하는 능력을 파괴할 수 있고, 그런 타격을 실행 가능한 옵션으로 만들어 결국 불안정화를 이끌 수 있다고 생각할 수 있습니다.

구체적인 사례: 이동식 미사일 발사 장치 추적

AI는 너무 잘 작동하기 때문이 아니라 불확실성을 줄 정도로 잘 작동하기 때문에 전략적으로 불안정을 이끌 수 있습니다. 이 점을 설명하기 위해 이 섹션에서 우리는 이동식 미사일 발사 장치의 표적화 문제에 대한 초기의 RAND 연구 결과를 설명합니다.

확실한 2차 타격에 대한 영향

대부분의 핵 보유국은 이동식 미사일 발사 장치가 추적 및 표적화하기 어렵고 따라서 생존이 가능하다고 생각하기 때문에 선호합니다. 이 미사일들은 정기적으로 도로나 철도를 통해 이동하며, 적에게 그 위치가 계속 알려지지 않는다면 이 미사일들을 위협하는 유일한 방법(미사일들이 야전에 배치되기 전에 무기를 파괴하는 선제 타격 제외)은 핵무기로 커다란 미사일 경비 지역을 표적화하려 하는 것입니다. 이런 폭격 전략 조차도 가능한 미사일 위치 탐지가 최소한 어느 정도는 좁혀질 수 있어야만 실질적인 효과가 있습니다. 소련의 이동식 ICBM 발사 장치를 표적화하려는 냉전 시기의 계획들은 폭격 전략과 소련 연방의 미사일 이동 방식의 패턴에 대한 정보를 결합한 것이었습니다.

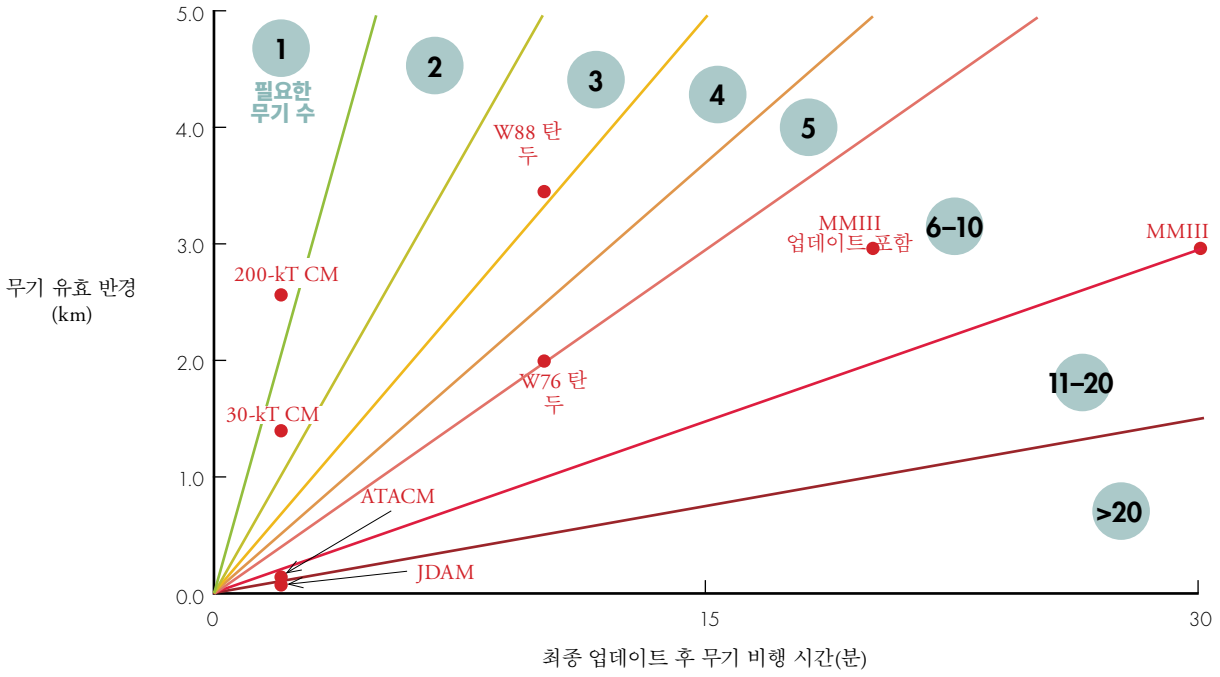
AI는 ISR 및 분석 시스템에 중대한 기여를 할 수 있었고, 이런 가정을 뒤집고 이동식 미사일 발사 장치를 선제 타격에 취약하게 만들수 있었습니다. 러시아와 중국은 이동식 ICBM의 억제력에 크게 의존하고 있었기 때문에 이런 가능성은 이들 국가의 방어 계획자들을 크게 놀라게 했습니다. AI가 적 미사일의 배치 데이터를 통합하는 능력만 꾸준히 향상시킨다 해도, 국가의 안보 감각과 위기 안정성이 실질적으로 약화될 수 있습니다. 현재 ATR, 센서 통합 및 신호 처리에 필요한 능력은 매우 어려운 상태로 남아 있습니다. 하지만 이런 난제들이 이런 무기들을 완전히 쓸모 없게 만들 정도로 충분히 잘 작동하는 것과 완전히 신뢰를 상실한 것 사이에서 불편한 중간 지점을 유지할 수는 있어 보입니다.

