

September 2026  
No. 409

# INSS

## 전략보고

### 전략적 비핵무기체계의 AI 전환과 위기안정성 확보 방안

윤정현  
yjh5791@inss.re.kr

- I. 문제 제기
- II. 전략적 비핵무기 체계의 AI 전환과 위기안정성
- III. SNNW 체계의 AI 전환에 따른 위험관리 현황과 쟁점
- IV. AI 기반 SNNW의 위기안정성 확보 방안

# 전략적 비핵무기체계의 AI 전환과 위기안정성 확보 방안

## I. 문제 제기

---

## II. 전략적 비핵무기 체계의 AI 전환과 위기안정성

---

1. 전략적 비핵무기(SNNW) 체계의 특징과 범위
2. 전략적 비핵무기 체계의 AI 전환 효과
3. 전략적 비핵무기체계의 AI 전환과 위기안정성 도전

## III. SNNW 체계의 AI 전환에 따른 위험관리 현황과 쟁점

---

1. SNNW 체계의 AI 전환에 대한 주요 위험관리 접근
2. 기존 군사 AI 위험관리의 한계와 위기안정성 과제

## IV. AI 기반 SNNW의 위기안정성 확보 방안

---

1. 위기안정성 위험 요인의 구조적 특성과 한반도 환경
2. 확산 요인 통제를 위한 설계 원칙
3. 위기안정성 확보의 이행 과제

# 전략적 비핵무기체계의 AI 전환과 위기안정성 확보 방안

저자 | 윤정현

## 국문 초록

최근 '전략적 비핵무기(Strategic Non-Nuclear Weapons, SNNW)'의 AI 전환에 따라 억제와 위기관리의 안보 구조가 근본적으로 변화하고 있다. 특히, 한국은 비대칭 전략구조와 시간·공간 압축이라는 복합적 조건에 노출되어 위기 고조 통제가 어려운 구조적 제약을 안고 있다. 이에 본 연구는 SNNW체계의 AI 전환이 초래하는 '위기안정성(crisis stability)' 측면의 도전 요인을 분석하고, 한반도에서의 위기관리·평화공존의 제도화를 위한 이행 과제를 제시하였다. 구체적으로는 다음과 같은 대응방향이 요구된다. 첫째, AI 권고와 인간 승인 사이의 최소 결심시간 보장과 운용 신호 규율을 통한 교차검증 절차를 마련해야 한다. 둘째, '실패-안전(fail-safe)' 설계에 기반하여 자동화 위협에 대응하는 운용 원칙을 수립해야 한다. 셋째, 한미, 남북 간 위험소통체계 수립을 위한 단계별 추진이 필요하다. △한미 연합 차원의 결심시간 기준 및 데이터 검증 거버넌스 정비, △남북 AI 관련 오경보 확인·사고 통보 절차와 관련한 '9.19 체제 2.0'의 추진, △SNNW 획득·검증 절차 내 전략적 효과 평가 도입, △책임있는 군사 AI 규범과 핵 인간통제 규범의 국제화 시도가 대표적이다.

주제어: 인공지능(AI), 전략적 비핵무기(SNNW), 위기안정성, AI의 인간 통제, 책임있는 군사 AI

## I 문제 제기

- 최근 인공지능(AI)의 군사적 활용이 전쟁 수행의 전 과정으로 확대되면서, 전략적 비핵무기체계의 AI 전환이 새로운 안보 쟁점으로 부상
  - 전략적 비핵무기(Strategic Non-Nuclear Weapons, SNNW) 체계는 상대의 핵전력·지휘통제 체계 등 핵심 전략자산의 운용에 중대한 영향을 미침으로써 전략적 효과를 창출할 수 있는 비핵 능력을 의미
  - 최근 AI가 기존 무기체계와 지원체계에 결합되면서 SNNW 체계는 단순한 무기의 성능 향상을 넘어 탐지-분석-표적화-결심-타격으로 이어지는 의사결정 과정 자체를 변화시키고 있음
- 이러한 변화는 전략적 비핵무기체계의 군사적 효과를 증진시키는 반면, 기존의 억제와 위기관리 환경에 새로운 불확실성을 야기
  - AI의 군사적 도입이 초래한 탐지·표적화 능력의 향상과 의사결정 시간의 압축은 위기 상황에서 상대의 취약성 인식과 선제행동 유인, 비의도적 확전에 광범위한 영향을 미칠 가능성
- 국방 분야 AI 전환은 주요 SNNW의 성능 뿐만 아니라 인간의 의사결정 방식까지 변화시킬 수 있다는 점에서 보다 확장된 접근이 필요
  - 기존 논의는 자율무기체계의 인간 통제나 개별 무기체계의 성능 향상에 집중되어 있어 전략적 비핵무기체계의 AI 전환이 '위기안정성(crisis stability)'에 미치는 영향에 대한 논의는 상대적으로 부족
- 본 연구는 SNNW체계의 AI 전환이 안보 환경에 미치는 도전적 영향을 분석하고, 한국의 위기안정성 확보를 위한 대응방향 모색
  - 이를 통해 AI 기반 SNNW 운용 환경의 불확실성 제고 및 남북 간 위기관리·평화공존 제도화의 실효성 확보를 위한 토대 제공

