“킬러 로봇”을 넘어:
자율적 군사로봇의 윤리적 문제들*

천현득 (서울대학교)

I. KAIST 보이콧 사건

2018년 4월, 30개국의 인공지능 연구자들이 한국의 주요 연구대학 중 하나인 한국과학기술원(이하, KAIST)의 총장에게 공개서한을 보냈다. 그들은 KAIST에 방문하지도 않고, KAIST 연구자를 초빙하지도 않으며, KAIST가 연관되어 있는 연구에 어떠한 기여도 하지 않을 것이라며 보이콧을 선언했다. 이러한 소위 “KAIST 보이콧” 사건은 KAIST가 한화시스템과 함께 국방인공지능융합연구센터를 개소한 데서 촉발되었다. 보이콧을 선언한 세계 여러 나라의 인공지능 연구자들은 연구센터의 목표가 “군사 무기에 적용될 수 있는 인공지능을 개발하여, 자율 무기를 개발하려는 세계적인 경쟁에 참여”하는 데 맞추어져 있음을 우려하고, KAIST의 우수한 연구진이 그러한 무기 개발의 군비 경쟁에 참여하는 것에 심각한 유감을 표시했다. 연구센터가 유의미한 인간의 통제를 받지 않는 자율 무기를 개발하지 않겠다고 총장이 확언하지 않는 한, KAIST의 모든 부분과 협업하지 않겠다는 선언이었다. 이러한 공개서한은 세계 유수 언론을 통해 보도되면서 큰 파장을 불러일으켰다. 가디언(*The Guardian*) 지는 “킬러 로봇: AI 전문가들이 한국의 대학 연구실에 보이콧을 선언하다”라는 제목의 기사를 게재하면서 영화 ‘터미네이터’의 살인 병기 로봇의 사진을 함께 실었다.(Haas 2018)

* 유익한 논평을 해주신 익명의 심사위원들께 감사드린다. 이 논문은 2016년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임.(NRF-2016S1A5A2A03927217)

CNN, Financial Times, Fortune, Reuter, Science Magazine 등의 유력 언론에서도 유사한 제목과 유사한 그림을 통해 비슷한 논조의 기사를 내보냈다. 대한민국의 유력 대학이 '킬러 로봇'을 만드는 데 동참하고 있으며, 이러한 이유로 전 세계 연구자들이 우려하고 있다는 내용이었다. 여론에 놀란 KAIST 신성철 총장은 킬러 로봇이나 치명적인 자율 무기 시스템을 만들려는 의도가 전혀 없으며 인공지능 기술을 적용하는 데 있어 윤리적인 우려들을 충분히 인지하고 있다는 성명을 발표했고, 과학자들이 이를 받아들여 보이콧을 철회함으로써 사건은 일단락되었다.

물론 KAIST 보이콧 사태는 갑작스럽게 발생한 것이 아니다. 군사 기술 분야에서 인공지능 기술이 점차 더 활발히 적용되고 이에 대한 사람들의 우려가 커져가는 상황 속에서 돌출된 것이다. 보이콧을 주도한 호주의 인공지능 연구자 월시(Toby Walsh) 교수는 인공지능과 로봇공학의 영향 연구센터(Centre on Impact of AI and Robotics, 줄여서 CIAIR)를 이끌고 있으며, CIAIR은 2015년부터 자율적 군사 무기의 개발을 반대하는 일련의 운동을 전개해왔다. 2015년 7월에는 인간의 유의미한 통제를 벗어난 자율적 공격 무기의 개발과 사용을 반대하는 공개서한을 UN에 제출한 바 있다. 서한에는 스튜어트 러셀 교수 등 인공지능 및 로봇공학 연구자 3724명을 포함하여 총 20,486명이 서명을 했는데, 여기에는 스티븐 호킹, 일런 머스크, 막스 테그마크, 다니엘 데넷, 노암 촘스키 등 저명한 과학자와 사상가들도 포함되어 있었다. 2017년 8월에는 "Killer robots: world's top AI and robotics companies urge united nations to act on lethal autonomous weapons"이라는 제목의 공개서한을 UN에 다시금 보냈는데, 여기에는 28개국 137명의 회사 창립자들과 최고경영자들이 서명을 했다.

물론, 월시 교수와 CIAIR의 활동은 완전히 자율적인 군사 로봇에 반대하는 국제적인 운동의 일부이다. 2009년 알트만(Juergen Altmann), 아사로(Peter Asaro), 샤키(Noel Sharkey), 스패로(Rob Sparrow)는 무인 로봇 무기의 개발과 사용을 우려하며 로봇 군대 통제를 위한 국제 위원회(ICRAC, International Committe for Robot Arms Control)를 발촉시켰고, 2012년

국제인권감시기구(Human Rights Watch)를 비롯한 여러 국제비정부기구들은, 완전히 자율적인 무기의 개발, 생산, 사용을 선제적으로 금지하도록 촉구하는 시민단체인 "킬러 로봇 반대 캠페인(Campaign to Stop Killer Robots)"을 조직했다.[1] 이들의 노력으로 인해, 2014년 5월에는 세계 87개국, IRCR, "캠페인"의 대표들이 "치명적 자율무기체계(lethal autonomous weapons systems, 통상 LAWS로 줄임)"에 대한 다자간 회의에 참여했는데, 이 모임은 국제연합(UN)의 재래식 무기 협약(Convention on Conventional Weapons)의 후원을 받았다. 이후 재래식 무기 협약에서 치명적 자율무기체계에 관한 논의는 매년 계속되었고, 2017년 11월에는 치명적 자율무기체계를 규제하는 방안을 놓고 정부전문가그룹(Group of Governmental Experts)의 첫 모임이 개최되었다.

보이콧 사태는 이러한 전반적인 흐름 속에서 이해될 필요가 있다. 그렇더라도 왜 하필 그 시점에서 KAIST가 보이콧의 대상이 되었는지에 관해서는 조금은 더 세부적인 논의가 필요해 보인다. 첫째, 의사소통 상의 문제가 없지 않았다. 해외 연구자들이 국방인공지능융합연구센터에 관해 알게 된 경로는 코리아 타임즈(Korea Times)의 한 기사였다.(Jun 2018) 기사에서는 연구소가 한국의 방위산업체 한화시스템과 공동으로 운영되며, 네 가지 과제 즉, AI 기반 지휘 시스템, 무인 해저선의 네비게이션을 위한 AI 알고리듬, AI 기반 항공기 훈련 시스템, AI 기반 물체추적 및 인식 기술에 초점을 맞출 것이라고 밝히고 있다. 연구센터의 전반적인 목표와 함께 이러한 중점 과제의 설정은, 당사자들의 부인에도 불구하고, 이 센터가 자율 무기, 혹은 일명 "킬러 로봇"을 개발할 것이라는 우려를 주기에 충분해 보였다. 물론 단지 기자의 보고 방식과 연구센터의 미디어 커뮤니케이션 상의 문제를 지적하는 것으로 충분치 않을 수 있다. 연구진과 학교 당국이 자율적 군사 무기의 개발이 불러올 수도 있는 윤리적 우려에 대해 민감했다면

1. 국제인권감시기구와 하버드대학 법학대학원의 국제인권 클리닉은 완전히 자율적인 무기의 금지를 촉구하는 보고서 *Losing Humanity: The Case Against Killer Robots* (2012. 11.19.)을 출판했다.

그와 같은 사태는 벌어지지 않을 수 있었기 때문이다. 즉, 겉보기에는 의사소통 상의 문제로 보일지라도 기술 개발의 사회적, 윤리적 영향에 관한 사고의 둔감성이 배경에 자리하고 있었다고 볼 수 있다. 윤리적 둔감성에 대한 지적은 정부 당국에도 적용될 수 있다. 융합연구센터의 개소식이 알려진 한 주 뒤, 컬러로봇 금지 캠페인은 외교통상부 강경화 장관에게 보내는 서신(2018.3.5.)에서 연구센터에 대한 우려를 전달하면서 완전히 자율적인 무기를 금지하도록 촉구했다. 이에 대해 대한민국 정부나 외교부가 어떻게 대응했는지는 알려진 바가 없다.