어려운 기술적 도전

RAND는 감지, 이미지 처리 및 무기의 속도와 살상 반경의 한계를 포함한 적대 세력 추적 및 표적화를 위한 모델을 개발했습니다.[9] 이런 한계 중 일부는 비교적 짧은 시간 내에 AI의 발전으로 극복될 수 있지만, 다른 일부는 2040년까지도 어려울 것으로 보입니다. 예를 들어, 표적 위치를 완벽하게 안다 해도 이동식 표적은 무기가 발사되는 시간과 도착하는 시간 사이에 이동할 수 있습니다. 이동식 시스템을 표적화하는 무기들은 더 빠르게 비행하고 경로를 더 잘 조정할 수 있겠지만, 이런 무기는 필요한 포의 양을 실질적으로 줄이기 위해 여전히 매우 정교한 종말 유도 능력이 필요합니다. 결과적으로, 이미지 처리와 표적 식별의 발전이 있더라도 많은 대형 무기나 근거리 발사를 위한 소형 무기가 필요합니다. 17페이지의 그림은 유효 반경이 0-5km인 무기로 이동식 표적을 파괴하는 데 필요한 다양한 유형의 탄두 수를 보여줍니다. 직경(km)으로 측정되는 거대한 "살상 반경"에도 불구하고 미사일 발사 장치를 파괴하는 데 커다란 확신을 주기 위해 탄도 미사일에 의해 운반되는 여러 개의 열원자 탄두가 필요합니다. 예를 들어, 475-kT W88 탄두의 경우 1개의 표적을 타격하기 위해 Trident II 미사일에 탄두 3개를 탑재하고 10분 간 비행하면 되지만, 100-kT W76 탄두의 경우에는 1개의 표적을 타격하는 데 5개의 탄두가 필요합니다. 그러나 분석에 따르면, 가까운 위치 표적(그림의 30-kT CM와 200-kT CM)에서 발사된 정확한 크루즈 미사일(CM)은 단 한 개 또는 두 개의 탄두로 이동식 미사일 발사 장치를 타격할 수 있었습니다. 초근거리에서 발사될 경우(비행 시간 단 몇 분), 재래식 무기라도 실행 가능한 옵션이 될

매우 근거리에서 발사될 경우 재래식 무기라도 실행 가능한 옵션이 될 수 있으며 따라서 선제 대항 타격의 신뢰성을 크게 증가시킬 수 있습니다.

표적 타격에 필요한 최소 무기 수



이 그림은 유효 반경이 0~5km인 무기로 이동식 표적을 파괴하는 데 필요한 다양한 유형의 탄두 수를 보여줍니다. 거대한 "살상 반경"은 직경(km)으로 측정되지만, 탄도 미사일에 의해 운반되는 여러 개의 열원자 탄두는 미사일 발사 장치를 파괴하는 데 커다란 확신을 주기 위해 필요합니다. ATACM = 육군 전술미사일시스템, JDAM = 합동 정밀직격탄, kT = 킬로톤; MMIII = 대륙간 탄도탄 III.

수 있고, 따라서 선제 대항 타격의 신뢰성을 크게 증가시킬 수 있었습니다.

이런 발견에 따르면 이동식이면서 자율적인 센서 플랫폼과 결합한 AI는 이동식 ICBM 발사 장치의 생존 가능성에 전략적으로 불안정한 위협을 가할 수 있지만, 군축협정이 위협을 미연에 방지할 수 있도록 도울 수 있다는 약간의 희망을 제공할 수도 있었습니다. 이동식 ICBM 발사 장치에 확실한 위협을 가하기 위해, 발사 장치를 추적하고 표적화하는 AI 시스템이 비교적 정교하다 해도 공격 부대는 발사 장치에 매우 근접한 곳에 주둔해야 합니다. 이런 조건에서도, 공격이 수행될 수 있는 "취약점들"은 단 몇 분 간만 지속될 수 있으며, 향후 공격 부대에게 이런 사태가 발생할 경우 반드시 기회를 잡으라고 압박할수 있습니다. 무

장 해제 타격을 우려하는 국가들은 주변에 이런 세력들이 나타나면 매우 놀랄 것이고, 아마도 "사용하지 않으면 잃게 되는" 상황에 처했다고 생각할 것입니다. 따라서 이런 이동은 불신의 악순환을 만들 수 있고, 실제로 어느 쪽도 의도하지 않은 갈등을 조장할 수 있습니다. 그러나 이처럼 바람직하지 않은 결과는 이동식 미사일 발사 장치와 일정한 거리 내에 그와 같은 무장 해제 타격에 사용할 수 있는 무기를 주둔 또는 배치하지 않는다는 검증 가능한 협정으로 피할 수 있습니다.

구체적인 사례 — 믿음직한 조언자 AI

2차 타격 부대에 대한 믿음을 잠재적으로 감소시키는 것 외에도, AI는 실수로 인한 전쟁 발발, 확전, 발사 결정을 처리하는 한 국가의 능력을 약화시킬 수 있습니다. 자율적인 통제는 국내 발사 장치나 명령 통제식 플랫폼에서 직접 실현되기가 어렵지만, 이런 것이 AI가 영향력을 행사하는데 필수적인 것은 아닙니다. 이런 영향은 컴퓨터 프로그램, 시뮬레이션 또는 데이터 분석 절차가 인간의 결정을 알리기 위해 사용되는 정도로 이미 일어나고 있습니다. AI는 의사 결정을 돕기 위해 더 광범위하게 사용될 것으로 기대됩니다(일반적으로 *의사 결정 지원 시스템*이라 불림).

의사 결정에 있어 AI의 가능한 역할

AI는 급격히 발전하면서 점점 복잡한 업무에서 인간을 능가하는 성과를 보이고 있습니다. 바둑 세계 챔피언을 이긴 Alphabet DeepMind의 AlphaGo는 AI와 전략 전문가조차도 놀라게 했습니다(Etherington, 2017). 확실히, 바둑의 의사 결정은 핵 전쟁에서보다 처리하기가 훨씬 더 단순합니다. 움직임이 순차적이고 명확히 제한되어 있기 때문입니다. 하지만 DeepMind 개발자들은 군수관리, 기반 시설 및 지정하기 어려운 움직임과 전략의 범

AI는 급격히 발전하면서 점점 복잡한 업무에서 인간을 능가하는 성과를 보이고 있습니다.