## II 전략적 비핵무기 체계의 AI 전환과 위기안정성

### 1. 전략적 비핵무기(SNNW) 체계의 특징과 범위

가. ‘전략적 비핵무기(Strategic Non Nuclear Weapons)’의 개념

- SNNW는 핵무기를 사용하지 않고도 상대의 핵전력과 지휘통제체계, 주요 전략가치 시설 등에 중대한 영향을 미침으로써 전략적 효과를 창출할 수 있는 비핵 무기를 의미<sup>1</sup>

- SNNW는 특정 무기군으로 고정되어 있지 않으며, 장거리 정밀타격, 미사일방어, 대우주, 사이버 능력 등 상대의 전략자산과 핵심 기반에 영향을 미칠 수 있는 다양한 비핵 자산들을 포함<sup>2</sup>

※ 최근 신형기술과 비핵 군사기술의 상호 결합이 전략적 억제관계에 영향을 미칠 잠재력이 부각되면서 SNNW의 안보적 중요성이 재조명되기 시작<sup>3</sup>

나. 전략적 비핵무기(SNNW) 체계의 구성과 전략적 의미

- 반면, ‘전략적 비핵무기체계’는 개별 SNNW와 함께 전략적 운용을 가능하게 하는 감시·정찰(ISR), 정보 융합, 표적화, 결심지원, 지휘통제 등 지원체계가 상호 연계된 전체 구조를 의미

- SNNW의 전략적 효과는 개별 무기의 파괴력이나 자체 성능뿐만 아니라, 탐지·분석·표적화·지휘통제·타격 기능을 유기적으로 연계하는 체계적 운용 능력에 의해 결정

- 특히, 상호 연계된 비핵 능력을 통해 상대 핵전력과 핵심 전략자산의 생존성에 영향을 미치고, 기존 억제 구조의 변화를 초래할 수 있다는 점에서 SNNW 체계의 전략적 중요성 증대

- SNNW가 창출하는 전략적 효과는 발현 경로에 따라 다음의 세 가지 부문의 층위로 분류할 수 있음

- 물리적 효과(파괴·무력화): 직접적 타격으로 표적의 군사적 가치를 제거하는 전통적 작용 경로<sup>4</sup>

- 지휘통제 효과(C2·NC3·전략자산 마비): 상대 운용 여건을 구조적으로 제약하여 의사결정과 작전 수행 자체를 저해하는 작용 경로<sup>5</sup>

1 Andrew Futter and Benjamin Zala, “Strategic Non-Nuclear Weapons and the Onset of a Third Nuclear Age” European Journal of International Security (2021), 6, pp. 7.

2 Fabian Hoffmann, “Strategic Non-nuclear Weapons and Strategic Stability: Promoting Trust through Technical Understanding” (November 2021), p. 5.

3 Hoffmann(2021), p. 9.

4 Fabian Hoffmann, “The Strategic-Level Effects of Long-Range Strike Weapons: A Framework for Analysis,” Journal of Strategic Studies, Vol. 47, No. 6-7 (2024).

5 James Johnson, AI and the Bomb: Nuclear Strategy and Risk in the Digital Age, (Oxford: Oxford University Press, 2023).

- 심리적 · 인지적 효과(use-it-or-lose-it 압박 · 자동화 편향 · 결심지연):상대의 핵 · 전략자산 생존성 인식과 결심 압박을 변화시키는 작용 경로<sup>6</sup>

## 2. 전략적 비핵무기 체계의 AI 전환 효과

### 가. AI 결합에 따른 SNNW 운용체계의 변화

- AI의 군사적 도입·확대는 대규모 전장정보를 처리하고 이를 실제 작전 수행에 연결하는 능력을 고도화하는 방향으로 전개 중
  - 위성 · 레이더 · 무인체계 등 다양한 센서에서 수집되는 정보를 실시간으로 융합 · 분석하고, 다수의 표적 후보를 식별 · 추적하여 우선순위를 제시하는 과정에서 AI의 활용이 대폭 증가<sup>7</sup>
    - ※ 2023년 가자전쟁에서 이스라엘군 정보부는 하마스 · 팔레스타인 지하드(PIJ) 무장세력의 잠재적 구성원을 대규모로 식별 · 분류하여 표적화하는 과정에서 AI 기반 시스템 'Lavender'를 활용하여 공습 작전을 수행<sup>8</sup>
    - ※ 2026년 이란전에서 미군은 팔란티어의 AI 기반 'Maven Smart System (MSS)'을 활용해 전장정보 분석 · 표적화를 지원하고 탐지-결심-타격 사이클 단축에 활용<sup>9</sup>
  - AI가 서로 분절되어 있던 정보수집 · 분석 · 표적화 및 작전계획 간 연계성을 강화하면서, SNNW 체계의 전략적 효과를 좌우하는 주요 변수로 부상<sup>10</sup>

