ISSUE BRIEF

이슈브리프 600호
(2024. 9.20)

제2차 REAIM 고위급회의 서울 개최의 의미와 시사점

제600호

윤정현 신안보연구실

국가안보전략연구원
Institute for National Security Strategy

INSS
www.inss.re.kr

국문초록

지난 9월 9일~10일, 「제2차 인공지능의 책임있는 군사적 이용에 관한 고위급회의(REAIM 2024)」가 서울에서 개최되었다. 외교부와 국방부의 공동주최로 열린 이번 2차 REAIM 고위급회의에서는 군사 분야에서 책임있는 AI의 구현을 위한 기본 원칙과 도전 과제, 국제 협력 방안 등이 폭넓게 논의되었다. 최종 결과물로 61개국이 채택한 '행동을 위한 청사진(Blueprint for Action)'은 지난 1차 회의의 결과물인 '행동 원칙(Call to Action)'에서 한층 진일보한 실천 방향을 제시하였다고 평가된다. 책임있는 AI의 군사적 이용을 위해 국제법 준수, 적절한 수준의 인간 통제, 신뢰도와 설명가능성 증진 방안들이 보다 구체적으로 명시되었기 때문이다. 우리로서는 산업 분야의 AI 통제 원칙을 논의한 「서울 AI 안전 정상회의('24. 5)」에 이어 이번 회의를 통해 군사 분야에서도 AI 규범 마련을 주도하였다는 의미를 부여할 수 있을 것이다. 그러나 공동선언문의 실효성을 보다 강화하고 한국의 지정학적 현실과 안보 이익에 부합할 수 있도록 추진해야하는 숙제가 남게 되었다. 이를 위해 우리는 세 가지 실천 방향을 염두에 둘 필요가 있다. 첫째, 기술력과 소프트파워를 갖춘 우리의 전략적 위치를 활용하여 AI 선도국과 후발국을 이어주는 가교역할을 적극적으로 수행해야 한다. 둘째, REAIM의 핵심 윤리 기준과 원칙을 선별·적용함으로써 책임있는 AI 의제와 우리의 국방혁신 역량을 조화시켜야 한다. 셋째, 긴밀한 안보 동맹인 미국과 책임있는 AI 구현을 위한 협력을 공고히 해야 한다. 이를위해 한미 AI 안보협력 대화체를 AI 책임성 구현 의제로 확대하고 정례화하는 노력이 필요할 것이다.

핵심어 : 인공지능(Artificial Intelligent), 책임있는 AI(responsible AI), 인공지능의 책임있는 군사적 이용에 관한 고위급회의(REAIM), 행동을 위한 청사진 (Blueprint for Action)

최근 군사 분야에서 인공지능(AI)의 도입 이슈는 안보적 측면뿐만 아니라 윤리적, 법적 관점에서도 매우 중요한 사안으로 간주되고 있다. 이와 관련하여 2023년 출범한 「인공지능의 책임있는 군사적 이용에 관한 고위급회의(REsponsible AI in the Military domain (REAIM) Summit」는 정부 관계자 뿐만 아니라 산학연 시민단체가 함께하는 대표적인 1.5트랙 다자회의체이다. 지난 2월, 헤이그에서 열린 1차 회의에 이어 올해 9월 9일~10일 서울에서 외교부와 국방부의 주최로 2차 REAIM 고위급회의가 개최되었다. 한국 뿐만 아니라 네덜란드, 싱가포르, 케냐, 영국이 공동으로 주최한 이번 회의는 90여개국 정부대표단을 비롯하여 기업, 군, 시민단체 2,000여 명이 참가하여 진행되었다.