위로 완성되는 군사 개입을 반영하는 스타크래프트(Starcraft)라는 컴퓨터 게임을 할 수 있는 AI를 만들고 있습니다(Woyke and Kim, 2017). 스타크래프트 역시 핵 전쟁보다 훨씬 더 단순하지만, 2040년이 되면 AI 시스템이 군대 전쟁 게임의 측면들이나 단계들을 플레이하거나 인간을 능가하는 수준으로 활동할 수 있다고 기대하는 것이 비합리적으로 보이지는 않을 것입니다. 그런 능력이 일단 증명되면, 명령 결정을 내리는 인간들은 AI 시스템의 제안들을 인간 조언자들의 제안과 같거나 더 훌륭한 것으로 취급할 것입니다. 이와 같이 잠재적으로 정당화되지 않는 신뢰는 고려해야 할 새로운 위험을 제기합니다.

일부 워크샵 참가자들은 컴퓨터가 핵 전쟁에 대한 결정에 영향을 주도록 인간이 내버려둘 리가 없다고 확신했지만, 다른 참가자들은 이런 생각을 점점 더 쉽고 편안하게 받아들일 수 있었습니다. 입증되지 않은 이야기지만, 관점의 차이는 세대 간 차이였고, 특히 AI가 향후 몇 십 년 동안 점점 복잡해지는 일상적인 업무에서 스스로를 계속 입증할 것이므로 이는 2040년 쯤 고삐를 물려 받게 될 사람들은 어느 정도의 통제력을 더 편안하게 넘겨주리란 것을 시사합니다. 미국인들의 경우 운전을 할 때 길을 찾고 일정관리를 용이하게 하고 간단한 이메일에 답장하는 것을 AI에게 의존하는 것은 이미 흔한 일이 되었습니다. 그러나 이런 성공들은 일상적인 결정과 핵 전쟁 사이의 간격을 고려할 경우 보증되지 않은 자신감을 구축할 위험이 있습니다.

적대적 행위로 인한 제한적 효과

핵 전쟁 관련 업무를 위한 AI 개발 시 두 가지 중요한 문제가 있습니다. 첫째, 1945년 일본에 대한 미국의 폭격이 무조건적인 항복을 이끌어낸 이후, 핵무기는 사용된 적도 교환된 적도 없었습니다. 따라서 실제 교육 데이터가 완전히 부족합니다. 시뮬레이션, 전쟁 게임, 연습이 이런 문제를 완화시킬 수 있지만, 실제 데이터의 부족이 인간의 학습과 의사 결정 역시 제한한다는 점을 결코 잊어서는 안 됩니다.

두 번째 커다란 특징은 길찾기나 일정 조정과 같은 일반적인 업무에 관련된 모든 당사자는 업무를 성공적으로 완수하기 위해 동일한 장려책을 갖게 되지만, 핵 전쟁은 근본적으로 적대적이라는 점입니다. AI 시스템을 파괴하는 다양한 방법들이 있고, 그런 파괴는 앞으로 오랫동안 효과적인 옵션이 될 수 있을 것처럼 보입니다. 이제부터는 현재 존재하는 문제 유형의 명확한 예로 해킹, 교육 데이터 공격, 입력 조작에 대해 간략하게 논의할 것입니다.

해킹

해킹이 AI에 특정한 것은 아니지만, 컴퓨터와 관계가 있는 한 AI는 해킹에 취약하다고 간주되어야 합니다. 나중에 논의되겠지만, 지능 자체가 해킹될 수 있습니다. 하지만 데이터 역시 입력 및 출력 시, 또는 디스플레이로 출력하는 과정에서 변경될 수 있습니다. 물론 핵 사업에서 어떤 역할을 수행하는 모든 AI는 신중하게 보호되겠지만, 또한 높은 가치의 표적이 될 수 있습니다.

교육 데이터 공격

AI를 파괴하는 또 다른 방법은 교육 데이터를 조작하는 것입니다. 이것은 공개적으로 사용 가능한 데이터 내의 오류 샘플을 포함하여 내부자가 데이터를 교체하거나 해킹으로 데이터를 전환하는 것 또는 잘못된 선례를 설정하는 방식으로 적이 자신의 행동을 신중하게 선택하는 것과 같이 여러 가지 방법으로 행해질 수 있습니다.

연구 범위는 다양한 머신 러닝 알고리즘을 위한 교육 데이터 오염 전략 및 그 효과를 약술하는 것부터 시작했지만(Anderson et al., 2017; Biggio, Nelson, and Laskov, 2012; Kearns and Li, 1992), 해야 할 작업이 많이 남아 있어서 더 많은 발견이 추정됩니다. 이 작업의 대부분은 본질적으로 대립 관계에 있는 소수의 다른 응용 공간인 바이러스 백신 공동체에서 이끌고 있습니다. 최근 몇 년 사이에 이 공동체는 더 전통적인 신호 기반 방식과 대립되는 머신 러닝으로 방향을 틀었습니다. 일부는 데이터 조작 공격에도 불구하고 머신 러닝이 효과적으로 남아 있도록 하는 방법들을 추구하고 있지만, 이런 노력은 초기 단계에 불과합니다(Kegelmeyer et al., 2015). 데이터 조작은 앞으로 오랫동안 위협이 될 것으로 예상됩니다.