### 나. AI 전환에 따른 인간 의사결정 구조의 변화

- 이 같은 변화는 기존 SNNW 체계가 전제해왔던 인간의 역할의 분담과 각 기능 간 연결 방식의 전환을 의미
  - 과거 인간이 직접 수행하고 판단하던 정보의 분류 · 비교 분석 및 표적 우선순위 설정 기능 등을 AI가 대체하게 되며 인간은 AI가 제시한 결과 내에서 검토 · 판단하는 역할에 집중<sup>11</sup>

6 Michael C. Horowitz, "Artificial Intelligence and the Future of Strategic Stability," Texas National Security Review, Vol. 9, No. 2 (2026).

7 U.S. Department of Defense, "AI Aided Sensor Data Fusion," Small Business Innovation Research (SBIR) Phase II, (2025), <https://www.sbir.gov/awards/218485?utm> (검색일: 2026. 7. 1).

8 Yuval Abraham, "'Lavender': The AI machine directing Israel's bombing spree in Gaza" (April 3, 2024), <https://www.972mag.com/lavender-ai-israeli-army-gaza/> (검색일: 2026. 4. 23).

9 Reuters, "Pentagon to adopt Palantir AI as core US military system, memo says" (March 20, 2026), <https://www.reuters.com/technology/pentagon-adopt-palantir-ai-as-core-us-military-system-memo-says-2026-03-20/> (검색일: 2026. 9. 9)

10 John E. Wissler III, "Artificial Intelligence and Cyber Warfare: Evolution or Revolution?," The Cyber Defense Review, Vol. 8, No. 1 (Spring 2023), pp. 53-64.

11 James Johnson, AI and the Bomb: Nuclear Strategy and Risk in the Digital Age, (Oxford: Oxford University Press, 2023), pp. 1-36.

- 결과적으로 SNNW 체계의 AI 전환은 정보의 획득에서 타격에 이르는 일련의 작전 수행과 인간의 의사결정 구조를 재편하는 방향으로 진화

### 3. 전략적 비핵무기체계의 AI 전환과 위기안정성 도전

#### 가. 전략적 안정성과 위기안정성의 개념적 구분

- ‘위기안정성(crisis stability)’은 위기 상황에서 선제공격 가능성이 억제되어 있는 상태를 의미하며, 상대보다 먼저 무력을 사용해야 할 유인이 낮은 단기적 의사결정의 안정성에 방점을 둔 개념<sup>12</sup>
  - 이는 억제 요소들이 안정적으로 유지되어 핵무기 사용의 유인이 최소화되고, 군비경쟁의 동기가 억제된 상태를 의미하는 ‘전략적 안정성(strategic stability)’ 개념과 차이<sup>13</sup>
  - 최근의 위기안정성 연구는 선제공격 가능성 뿐만 아니라 오경보·오인 및 지휘통제 실패에 따른 비의도적 확전 가능성도 중요 고려 대상

〈표 1〉 AI 기반 SNNW 체계와 관련 개념 간 비교

| 관련 개념         | 핵심 특징          | 주요 쟁점                                                            | 분석 초점                     |
|---------------|----------------|------------------------------------------------------------------|---------------------------|
| 자율무기체계 (AWS)  | 무력사용의 자율성      | 활성화 이후 인간의 추가 개입 없이 표적선정·교전 기능을 수행할 수 있는가?                       | 인간통제·책임성 (단기·전술적)         |
| SNNW          | 비핵수단의 전략적 효과   | 핵전력·NC3·전략자산의 생존성이나 억제관계를 바꿀 정도의 효과를 내는가?                        | 억제·군비경쟁 전략 안정성 (중·장기·구조적) |
| AI 기반 SNNW 체계 | AI와 전략적 효과의 결합 | AI가 ISR·융합·표적화·결심지원·임무조정 등을 통해 SNNW의 전략적 효과나 운용방식을 실질적으로 변화시키는가? | 위기안정성 (단기·위기상황)           |

출처: U.S. Department of Defense, Directive 3000.09: Autonomy in Weapon Systems, 2023; Vincent Boulanin et al., Artificial Intelligence, Strategic Stability and Nuclear Risk, SIPRI, 2020; Johnson(2023), 이성훈, “첨단기술이 억제 효과성에 미치는 영향”, 『국방연구』, (2025 12월), 제68권 4호, pp. 30-31을 토대로 저자 작성.