'더 안전한 내일을 위한 책임있는 AI(Responsible AI for Safer Tomorrow)'라는 슬로건과 같이 2차 REAIM 회의는 AI가 국제평화와 분쟁, 대량살상무기(WMD) 확산에 미치는 잠재적 영향에 주목하였으며 군사 AI의 책임 있는 이용을 위한 미래 거버넌스 방향에 대한 논의가 이루어졌다.[1] 특히, 34개국 외교, 국방의장·차관급 참석자들이 참여한 고위급 라운드테이블에서 도출된 '행동을 위한 청사진(Blueprint for Action)'은 군사 분야의 책임있는 AI 이용을 위해 국제사회가 준수해야할 최소한의 가이드라인과 원칙이 담긴 것으로 평가할 수 있다. 동 원칙들은 향후 유엔총회에서 실효성을 갖추기 위한 후속 의제로 다뤄질 전망이다.

1차 회의의 한계를 넘어: 행동원칙의 실효성 강화를 위한 노력

지난 1차 REAIM 회의에서는 AI의 군사 분야 도입이 수반하는 문제들의 심각성과 다양한 이해관계자들의 책무가 중점적으로

1) 외교부 보도자료, "2024 인공지능의 책임있는 군사적 이용에 관한 고위급회의(REAIM) 결과", (2024. 9. 10.)

논의된 바 있다.[2] 그러나 AI의 군사적 이용에 대한 국제사회의 참여와 협력의 필요성을 공론화 하였음에도 불구하고, 책무성의 적용 범위에 대한 구체성은 부족한 한계를 가지고 있었다. 특히, 생성형AI와 함께 비약적으로 증가한 알고리즘 모델에 책임을 부여할 방안은 무엇인지, AI 규범 선도국과 후발국, 기술 격차를 고려한 AI의 군사적 책임의 간극을 어떻게 조정할 것인지는 미해결된 숙제로 남은 측면이 있다.[3]

반면, 이번 2차 회의는 행동원칙의 실효성을 보다 강화하였다는 진일보한 측면이 있다. 특히, AI 기술의 군사적 이용을 위한 원칙과 규범을 구체화하고 그 실효성을 담보하기 위해 이해당사자 간의 협력과 조율을 추진해나가는 방식이 좀더 구체적으로 논의되었다. 공동선언문의 AI 관리 의무 적용 범위 측면에서 볼 때, 군사 AI는 인도주의적, 법적, 안보적, 기술·사회적, 윤리적으로 다양한 위험과 연계될 수 있으며, 반드시 사전에 식별·평가·관리되어야 한다는 것이 명시되었다. 그러나 동시에 이러한 수단이 여타 비군사적 분야에서 AI가 제공하는 혜택에 대해 평등한 접근을 저해해서는 안된다는 점 또한 첨언하고 있다. 이는 군사 AI 시스템에 대한 통제와 활용 기회가 특정국이나 소수 집단에만 향유되는 것이 아닌, 보편적 원칙으로 기능해야 함을 시사한 것이다. 뿐만 아니라 무기화 전 단계의 기획·개발·배치·유지보수·폐기에 이르는 전주기에 걸쳐 책임성 있는 방식으로 적용되어야하며, 유엔 헌장 및 국제인도법 준수와도 궤를 같이해야 한다는 점 역시 강조되었다.[4]

특히 주목할 부분은 이 같은 REAIM 원칙들이 일회적 선언에

2) Government of the Netherlands, “REAIM Call to Action” (Feb 16, 2023), https://www.government.nl/documents/publications/2023/02/16/reaim-2023-call-to-action
3) ibid.
4) Readable, “Full statement: REAIM Blueprint for Action” (Sep. 10, 2024), https://thereadable.co/full-script-reaim-blueprint-for-action/