윤리적 둔감성이라는 배경 위에 의사소통 문제가 이 사건을 촉발했다고 하더라도 보이콧이 이루어진 시점에 관해서는 조금 더 생각해볼 여지가 있다.[2] 사건의 시점에 관해서는 월시 교수를 비롯한 CIAIR의 활동과 유엔의 활동을 연동시켜 봄으로써 힌트를 얻을 수 있다. 유엔은 2016년 특정 재래식 무기 협약(CCW) 체약국의 5차 검토회의에서 치명적 자율무기시스템(LAWS)에 관한 정부전문가그룹(GGE)를 설립하였고, 이 정부전문가그룹은 2017년 11월 13일부터 17일까지 제네바에서 첫 번째 회의를 가졌다.[3] 이 회의에서 체약국들은 치명적 자율무기시스템에 대한 정부전문가그룹이 2018년에 다시 모임을 가지는데 동의했으며, 그 첫 회의는 2018년 4월 9일부터 13일까지 제네바에서 열렸다. 이 회의 바로 직전에 KAIST 보이콧 사태가 일어났다는 점과 이 사건이 총장의 성명으로 싱겁게 일단락되었다는 점은, 제네바 회의를 앞두고 이슈를 환기하려는 의도가 다분히 있었다고 의심할만한 이유가 있었음을 보여준다.

왜 하필 한국의 유력 대학 KAIST가 대상이 되었는지에 관해서는 또

2. 예컨대, KAIST 보이콧 사태를 소개하는 국내 언론의 기사에는 강대국에서는 이미 개발하고 있으면서 우리처럼 힘이 약한 나라에는 엄격하게 기술 개발을 통제하는 것 아니냐는 논조의 댓글들이 많이 달렸다. 물론 이러한 댓글은 이 사안을 단지 국가 간 정치경제적 문제로 축소하는 경향이 있지만, 왜 그 시점에서 KAIST의 특정한 연구센터가 문제가 되었는지 생각해보도록 만든다.
3. Report of the 2017 Group of Governmental Experts on Lethal Autonomous Weapons Systems(LAWS). 22 December 2017.(http://undocs.org/CCW/GGE.1/2017/3)

다른 설명이 요구된다. 국방인공지능융합연구센터는 연구소가 아니라 연구소 산하의 한 연구센터에 불과하다. 그러나 최대 군수업체 중 하나인 한화시스템과 우수 인공지능 연구진을 보유한 KAIST가 결합했다는 점은, 특히나 군사적 긴장이 높은 한반도의 상황에서 국제 사회에 실제적인 우려를 자아내기에 충분해 보인다. 대한민국은 미국, 러시아, 중국과 함께 대인지뢰금지협약(오타와 협약)과 집속탄금지협약에 서명하지 않은 몇 안 되는 나라이다. 우리나라는 이들 비서명국들과 함께 주요 집속탄 생산 및 보유국으로 인식되고 있으며, 집속탄 개발 투자 규모 세계 2위로 전세계 생산량의 1/4을 한화시스템과 풍산이 담당하고 있는 것으로 알려져 있다.[4] 그밖에 비무장지대에서 삼성중공업이 개발한 자동 무기가 배치된 것으로도 알려져 있다. 이러한 실정을 감안할 때, 한화시스템과 KAIST가 공동으로 자율 무기를 개발하는 프로젝트에 참여한다는 것은 실질적인 우려를 낳을 소지가 충분하다.

KAIST 보이콧 사태로부터 우리는 무엇을 배워야하는가? 인공지능과 관련해 논의해야할 심오한 철학적 쟁점들이 많이 있지만, 때로는 긴급한 쟁점들을 우선적으로 다룰 필요가 있다. 인공지능의 군사적 사용이 제기하는 윤리적 문제는 그러한 긴급한 쟁점들 가운데 하나이다. 인공지능과 로봇공학 분야에 투자되는 엄청난 연구비 가운데 많은 부분은 국방예산에서 나오며, 이를 바탕으로 관련 분야들이 빠른 속도로 발달하고 있을 뿐만 아니라 발달된 인공지능 기술이 다시금 군사적으로 응용될 가능성이 커지고 있다. 인공지능 시대를 맞아, 일자리의 위협, 인간과 기계의 새로운 관계 모색, 포스트자본주의 등이 논의되고 있지만, 긴급한 논의가 필요한 분야 중 하나는 3차 군사혁명이다. 1차 군사혁명이 화학에, 2차 군사혁명이 핵무기에

4. 집속탄이란 어미 폭탄 속에 많은 새끼 폭탄들이 들어있어, 새끼 폭탄들이 표적 주변에 흩어져 폭발하면서 무차별적으로 살상하는 무기이다. 2008년 5월 아일랜드 더블린에서 107개국이 집속탄 금지 협약을 채택하고 서명했으며, 2010년 8월 1일 발효되었다. 협약은 집속탄의 사용, 생산, 비축, 이동을 금지함으로써 집속탄을 포괄적으로 금지하고, 국가가 협약의 규정에 의해 금지된 활동을 수행하도록 지원, 장려 또는 유도하는 행위를 역시 금지한다. 협약체결국은 금지 규정 외에도 비축 집속탄 폐기, 잔존 집속탄 제거, 피해자 지원의 의무를 가진다.

의존했다면, 이제 3차 군사혁명은 치명적 자율무기시스템(LAWS)에 달려있다고 말해진다. 로봇공학자이자 로봇윤리학자인 아킨(Ronald Arkin)은 이렇게 말한다. "흐름은 분명하다. 앞으로도 전쟁은 일어날 것이고, 자율적 로봇이 언젠가는 전쟁을 수행하는 데 사용될 것이다."(Arkin 2009) 따라서 자율무기시스템, 혹은 군사 로봇이 야기할 수 있는 윤리적 문제에 철학자들의 진지한 관심이 필요하다.

II. "살인 로봇 반대"를 넘어서

기계가 사람을 무차별적으로 살상하는 일을 허용해도 좋다고 생각하는 사람은 아무도 없을 것이다. 따라서 "살인 로봇 반대(Stop Killer Robots!)"는 상당한 직관적 설득력을 가진다. 이 같은 구호는 유엔의 특정재래식무기협약에 관한 논의에서 특정한 방향의 여론을 환기하고 군수 업체들이 자동살상무기를 제작하려는 움직임에 대해 압박을 가하기 위해 활동가들이 손쉽게 사용할 수 있는 효과적인 의사표현 방식일 수도 있다. 그럼에도 단지 '킬러 로봇'의 개발을 멈추고 사용을 금지하라는 구호를 외치는 것만으로는 군사 로봇의 윤리적 문제를 제기하고 이를 진지하게 다루는 일을 대신할 수 없다.

첫째, 무엇에 반대하려면 당신이 반대하려는 대상이 무엇인지 분명해야한다. 살인자 로봇을 반대하려면 그것의 개념과 범위부터 확정해야한다. "킬러 로봇 반대"를 외치는 많은 사람들은 부지불식간에 전쟁에서 사용될 목적으로 개발되는 모든 인공지능 로봇을 킬러 로봇으로 지칭하는 경향이 있다. 그러나 살인 로봇을 이렇게 광범위하게 적용하는 일은 적절치 않다. 문제는 무엇이 로봇인지에 관해서는 손쉽게 합의될 수 있는 반면, 살인자는 그 자체로 규범적인 개념이어서 윤리적, 법적 차원을 가진다는 데 있다. 통상, 정당하지 않은 이유로 혹은 정당하지 않은 수단이나 방법으로 사람의 목숨을 해친 경우, 그 행위를 수행한 사람을 살인자로 부른다. 결과적으로

사람의 목숨을 해쳤더라도 정당한 이유가 있다면, 그 행위는 살인이 아닐 수 있다. 살인과 살인자에 관한 이러한 통상적인 용법과 일치하여, 우리는 전쟁에서 사용되는 모든 무기를 살인자(killer)라고 부르지 않고, 참전한 모든 군인을 살인자로 부르지도 않는다.[5] 그렇다면 전쟁에서 잠재적으로 사용될 수 있는 모든 인공지능 기술을 "살인 로봇"으로 이름붙이는 일은 (그 수사적 효용을 떠나) 성급하다.

"킬러 로봇 반대"라는 슬로건에 만족할 수 없는 또 다른 이유는 그런 구호가 때로 합리적인 대화를 가로막을 수 있다는 우려에 있다. 군사 로봇의 개발을 "킬러 로봇"으로 낙인찍어 비판하기는 쉽지만, 왜 군사 분야에서 인공지능 기술 개발에 엄청난 재원을 투자하고 있는지 이해하는 일이나, 전쟁에서 군사 로봇은 윤리적인 임무 수행을 하는 것이 불가능한지, 로봇이 인간을 대체함으로써 인명의 피해를 줄일 가능성은 없는지 등을 검토하는 일이 간과되기 쉽다. 왜 인공지능 기술을 군사 분야에 응용하려는지, 그리고 그것이 어떻게 전쟁의 풍경을 바꿀 것인지, 전쟁에서 군사 로봇의 윤리적 사용이 가능한지를 살펴보기 위해 우리는 "살인 로봇"을 금지하자는 구호에서 멈출 수 없다.

