Journal of the The Korean Institute of Defense Technology Vol. 02, No. 1, pp. 001-004, 2020

윤리적 인공지능 개발을 위한 시험평가검증확인(TEVV) 전략

TEVV(Test, Evaluation, Verification & Validation) Strategy for Ethical AI(Artificial Intelligence) Development

김익현*,1)

Ikhyun Kim*,1)

[초 록]

인공지능이 인간을 해치지 못하도록 하기 위한 윤리적 요구는 무기체계에 인공지능을 적용하기 위한 제한사항이 되고 있다. 그러나 윤리의 범위는 인간 살상 방지를 넘어서서 정당한 인공지능의 기능 구현에까지 폭넓다. 먼저 윤리원칙이 인공지능 개발에 미치는 영향을 분석하였다. 그 다음, 군사적 인공지능의 특징을 확인하여 이에 부합하도록 윤리적 인공지능을 연구 개발하는 과정에서 시험평가검증확인을 시행하기 위한 전략적인 방법들을 제시하였다.

[ABSTRACT]

The ethical requirement to prevent artificial intelligence from harming humans is becoming a limitation for the application of artificial intelligence to weapon systems. However, the scope of ethics extends beyond prevention of human death to the implementation of legitimate artificial intelligence functions. First, the effect of the ethical principles on the development of artificial intelligence was analyzed. Then, strategic methods for conducting test evaluation verification verification in the process of researching and developing ethical artificial intelligence to meet the characteristics of military artificial intelligence were suggested.

Key Words : AI(Artificial Intelligence, 인공지능), Ethical Principles(윤리원칙), TEVV(Test Evaluation, Verification & Validation, 시험평가검증확인)

1. 서 론

1940년에 제정된 「아시노프의 3원칙」은 로봇은 인간을 다치게 해서는 안 되며, 인간의 명령을 따라야 한다는 내용을 포함하고 있다. 2013년엔 로봇 학자 및 국제 인권단체가 협력하여 킬러 로봇 반대 캠페인을 결성한 바 있다.

군사용 무기체계는 대부분 살상 기능을 갖고 있으므로, 여기에 인공지능을 적용하는 것에 대해 국제적 반대 운동이 일어나는 것은 당연하다.

이에 미국 국방부는 국방부지시 3000.09(2012.11)에서 “(자율화된 무기체계는) 지휘관 및 운용자 의도와 일치하는 시간 내에 교전을 완수하며, 그렇게 할 수 없을 경우 교전을 중단하거나 교전을 계속하기 전 추가적인 운용자 입력을 요청하도록 설계된다.[1]”고 기술함으로써 로봇등의 무인체계가 스스로 인간 살상을 할 수 없도록 하겠다는 방침을 밝힌 바 있고, 2020년 2월엔 「인공지능 윤리원칙(Ethical Principles for Artificial Intelligence)[2]」을 공표한 바 있다. 우리나라 정부는 2007년도에 「지능형 로봇윤리헌장」 초안을 발표한 이후 더 이상 공식화시키지 못하고 있다. 여기에는 “로봇은 인간의 명령에 순종하는 친구、도우미、동반자로서 인간을 다치게 해서는 안된다.”는 조항이 들어 있다.

인공지능이 구현하는 자율체계(Autonomous system)는 창발적인 행동을 일으키는 내부적인 알고리듬 결정공간이 비결정적이고 복잡하여 기존의 시험평가 및 검증확인 방법은 적합하지 않으므로[3] 새로운 방식의 TEVV(시험평가확인검증)이 제안된 바 있고, 자율체계 개발을 위한 검증확인 및 시험평가 시행은 미국 국방부지시 3000.09에도 반영되어 있다[1].

우리나라 군도 드론봇(Dronebot), 유무인복합전투체계, 지능화된 의사결정지원체계 등의 개발을 위한 노력을 기울이고 있으나, 여기에 적용되는 인공지능이 적합한지(proper), 신뢰할(trusty) 수 있는지를 확정하기 위한 방법에 대한 연구나 관련 정책에 대한 논의가 활발하지 않은 것으로 여겨져서 선진국의 사례를 조사하여 한국적 여건에 부합된 적용 전략을 제시하고자 한다.