머물지 않도록 유엔 내 주요 다자간 협의체와 연계, 발전시켜 나갈 것임을 명시한 것이다. AI 시스템이 핵무기와 같은 대량살상무기에 활용되거나 무분별한 기술경쟁으로 치닫는 것을 막기 위해 가장 보편적이고 포괄적 협의 채널인 유엔을 중심으로 해당 군사 규범 이슈의 논의가 지속되어야 한다고 보았기 때문이다. 공동선언문의 "군사분야의 AI 미래 거버넌스 구상" 19항에서는 △특정재래식무기금지협약(Convention on Certain Conventional Weapons, CCW), 자율살상무기체계(Lethal Autonomous Weapon Systems, LAWS)에 관한 정부전문가그룹(Group of Governmental Experts, GGE), 유엔군축위원회(United Nations Disarmament Commission, UNDC) 등에서 추진해왔던 다수의 군사 AI 통제에 대한 결의사항을 인정하고, △REAIM의 원칙들이 해당 이니셔티브들과 상호 시너지 효과를 얻을 수 있도록 하며, △유엔 헌장을 비롯하여 적용 가능한 국제법에 합치되는 방식으로 개발, 배치, 이용되어야만 한다는 점을 확인하고 있다.[5)]

이밖에도 AI 규범 논의에서 상대적으로 소외되어 있었던 다양한 집단의 입장을 적극적으로 반영하고자 노력했던 점도 이번 2차 회의의 진일보한 특징을 보여준다. AI 시대의 주역인 청년들의 아이디어를 선보인 'AI 시대의 청년과 평화(Youth and Peace in the age of AI, Y.PAI)' 세션이 대표적이다. 이들은 AI 시대의 군축과 비확산 이슈들을 청년 세대의 시각에서 이해하고, 대응 방안의 아이디어를 SF, 웹툰과 같은 다양한 형태로 발신함으로써 대중의 관심 제고와 소통에도 기여할 것으로 기대를 모았다.

5) Ministry of Foreign Affairs, "Outcome of Responsible AI in Military Domain (REAIM) Summit 2024" (Sep. 10, 2024) https://overseas.mofa.go.kr/eng/brd/m_5676/view.do?seq=322676

책임있는 AI 실현을 위한 3가지 포인트: 평가, 적용, 거버넌스의 정착

무엇보다도 이번 REAIM 2024는 AI 규제에서 예외적으로 다뤄졌던 국방 분야 역시 엄격한 윤리적 기준이 필요하며 규범을 준수해야한다는 공동의 인식이 반영된 자리였다. 행동을 위한 청사진의 실효성을 담보하기 위해 AI 기술의 '평가(assessment)', '적용(application)', '거버넌스 정착(anchoring governance)'이라는 세 가지 핵심 테마를 바탕으로 심도있는 논의가 전개되었기 때문이다.[6)]

첫 번째로, 평가 측면에서 보면, AI의 군사적 이용이 미칠 수 있는 긍정적·부정적 효과를 균형있게 진단하고, 무기시스템에 통합될 때 그 파급력을 고려하도록 하는 국제평가기준의 필요성을 강조하였다. 즉, 관리가능한 AI 위험과 수용불가한 AI 위험을 분류함으로써 군사 분야에서의 AI 위험평가 프레임워크를 보다 체계적으로 발전시키기 위한 접근 방향을 제시한 것이다. 또한 각기 다른 지역별·국가별 활용 시나리오 수립을 권고함으로써 다양한 안보적 맥락에서도 큰 틀의 행동 원칙이 이행될 수 있도록 하였다.

두 번째로, AI 적용 측면에서는 책임있는 AI 규범의 실효성을 담보하기 위한 인간과 기계의 책임성을 둘러싼 법제도적 보완 방안이 논의되었다. 기본적으로 AI 무기 시스템이 인도적 원칙을 위반하거나 인간의 감독 없이 살상 결정을 내리지 않도록 하며, 이 원칙을 준용할 수 있는 국제법의 토대를 마련해야 한다는 것이다. 이는 궁극적으로 전장에서 국제인도법과 인권법 위반 책임을 AI에게 전가하지 못하도록 한 것이다. 또한, 기술적으로도 각국의 AI 정책이 인간통제의 공통된 기준 요건을 충족하는

6) 외교부 보도자료, "인공지능(AI) 시대 국제평화와 안보를 고민하는 전 세계 청년들의 연결!- REAIM 고위급회의 청년행사 'AI 시대의 청년과 평화(Y.PAI)'개최", (2024. 9. 10.).