우리가 첫 번째로 착수해야 할 과제는 분류 작업이다. 비교적 중립적인 표현인 군사 로봇(혹은 자율무기시스템)을 사용해 관련된 무기들의 범주화를 간략히 시도해 보자. 먼저, 임무의 성격에 따라 전투 로봇과 비전투 로봇으로 구분해볼 수 있다. 비전투 로봇은 전쟁 중에 운송, 탐지, 사상자 후송 등에 사용됨으로써 전장에 나가있는 병사의 수를 줄일 수 있고, 폭발물 제거와 같은 위험한 작업을 대신할 수도 있다. 예컨대, 2003년 출시된 iROBOT의 Packbot은 폭탄을 탐지하고 처리하며 정찰 및 감시 임무를 수행할 수 있고, 2004년 개발된 MARCbot은 이라크와 아프가니스탄에 1,000대 이상이 정찰 및 폭탄 제거 임무에 투입되었는데 팔 부분 위에

5. 물론 "살상무기"라는 표현은 자주 사용되지만, 살상무기는 인명을 살상하는 데 사람에 의해 사용될 수 있는 무기라는 뜻이지, 그 자체로 살인자라는 뜻은 아니다.

카메라를 탑재해 수 십 미터 밖에서 원격조정할 수 있고, 땅 밑에 매설됐거나 숨겨진 폭탄 탐지할 수 있다. 둘째, 모든 무기는 공격 무기와 방어 무기로 구분할 수 있고, 이에 따라 군사 로봇도 공격 로봇과 방어 로봇으로 구분할 수 있다. 예컨대 상대방의 미사일에 대해 자동으로 대응하는 자율 무기와 상대방을 인간의 제어 없이 자동으로 공격하는 무기는 다른 지위를 가질 수밖에 없다. 이때, 자동 방어 로봇을 킬러 로봇이라고 부르는 일은 부자연스럽다. 셋째, 군사 로봇은 자율성의 정도에 따라 구분될 수 있다. 현재 사용되고 있는 군사 로봇은 주로 원격조종되는 것들이지만 점차 반자율적인 것으로, 그리고 궁극적으로는 자율적인 것으로 발달할 것이다. 자율성의 정도란 결국 사람의 개입과 제어에 달려있으므로, 사람이 루프 안에/위에/밖에 (man in/on/out of the loop) 있는 것으로 구분해볼 수 있다. 원격조종은 사람이 루프 안에서 결정권을 가지는 경우이되 물리적인 거리를 늘리는 기술이라면, “루프 위의 사람”이란 무기시스템이 자율적으로 판단하고 실행하는 것을 허용하되 모든 과정을 사람의 감독 하에서 하도록 하는 반자율적 체계를 말하고, “루프 밖의 사람”이란 완전히 자율적인 무기시스템을 뜻한다.[6]

인공지능 기술이 적용된 군사 로봇이 개발되는 데에는 다양한 동기들이 있다. 단순하게는, 인간 전투원 대신 군사 로봇을 전장에 내보낼 수 있다면 전쟁을 수행하는 병력의 수를 줄일 있을 것으로 기대할 수 있다. 군사 로봇을 전장에 내보냄으로써 더 적은 인원수로 전쟁을 수행하고 마무리할

6. 현재의 군사 기술이 주로 원격조종에 머물러 있는 것으로 보이지만 반자율적 기술로 가기 위한 기술 개발이 활발하다는 것도 부인할 수 없다. 원격조종되는 드론은 이미 위협적인 살상 기술이 되었다. 예컨대, MQ-1 Predator는 애초에 정찰 목적으로 개발되었으나 이후 강력한 무기를 장착하여, 아프가니스톤, 파키스탄, 보스니아, 그리고 이라크 전쟁에 투입되어 다양한 지역을 공습한 것으로 알려져 있다. 전장에서 12,000km 떨어진 미국 네바다 주에서 마치 비디오게임기 같은 조종기로 조종하며 hellfire 미사일을 발사할 수 있다. 더 개량된 무인 전투항공기(Unmanned Combat Air Vehicle)는 MQ-9 Reaper인데, hellfire 미사일을 두 개 탑재할 수 있었던 Predator와 달리, 14개까지 미사일을 탑재하여 2007년 아프가니스탄 전쟁에서 사용된 것으로 알려져 있다.

수 있다면 군사 로봇을 개발할 이유가 있어 보인다. 둘째, 인간 전투원이 꺼려하거나 수행하기 어려운 임무 수행을 군사 로봇이 대신할 있다.(Lin, Bekey, and Abney 2009) 예컨대, 지형적으로 제약이 심한 지역의 정찰 임무를 수행하거나, 핵무기나 생화학무기 공격 이후 환경 표본을 수집하거나, 급조폭파물(IED)을 무력화하는 등 힘들고 더럽고 위험한 임무를 수행하는 데 군사 로봇이 더 적합할 수 있다. 따라서 병력의 수 자체를 줄이거나 인간이 수행하기 어려운 임무를 대신 수행함으로써 군사 로봇은 아군의 사상자를 감소시킬 수 있다. 셋째, 전쟁 수행의 범위와 방식을 획기적으로 개선할 수 있다. 군사 기술은 거리를 늘리는 방식으로 발전해왔다. 근접 거리에서 창이나 칼로 상대방을 공격하는 데에서 총기와 대포를 사용하는 데에로, 그리고 전투기로 폭격하거나 원거리에서 미사일을 발사하는 방향으로 발전해왔다. 군사 기술의 발전 과정은 대상과의 물리적 거리뿐 아니라 심리적 거리도 늘려왔다. 인공지능 시대 군사 기술은 증대된 자율성 덕택에 더 넓은 지역에서 더 정교한 작전 수행이 가능한 방향으로 발전하고 있다.[7] 자율적 무기의 한 가지 잠재적 가능성은 군집 기술(swarm technology)이다. 한 명의 운용자가 하나의 원격조정 로봇을 조종하는 대신, 자율화된 여러 로봇들은 서로 통신하면서 오직 소수, 심지어 한명에 의해 관리될 수 있고, 이를 통한 동시다발적 대규모 로봇 공격도 가능해진다.

군사 로봇이 인간 병사를 대체하는 것은 단지 아군의 사상자를 줄이는 것만 목표로 하고 있지 않다. 전장에 참여하는 전투원은 배고픔, 갈증, 피로, 수면부족 등으로 인해 극심한 스트레스에 시달리거나, 두려움, 망각, 사기저하 등에 노출되기도 하고, 심지어 전쟁 이후에는 PTSD로 고통받는다.(Sharkey 2012) 군사 로봇의 활용은 효과적인 임무 수행을 방해하는

7. 현재 많이 활용되고 있는 원격조종 기술은 제조 및 운용에 큰 비용이 소요되는 것으로 알려져 있다. 게다가 일과 시간에 원격으로 살상 무기를 조종하고 저녁에는 집에 돌아가 가족과 함께 시간을 보내는 군인들의 심리적 상태를 살피는 일도 결코 쉽지 않다. 이렇게 심리적 상태에 대한 관리도 운용비용에 포함된다. 군사 기술의 자율성을 증대시키는 것이 오히려 비용을 줄일 수 있다.