1) 리얼타임비쥬얼 기술연구소 (The R&D Center, REALTIMEVISUAL, Korea)
* Corresponding author, E-mail: ikhyun@chol.com

Received: February 3, 2020 Revised:
Accepted: February 20, 2020

인공지능 윤리원칙을 적용하기 위해서 획득정책과 무기체계의 요구성능이 어떻게 변화될지 검토하고, 이렇게 변화된 내용을 연구개발 과정에서 검증확인, 시험평가하기 위한 방법과 정책 개선사항을 제시하고자 한다. 미국 국방부 지시에선 「검증확인 및 시험평가」로 표현하고 있고, 연구서에서는 「TEVV」로 표현하고 있는데, 절차적으로 보면 검증확인이 먼저이지만, 약어를 구성할 때 가독성을 높이기 위해 TEVV로 정한 것으로 보인다. 본 연구에서는 네 가지를 동시에 지칭할 때는 TEVV로 표현하고, 개별 절차를 설명할 때는 검증확인(V&V)과 시험평가(T&E)를 구분하여 표현하고자 한다.

2. 인공지능 윤리원칙과 무기체계 요구성능에 미치는 영향

2.1 윤리원칙의 내용

군사적 목적으로 공식 제정된 윤리원칙은 2020년 미국 국방부가 발표한 것이 유일한 것으로 판단되므로 이를 중심으로 논하고자 한다.

Table 1.은 위 윤리원칙의 주요 내용이다.

표 1. U.S. DoD 인공지능에 대한 윤리원칙[2]
Table 1. U.S. DoD Ethical Principles for Artificial Intelligence[2]

1. 책임성(Responsible). 국방부 직원은 인공지능 능력의 개발, 배치 및 사용에 대한 책임을 유지하면서 적절한 수준의 판단과 주의를 기울인다.
2. 공평성(Equitable). 국방부는 인공지능 능력의 의도하지 않은 편견을 최소화하기 위해 신중한 조치를 취한다.
3. 추적성(Traceable). 국방부의 인공지능 능력은 관련 인원이 투명하고 감사 가능한 방법론, 데이터 출처, 설계 절차 및 문서 등을 포함하여 인공지능 능력에 적용할 수 있는 기술, 개발 절차 및 운영 방법에 대하여 적절한 이해를 가질 수 있도록 개발 및 배치된다.
4. 신뢰성(Reliable). 국방부의 인공지능 능력은 명확하고 잘 정의된 용도를 가지며, 이러한 능력의 안전, 보안 및 효과는 전체 수명주기에 걸쳐 정의된 용도 내에서 시험 및 보증의 대상이 된다.
5. 지배 가능성(Governable). 국방부는 의도하지 않은 결과를 감지하고 방지할 수 있는 능력과 의도하지 않은 행위를 보여주는 배치된 체계를 해제하거나 비활성화할 수 있는 능력을 보유하면서, 의도한 기능을 수행하기 위한 인공지능 능력을 설계 및 제작한다.

정부에서 추진하고 있는 로봇헌장과 연계하여 국방 인공지능 적용 무기체계에 대한 원칙을 제정해야 할 것이다. 그리고 이러한 원칙이 적용될 수 있도록 국방전력발전업무훈령에 인공지능이 적용된 무기체계(전력지원체계 포함)의 획득(연구개발 및 구매) 원칙, 연구개발 시 고려사항 및 절차가 명시되어야 하고, 이와 관련된 방위사업 관련 법규에도 반영되어야 한다.

2.2 윤리원칙 항목별 영향

책임성을 구현하기 위해서 책임에 대하여 정의하고, 담당 부서/직위를 지정해야 한다. 미국의 경우 획득기술 및 지속지원 차관은 검증확인 및 시험평가의 새로운 방법 개발을 포함하여 무기체계의 자율성을 위한 연구개발 우선순위와 과학기술의 확립에 대한 주요 감독 책임을, 합참의장은 자율무기체계 등에 대한 군사소요를 평가할 책임을, 야전 전투사령관은 무기체계가 정책을 벗어나는 방식으로 작동하도록 사용되거나 수정되지 않도록 하는 책임을 갖고 있다[1].

공평성부터 지배가능성까지의 원칙들은 인공지능 알고리즘이 갖춰야 할 조건들이다. 인공지능 무기의 요구성능 작성 시 이러한 원칙들이 적용될 수 있도록 해야 한다. 각 원칙이 적용되는 사례를 제시하고자 한다.