한편, 부작용과 오작동 위협을 최소화하기 위한 표준 가드레일(guardrail)과 보호 장치(safeguards)를 적용하게 한 점도 실효성을 제고하기 위한 노력으로 평가할 수 있다.

세 번째로 거버넌스의 정착 측면에서 이번 2차 회의는 향후의 AI 규범 체계가 다양한 이해당사자들과 각국의 특수한 환경을 고려한 포용적인 방향으로 나아가야 함을 강조하였다. 이 같은 접근이 중요한 이유는 최근 급속히 발전하는 AI 기술이 경직된 관리 조직보다는 유연한 거버넌스 형태를 요구하기 때문이다. 뿐만 아니라 민간 AI 기업들이 무력 분쟁에 관여될 가능성이 점점 더 높아짐에 따라 전술·전략 측면에서 군과의 협력 파트너십 또한 중요해지는 추세이다. 이를 고려하여 2차 REAIM 회의는 정부 외 행위자들 또한 AI 거버넌스 수립의 주체로서 적극적으로 참여해야 한다는 점을 확인한 것이다. 이 과정에서 디지털 불평등이나 데이터 종속을 심화시키지 않도록 주의함으로써 국가 간 협력을 제약하는 AI 기술 격차나 디지털 격차 해소 노력을 이어갈 수 있도록 한 점도 주목할만한 부분이다.

향후 쟁점 및 시사점

결국 2차 REAIM 고위급회의의 가장 중요한 의미는 실천적 차원에서 군사 분야의 AI 규범과 거버넌스 논의 기반을 마련한 것이라 할 수 있다. AI의 책임있는 군사적 이용 조치와 미래 거버넌스를 위한 방향을 공동선언문에 구체적으로 명시했기 때문이다. 특히, 포괄적 다자군사협의기구인 유엔 산하기구를 중심으로 AI의 법제화를 지속적으로 이어갈 수 있는 근거를 마련한 점도 중요한 성과라 할 수 있다. '행동을 위한 청사진'이 권고에 그치는 것이 아닌, 보편적 의제를 시작으로 군사 분야의 국제 AI 규칙으로 발전할 수 있는 경로를 열어둔 셈이다.

그러나, 이번 공동선언문이 제시한 체계적인 미래 AI 거버넌스 방향으로 나아가기 위해서는 여전히 갈길이 먼 것도 사실이다. 당장 현재보다 더 긴밀한 국가 간 협력과 정보 공유가 필요하나, 이는 활용 데이터를 얼마나 폭넓게 공유할 수 있으며, 생성형 AI와 같은 대규모 언어모델 소스에 대한 개방성을 유지할 수 있는가의 문제와도 맞닿아 있다. 해당 이슈는 주요 플랫폼 기업 간에도 첨예하게 다투고 있는 사안인데, 이를 안보적 관점에서 국가 간 교류·협력에 어디까지 허용, 강제할 수 있을지가 불명확한 상황이다. 더욱이 민간보다 훨씬 더 엄격한 보안수준이 요구되는 군사 영역에서는 그 장벽이 더욱 높을 수밖에 없다. 책임성을 보장하기 위한 가드레일과 보호장치의 표준 수립 이슈만 하더라도 누가, 어떤 그룹이 이를 주도해 나갈 것인지를 두고 불신과 경쟁이 벌어질 수도 있다.

이는 자칫 국가간 협력이 이루어지기도 전에 갈등의 소지로 작용할 수도 있을 것이다. 지난 해 「AI 안전 정상회의(AI Safety Summit)」 당시 AI 안전 평가 주도권을 둘러싸고 벌어졌던 회의 개최국 영국과 미국의 경쟁이 대표적이다. 영국이 자국 내 국제 AI 안정성 평가기구를 유치하겠다고 하자, 미국 역시 독자적인 AI 안전연구소 설립을 발표함으로써 양국 간에는 미묘한 긴장이 흐르기도 하였기 때문이다. 이는 규범의 법제화 뿐만 아니라 향후 기술개발 및 표준 수립 과정에도 다양한 국가들과 이해관계자의 참여와 소통이 필수적임을 시사한다. 학습훈련 프로그램, 시험 및 평가 체계, 검사·검증·인증 절차, 모니터링 매뉴얼 수립 사안에서는 공동개발과 적용의 원칙을 강화함으로써 책임있는 AI 표준을 만드는 과정에서 국가 간 상호신뢰를 구축해야할 것이다.