인간 전투원의 심리적 문제를 우회해 이 같은 문제를 완화하는 데 도움을 줄 것으로 기대된다. 인간은 정서적 동물이고 전장에서 죽을 수 있다는 공포감으로 겁에 질리기도 하지만, 반대로 적군이더라도 사람을 죽이는 일을 꺼리기도 한다. 2차 대전에 사용된 총알에 대한 한 분석에 따르면, 보병들이 사용한 대부분의 총알은 적군을 겨냥하지도 않았다.(Sharkey 2012, 111-112) 전투원들은 여러 심리적 요인으로 인해 살상을 꺼리기도 하지만 때로는 지나친 행동을 보이기도 한다. 만일 전투원이 임무 수행을 꺼리면 효과적인 전쟁을 수행할 수조차 없다. 반대로, 교전 수칙을 위반하는 과도한 행동들은 전쟁 범죄로 발전하기도 한다. 예컨대, 전우의 사망으로 인해 복수심에 사로잡히면 적군이나 비전투원을 학대하고 고문하기도 하며, 동료의 잘못은 사소한 것으로 치부하고 은폐하기도 하고, 때로는 일부러 살생을 즐기는 경우도 생겨나게 된다.[8] 전쟁에서 필요한 것은 "효과적인" 상대방에 대한 제압이다. 필요한 경우 필요한 만큼의 살상을 통해 추가적인 불필요한 살상을 막는 것이 필요하다. 우리는 전투에서 군인이 냉철하고 효과적으로 작전을 수행하길 바라고, 평상시 그런 임무 수행이 가능하도록 군인들을 훈련시킨다. 그러한 훈련을 통해 우리는 군인을 효과적인 살인 기계로 만드는 것은 아닌지, 그렇다면 오히려 그러한 일은 기계가 더 잘할 수 있는 것 아닌지 생각해봄직하다.[9]

우리는 군사 로봇을 여러 차원에서 분류해보고 군사 로봇의 자율성을 증대하려는 기술적 시도 뒤에 숨겨진 동기들을 살펴보았다. 이 절의 앞부분

8. 적에게 사로잡힌 동료를 구하기 위해 고문이 허용되어야 한다고 생각하는 군인들이 다수이며, 이라크 전에서 미군들은 실제로 고문이 필요하지 않은 상황에서도 이라크의 비전투원들을 학대하기도 했다.
9. 이 글에서는 현재 전쟁수행이나 훈련의 행태에 관해서는 자세히 논의하지 않겠다. 전쟁이 지속적으로 벌어지고 있고 앞으로도 일어날 것이라는 현실적 상황 판단 하에서, 로봇의 윤리적 전쟁 수행이 가능한지에 초점을 맞추어 논의한다. 자율무기의 개발을 통해 전쟁 없는 세상을 상상하려는 시도는 불가능하지 않지만, 특정한 기술이 우리는 더 나은 세계로 데려갈 것이라는 단순한 낙관주의는 기술의 개발을 둘러싼 경제적, 사회적, 정치적, 윤리적 쟁점들을 간과할 우려가 있다.

에서 나는 "킬러 로봇 반대"만을 외치는 것으로는 군사 로봇의 윤리적 탐구를 위해 충분치 않다고 지적했다. 전투에 참여하는 모든 군인이 살인자가 아닌 이유는 전쟁에서 정당한 무력의 사용과 부당한 사용을 구분할 수 있기 때문이다. 그렇다면 전쟁에서 자율적 무기시스템의 정당한 사용과 부당한 사용을 구분할 여지가 있는지, 언제 군사 로봇의 사용이 정당할 수 있는지 세심하게 검토하는 일이 필요하다.

Ⅲ. 정당한 전쟁과 군사 로봇의 윤리

군사 로봇의 윤리적 쟁점은 다양하다. 로봇이 전쟁에 사용될 때 인간과 로봇이 맺는 관계는 어떠하며 또 무엇이 바람직한 인간-로봇 관계인지를 탐구하는 것도 한 가지 주제일 수 있다. 예컨대, 로봇이 자기 주위에서 벌어지는 모든 세부사항을 기록하도록 설계되어 있다면, 이것이 인간 분대원들 사이의 관계, 인간 분대원과 군사 로봇 사이의 관계에 어떤 영향을 미칠지 탐구할 수도 있다. 여러 논점에도 불구하고, 핵심적인 쟁점은 군사 로봇을 개발하고 전쟁에서 실제로 사용하는 것이 과연 윤리적인가 하는 것이다. 로봇의 개발자와 군사 관계자 그리고 일부 철학자를 제외하면, 일반 대중의 직관적인 반응은 꽤 분명해 보인다. 많은 사람들은 스스로 판단하고 움직이는 로봇에 의해 사람의 인명이 살상되는 것을 공포스럽게 여기며, 아마도 그러한 강한 심리적 거부감으로 인해 군사 로봇의 옹호자는 소수에 불과할지도 모른다. 그러나 중요한 윤리적 문제가 다수결이나 강한 직관적 반응에만 의존하여 결정될 수는 없다. 많은 사람들은 자율적인 무기의 도입으로 전쟁이 쉬워질 가능성을 우려하기도 한다. 군사 무기의 자율성이 점점 증대될수록 전쟁의 개시, 참전 결정, 교전 등의 문턱이 낮아질 수 있다는 것이다.

자율적 무기 시스템 혹은 킬러 로봇의 개발과 제한적 사용을 옹호하는 논변들도 존재한다. 우선, 어떤 기술이든 한번 개발되기 시작하면 지속적인

발전과 사용을 결코 막을 수 없으며, 우리가 할 수 있는 최선은 그것을 선점하여 충분한 (국가적) 이득을 누르고 오용을 막는 것뿐이라는 생각이 팽배하다. 군사 로봇 개발의 불가피성에 관한 이러한 생각은 기술결정론적 시각을 가정한 것으로 보인다. 그러나 기술적 가능성은 그것의 사회적 실현가능성도 동일하지 않고, 기술 발전의 경로가 사회적 맥락 속에서 정치, 경제, 문화, 젠더 등 여러 요소들과 상호작용한다는 이론은 잘 확립되어 있다. 군사 무기의 경우에도 예외가 아니다.(Bijker, Hughes, and Pinch 1987) 재래식 무기에 관한 국제적 조약을 통해 국제사회는 전쟁에 사용해서는 안 되는 무기의 종류들을 규정하고 있으며, 이를 위반하지 못하도록 국제법에 의해 압력을 가하기도 한다. 자율적 무기 체계를 규제하려는 국제적 노력이 통상적인 무기들의 경우와 달리 작동하지 않는다고 보아야할 이유는 없다.

일군의 옹호자들은 결과주의적 논변을 제시하기도 한다.(Arkin 2009; Sullins 2010) 단적으로 말해, 군사 로봇을 전투에 내보내 사람의 목숨을 덜 희생시킬 수 있다면, 군사 로봇의 사용은 윤리적으로 정당화될 수 있다는 것이다. 물론 전쟁에서 희생되는 인명을 조금이라도 줄일 수 있다면 바람직한 일일 것이다. 그러나 결과주의가 반직관적인 귀결을 산출하는 많은 사례들을 우리는 알고 있다. 예컨대, 무고한 한 사람의 장기를 이식해 다섯 사람을 살릴 수 있다고 해도, 우리는 무고한 이에 대한 정기적출이 정당화될 수 있다고 여기지 않는다. 우리의 논의에서 과연 로봇이 사람을 해치는 것이 허용될 수 있는지가 관건이 된다. 물론 로봇에 의한 인명 살상은 무조건적으로 금지해야한다는 주장할 수도 있다.[10] 그러나 전투에서 이미 인간들 사이의 살상이 무자비하게 벌어지고 있을 뿐 아니라 고도로 발달한 수많은 전투 장비들이 사용되고 있는 상황에서 로봇이 참여하지 말아야할 원칙적인 이유가 있는지 의문이다. 따라서 이 절에서는 군사 로봇이 정당한 전쟁 수행에 참여할 수 있는지, 즉 전쟁법과 교전수칙에

10. 아이작 아시모프의 로봇 3원칙은 로봇이 인간을 해치지 못하도록 금지하고 있다.

따라 행동하며 전쟁범죄를 피할 수 있는지에 초점을 맞춘다. 이를 위해서는 완벽하게 윤리적인 로봇이나 모든 상황에서 올바른 결정을 내리는 판단 시스템이 요구되지 않는다. 다만, 전쟁에서의 적법 행위와 불법 행위를 구분하고 불법적이거나 비윤리적 전투 수행을 억제하고 전쟁 범죄를 감소시킬 수 있어야한다.