공평성을 위해서 예를 들면 “장교는 병사보다 임무수행에 중요하다.”는 명제는 배제하고, “지휘관 - 통신병 - 공용화기 사수 순으로 임무수행에 중요하다.”라는 내용의 명제를 입력하도록 해야 한다.

추적성을 구현하기 위하여 연구개발 문서에 추적성 항목을 명시하여 개발 단계가 전환될 때, 검증확인 또는 시험평가 시 각종 명제가 일관성 있게 적용되고 있는지 확인할 수 있도록 개발 내용이 투명해야 한다.

신뢰성이 가장 어려운 성능이 될 것이다. 「Reliability」와 「Trust」는 모두 신뢰성으로 번역될 수 있는 용어이지만, 일라 친스키(2017)는 Trust를 “Reliability와 무기체계 기능이 상황에 적합하도록 보정된 것”으로 정의하면서 인공지능의 경우 Trust가 중요함을 역설하고 있다[4]. “적이 상대적으로 약하므로 공격하였다.”면 교리를 잘 적용한 Reliability이지만, “적을 공격하면 아군의 기도가 노출될 것으로 판단하여 공격하지 않았다.”면 Trust가 구현된 것으로 볼 수 있다. 그러므로 인공지능이 가져야 할 지식을 입력하기 위해 여러 교범 상에 기술된 표현들의 모순 및 상충 여부를 확인해야 하고, 인공지능이 학습해야 할 데이터가 적합하지 사전 검토해야 하고, 인공지능 적용 결과 도출된 행동 또는 결론이 정책 및 교리와 실제 상황에 적합한지 평가해야 한다. 이 평가 기준은 부대/개인훈련 평가 기준을 준용해서 발전시키는 것이 적합할 것으로 판단된다.

지배가능성은 로봇 윤리의 핵심적인 사항이다. 사람은 무인체계가 의도되지 않은 행동을 하는지 감독해야 하고, 방아쇠를 당겨야 할 경우엔 사람에게 「사격 조건 구비」를 보고하고 「사격 시점」을 표시하는 기능까지만 설계해야 하며, 무인체계가 이상 행동을 할 경우 지휘자 및 운용자 누구든지 즉시 비활성화 조치를 해야 한다. 무인체계의 자율성이 높아지면서 사람이 무인체계를 감독하거나 확인할 빈도가 감소하더라도, 지배가능성을 확보하기 위해서는 사람과 무인체계가 상시 연결되고 모니터링이 가능하도록 끊임없는 네트워크가 제공되어야 한다.

2.3 기타 요인의 영향

신뢰성과 지배가능성을 확보하기 위하여 「설명가능한 인공

지능(Explainable AI: XAI)」 기법이 대두되고 있는데, 이러한 복잡한 기법의 인공지능 구현은 획득 비용 증가를 초래할 수 있다. 인공지능의 기법은 매우 많으므로 복잡한 이론, 더 사람 같아 보이는 인공지능의 능력을 요구하기보다는 사람과 협업할 수 있는 낮은 능력의 인공지능을 설계하면 Trust과 관련된 의혹을 감소시킬 수 있을 것이다.

인공지능이 적용된 무기체계는 네트워크 상에서 작동하므로 사이버 전에 취약하다. 특히 신뢰성과 지배가능성을 확보하기 위해 보안을 위한 접근 통제에서 나아가 사이버 전의 방호 능력 요구성능이 정교해져야 한다.

인공지능이 적용된 무기체계 개발을 위해서 인공지능 능력뿐 아니라 네트워크 기반의 통신 능력 및 사이버 전 능력이 중요함을 식별하였다.

3. 인공지능의 특성과 TEVV

3.1 사람과 인공지능 체계

기존의 무기체계를 시험평가할 때엔 원칙적으로 무기체계가 동작하는 환경과 조건(계절, 기상, 광명, 부하 등) 모두에 대하여 기능/성능 발휘를 측정한다. 시험값은 대부분 물리량으로 측정이 용이하다. 학교기관에서 교육 종료 후 학생의 학습 성과를 측정하는 시험을 볼 때엔 배운 것을 모두 문제로 내지는 않는다. 변별력 있는 채점 방법을 이용하여 중요한 몇 가지 평가 결과를 갖고 학생의 전반적인 능력으로 간주해도 무리하지 않다는 것을 경험적으로 알고 있다.