- 특히 SNNW 체계의 AI 전환은 군사적 의사결정의 속도와 방식을 변화시켜 위기 상황에서 행위자의 판단과 대응에 직접적인 영향을 미친다는 점에서 위기안정성과 밀접하게 연계

12 Thomas C. Schelling, The Strategy of Conflict (Harvard University Press, 1966).

13 Lawrence Freedman, The Evolution of Nuclear Strategy, (London: Palgrave Macmillan 2004), p. 165.

#### 나. AI 전환이 위기안정성에 미치는 양면적 효과

- SNNW 체계의 AI 전환은 군사적 불확실성을 제거하여 대응 능력을 향상시키는 안정화 효과와 새로운 통제 불확실성을 증대시키는 불안정화 효과를 동시에 내재<sup>14</sup>

※ Johnson은 AI가 ISR과 데이터 분석, 상황인식 등을 개선함으로써 전략적 의사결정을 지원할 가능성과 새로운 위협을 발생시킬 가능성을 함께 분석<sup>15</sup>

- 안정화 측면에서 AI를 활용한 정보분석 및 상황인식 능력의 향상은 상대의 군사행동과 전력 변화를 보다 신속·정확하게 파악하도록 지원함으로써 판단 오류를 줄이고 위기 상황에서의 예측가능성을 제고

※ 특히 조기경보·미사일방어 등 방어적 영역에서 AI를 활용할 경우 위협 탐지와 대응수단 배분 능력을 향상시켜, 상대의 공격 효과를 감소시키고 거부적 역제의 신뢰성을 강화하는 방향으로 작용<sup>16</sup>

- 반면, AI가 SNNW 체계가 탐지·추적 및 표적화 능력을 향상시키고 의사결정 속도를 가속화할 경우, 위기 시 상대의 위협 인식을 증대시킴으로써 “use it or lose it”의 압박 초래 가능<sup>17</sup>

#### 다. AI 전환에 따른 위기관리의 시간적·인지적 제약

- 특히 위기 상황에서 신속한 판단 압박과 AI에 대한 의존이 결합될 경우, 인간이 AI의 분석·권고를 충분히 검증하지 못한 채 수용하는 ‘자동화 편향(automation bias)’이 발생할 수 있음

- AI에 의해 탐지·판단·대응의 속도가 빨라질수록 상대의 실제 의도를 확인하거나 상호 신호를 교환할 시간 역시 제한되어, 핵·비핵 전력 간 ‘얽힘(entanglement)’에 따른 확전 위험 증폭 가능성<sup>18</sup>
- 이는 전통적 억제 이론이 전제해 온 ‘결심 지연(decision latency)’과 정치·외교적 조정 기회를 축소시키며 비의도적 확전 가능성을 증대

14 이성훈, “첨단기술이 억제 효과성에 미치는 영향”, 『국방연구』, (2025 12월), 제68권 4호, pp. 30-31

15 James Johnson, *AI and the Bomb: Nuclear Strategy and Risk in the Digital Age*, (Oxford: Oxford University Press, 2023), pp. 37-71.

16 Vincent Boulanin et al., *Artificial Intelligence, Strategic Stability and Nuclear Risk* (Stockholm: SIPRI, 2020).

17 Johnson(2023), pp. 111-153.

18 ‘얽힘’은 핵·비핵 전력이 감시정찰·통신·지휘통제 등 일부 기반을 공유하는 상황에서 재래식 목적의 활동이 의도치 않게 핵전력의 운용에 영향을 미치거나 핵공격의 전조로 오인될 수 있는 상태를 의미한다. James M. Acton, “Escalation through Entanglement: How the Vulnerability of Command-and-Control Systems Raises the Risks of an Inadvertent Nuclear War,” *International Security*, Vol. 43, No. 1 (Summer 2018), pp. 56-99.

### III SNNW 체계의 AI 전환에 따른 위험관리 현황과 쟁점

#### 1. SNNW 체계의 AI 전환에 대한 주요 위험관리 접근

##### 가. 전략적 군사행동에서 인간 판단과 통제의 유지

- 최근 군사 AI의 인간 통제에 관한 논의는 자율무기의 최종적인 무력사용 여부뿐 아니라, AI가 의사결정을 지원하는 과정에서 인간의 실질적인 판단능력을 어떻게 유지할 것인가에 초점<sup>19</sup>
  - 미국 국방부는 자율·반자율 무기체계의 운용에서 지휘관과 운용자가 무력사용에 대해 ‘적절한 수준의 인간 판단(appropriate levels of human judgment)’을 행사할 수 있도록 설계할 것을 요구<sup>20</sup>
  - 중국 역시 「군사 분야 AI 활용 규제에 관한 입장문」을 통해 군사 AI의 안전성·신뢰성·통제가능성 향상, 무기체계에 대한 인간 통제와 중지 가능성, 인간의 최종 책임 등을 명시<sup>21</sup>
- 이는 AI가 감시·정찰(ISR)·정보융합·표적화·결심지원 등 무력사용 이전 단계에 광범위하게 활용되는 상황에서 최종 승인권의 형식적 보유를 넘어 인간 판단의 실질성을 확보하는 문제가 중요해지고 있음을 의미