나아가 핵·대량살상무기 운용체계와 같이 적용불가한 영역으로의 확산을 철저히 금지하고, 의사결정지원체계, 사이버·전자전,

정보전 시스템 등 전략적 중요성이 강화되고 있는 영역에서 AI의 파국적 영향을 막도록 설명가능성과 투명성 원칙을 엄격히 적용해 나가야 한다.

우리로서는 산업 분야의 AI 통제 원칙을 논의한 「서울 AI 안전 정상회의('24. 5)」에 이어 이번 회의를 통해 군사 분야에서도 AI 규범 마련을 주도하였다는 의미를 부여할 수 있을 것이다. 그러나 공동선언문의 실효성을 보다 강화하고 한국의 지정학적 현실과 안보 이익에 부합할 수 있도록 의제들의 추진해야하는 숙제가 남게 되었다. 이를 위해 우리는 세 가지 실천 방향을 염두에 둘 필요가 있다.

첫째, AI 검증 체계 및 표준을 개발하는데 있어 AI 선도국과 후발국을 이어주는 가교역할을 보다 적극적으로 수행해야 한다. 최근 한국은 2차 「AI 안전 정상회의」, 「AI 글로벌 포럼」 등을 개최하며 기술력과 AI 의제 선도의 소프트파워를 갖춘 중견국으로 자리매김 하였다. 특히, 기술 종속이나 데이터 주권 위협 우려를 불식시킬 수 있는 위치에 있으며 협력을 선도할 수 있는 역량 또한 갖추고 있다. 우리는 이러한 입장을 적극 활용하여 공동의 AI 국제규범 프레임워크 수립 및 유사입장국과 최적 사례 공유에 적극 나서야할 것이다. 이를 토대로 신뢰 기반의 책임있는 AI 규칙을 선도해가는 글로벌 중추국가로서 역량을 발휘해야 한다.

둘째, 책임있는 AI 의제와 국방혁신 역량을 조화시킬 수 있는 방안을 모색해야 한다. 먼저 범정부 차원에서 발표된 “사람이 중심이 되는 AI 윤리기준”과의 정합성을 고려한 국방 AI 윤리 기준 마련이 필요하다. 예를 들어 ‘인간의 존엄성’, ‘사회의 공공선’, ‘기술의 합목적성’이라는 3개 기본 윤리원칙은 그대로 수용하되, 국방 혁신과 충돌하지 않는 5대 우선 핵심 요건(책임성, 통제

가능성, 신뢰성, 추적 가능성, 데이터 관리)을 선별·적용할 필요가 있다. 동시에 편향성 방지, 투명성, 강건성 등 책임있는 AI 세부 원칙 구현을 위한 기술개발과 생태계 육성 노력도 병행해야 할 것이다.

셋째, 긴밀한 안보 동맹인 미국과 책임있는 AI 구현을 위한 협력을 공고히 해야 한다. 그간 전투력 강화 측면에서 추진되었던 양국의 훈련 프로그램 등을 AI 시스템의 운용과 책임있는 행동 원칙의 준수 측면으로 확대할 필요가 있다. 자율무기성능 및 거대언어모델 검증 도구의 공동개발 등이 대표적이다. 동시에 국제표준을 위한 AI 모델의 세부 보완 지침도 함께 마련해야 하며, 한미 AI 안보협력 대화체를 AI 책임성 구현 의제로 확대하고 정례화하는 노력이 필요할 것이다.

//끝//