국제인도법(International Humanitarian Law)은 무력충돌 시 인간의 고통을 예방하고 최소화하기 위한 목적으로 제정되었으며, 무력충돌 시 적대행위에 가담하지 않거나 할 수 없는 사람들을 보호하고 전투의 수단과 방법을 규제하기 위한 내용을 담고 있다. 이 법은 정부와 군대뿐 아니라 무장단체 등 무력충돌 당사자 모두가 준수해야 하는 국제법으로, 그 적용대상에 있어 보편성을 지닌다. 국제인도법의 기원은 1864년 최초의 제네바 협약인 "육전에 있어서의 군대 부상자의 상태 개선에 관한 협약"으로, 이후 체결된 총 4개의 제네바 협약과 2개의 추가 의정서로 구성되어 있다. 1906년 제정된 제2협약은 "해상에 있어서의 군대의 부상자, 병자 및 조난자의 상태 개선"에 관하여, 제3협약(1929)은 포로 대우에 관하여, 제4협약(1949)은 민간인 보호에 관한 내용으로 되어있고, 1977년 "국제적 무력충돌 피해자들의 보호를 강화"와 "비국제적 무력충돌 피해자들의 보호를 강화"라는 두 개의 추가 의정서가 채택되었다. 국제인도법은 적대행위에 가담하지 않거나 할 수 없는 자, 투항한 자, 부상자와 병자 등의 육체적, 정신적 보전에 대한 권리를 옹호하며, 신체적 정신적 고문이나 학대 행위를 금지하고 있고, 특히 민간인과 전투원의 구분을 강조하고 하면서 공격이 전적으로 군사 목표물에 국한되어야 함을 강조한다. 또한, 화학 무기, 생물학적 무기, 대인 지뢰 등의 무기 사용도 금지하는데, 금지의 원칙은 다음과 같다. 전투에 참여하는 전투요원과 그렇지 않은 자를 구별하지 못하거나, 불필요한 살상과 고통을 초래하거나, 환경에 심각하고 장기적인 손해를 야기하는 무기는 금지된다.

국제인도법이 적대행위에 가담하지 않는 사람들을 보호하고, 전쟁에 사용되는 수단과 방법을 규제하지만, 그러한 규제의 원칙을 제공하는 한편

전쟁의 윤리를 포괄적으로 다루는 가장 영향력 있는 이론은 정당한 전쟁 이론(Just War Theory)이다.(Orend 2008) 이 이론에 따르면 전쟁의 정당성은 세 가지에 달려있다. 전쟁 개시의 정당성을 다루는 개전법(jus ad bellum), 전쟁 수행의 정당성을 다루는 교전법(jus in bello), 그리고 평화조약 및 전쟁 종식에 관한 전후법(jus post bellum)이다. 이 가운데 군사 로봇의 정당한 전쟁 수행을 다루는 우리의 맥락에서 교전법이 가장 유관한 범주이다. 교전법에서 가장 중요한 두 원칙은 식별 원칙(the principle of discrimination)과 비례성 원칙(the principle of proportionality)이다.

전쟁 등 무력충돌에서는 공격의 목표물이 명확히 결정되어야 하며, 적절한 공격 대상에 한정하여 수행된 공격 행위만이 정당하다. 식별 원칙은 이를 위해 민간인과 전투원을 구별할 것을 요구한다. 또한, 민간인뿐 아니라 더 이상 전투의지가 없거나 능력을 상실한 부상자, 투항자, 정신이상자에 관해서도 민간인에 준하여 대우하고 전투원과 구별할 것을 요구한다. 두 번째 원칙은 비례성의 원칙이다. 아무리 적군의 전투원을 대상으로 하더라고 무의미하고 무차별적인 살상은 용인되지 않는다. 공격에 따라 예상되는 사상과 재산의 손실은 구체적이고 직접적인 군사적 이득에 비추어 과도해서는 안 된다.(Petraeus and Amos 2006) 군사적 중요성이 무고한 시민의 희생보다 큰 정도에 비례해서만 공격은 허용될 수 있다. 이제 문제는 전장에 내보낼 군사 로봇이 이러한 원칙을 내장하거나 학습하고 이를 준수함으로써 정당한 군사 행동의 주체가 될 수 있는가 하는 것이다. 식별 문제와 비례성 문제를 차례로 다루되, 더 중요하게 생각되는 식별 문제에 논의를 집중하자.

먼저, 정당한 전투 수행을 위해 군사 로봇은 아군과 적군, 전투원과 민간인을 식별할 수 있어야한다. 군사 로봇이 적군의 전투원과 비전투원을 구분하는 능력을 가지도록 설계될 수 있을까? 이 문제는 일차적으로 기술적 문제처럼 보인다. 일부 연구자는 이 물음에 대해 긍정적이다. 로봇의 행동은 컴퓨터 프로그래밍에 의해 좌우되기 때문에, 교전 규칙을 따라 행동하도록 만들 수 있다면 필요한 식별 능력을 갖출 수 있다는 것이다.(Powers 2006) 그러나 많은 이들은 이런 추정에 회의적이다. 규칙 따르기를 윤리적 행동의

수행과 동일시할 수 없다. 규칙이란 늘 여러 해석의 가능성에 열려 있기 마련이며, 로봇이 주어진 규칙을 구체적인 상황에서 올바로 해석하리라는 보장이 없다. 이때 문제가 되는 것은 결국 로봇이 인간 병사가 가지는 방대한 배경 지식을 결여한다는 사실이다. 적군의 전투원으로 분류하는 것이 매우 확실해보여도, 때로는 그가 항복의 몸짓을 표시하고 있으며 그래서 교전의 의지가 없음을 알아차리는 일이 필요할 수도 있고, 때로는 그 상대가 심각한 부상으로 교전 능력이 없음을 확인해야할 수도 있다. 이를 위해서는 때로는 감정 표현, 기만적 의도나 속임수 파악 능력이 필요하고, (구조화되지 않은) 다양한 환경 하에서 정보처리가 가능해야하며, 전쟁 상황에 대한 상당한 배경 지식도 필요할 수 있다. 그런데 실전에 배치하기 전에 이를 시험해보는 일은 매우 어렵다. 시험 환경에서 미리 점검해보는 경우에도, 시험 환경이란 복잡하고 역동적이며 불확실한 전투 환경과 다를 수밖에 없기 때문이다. 결국 로봇의 실전 배치는 매우 높은 기준점을 통과한 경우에만 허용되어야 할 것이다.

물론 얼마나 높은 기준점을 통과해야하는지가 관건이다. 인간과 마찬가지로 군사 로봇의 식별 능력이 100% 정확할 수 없다. 인간의 식별 능력이 완벽하지 못하다고 해서 전쟁 참여가 불가능한 것은 아니듯, 군사 로봇의 식별 능력이 어느 정도가 되어야 만족할 수 있는지에 관해서는 추가적인 논의가 필요하다. 이는 식별 문제가 단순히 기술적 문제에 국한되지 않음을 보여준다. 어떤 이는 군사 로봇의 식별 능력에 대한 평가는 인간의 식별 능력에 상대적으로 이루어져야 한다고 주장할 수 있다. 이는 상당히 타당한 주장으로 들리지만, 현재 기술 수준을 짐작컨대 군사 로봇의 식별 능력은 아직 충분한 수준에 이르지 못한 것 같다. 그러나 현재의 낮은 기술 수준을 감안하더라도 군사 로봇을 전장에 내보낼 수 있다고 주장할 수 있다. 슐츠케(Schulzke 2011)는 군사 로봇에게 상당히 제한적인 경우에만 공격하도록 교전 규칙을 부여함으로써 문제를 비교적 간단히 해결할 수 있다고 제안한다. 공격대상이 전투원이라는 판단이 확실하고 그로 인해 근방의 민간인에게 해를 끼치지 않을 경우에만 공격하도록 하고, 그렇지 않은 경우에는 무력

사용을 자제하도록 하면 비교적 낮은 식별 능력을 가진 군사 로봇의 사용이 가능하다는 것이다. 이러한 제안에 대한 조금은 상세한 검토가 필요해 보인다.

통계학적 가설 검정에서 사용되는 1종 오류와 2종 오류의 개념을 차용해서 슐츠케의 제안을 이해해 볼 수 있다. 1종 오류는 거짓 양성, 즉 해당 가설이 참이 아닌데 채택하는 경우를 말하며, 군사적 식별이 필요한 상황에 대입해 보면 민간인이거나 전투 의사가 없는 사람을 전투요원으로 오인하는 경우이다. 반대로, 2종 오류는 거짓 음성으로 해당 가설이 참인데 채택하지 않고 기각하는 오류를 말한다. 현재 맥락에서 2종 오류는 적군의 전투원을 민간인으로 오인한 경우를 뜻한다. 통상 2종 오류가 더 치명적이며, 군사적 맥락에서 2종 오류는 그 오류를 저지르는 사람 자신에게 매우 치명적이다. 따라서 우리는 통상 2종 오류를 피하려는 경향을 가지며, 이와 연동하여 때때로 1종 오류를 저지르게 된다. 군사적 맥락에서 1종 오류는 민간인을 살상하는 경우를 가리키며, 현대 전쟁 양상에서 점차 그 비중이 커져가는 게릴라전에서는 1종 오류를 저지를 가능성이 더 커지게 된다.