##### 나. AI 기반 정보·결심지원의 신뢰성 및 책임성 확보

- AI 기반 SNNW 체계에서는 개별 알고리즘의 정확성 뿐만 아니라 데이터의 신뢰성, 기능 간 연계 과정의 오류 전파 가능성, 실제 전장에서의 성능 및 인간 검증 가능성을 포함한 신뢰성 확보가 중요
  - 이에따라 주요국의 AI 위험 관리 접근은 추상적 윤리원칙 제시를 넘어 실제 운용환경에서의 신뢰성과 책임성 확보를 위한 방향으로 전환 중
    - ※ 미국은 트럼프 2기 출범 이후 AI 전력화의 속도와 군사적 우위 확보를 최우선 목표로 제시, 작전개념과 업무절차 자체를 AI 중심으로 재설계<sup>22</sup>
    - ※ NATO 역시 책임있는 AI 활용 원칙만을 강조하기보다는 이를 실제 군사능력으로 구현하기 위해 동맹 차원의 데이터 품질·시험평가검증 체계 구축과 상호운용성 확보에 중점<sup>23</sup>

19 U.S. Department of State, “Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy”, (2023).

20 U.S. Department of Defense, Directive 3000.09: “Autonomy in Weapon Systems”, (January 25, 2023).

21 Ministry of Foreign Affairs of the People’s Republic of China, “Position Paper of the People’s Republic of China on Regulating Military Applications of Artificial Intelligence (AI)”, (December 13, 2021).

22 US Department of War, “War Department Launches AI Acceleration Strategy to Secure American Military AI Dominance” (Jan. 12, 2026).

23 2025년 「NATO Data Strategy」와 「Data Quality Framework」는 동맹 간 데이터 공유체계를 구축하는 한편, 정확성·완전성·일관성·적시성 등 데이터 품질 관리와 출처, 변경 이력을 추적하는 기반을 강화했다. NATO OTAN, “Summary of NATO’s revised Artificial Intelligence (AI) strategy” (July 10, 2024).

- 이러한 변화는 SNNW 체계의 AI 전환에서 신뢰성의 관리대상 역시 개별 AI 모델에서 데이터-센서-분석-결심지원-타격수단이 연계되는 전체 운용체계로 확대 중
  - 특히, NATO 「Alliance Digital Strategy」(2026)는 데이터 흐름을 통합하고, 실시간 예측분석과 인지적 의사결정 지원을 구현함으로써 AI 기반 군사 의사결정체계의 통합은 더욱 가속화될 전망<sup>24</sup>

## 2. 기존 군사 AI 위험관리의 한계와 위기안정성 과제

### 가. 개별 체계 중심 위험관리의 한계

- 현행 군사 AI 위험관리는 개별 체계의 기술적 신뢰성 확보를 중심으로 구성되며, ‘인간의 판단 유지’와 ‘시협·평가·검증·확인(TEVV) 및 상호운용성’의 두 축에 기반
  - 이는 AI의 오류·오작동이 무력사용으로 이어질 가능성 감소시키나, 위험관리의 분석단위를 개별 체계로 접근하는 구조적 한계 내포
- AI 체계의 기술적 안전성 확보와 SNNW 체계가 초래하는 전략적 위험 감소는 반드시 일치하지 않으며, 개별 체계의 성능과 안전성을 넘어 AI 결합에 따른 전략적 효과를 함께 고려할 필요
  - ※ 개별 알고리즘의 정확성이 검증되더라도, 기능 간 연계 과정에서의 오류 전파와 상대의 대응에 따른 체계 전체 효과의 변동은 검증에서 제외될 수 있음
  - 자율무기체계(AWS)에서의 인간통제 원칙과 같이 단순히 ‘무력사용의 최종 결정권’ 관점으로 접근할 경우, 상호적 특징을 가진 AI 기반 SNNW 체계의 불확실성을 관리하기에 불충분

### 나. 자국 중심 위험관리와 상호작용적 위험관리의 공백

- 기존 군사 AI 규범은 위험을 ‘각국이 자국의 개발·배치·운용 과정에서 통제할 수 있는 대상’으로 상정하는 자국 중심 구조로 설계
  - 그러나 SNNW에서의 위기안정성은 상호작용의 관계에서 ‘상대가 나의 체계를 어떻게 인식하고 어떻게 대응하는가’에 의해 결정
- 자국의 내부 통제장치는 상대의 위협 인식을 변경할 수단을 갖지 못하며 ‘인간의 최종 통제’ 역시 관찰 불가능한 내부 속성에 기인하는 한계
  - 각국이 상대보다 빠르고 정확한 AI 기반 의사결정 능력을 추구하는 행위 자체가 상대에게는 자국 취약성의 증대로 인식시킬 수 있으며 이는 AI 전력화·배치 경쟁의 가속화라는 안보딜레마를 형성 가능
    - ※ AI가 상대 행동의 예측을 고도화할수록 의도 해석의 여지가 구조적으로 축소될 수 있으며 불확실한 행동은 ‘이상 징후’로 자동 분류되어 외교적 판단 여지를 감소시킴<sup>25</sup>

24 NATO OTAN, “Alliance Digital Strategy” (Jan. 13, 2026).

25 Avi Goldfarb and Jon R. Lindsay, “Prediction and Judgment: Why Artificial Intelligence Increases the Importance of Humans in War,” *International Security*, Vol. 46, No. 3 (2021/2022), pp. 7-50.