입력 조작

AI를 파괴하는 세 번째 기회는 완전히 교육 받은 후에 옵니다. 미묘하게 데이터를 조작하면 고성능 AI 시스템조차도 공격자가 선호하는 결론에 도달할 수 있습니다. 이는 이미지 식별에서 증명되었습니다. 이 경우 인간이 감지할 수 없을 정도로 굉장히 작은 변화가 이미지에 적용되어, AI가 변경된 이미지를 공격자가 선택한 범주로 분류하게 했습니다(Karpathy, 2015). 인간이 모든 입력 또는 분류에 대해 정밀한 지식을 갖추고 있지 않을 수 있는 핵 문제에서는 이것이 더 어려울 수 있습니다. 이미지 인식의 경우, 입력의 적대적 범위는 단순했고 픽셀로 제한되었습니다. 다른 작업의 경우, 적은 특정한 패턴으로 부대를 동원하거나 특정한 순서로 특정 메시지가 담긴 성명을

발표해야 할 수도 있지만, 최소한 원칙적으로 완전한 교육을 받은 AI 시스템을 "속이는" 것도 여전히 가능합니다. 중요한 것은, 입력 조작 공격은 적이 교육 받은 시스템에 접근하도록 요구하지 않으므로, 잘 보호된 AI조차도 여전히 취약할 수 있습니다(Papernot, McDaniel, and Goodfellow, 2016). 취약성의 정도를 이해하고 입력, 출력 및 데이터의 어떤 부분들에 보안을 유지해야 하는지를 이해하려면 더 많은 연구가 필요합니다.

핵 안보에 대한 제한된 효과의 영향

이전의 추적 및 표적화의 경우에, AI의 위협은 그 효과에 대해 적이 과도한 믿음을 갖고 있기 때문에 전략적 안정성을 해치지만, 그 반대의 경우도 역시 가능합니다. 결정 지지 시스템의 경우, AI를 채용한 부대가 효과가 없는 데도 효과적이라고 믿는 것이 더 치명적입니다. 또한 적이 AI를 파괴하고 보복을 피할 수 있다고 믿게 되고, 그렇지 않았다면 자연스럽게 확전이 되었을 것을 선제 타격을 포함하는 길을 쫓도록 하게 할 수 있습니다. 예를 들어, 적은 미사일이 표적으로 향하고 있을 때도 AI가 보고 그 패턴이 안전하다고 결론을 내리는 발사 및 궤적을 발견했다고 확신할 수 있습니다.

AI는 실시간으로 감지하기 어려운 새로운 취약점들을 제시합니다. 하지만 이 점은 전쟁으로 가는 길, 확전, 심지어 발사 결정에서 거의 확실히, 궁극적으로 또는 점차적으로 더 중요해질 것입니다. 이런 책임을 가진 모든 시스템은 적대적 접근법을 포함하는 엄격한 테스트를 거쳐야 합니다. 테스트 시 적에 대한 시뮬레이션은 적이 생성할 수 있는 모든 범위의 공격을 테스터가 구상할 수 있을 때만 완전한 효과를 거둘 수 있습니다. 그렇지만 불가능할 정도로 무리한 이런 명령은 배치된 모든 군사 시스템에 직면하게 됩니다.

AI의 발전은 거침 없어 보입니다. 기업과 정부들은 공격적인 사용과 방어적인 사용을 모두 포함하여 계속 확대되는 응용 범위에 앞다퉈 AI를 채용하고 있습니다.

일부분 가능한 AI의 안정성 향상 효과

핵 공격 없이 수십 년이 지날 경우, 전략적 안정성을 당연한 것으로 여기기 쉽습니다. 이전 섹션에서는 AI 발전이 전략적 안정성을 약화시킬 수 있는 방법들을 대략 기술했지만, 반드시 그럴 필요는 없습니다. AI의 발전은 거침 없어 보입니다. 영화와 정부들은 공격적인 사용과 방어적인 사용을 모두 포함한 계속 확대되는 적용 범위에 앞다퉈 AI를 채용하고 있습니다. 이런 전략적 적용에 미치는 AI의 효과는 시간이 지나야만 분명해질 것입니다. AI는 핵 전략의 다양한 측면에서 긴장을 악화시킬 수 있지만, 유리한 환경에서는 이런 긴장을 완화하고 전략적 안정성을 향상시킬 수 있습니다. 상호 불신에도 불구하고 핵 보유국들은 자국의 이익을 위해 이런 목적을 향해 협력할 수 있습니다.

높은 오류 발생을 넘어서는 시기

워크샵 참가자들은 AI가 추적 및 표적화 또는 확전 결정 지지와 같은 새로운 능력을 갖춘 직후에 가장 위험한 시기가 닥칠 거라는 데 동의했습니다. 이 시운전 시기 동안에는 오류와 오해가 발생할 가능성이 비교적 높습니다. 시간이 흐르고 기술적 발전이 증가하면 이런 위험은 줄어들 것으로 예상됩니다. 평화로운 시기 동안 사용 가능한 중요한 능력들이 개발되면, 처음에 시작한 시점 이후로 발전이 계속되고, 시간이 지나면서 신뢰성이 증가하

고, 제한 사항이 잘 이해될 거라고 기대할 수 있습니다. 결국 AI 시스템은 오류가 있긴 하지만 인간 대안보다 오류를 덜 범하며, 따라서 장기적으로 안정적인 능력을 개발할 것입니다.

전략적 안정성을 위한 잠재적 협력

핵 전쟁의 위험을 고착시키는 요인들 중 하나는 정부들로 하여금 "공격 시 발사" 태도를 채택하라고 부추기는 확실한 보복의 선제 타격 안정성 요건과 사고 또는 오작동 가능성 사이의 모순입니다. 예를 들어, 1983년에 오작동으로 발생한 소련 연방의 조기 경보 시스템은 존재하지 않는 미국 공격을 "감지" 했습니다(Hoffman, 2009, pp. 6–11). 특히 위기 상황에서, 이런 사건은 관료들로 하여금 유령 공격에 대해 보복성 공격을 명령하게 만들 수 있습니다. AI는 보다 신뢰할 수 있는 조기 경보 시스템을 만들어 냄으로써 이러한 모순을 완화시킬 수 있습니다. 결과적으로 더 큰 선제 타격 안정성이 위기 시 우발적인 확전의 위험을 줄이는 데 도움이 됩니다. 그렇기는 하지만 이런 식의 믿음은 복합적인 축복이 될 수 있습니다. 확전을 예측할 능력이 있다고 믿는 침략 국가는 그렇지 않으면 불확실성으로 인해 단념할 수 있는 도발 행위를 과감하게 저지를 수 있다고 느낄 수 있습니다.