이러한 분석틀에 비추어볼 때, 슐츠케의 제안은 식별 능력이 충분치 좋지 않은 로봇에게 자신보다 민간인 보호를 우선적으로 고려하는 가치를 입력함으로써 1종 오류의 가능성을 낮추자는 제안으로 이해된다. 이러한 제안은 이론적으로 가능하고 또 매력적으로 들리지만, 숨겨진 가정은 비교적 분명하다. 2종 오류를 저지르는 경우에도 로봇 자신이 받는 피해는 치명적이지 않고, 1종 오류의 확률을 낮춤으로써 더 윤리적인 전쟁 수행이 가능하다고 가정하는 것이다. 사실 이러한 제안은 윤리 이론의 차원에서 나쁠 것이 없어 보인다. 문제는 그러한 제안이 과연 실현가능한지 의구심이 든다는 데 있다. 자율적으로 움직이고 적을 식별하고 공격 여부를 판단하여 실행하는 로봇은 아직 현실에 존재하지 않겠지만 그러한 로봇의 제작에 엄청난 비용이 필요하다는 것은 구태여 강조할 필요가 없다. 최소 수억 단위의 돈을 들여 제작된 로봇이 전장에서 매우 소극적이고 보수적으로 임무를 수행한다는 것은 경제적 관점에서 전혀 현실적인 제안이 아니다. 게다가

그렇게 수동적으로 대응하는 경우라면, 위장한 적에 의해 탈취될 가능성이 높아지고 적군이 이를 개량하게 된다면 그 결과는 오히려 예상과 반대일 수 있다. 실현가능성 외에도 책임의 문제가 남아있다. 군사 로봇이 민간인을 적군 전투원으로 오인하여 전쟁 범죄를 저지를 가능성을 낮출 수 있다고 하더라도 완전히 없앨 수는 없다면, 벌어진 불행한 사태에 대해 누군가는 어떤 식으로든지 책임을 져야 한다. 즉, 책임의 문제에 관한 논의가 이루어지지 않는 한 슐츠케의 제안은 부분적이고 비현실적인 대안으로 머물고 만다.

식별 원칙과 더불어 비례성 원칙에 관해 생각해보자. 구체적이고 직접적인 군사적 이득에 비례해서만 상대방을 공격하는 것이 허용된다면, 군사로봇은 "비례하는 대응"에 관해 고려하고 계산할 수 있어야 한다. 무엇이 공격으로부터 얻을 것으로 기대되는 구체적이고 직접적인 군사적 이득인지, 무엇이 비례하는 대응인지를 판단하려면 고도의 지식이 필요해 보인다. 이런 문제에 관해서는 인간의 판단이 필요한 부분이 많고, 특히 불확실성에 관한 고려가 필수적이다. 직접적인 군사적 이득은 공격으로 인한 적군의 사상자 수를 포함하지만 공격을 감행했을 경우 아군과 상대편에게 끼치는 영향들을 두루 살펴야 계산될 수 있다. 무엇을 손실로 간주할 것인지도 가치 판단이 불가피하게 개입될 것이다. 현존하는 로봇이 그러한 판단을 독자적으로 내릴 수 있다고 기대하는 것은 지나치다. 물론 인간의 판단 능력이 필요하다고 해서 인간들 사이의 공감능력이나 고통 인지 능력을 필요로 하는 것은 아니다. 그리고 불확실성을 확률적, 통계적 정보로 취급함으로써 어느 정도 계산에 고려하는 것이 원칙적으로 불가능한 것은 아닐 것이다. 그렇다면 식별 원칙에서 취했던 보수적 접근을 비례성 원칙에서도 시도해볼 수 있을 것이다. 즉, 추정되는 비례성을 감안할 때, 그 가운데에서 최소한의 물리력만을 행사하도록 로봇을 프로그래밍할 수 있을 것이다. 그리고 오히려 로봇이기에 무의미한 공격이 아니라 정확하고 효율적인 공격이 가능하다고 주장할 수도 있다. 그러나 이러한 대응은 식별 문제에서 대해 지적된 동일한 문제에 직면하게 된다. 그러한 보수적 접근은 현실적이지 않으며, 그 전에 해결해야할 책임의 문제가 남아있다.

IV. 자율군사로봇과 책임의 문제

아무리 정교한 군사 로봇이 개발되고 배치될지라도 때때로 나쁜 결과가 생길 수 있다. 실제로 군사 로봇공학 분야에서 실패의 사례들이 없지 않았다. 2008년 4월, 이라크에서 탤론 스워드는 오작동을 일으켰으며, 2007년 10월 준자율 로봇 포의 오작동으로 9명 사망하고 14명 부상한 경우도 있다. 점차 기술의 복잡성이 증대됨에 따라서 (실험실 내에서 여러 차례의 시험을 통과했다고 하더라도) 예측할 수 없는 사건이 발생하거나 프로그램들이 검증되지 않은 방식으로 상호작용할 가능성도 배제할 수 없다. 물론 이러한 오류의 가능성은 인간에게도 마찬가지로 적용된다.

전쟁에서 무언가 잘못될 수 있음을 전제로 할 때, 누가 나쁜 결과에 대해 책임질 수 있는지는 군사 로봇의 윤리학에서 근본 문제이다. 책임질 수 있는 능력 혹은 구조가 정당한 전쟁 수행의 전제 조건이기 때문이다. 이러한 조건에서 출발하여 스패로(Sparrow 2007)는 자율적 군사 로봇의 사용이 비윤리적이라고 주장한다. 논증의 핵심은 다음과 같이 요약될 수 있다.

전제1. 책임질 수 있음은 교전법의 선제조건이다.
전제2. 전쟁에서 군사 로봇을 사용할 때 그리고 그것이 해로운 결과를 야기했을 때, 책임을 질 수 있는 주체는 셋 중 하나이다. 바로, 군사 로봇의 설계자, 군사 로봇을 전장에 내보내고 임무를 준 지휘관, 그리고 로봇 자신이다.
전제3. 세 후보 가운데 어느 쪽도 군사 로봇이 발생시킨 해로운 결과에 대해 온전한 법적, 혹은 윤리적 책임을 질 수 없다.
결론. 따라서, 자율적 군사 로봇의 사용은 비윤리적이다.

스패로의 트릴레마 논변은 언뜻 보기에 타당한 연역 추론으로 보인다. 만일 세 전제를 모두 받아들이면 결론을 피할 수 없기 때문이다. 먼저

우리는 첫 번째 전제에 동의할 수 있을 것이다. 전제2에 관해서는 약간의 논란이 있을 수 있겠으나, 세 후보 외에 새로운 후보군을 추가한다고 해서 논증의 구조를 약화하지는 않을 것 같다. 핵심적인 단계는 전제3의 참을 입증하는 것이다. 그래서 스패로는 설계자도 지휘관도 로봇 자신도 책임을 질 수 없음을 입증하는 데 노력을 기울인다.

먼저, 스패로에 따르면, 로봇의 설계자는 자율 로봇이 일으킨 결과에 책임을 질 수 없다. 물론 설계상의 오류가 있다면 설계자에게 책임이 없다고 할 수 없다. 그러나 설계상의 문제가 없는 데에도, 군사 로봇의 "자율적인" 판단과 실행이 나쁜 결과를 일으켰다면 그리고 그러한 결과를 설계자가 이미 예측할 수 없는 경우라면 설계자에게 책임을 묻는 것은 온당치 못하다. 자신이 예측할 수도 통제할 수도 없는 사태에 관해 어느 누구도 온전히 책임을 질 수 없기 때문이다. 이러한 고려는 지휘관의 경우에도 마찬가지로 적용된다. 군사 로봇의 가진 나름의 자율성으로 인해 그것을 전장에 내보낸 지휘관이 자율 로봇의 행동을 완전히 통제할 수 없는 상황이었다면 그것이 가져온 결과에 관해 지휘관에게 책임을 지우는 것은 부당할 것이다. 그렇다면 로봇에게 책임을 물어야하지 않을까? 스패로에 따르면, 기계 자체는 책임을 질 수 없다. 어떤 것이 책임일 질 수 있으려면 그것에 대한 칭찬과 비난이 가능해야하고, 칭찬을 위해서는 보상이 또한 비난을 위해서는 처벌이 가능해야한다. 일단 처벌에 초점을 맞추어보면, 어떤 대상을 처벌하려면 그 대상은 고통을 받고 괴로워할 수 있는 능력이 있어야한다. 그렇지 않다면 처벌은 무의미한 이야기가 되기 때문이다. 문제는 현재 상태로, 그리고 가까운 미래에 기계가 고통을 느낀다는 생각은 실현되기 어렵다는 데 있다. 기계가 고통을 느낄 수 없다면 기계를 처벌한다는 생각은 이치에 맞지 않게 되고, 처벌이 원칙적으로 불가능하다면 그것이 책임질 수 있는 주체일 수 없다. 따라서 군사 로봇이 야기하는 나쁜 결과에 관해 책임을 질 수 있는 세 후보, 설계자와 지휘관 그리고 로봇 어느 쪽도 온전히 책임을 질 수 없기에 자율적 군사 로봇의 사용은 비윤리적이며 전쟁에 사용되지 말아야한다.