## IV AI 기반 SNNW의 위기안정성 확보 방안

### 1. 위기안정성 위험 요인의 구조적 특성과 한반도 환경

#### 가. 비대칭 전략구조와 상호 취약성

- 핵·재래식 전력이 혼재하는 한반도 안보 환경에서 북한의 AI 적용은 드론·사이버·전자전 등 진입장벽이 낮은 영역을 중심으로 추진 중
  - 북한은 9차 당대회를 계기로 핵무기를 ‘보유’에서 ‘운용’ 단계로 전환, AI 기반 SNNW 역시 감시·정찰, 지휘통제, 무인체계 등에서 핵·재래식 전력의 운용을 보조하는 방향으로 활용될 가능성<sup>26</sup>
  - 동시에 AI 기반 사이버 역량을 활용하여 상대의 정보수집·지휘통제 체계에 혼란을 야기하는 비대칭 방식의 역량을 확대 중

#### 나. 시간·공간의 압축과 결심 지연

- 한반도 전역은 단거리 미사일의 비행 도달 시간이 수 분 단위에 불과, ‘감지-식별-결심-대응’의 루프가 극히 제한된 시간 폭 안에서 완결
  - AI가 각 단계의 결심지원에 결합되면 작전 루프가 분 단위 아래로 더욱 압축될 수밖에 없으며 외교적 조정 기회 역시 줄어들 가능성
- 위기 상황에서 핵·지휘통제 자산의 생존성에 대한 불확실성이 커질 경우, 북한 지도부가 ‘use-it-or-lose-it’ 압박을 인식할 가능성을 위기관리 차원에서 분석할 필요<sup>27</sup>
  - AI 기반 ISR·정밀타격 능력의 고도화가 가져오는 한국의 억제력 강화 효과와 북한의 위협인식·대응에 미치는 이중적 영향 고려

26 김보미, 2026, “9차 당대회와 북한 군사전략의 전환: 핵무기 보유에서 운용으로”, 『INSS ISSUE BRIEF』 제801호.

27 James Johnson, “Revisiting the ‘Stability-Instability Paradox’ in AI-Enabled Warfare,” Review of International Studies, (2024).

#### 다. 정보·인지 비대칭과 오인 위험

- 한반도에서는 한미 연합의 ISR 우위와 북한의 감시 열세가 뚜렷하며, 북한은 자체 정보자산 부족을 정찰위성 확보로 보완하는 과도기적 단계
  - 북한은 2023년 11월 첫 군사정찰위성 ‘만리경-1’을 발사해 궤도 진입에 성공했으나, 이후에도 정상 작동 여부는 명확히 확인되지 않음<sup>28</sup>
- 한반도에서는 경보-결심 단계에서 AI 판단의 인간 검증이 생략될 경우, 자동화 편향이나 식별 오류·오경보 노출 사고가 발생할 가능성
  - 북한의 의사결정 구조는 단일 지도자 중심의 집중형으로, 의도·신호 판독이 구조적으로 어려우며, 내부 숙의·신호 발신 과정이 외부에서 관측 불가능한 ‘블랙박스’로 작동, AI 판단의 불확실성 증대<sup>29</sup>
  - 이러한 불투명성은 AI 기반 의도 해석의 불확실성을 높일 수 있으며, 판단 과정에서 오인·오판의 가능성을 증대시키는 요인으로 작용
- 정보 공유 범위·기준·의사결정 속도의 차이가 존재하는 한미 연합 작전 환경에서는 AI 기반 SNNW 운용에서의 공동 상황인식이 지연될 가능성
  - 한국은 대북 정보의 상당 부분을 위성 등 미국의 ISR 자산에 의존해 왔으며, 위기 시 공동 판단을 분절시키고 단일한 대응 신호를 제약하는 구조적 약점으로 작용할 수 있음<sup>30</sup>
- 이상의 구조적 특성은 △시간 압축에 결심 시간의 축소 압박, △정보·인지 비대칭에서 비롯된 오인·오판 경로, △전략자산·지휘통제 체계의 상호의존에서 파생된 자동화 얽힘 등, 불안정한 확전 요인으로 수렴 가능

28 38 North, “Modest Beginnings: North Korea Launches Its First Reconnaissance Satellite,” (2023. 11),

29 Nick Kodama, “Strategic Stability and the Credibility of North Korea’s Nuclear Threats,” *Journal of Global Security Studies*, Vol 6, No. 1, (2021).

30 Gabriela Bernal, “The US-ROK Intelligence Rift and Seoul’s Efforts for Greater Defense Autonomy”, *Peninsula Dispatch*, (Apr 29, 2026).