정보 수집과 분석의 향상된 정확성 역시 억제력, 확신, 안심에 신뢰성을 더함으로써 전략적 안정성을 강화할 수 있습니다. 잠재적 적국이 은밀하게 공격을 준비할 기회가 더 적어진다면, 자국 또는 그 동맹국에 대해 무력을 사용하는 위협이 덜 실행될 수 있습니다. 전략적 파트너들이 더 포괄적인 정보와 분석에 접근한다면, 더 쉽게 안심할 수 있습니다. 확신을 주기까지 더 적은 군사력이 필요한 미국과 같은 핵 보유국은 핵 군비의 크기를 줄일 수 있고, 이로 인해 적을 더 안심시킬 수 있습니다. 이 과정이 선순환으로 바뀔 수 있고, 궁극적으로 전쟁의 위험을 크게 줄일 수 있습니다. 하지만 이런 결과는 AI 기술의 상태와 상관 없이 실현하기 어려운 조건을 요구합니다. 우선, 모든 당사국들은 정보 및 분석 능력에 똑같이 접근해야 합니다. 새로운 비대칭 정보에 약한 국가는 아마도 스스로를 받아들이기 어려울 정도로 취약하다고 여기고 적을 더 깊이 의심할 수 있습니다. 또한 경쟁 국가들의 의도가 진정 우호적이어야 합니다. 마지막으로, 정보 수집 및 분석 시스템(비 AI 구성요소 포함)에 대한 관료들의 신뢰가 확실히 정당화 되어야 합니다. AI의 잠재력을 현실화하여 전략적 안정성을 강화하려면, 각 국가는 이런 함정을 피하기 위해 기술이 성숙함에 따라 조정을 시작해야 합니다. 이런 논의에는 외교, 군사 관료들 뿐만 아니라 기술 전문가도 포함되어야 합니다.

철저한 투명성

매우 낙관적인 한 가지 가능성을 들면, 확전 결정을 지원하는 데 사용되고 있는 AI 알고리즘을 적과 공유할 수 있습니다. 그와 같은 철저한 투명성은 많은 위험을 동반할 수 있습니다. 적은 확전 임계값 끝까지 바람직하지 않은 행동을 추구할 수 있습니다. 또한 그들은 AI를 파괴할 수도 있습니다. 동시에 그런 의사 결정 도우미로 사용되는 AI는 적대적인 속성을 포함한 다양한 테스트를 받아야 합니다. 적이 알고리즘을 얻게 된다 해도 안전한 상태를 유지할 수 있는 방식으로 AI를 설계하려는 시도는 어떤 경우든 매우 실용적입니다. 적이 그것을 얻을 수 없다고 가정하는 것은 위험합니다(Kerckhoff, 1883). AI 컴퓨터 시스템이 수비에 앞서 그토록 높은 견고성 기준을 충족시켜야 한다면, AI 컴퓨터 시스템을 널리 퍼트리는 것은 두려움을 완화하고 계산 착오를 거의 불가능하게 만들 수 있습니다.

결론

대부분의 워크샵 참가자들은 AI가 핵 안정성의 토대를 흔들고, 특히 점점 다극화되어가는 전략적 환경에서 2040년까지 억제력을 약화시킬 커다란 잠재력이 있다는 데 동의했습니다. 핵무기로 인류를 파괴하려는 악의적인 AI들의 헐리우드 악몽을 일축하는 전문가들은 그 대신 향상된 능력으로부터 발생하는 더 일상적인 문제에 관심을 가졌습니다. 논의된 AI 적용에는 대항 표적화를 위해 적의 발사 장치를 추적하고 표적화하는 능력과 핵무기의 사용에 대한 선택을 알리는 결정 지원 시스템에 AI를 포함하는 것이 포함되었습니다.

일부 전문가들은 AI에 대한 의존성이 증가하면 새로운 유형의 치명적인 실수를 범할 수 있다는 점을 우려합니다. AI가 기술적으로 성숙하기도 전에 사용을 강요받을 수 있습니다. AI는 적대적인 파괴에 영향을 받을 수 있습니다. 아니면 적이 AI가 더 능력이 있다고 믿음으로써 그들이 치명적인 실수를 하도록 이끌 수 있습니다.

반면에, 핵 보유국들이 AI가 제공할 수 있는 새로운 능력과 양립할 수 있는 전략적 안정성의 형태를 그럭저럭 수립한다면, 기계들은 불신을 줄이고 대외적 긴장을 완화시켜 핵 전쟁의 위험을 감소시킬 수 있습니다.

지금 우리는 이런 시나리오들 중 어떤 것이 발생할지 예견할 수 없지만, 이런 난제가 첨예해지기 전에 핵 안보에 대한 AI의 잠재적인 영향을 고려하기 시작해야 합니다. 앞으로 몇 십 년 동안 전략적 안정성을 유지하는 것이 매우 어려울 수 있으므로, 모든 핵 보유국들은 핵 위험을 제한하는 기구의 설립에 참여해야 할 것입니다. 이 목표는 경쟁국들의 협력을 요구하는 기술적, 군사적, 외교적 조치들의 결합을 요구할 것입니다. 우리는 이 전망이 그와 같은 논의를 시작하게 하며 이처럼 논란이 분분하고 자주 엇갈리는 주제에 대해 실용주의 및 현실주의로 향하는 길을 열기를 희망합니다.