전쟁에서 자율 로봇의 사용을 금지해야한다는 스패로의 논변은 여러 흥미로운 논점들을 제기한다. 하나는 그가 말하는 자율 로봇의 "자율성"에 관한 것이다. 완전히 자율적인 존재는 (그것이 통증의 감각질을 가지는지 여부를 떠나) 자신이 야기한 의도된 결과에 대해 비난받을 수 있고 또 응당 책임을 져야한다. 반면, 자율적이지 않은 로봇의 작동으로 인해 발생한 결과에 대해서는 설계자와 감독자가 전적으로 책임을 져야한다. 스패로의 논변은 완전히 자율적이지 않지만 그렇다고 전적으로 통제 아래 있지도 않은 "부분적 자율성"을 가진 로봇의 존재를 가정함으로써 성립한다.(Simpson and Muller 2016) 다른 논점은 세 행위자 가운데 어느 누구도 온전히 책임질 수 없다는 스패로의 논변이 하나의 행위자가 온전히 책임을 져야한다는 강한 가정 위에 서있다는 점이다. 사실 두 논점은 중첩되어 있다. 완전히 자율적인 존재는 자신의 행위가 야기한 결과에 대해 온전히 홀로 책임을 질 수 있어야 한다고 요구할 수 있다. 반면, "부분적 자율성" 만을 가진 존재의 행위에 관해서, 혹은 자율적인 존재이지만 외부의 제약에 의해 자신의 자율성을 부분적으로만 발휘한 경우, 책임질 수 있는 방식이 없다고 스패로는 가정하고 있다. 나쁜 결과는 발생했지만 책임지는 사람은 없는 책임의 간격(responsibility gap)이 생길 수도 있다는 것이다.

부분적 자율성이 책임의 간격(responsibility gap)을 낳는다는 이러한 가정은 다소 비현실적이며 지나치게 강하다. 부분적 자율성을 가진 개체의 행동에 관해 책임을 물을 수 있는 여러 방식이 존재한다. 예컨대, 반려견은 그 행동을 주인이 완전히 통제할 수 없다는 점에서 부분적으로 자율적이다. 반려견이 지나가는 행인을 물어 상해를 입히거나 타인의 소유물을 손상시켰다면, 그가 가진 부분적 자율성에도 불구하고 반려견의 주인은 책임을 져야한다. 법적으로 완전한 자율성을 가진 것으로 인정되지 않는 10대 청소년의 경우를 생각해보자. 내전으로 혼란스러운 지역에서는 미성년인 소년들을 훈련시켜 군사 작전에 가담시키는 경우가 많다. 내전에 참전한 소년병이 살인을 했다면 그 행위에 대해 책임져야하는 사람은 명확하다. 바로 그를 징집하고 훈련시켜 전장으로 내보낸 지휘관이다. 소년의 손에

총을 들려주고 그를 극심한 심리적 스트레스를 받도록 만들어, 정당하지 않은 살상이 발생할 수도 있는 상황으로 그를 내몰았다면, 소년병이 아니라 지휘관이 책임을 져야한다. 따라서 부분적 자율성이 책임의 간격을 발생시킨다는 가정은 너무 강하다.

게다가, 단지 군사적 맥락 뿐 아니라 많은 일상적 맥락에서 책임의 귀속은 단일 행위자에 국한되지 않는다. 어떤 행위자의 행동으로 인해 나쁜 결과가 발생한 경우라고 하더라도, 그 결과에 대한 책임이 여러 주체와 행위자에 분산되는 경우가 많고, 특히 이는 전쟁에서 군대의 활동과 그것의 책임 실행에도 부합한다. 스패로는 분산된 책임 혹은 집단 책임이라는 현실을 간과하고 단일 행위자의 책임이라는 비현실적 가정 위에 자신의 논변을 세우려했다는 비판을 피하기 어렵다.

병사 한 명의 행동이 나쁜 결과를 일으켰다고 하자. 이것은 온전히 자율적인 한 명의 병사가 의도하고 실행한 결과일 수도 있지만, 많은 경우 군사적 의사결정은 명령 계통을 따라 여러 차원에서 이루어진다. 따라서 우리는 군대의 명령 계통의 위계적 구조에 주목할 필요가 있다. 전쟁의 개시와 관련된 결정은 군대 조직 자체가 아니라 주로 정치인들에 의해 이루어진다. 대통령이나 의회는 군대가 수집한 정보에 기초해서 상층부 군인들의 조언을 받아 전쟁의 개시 여부와 시점을 결정한다. 군대의 상층부는 전쟁의 목표와 전략을 수립한다. 실제 전쟁의 수행은 다양한 수준의 지휘권자로부터 부대장, 그리고 개별 군인에 이르기까지 위계적 구조 속에서 이뤄진다. 개별 군인들은, 비록 그들이 자율적인 존재이지만, 본인의 의지와 의도가 아니라 상층부에서 수립한 전쟁의 방법과 전략, 전술에 입각해 그리고 지휘관이 하달하는 교전 규칙에 따라 임무를 수행한다. 이러한 위계적 명령 계통의 의미를 들여다볼 필요가 있다. 명령 계통이 위계적이라는 것은 상위 수준의 결정에 의해 하위 수준의 결정과 수행이 제약을 받는다는 뜻이고, 이에 따라 "자율적" 군인의 자율성이 상부의 명령에 의해 제약됨을 뜻한다. 말단 단위의 행동이 상부의 지휘를 받고 제약되는 만큼 그것의 책임은 가벼워진다. 책임은 명령 계통을 따라 분산되며 어느

단일 행위자가 온전히 떠안지 않는다. 자율성을 제약받는 쪽은 그만큼 책임도 경감되며 자율성을 제약하는 쪽은 그만큼의 책임을 더 부담해야한다.

군대의 위계적 의사결정 구조를 포함한 위의 사례들은 부분적 자율성이나 제약된 자율성이 곧바로 책임의 간격을 발생시키지 않음을 보여준다. 그렇다고 책임의 간격이 전혀 존재하지 않음을 뜻하지는 않는다. 예를 들어, 미성년인 10대 초반의 소년이 저지른 범죄는 많은 경우 형사적 처벌의 대상이 되지 않는다. 그에게 완전한 자율성이 있지 않다고 간주되기 때문이다. 혹은 예측할 수 없는 자연재해로 (500년 만에 가장 강력한 태풍이 왔다고 가정해보라) 교량이 무너져 많은 인명 피해가 발생했다고 해도, 교량의 설계자와 시공자, 감독 당국을 비난할 수 없다. 이 같은 사례는 책임의 간격이 때때로 발생할 수 있음을 보여준다. 이제 남은 과제는 관련 당사자들의 자율성이 부분적이거나 제약되었음에도 불구하고 책임 소재가 분명한 경우와 책임의 간격이 발생하는 경우는 어떻게 다른지를 밝히고, 자율적 군사 로봇의 경우 어떠한 경우에 해당하는지를 보이는 것이다.

논의를 지나치게 확대하지 않기 위해, 로봇이나 군사 기술을 포함한 기술적 인공물로 대상을 한정해 보자. 우리 논의에서 "군사 로봇"이든 "자율무기체계"이든 그것이 스스로 책임을 질 수 있을 정도로 충분히 자율적이지는 않다. 그것은 부분적 혹은 제약된 자율성을 가진다. 그럼에도 그것이 자율적인 것으로 간주되는 이유는 그 행동을 완전히 예측하거나 통제할 수 없기 때문이다. 완전한 자율성이 아닌 이상 자율성 개념 자체를 분석하는 것은 큰 의미가 없으므로, 예측불가능성과 통제불가능성이라는 다소 완화된 의미의 자율성의 지표를 활용하여 논의를 전개할 수 있다. 이제 새삼스럽게 깨닫게 되는 것은, 거의 대부분의 기술적 인공물이 (정도의 차이는 있지만) 어느 정도는 예측불가능하다는 것이다. 달리 말해, 대부분의 기술은 위험(risk)을 내포하기 마련이다. 예컨대, 우리는 감기약이 인체 내에서 작동하는 방식을 완전히 예측할 수 없고, 때로는 예상치 못한 부작용을 발생시킬 위험성이 있음을 알고 있다. 아주 튼튼하게 지은 건물도 때로는 예측하지 못한 상황에서 붕괴할 위험성(risk)가 있다. 그러나 책임은 규범적 개념이다.