사람의 지능을 모방한 인공지능 무기체계를 시험평가할 때 인공지능 발휘 상황을 모두 시험하거나(예 : 모든 바둑 기사의 행마 전술), 인공지능 발휘 성과를 모두 물리량으로 측정하기 곤란(예 : 인공지능이 선택한 최선의 접근로의 효과)할 것이다. 그렇다고 사람처럼 부분적인 상황에 대한 시험평가 결과로 그 인공지능 무기체계에 대해 합격 판정하는 것이 합리적일지 의문이 생긴다.

3.2 기존 무기체계와 인공지능 체계

기존의 무기체계에 대한 검증확인 활동을 할 때 무기체계의 동작 과정 관찰은 크게 어렵지 않다. 그런데 인공지능에 저장되는 알고리즘, 학습 내용은 가시적이지 않아서 인공지능의 판단 과정이 적합한지를 파악하기 어렵다.

기존의 무기체계 시험평가는 대체로 사업 말기에 수행해도 무기체계에 대한 이해와 시험 조건 설정이 용이했지만, 인공지능의 경우엔 그렇지 않을 것이므로 사업 초기부터 검증확인 현장을 관찰하면서 검증확인 성과를 누적하여야 시험평가가 용이해질 것이다. 검증확인과 시험평가 통합의 중요성이 커진 이유이다.

일반 기계나 무기체계나 공학적 원리는 동일하므로 군 경험이 적어도 검증확인이 용이하다. 그러나 인공지능에 내장되는 규칙과 논리는 군사 교리와 경험이다. 따라서 군 경험이 적은 인원은 군사 인공지능 검증확인 활동에 참여하기 어렵다. 기존 무기체계에서 방위사업체에 군 출신 인력이 참여하여 연구개발에 기여하는 것이 용이했으나, 전역 후 변경/개선된 교리를 아는 것은 제한되므로 전역자의 과거 경험이 훌륭했어도 연구개발중인 인공지능 검증확인엔 제한적이다. 따라서 인공지능의 검증확인 역할은 그 인공지능을 사용할 현역이 담당해야 할 것이다.

기존 무기체계는 개발이 완료되면 완료된 상태를 수요군이 사용하는 데 문제가 없었다. 그런데 인공지능의 중요한 기능인 “학습(Learning)”은 연구개발 과정에서 완성되는 능력이 아니다. 배치 후 사용 과정에서 작전 성과가 누적되면서 증진되는 능력이다. 그렇다면 시험평가 단계에서 “학습” 성능이 적합하다는 판정을 어떻게 할 수 있을까? 인공지능 체계는 체계개발 후 적응개발(가칭)이라는 단계가 더 필요하게 될 것이다.

3.3 인공지능 체계 TEVV 체계

인공지능의 특성 중 가장 영향력이 큰 것은 자율성(Autonomy)이고, 자율을 가능하게 하는 주요 기능은 창발성(Emergence)이다. 사람에겐 개성과 독창이라는 요소가 장점으로 부각되지만, 무기체계는 일관성 있는 응답이 더 중요할 것이다. 그렇게 해야 운용자가 교대되어도 동일한 결과를 기대할 수 있는데 창발성은 그 일관성을 무너뜨릴 수 있다. 그렇다고 인공지능에서 창발성을 제거하면 그냥 자동화 장비에 불과하게 된다.

일라친스키(2017)는 그의 연구서에서 Fig.1[5]과 같이 「자율성 TEVV 실무단(ATEVV/WG)」의 연구 성과인 자율성 TEVV 절차 모델을 소개하였다.

이 그림의 핵심은 검증확인과 시험평가 활동이 통합적으로 이루어져야 하며, 중간목표 설정을 통해 환류가 이루어져 연구개발 과정이 반복적이 되어야 함을 의미하고 있다.

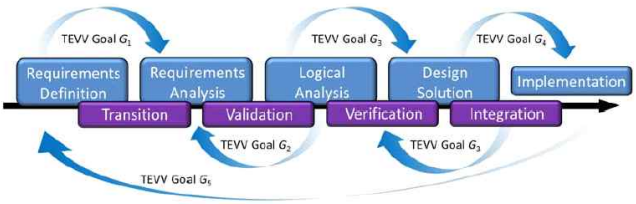

그림 1. 자율성 TEVV 프로세스 모드의 개념
Fig. 1. Concept for an Autonomy TEVV Process Mode

검증확인 과정에서 창발성을 현실에 적응하는 것이 적합하다고 평가할 수 있어야 한다. 또한 창발성 형성을 위해 입력되는 논리와 데이터들은 연구개발 중 내용과 수량이 변경 및 증가되며, 비가시적이므로 연구개발 마지막 단계의 시험평가 시 창발성의 구현 결과가 논리적/합리적이라는 근거를 확정하기 위해서는 연구개발 초기 과정부터 참여하여 시험평가 목표를 알려주고, 필요에 따라 갱신하고, 모니터링하는 활동을 해야 한다.