참고

[1] 이 전망에서, 이런 업적들이 궁극적으로는 그 자체로 인간의 지성을 모방하는 것과 아무런 관련이 없다 해도, 우리는 AI와 광범위하게 관련된 연구 프로그램에서 많은 컴퓨터 공학의 업적들을 포함하는 *인공 지능*이란 용어를 비공식적으로 채용합니다. 그런 프로그램들은 패턴 인식 알고리즘, 새 프로그래밍 언어, 자연 언어 처리, 지난 수십 년 동안 AI로 언급된 다수의 다른 기능들로 연결되었지만, 주류 컴퓨팅에 편입된지 오래되었습니다.

[2] 러시아 언론들은 Status-6가 자체의 자율적 능력에 도달하기 위해 *인공 지능*(AI)을 채용했다고 밝혔습니다. Tuchkov(2016)와 "러시아 프로젝트 'Status-6'가 세계 핵 전력의 균형을 변화시킨다(Rossiiskii proekt 'Status-6' meniaet sootnosheniia iadernykh sil v mire)" (2016)를 참조하십시오. 두 번째 기사는 "인공 지능이 장착될 예정"인 Status-6는 "그렇지 않으면 도달할 수 없는 궤도를" 따라 "예상할 수 없는 곳"에 있는 적을 공격함으로써 대잠수함 작전을 피할 수 있다고 주장합니다.

[3] 상투적으로, 우리는 미래와 최근의 발전을 흔히 AI의 정의에 포함하지만, 과거 어느 시점에는 인간을 필요로 했던 과제에 오랜기간 응용되어왔던 어떤것들은 포함하지 않습니다.

[4] 냉전이 끝나기 전 몇 년 동안, 소련 연방은 미국의 미사일 방어 체제를 공격하기 위한 미사일 기술을 개발함으로써 곧 다가올 미국 미사일 방어 체제에 대응하기로 결정했습니다. Ronald Reagan의 전략 방위 구상(Strategic Defense Initiative)에 대한 소련의 "비대칭적 대응"과 관련된 러시아 설명은 Oznobishev, Potapov, and Skokov(2008)을 참조하십시오. Vladimir Putin은 같은 말을 반복하고 있습니다. 예를 들어, 2012년 선언에서 "미국의 전세계 미사일 방어 및 유럽의 구성요소에 대한 러시아의 군사적-기술적 대응은 효과적이고 비대칭적일 것" (Putin, 2012)이라고 말했습니다.

[5] 2010 핵태세 검토 보고서(Nuclear Posture Review)는 미국 핵 전략의 목표는 러시아와 중국에 대한 "전략적 안정성을 강화" 하면서 북한과 같은 "지역적 세력에 대한 억제력을 강화하는" 것이라 명시하고 있습니다. 그러나 이 보고서는 "전략적 안정성"에 대한 간결한 정의는 제공하지 않습니다(U.S. Department of Defense, 2010).

[6] 미국과 소련은 각각 긴급한 공격시 상대국을 무력화하는 데 사용할 수 있는 선제 타격 능력을 개발하려 했지만, 이것은 "마른 하늘에 날벼락" 같은 공격을 하기 위해 고안된 선제 타격 능력과는 구별해야 합니다. 선제 타격을 위해 조성된 전략적 군사력과 선제 공격을 위해 조성된 전략적 군사력을 구별하기가 실제로 어렵기 때문에 두 초강대국의 관료들은 상대국이 핵 전쟁 시작을 준비하고 있지 않은지 두려워했습니다.

[7] 페리메트르(Perimetr)가 일종의 인공 지능(AI)을 채용한다는 주장이 러시아 국영 매체에 반복해서 등장했습니다. 이에 대한 예는 Timoshenko(2015)와 Valagin(2014)을 참조하십시오.

[8] 이에 대한 예는 Akhmerov, Akhmerov, and Valeev(2016)을 참조하십시오.

[9] Brien Alkire and Jim Powers의 미간행 RAND 연구.

참고 문헌

Akhmerov, D. E., E. N. Akhmerov, and M. G. Valeev, "'Uiazvimost' kontseptsii neiadernogo razoruzheniia strategicheskikh iaderniyi sil Rossii" ["The Dubiousness of the Concept of a Non-Nuclear Disarming Strike Against Russia's Strategic Nuclear Forces"], *Vestnik akademii voennykh nauk*, Vol. 54, No. 1, 2016, pp. 37–41.

Anderson, H. S., A. Kharkar, B. Filar, and P. Roth, *Evading Machine Learning Malware Detection*, blackhat.com, July 2017. As of August 15, 2017: https://www.blackhat.com/docs/us-17/thursday/us-17-Anderson-Bot-Vs-Bot-Evading-Machine-Learning-Malware-Detection-wp.pdf

Biggio, B., B. Nelson, and P. Laskov, "Poisoning Attacks Against Support Vector Machines," *Proceedings of the 29th International Conference on Machine Learning*, July 2012, pp. 1467–1474. As of August 15, 2017: https://arxiv.org/pdf/1206.6389.pdf

Bostrom, Nick, *Superintelligence: Paths, Dangers, Strategies*, Oxford: Oxford University Press, 2014.

Bracken, P., "The Intersection of Cyber and Nuclear War," *The Strategy Bridge*, blog post, January 17, 2017. As of August 15, 2017: https://thestrategybridge.org/the-bridge/2017/1/17/the-intersection-of-cyber-and-nuclear-war

Buchan, G., D. Matonick, C. Shipbaugh, and R. Mesic, *Future Roles of U.S. Nuclear Forces: Implications for U.S. Strategy*, Santa Monica, Calif.: RAND Corporation, MR-1231-AF, 2003. As of March 8, 2018: https://www.rand.org/pubs/monograph_reports/MR1231.html

Cimbala, S. J., *The Dead Volcano: The Background and Effects of Nuclear War Complacency*, Westport, Conn.: Praeger, 2002.