## 2. 확전 요인 통제를 위한 설계 원칙

### 가. 결심시간 보장과 운용 신호 규율

- AI 결심지원 산출물(권고안 · 표적 후보)과 최종 승인 사이에 인간 심의를 위한 절차적 최소 시간을 규정
  - 다만, 결심시간 보장을 일률적으로 의무화할 경우 방어 영역에서의 군사적 열세를 초래할 수 있으므로 최소 결심시간은 위기 단계와 임무 위험도에 따라 차등 적용
    - ※ 기동 · 요격 등 속도가 생존에 직결되는 방어적 대응 영역은 예외로 하여 ‘속도와 안정성’의 우선순위를 체계별로 사전 규정
- 한미는 AI 기반 SNNW체계의 인간 승인 · 검증 절차를 명확히 하고, 정상적인 훈련 · 감시 활동이 북한 전략자산에 대한 이상 공격 신호로 오인되지 않도록 운용 투명성 · 신호 규율 원칙을 마련할 필요
  - △예측 가능한 운용 패턴 유지, △정찰 · 훈련 등 일상 활동이 선제 공격 준비로 오독될 소지의 최소화, △위기 시 AI 운용 태세의 신호화 · 사전 통보 등이 대표적
  - 이러한 원칙은 위기 상황에서 신호의 오인과 오판을 줄이고, 북한의 불필요한 결심 압박을 완화하는 안전장치로 기능

### 나. 오인 · 오판 차단을 위한 인간 판단 · 교차검증 절차

- 인간 개입을 ‘최종 승인’이라는 형식적 단계가 아닌 정보 선별 · 표적화 · 결심지원 등 선행 단계 전반에 배치하는 실질적 검증 절차로 전환할 필요
  - AI 권고의 신뢰도 표시를 의무화하고, 한미 공통 기준의 ‘수용-기각-불인정-이탈(override: 자동 판단 루프 전체에서 벗어나 인간이 직접 판단)’에 이르는 층위별 경로 단계를 표준화할 필요
  - 단일 AI 모델의 산출물을 독립된 제2모델과 인간 분석관이 교차 확인하는 이중 검증 절차를 마련하고, 신뢰도가 임계값 이하로 하락할 경우 인간 판단으로 전환
  - 검증 절차는 기록 가능한 형태로 제도화하여 사후 감사와 책임 소재 확인이 가능하도록 설계
    - ※ 이는 미 국방부 지침 3000.09가 요구하는 ‘적절한 수준의 인간 판단’ 및 NATO의 ‘책임사용 원칙의 인간 통제’ 요건 방향과도 부합<sup>31</sup>

31 U.S. Department of Defense(Jan. 25, 2023); NATO, “Summary of the NATO Artificial Intelligence Strategy,” (July. 10, 2024),

#### 다. 자동화 얽힘 대응을 위한 실패-안전 설계

- AI와 SNNW · 핵심 전략자산의 결합은 ‘얽힘’을 통해 단일 오류를 연쇄 · 비선형적으로 증폭시킬 수 있으므로 실질 통제수단이 요구됨
  - SNNW 체계 설계에 ‘실패-안전(fail-safe)’ 원칙을 기본 전제로 적용하고 핵심 의사결정의 AI 완전 자동화를 배제하는 이중 전략 필요
    - ※ 실패-안전 설계는 오작동 · 통신 두절 · 데이터 오염 발생 시 체계가 ‘비작동’ 상태로 수렴하도록 기본값을 구성하는 접근을 의미<sup>32</sup>
  - 핵무기의 허가연동장치(PAL)와 단독 조작을 방지하는 2인 규칙, 이지스 체계의 자동화 운용 시 인간 개입 절차 등의 인간통제 사례를 참고
- AI-NC3 결합은 억제 신뢰를 위한 공개와 능력 보호를 위한 모호성의 긴장을 초래하므로 ‘무엇을 · 언제 · 얼마나 공개할 것인가’를 규율하는 통제 원칙을 체계 설계와 구분되는 별도 정책으로 명시할 필요<sup>33</sup>

### 3. 위기안정성 확보의 단계별 이행 과제

#### 가. 단기: 한미 차원의 연합 거버넌스 및 위기소통 강화

- 2024년 7월 한미가 서명한 「한반도 핵억제 · 핵작전 공동지침」은 동맹의 핵 운용 원칙과 절차를 제도화한 기반이나, AI 데이터 검증 · 오류 대응 규정은 포함하지 않은 공백이 존재<sup>34</sup>
  - 핵협의그룹(NCG)이 정보공유 보안 · 핵 지휘통제 협조 등을 의제로 운영되어 온 만큼, AI 결합 단계의 상호 검증은 기존 NCG 의제와 관련한 연장선으로 논의 가능<sup>35</sup>

32 해당 설계는 위험 완화 수단의 우선순위 계층(제거→대체→공학적 통제→경고→절차)에서 ‘안전상태 수렴’을 최상위군에 설정한 미 국방 시스템안전 표준(MIL-STD-882E) 관행으로 정립됨. HIMA Paul Hildebrandt GmbH, “Structures of Safety Systems”, (Jan. 2, 2024), [https://www.hima.com/en/shareandexplore/structures\\_of\\_safety-critical\\_systems](https://www.hima.com/en/shareandexplore/structures_of_safety-critical_systems) (검색일: 2026. 8. 12); MIL-STD-882E, “DoD Standard Practice: System Safety,” <https://www.safetyartisan.com/category/mil-std-882e/> (검색일: 2026. 8. 12).