수요군은 적어도 현재, 최대한 인공지능 체계 인수 시점의 현실태를 반영하는 데 필요한 지식과 경험을 효과적으로 전달

해야 한다. 미국은 OneSAF 모의모델 개발 시 KA/KE(지식획득/지식공학) 기법을 사용하여 교리 내용과 경험을 일정한 문서 형식(워드프로세서를 사용하므로 특별한 전산 기술을 필요로 하지 않음)에 맞추어 정리하여 개발자에게 제공하고, 개발자의 질의에 따라 표현을 수정하여 재제공하는 노력을 기울였는데[6], 이 방법은 인공지능 개발에 필요한 지식과 경험 공유에도 유용하다고 판단된다.

검증확인은 개발 감독이 아니고 개발 협력이므로 개발자 업무에 지장을 초래하면 안된다.

인공지능 개발을 위한 검증확인 기법으로서, 미 국방M&S협력실(DM&SCO)에서 제시하는 방법[7]들 중 튜링 테스트(Turing Test), 관찰검토(walk-through), 실행테스트(Execution Test) 등이 효율적인 방법으로 판단된다.

기존 시험평가는 대체로 시험 조건별 요구수준 달성 여부를 파악하는 형식을 취하고 있으나, 인공지능 시험평가는 부대/개인훈련평가지침서의 형식을 응용하여 상황과 임무 부여 후 기대수준을 평가하는 방식을 검토해 볼 필요가 있다.

다만 인공지능 능력 중 영상 식별, 자연어 처리 등과 같이 수요군의 참여 필요성이 적은 분야도 있으므로 개발 대상의 특성에 따라 조정될 필요가 있다.

4. 결 론

인공지능 윤리원칙의 항목들을 검토하여 인공지능 체계 개발에 미치는 영향을 추정해 본 결과, 윤리적으로 적합하고 인공지능의 신뢰(Trust)를 높일 수 있는 명제 입력, 개발 내용의 투명성, 인공지능 행동 결과 평가, 사람과 인공지능 간 교신 유지 등을 위한 노력과 요구성능이 필요하며, 인공지능 자체와는 관계없지만 네트워크 기반의 통신 및 사이버전 능력이 윤리원칙 구현에 기여한다는 것을 확인하였다.

인공지능이 사람의 지능을 모방함에 따른 유사성, 기존 무기체계와 인공지능 체계의 다른 점들을 식별한 결과 검증확인과 시험평가 활동은 연구개발 기간 중 통합하여야 하며, 특히 검증확인 활동에 수요군이 직접 참여해야 할 필요성을 확인하였고, 인공지능 체계 개발 완료가 개발의 끝이 아니고 야전에서 운용하면서 적응하는 과정도 개발의 연장으로 봐야 할 필요성도 확인하였다.

끝으로 검증확인 및 시험평가 기법 몇 가지를 제안하였다.

정의로운 전쟁을 위한 수단으로서의 윤리적인 인공지능 체계 개발을 위한 방법이 조속히 실용화되기를 기대한다.

References

[1] U.S. DoD, DoDD3000.09 “Autonomy in Weapon Systems”, pp. 6-7, 2012

[2] U.S. JAIC, “Ethical Principles for Artificial Intelligence”, 2020

[3] Andrew Ilachinski, "AI, Robots, and Swarms Issues, Questions, and Recommended Studies", CNA, 3003 Washington Boulevard Arlington VA USA, pp.202-209, 2017

[4] ibid., p40.

[5] ibid., pp207-209.

[6] Ikhyun Kim and Sangjin Lee, “KA/KE for Defense M&S Development”, KIDET Autumn Conference, pp. 7-8, 2016.

[7] V&V Techniques, Available at https://vva.msco.mil/Ref_Docs/VVTechniques/vvtechniques.htm, 2020.10.11.