Etherington, D., "Google's AlphaGo AI Beats the World's Best Human Go Player," TechCrunch, May 23, 2017. As of August 15, 2017: https://techcrunch.com/2017/05/23/googles-alphago-ai-beats-the-worlds-best-human-go-player/

Garthoff, R. L., "Refocusing the SDI Debate," *Bulletin of the Atomic Scientists*, Vol. 43, No. 7, September 1987.

Geist, E., "Was There a Real 'Mineshaft Gap'? Bomb Shelters in the USSR, 1945–62," *Journal of Cold War Studies*, Vol. 14, No. 2, Spring 2012, pp. 3–28.

Goodfellow, I., Y. Bengio, and A. Courville, *Deep Learning*, Cambridge, Mass.: MIT Press, 2016.

Healey, D., *The Time of My Life*, London: Michael Joseph, 1989, p. 243.

Hoffman, D. E., *The Dead Hand*, New York: Doubleday, 2009.

Holmes, J. R., "Sea Changes: The Future of Nuclear Deterrence," *Bulletin of the Atomic Scientists*, Vol. 72, No. 4, 2016, pp. 228–233.

Kahn, H., *On Thermonuclear War*, Princeton, N.J.: Princeton University Press, 1960.

Karpathy, A., "Breaking Linear Classifiers on ImageNet," *Andrej Karpathy blog*, March 30, 2015. As of August 15, 2017: http://karpathy.github.io/2015/03/30/breaking-convnets/

Kearns, M., and M. Li, "Learning in the Presence of Malicious Errors," *SIAM Journal on Computing*, Vol. 22, No. 4, March 1992, pp. 807–837. As of August 15, 2017: https://doi.org/10.1137/0222052

Kegelmeyer, P., T. M. Shead, J. Crussell, K. Rodhouse, D. Robinson, C. Johnson, D. Zage, W. Davis, J. Wendt, J. Doak, T. Cayton, R. Colbaugh, K. Glass, B. Jones, and J. Shelburg, *Counter Adversarial Data Analytics*, Albuquerque, N.M.: Sandia National Laboratories, SAND2015-3711, May 2015. As of August 15, 2017: http://www.sandia.gov/~wpk/pubs/publications/cada-full-uur.pdf

Kerckhoff, A., "La Cryptographie Militaire," *Journal des Sciences Militaires*, January 1883.

Long, A., *Deterrence from Cold War to Long War: Lessons from Six Decades of RAND Research*, Santa Monica, Calif.: RAND Corporation, MG-636-OSD/AF, 2008. As of March 8, 2018: https://www.rand.org/pubs/monographs/MG636.html

Long, A., and B. R. Green, "Stalking the Secure Second Strike: Intelligence, Counterforce, and Nuclear Strategy," *Journal of Strategic Studies*, Vol. 38, Nos. 1–2, August 2012, pp. 38–76.

Mack, C. A., "Fifty Years of Moore's Law," *IEEE Transaction on Semiconductor Manufacturing*, Vol. 24, No. 2, May 2011, pp. 202–207. As of August 15, 2017: https://doi.org/10.1109/TSM.2010.2096437

Mahan, A. T., *Armaments and Arbitration: Or, The Place of Force in the International Relations of States*, New York: Harper & Brothers, 1912.

Minsky, M., *Semantic Information Processing*, Cambridge, Mass.: MIT Press, 1968.

Newell, A., J. C. Shaw, and H. A. Simon, *Report on a General Problem-Solving Program*, Santa Monica, Calif., RAND Corporation, Report P-1584, revised February 9, 1959.

Oznobishev, S. K., V. Ia. Potapov, and V. V. Skokov, *Kak gotovilisia "asymmetrichnyi otvet" na "Strategicheskuiu oboronnyiu initsiativu" R.Reigana. Velikhov, Kokoshin i drugie* [*How the "Asymmetric Response" to R. Reagan's "Strategic Defense Initiative" Was Prepared*], Moscow: Legand, 2008.

Papernot, N., P. McDaniel, and I. Goodfellow, "Transferability in Machine Learning: from Phenomena to Black-Box Attacks Using Adversarial Samples," arXiv, May 24, 2016. As of August 15, 2017: https://arxiv.org/abs/1605.07277

Putin, V., "Byt' sil'nymi: garantii natsional'noi bezopasnosti dlia Rossii" ["Being Strong Is the Guarantee of National Security for Russia"], *Rossiiskaia gazeta*, February 20, 2012. As of December 5, 2017: https://rg.ru/2012/02/20/putin-armiya.html/

Roland, A., and P. Shiman, *Strategic Computing: DARPA and the Quest for Machine Intelligence, 1983–1993*, Cambridge, Mass.: MIT Press, 2002.

"Rossiiskii Proekt 'Status-6' Meniaet Sootnosheniia Iadernykh Sil v Mire" ["The Russian Status-6 Project Is Changing the World's Nuclear Balance of Forces"], *Russkaia Politika*, November 14, 2016. As of December 4, 2016: http://ruspolitica.ru/post/rossiyskiy-proekt-status-6-menyaet-sootnoshenie-yadernyih-sil-v-mire/

Schelling, T., *Arms and Influence*, New Haven, Conn.: Yale University Press, 1966.

Shannon, C. E. "Programming a Computer for Playing Chess," *Philosophical Magazine*, Vol. 41, No. 7, 1950, pp. 256–275.