우리는 약품이나 건물과 관련된 설계자, 제조자, 관리자, 감독자, 혹은 규제 당국의 책임을 묻기도 하고 책임을 면제하기도 하는데, 이를 판단하는 데에는 관련 행위자의 과실 여부가 핵심적인 역할을 한다.[11] 과실이 인정되면 책임을 묻고 과실이 아니라면 책임을 면할 수 있다. 그런데 무엇이 과실인지는 어디까지를 지켜야할 의무로 규정하는지에 달려있고, 그 의무를 다하지 못하는 경우 책임을 져야한다. 여기서 중요한 것은 그 의무의 범위가 선험적으로 주어져 있지는 않다는 점이다.[12]

부분적으로 자율적인 군사 로봇도 예외가 아니다. 군사 로봇 스스로가 책임 능력이 없기에, 누군가는 책임을 질 수 있어야한다. 설계자, 제조사, 각 단계의 군사 지휘관, 함께 임무를 수행하는 병사, 그리고 규제 당국 등은 군사 로봇이 일으킬 수 있는 나쁜 결과에 대한 책임을 누가 어떤 방식으로 나누어질지, 그 책임의 종류와 무게에 관한 타협을 시작해야한다. 물론 이러한 책임 논의 자체는 규제 당국이 책임을 지고 수행해야한다. 자율적 군사 로봇과 관련된 당사자들의 책임 배분 문제를 논의하기 위한 새로운 협상이 필요하며, 그러한 협상의 결과가 군사 로봇을 전쟁에 사용하기 위한 필요조건이다. 이에 도달하기 위해 스패로가 제시한 논변의 전제들을 거부할 필요는 없다. 사실 이 절의 논의는 그의 전제들을 모두 수용하고 있다. 책임 문제가 우선적으로 해결되어야 하고, 이 문제의 해결 없이 자율적 군사 로봇은 전쟁에서 사용되지 않아야 한다. 그리고 로봇 자신이나 지휘관, 설계자 가운데 어느 누구도 로봇이 가져올 결과에 대해 온전히 책임을 질 수 없다는 점도 분명하다. 문제는 한 행위자가 어떤 결과에 관해 전적인 책임을 가진다는 숨겨진 가정이다. 이를 부정함으로써, 그리고

11. 예측하거나 통제할 수 없는 나쁜 결과가 발생한 경우이므로, 의도적인 결과로 보기는 어렵겠다. 그러나 어떤 것이 과실이기 위해 반드시 의도적일 필요는 없다.
12. 예컨대, 우리는 500년에 한번 찾아올지도 모르는 태풍을 대비해야할 책임을 건축물의 설계자와 시공자에게 부여하지는 않을 것이지만, 5년에 한번 찾아올지도 모르는 태풍에는 (그것이 언제 찾아올지 예측하거나 통제할 수는 없더라도) 대비해야한다고 요구할 수도 있다. 현대 기술사회에서 그러한 의무의 범위를 결정하기 위한 과정은 지속적으로 요구된다.

분산된 책임의 가능성을 긍정함으로써, 우리 사회는 새로운 과제를 직면하게 되었다. 자율적 군사 로봇이 일으키는 나쁜 결과에 대해서 누가, 어떤 방식으로, 얼마만큼의 책임을 질 것인가?

참고문헌

Arkin, Ronald (2009), *Governing Lethal Behaviour in Autonomous Robots.* Boca Raton, FL: CRC Press.

Bijker, Wiebe E., Thomas P. Hughes, and Trevor Pinch (1987), *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology.* Cambridge, Mass.: MIT Press.

Haas, Benhamin (2018), "'Killer robots': AI experts call for boycott over lab at South Korea university." *The Guardian* 2018.4.5. (https://www.theguardian.com/technology/2018/apr/05/killer-robots-south-korea-university-boycott-artifical-intelligence-hanwha)

Jun, Ji-hye (2018), "Hanwha, KAIST to develop AI weapons." *Korea Times*, 2018.2.25. (https://www.koreatimes.co.kr/www/tech/2018/02/133_244641.html)

Orend, Brian (2008), "War." *Stanford Encyclopedia of Philosophy.* ed. Edward N. Zalta. <http://plato.stanford.edu/archives/fall2008/entries/war/>

Lin, Patrick, Geogey Bekey, and Keith Abney (2009), "Robots in War: Issues of Risk and Ethics". in *Ethics and Robotics,* Eds. R. Capuro and M. Nagenborg. AKA Verlag Heidelberg.

Powers, T. (2006), "Prospects for a Kantian machine." *IEEE Intelligent Systems* 4(21), 46-51.

Schulzke, Marcus (2011), "Robots as weapons in just wars." *Philosophy and Technology* 24(3): 293-306.

Sharkey, Noel (2012), "Killing made easy: from joystick to politics", in Lin, Abney, and Bekey eds. *Robot Ethics: The Ethical and Social Implication of Robotics*. The MIT Press.

Simpson, Thomas W. and Vincent C. Mueller (2016), "Just War and Robot's Killings", *The Philosophical Quarterly* 66(263): 302-322.

Sparrow, Robert (2002), "The march of the robot dogs." *Ethics and Information Technology* 4: 305–318.

_____ (2007), "Killer robots." *Journal of Applied Philosophy*, 24(1), 62-77.

Sullins, John (2010), "RoboWarfare: Can Robots be More Ethical than Humans on the Battlefield?", *Ethics and Information Technology* 12/3: 263–75.

국문초록

카이스트 보이콧 사태는 군사 로봇의 윤리적 문제가 인문학이 다루어야 할 매우 긴급한 문제임을 보여주었다. “킬러 로봇”을 금지해야한다는 주장에 반대하는 사람은 없겠지만, 그러한 금지 주장이 자율적 무기체계의 개발과 사용에 대해 구체적으로 어떤 제약을 가하는지는 분명치 않다. 전쟁터의 모든 군인이 살인자가 아니듯, 모든 군사 로봇을 살인자 로봇으로 규정하기는 어렵다. 이 논문은 먼저 카이스트 보이콧 사태를 소개한다. 모든 군사 로봇이 똑같은 의미의 킬러가 아님을 보이기 위해 다양한 종류의 군사 로봇들을 분류한 후, 정당한 전쟁 이론을 통해 군사 로봇의 사용이 가질 수 있는 잠재적인 윤리적 문제들이 무엇인지 해명한다. 특히, 군사 로봇이 전쟁에서 사용하기 위해서는 식별의 원리와 비례성의 원리를 따르는 것이 중요하며, 현 단계에서 그러한 원리들을 충실히 만족하도록 로봇을 제작하기는 쉽지 않음을 밝힌다. 끝으로, 군사 로봇의 잘못된 수행에 관한 책임 문제를 다루면서, 관련된 집단들이 자율적 무기체계의 사용과 관련하여 책임을 어떻게 분배할 것인가를 놓고 새로운 협상이 필요함을 주장한다.

키워드: 킬러 로봇, 카이스트 보이콧 사건, 정당한 전쟁, 치명적 자율무기체계, 책임의 문제

Abstract

Beyond "Stop Killer Robots": The Ethical Issues of Autonomous Weapon Systems

Hyundeuk Cheon (Seoul National University)

KAIST Boycott Affair has made us take the ethics related to military robots seriously. No one would deny that 'killer robots' should be banned. However, it is not crystal clear what the slogan "stop killer robots" implies for the development of autonomous weapon systems and their uses in warfare. Just as all soldiers are not merely killers, so are robots on the battlefield. Many ethical discussion on 'killer robots' hangs on the definition of wrongful killings in warfare since there might be ethically permissible acts of killing. In this article, I begin with an attempt to classify various kinds of military robots. Then, I diagnose the core ethical problems of using military robots by adopting just war theory. It is claimed that the operation of military robots can be justified only when it meets the principle of discrimination and the principle of proportionality. Finally, I argue that a new negotiation is needed on how to distribute the responsibilities to relevant groups before the uses of autonomous weapons.

Keywords: killer robots, KAIST Boycott Affair, Just War, lethal autonomous weapon system, responsibility

Received :	19	January	2019
Reviewed :	20	February	2019
Accepted :	20	February	2019