Simon, H. A., and A. Newell, "Heuristic Problem Solving: The Next Advance in Operations Research," *Operations Research*, Vol. 6, No. 1, January–February 1958, pp. 1–10.

Slocombe, W., "The Countervailing Strategy," *International Security*, Vol. 5, No. 4, Spring 1981, pp. 18–27.

Sutyagin, I., "Russia's Underwater "Doomsday Drone": Science Fiction, but Real Danger," *Bulletin of the Atomic Scientists*, Vol. 72, No. 4, June 2016, pp. 243–246.

Timoshenko, M., 'Mertvaia ruka' na strazhe perimetra Rossii" ["The 'Dead Hand' Guarding Russia's Periphery"], *Telekanal "Zvezda,"* February 18, 2015. As of August 15, 2017: http://tvzvezda.ru/news/krasnaya_zvezda/content/201502181414-gskc.htm

Tuchkov, V., *Status-6: Oruzhie Vosmezdia, Vognavshee Pentagon v Stupor* [*Status-6: The Retaliatory Weapon That Drove the Pentagon into a Stupor*], Svobodnaia Pressa, December 11, 2016. As of December 4, 2016: http://svpressa.ru/war21/article/162378/

U.S. Department of Defense, *Nuclear Posture Review Report*, Washington, D.C., April 2010.

Valagin, A., "Garantirovannoe vozmezdie: Kak rabotaet rossiiskaia sistema 'Perimetr'" ["Assured Retaliation: How the Russian 'Perimetr' System Works"], *Rossiiskaia gazeta*, January 22, 2014. As of August 15, 2017: https://rg.ru/2014/01/22/perimetr-site.html

Woyke, E., and Y. Kim, "Starcraft Pros Are Ready to Battle AI," *MIT Technology Review*, May 19, 2017. As of August 15, 2017: https://www.technologyreview.com/s/607888/starcraft-pros-are-ready-to-battle-ai/

이 전망에 관하여

지구의 위기 및 안보를 위한 RAND 센터와 Andrew Parasiliti 소장에게 이런 노력을 시작해준 데 대해, 그리고 RAND Ventures에 대한 후원에 감사드립니다. 또한 Angela O' Mahoney와 Bill Welser에게 전반적인 지도 편달에 대해 감사드립니다. 또한 Sonni Efron, Doug Irving, Greg Baumann에게 스토리를 찾는 데 도움을 준 데 대해, Hosay Yaqub에게 행사를 성공적으로 마치게 해준 데 대해 감사드립니다. 마지막으로, 채텀 하우스 룰(Chatham House Rule)에 따라 익명으로 남아준 워크샵 참가자들에게 감사드립니다.

안보 2040

이 전망은 앞으로 몇 십 년 후 이런 안보 문제를 형성한 정치적, 기술적, 사회적, 인구학적 동향의 효과를 고려하면서 2040년 세계의 중요한 안보 문제를 구상하는 RAND Ventures의 사업이자 더 큰 노력의 일환입니다. 이 연구는 글로벌 위험 및 안보를 위한 RAND 센터 내에서 수행되었습니다.

지구 위험 및 안보를 위한 RAND 센터

지구 위험 및 안보를 위한 센터(CGRS)는 지구 안보의 체계적인 위험을 다루는 여러 전문 분야에 걸친 연구 및 정책 분석을 개발하기 위해 RAND Corporation 전체와 관련되어 있습니다. 이 센터는 RAND의 전문 지식에 의지하여 안보, 경제, 건강, 기술 등 많은 분야에서 RAND 연구를 보완 및 확장합니다. 저명한 비즈니스 리더, 자선가, 전 정책 결정자들로 구성된 이사회가 지구 안보 동향 및 위험과 안보에 대한 파괴적 기술의 영향에 점점 더 초점을 맞추는 센터의 활동을 자문하고 지원합니다. 글로벌 위험 및 안보를 위한 RAND 센터의 자세한 정보는 www.rand.org/international/cgrs를 참조하십시오.

RAND Ventures

RAND는 전 세계의 커뮤니티가 보다 안전하고 보안을 유지하며 건강하고 번영할 수 있도록 공공 정책 과제에 대한 솔루션을 개발하는 연구 기관입니다. RAND는 비영리 비당파 단체로서 공공의 이익을 위해 노력합니다.

RAND Ventures는 정책 솔루션에 투자하기 위한 수단입니다. 후원금은 우리가 긴 안목으로 어렵고 종종 의견이 분분한 주제를 파고 들어 혁신적이고 설득력 있는 방법으로 우리의 발견물을 공유할 수 있도록 지원합니다. RAND의 연구 발간물과 권고사항은 데이터와 증거에 기초하며, 따라서 고객, 기부자 또는 후원자의 정책 선호도나 관심을 반드시 반영하지는 않습니다.

이런 벤처에 대한 자금 지원은 RAND 후원자의 기부금과 운영 수입으로 제공됩니다.

저자 소개

Edward Geist는 RAND의 정책 연구원입니다. 이전에 스탠포드 대학교 국제 안보 및 협력 센터(CISAC)의 MacArthur Nuclear Security 펠로우였고 RAND 워싱턴 사무소의 Stanton Nuclear Security 펠로우였던 Edward는 2013년 5월 노스 캐롤라이나 대학교에서 러시아 역사 박사 학위를 취득했습니다.

Andrew J. Lohn은 RAND Corporation의 엔지니어입니다. 그는 사이버 전쟁, 인공 지능 또는 무인 항공기 운반과 같은 매우 기술적인 정책 문제에 새로운 통찰력을 제공하기 위해 광범위한 수학적 및 머신 러닝 기술을 적용합니다. Lohn은 캘리포니아 대학교 산타 크루즈 캠퍼스 (UCSC)에서 전기 엔지니어링 박사 학위를 받았습니다.