33 UNGA Resolution 80/23, (2025. 12); “Lessons from the UN’s First Resolution on AI in Nuclear Command and Control,” Bulletin of the Atomic Scientists, (2025. 12), <https://thebulletin.org/2025/12/lessons-from-the-uns-first-resolution-on-ai-in-nuclear-command-and-control/> (검색일: 2026. 8. 22); Reuters, (2024. 11).

34 U.S.-ROK, “Joint Statement on the Guidelines for Nuclear Deterrence and Nuclear Operations on the Korean Peninsula,” (2024. 7. 11), <https://kr.usembassy.gov/071224-joint-statement-by-president-biden-and-president-yoon-on-u-s-rok-guidelines-for-nuclear-deterrence-and-nuclear-operations-on-the-korean-peninsula/> (검색일: 2026. 8. 17).

35 CSIS, “The U.S.-ROK Nuclear Consultative Group’s Successful Launching,” <https://www.csis.org/analysis/us-rok-nuclear-consultative-groups-successful-launching> (검색일: 2026. 8. 17).

- ※ 다만, 알고리즘·원시 데이터의 전면 공유는 기밀 제약상 현실적이지 않으므로, 1단계는 데이터가 아닌 '결심시간 기준·인간 개입 절차·사고 보고 규범' 등 절차적 표준의 공동 정의에서 출발, 점진적으로 확대하는 단계적 검증체계 모색
- 결심시간 보장·'수용-기각-불인정-이탈' 표준화를 NCG 후속 의제로 채택하고, 한미 공통 기준으로 정립할 필요
- ※ 이는 5차(2025. 12. 11)·6차(2026. 6. 11) NCG 회의에서 의제가 지속 확장되는 추세와도 방향성을 같이하고 있음<sup>36</sup>
- 위기 시 AI 운용 태세의 신호화·사전 통보 절차를 연합 위기소통 체계에 통합, AI 결합 ISR·타격 자산의 운용 패턴 예측 가능성 제고
  - 평시 패턴과 위기 행동 간 정합성을 확보하고 한미 간 AI 산출물의 오류·불확실성 정보 공유 원칙을 사전 합의하여 북한의 선제 공격 오독 유인을 최소화
- 중장기적으로 남북 간 군사적 소통 여건이 조성될 경우, 상호주의를 전제로 AI 관련 오경보 확인·사고 통보 기준 등 위기 소통을 위한 「9.19 체제 2.0」의 단계적 추진도 검토
  - 기존 9. 19 체제가 남북 간 지리적 완충과 재래식 군비통제에 집중했다면, 「9. 19 체제 2.0」은 AI 안보 환경에 부합하는 '정보·인지 영역의 신호 규율'과 '자동화 위험 감축 논의'로 확장 가능

#### 나. 중기: AI 기반 SNNW 획득·검증 절차의 전략적 효과 평가 도입

- AI-SNNW 획득·검증 절차는 파편적 체계의 자율성 검토 수준에서 상호적 전략적 효과 평가로 확장될 필요
  - 정확도·신뢰도 기술 성능 검증 뿐만 아니라 해당 기능이 위기 시 상대 위협 인식·결심 시간·행동 유인에 미치는 전략적 파급을 상호 관계의 맥락에서 평가 대상에 포함
  - ※ 위계임·레드팀 훈련을 통해 '상대가 이 행동을 어떻게 해석할 것인가'를 검증 항목으로 명시하고, 통신 두절·데이터 오염·전자전 교란 시 체계가 어떤 기본값으로 수렴하는지를 사전 검증

36 U.S. DoD, "Joint Press Statement on the Fifth Nuclear Consultative Group Meeting," (2025. 12. 11), <https://www.war.gov/News/Releases/Release/Article/4358035/>; USFK, "Joint Press Statement on the Sixth Nuclear Consultative Group Meeting," (2026. 6. 11), <https://www.usfk.mil/Media/Press-Products/Press-Releases/Article/4515329/> (검색일: 2026. 8. 23).

- 표적화 · 조기경보 · 정밀타격 연계 등 결심시간에 직접 영향을 미치는 고위험 기능을 등급화하고, 차등화된 인간 검증 · 교차검증 요건 적용
  - 상대 결심 구조로의 파급이 상대적으로 낮은 ISR · 정보융합 등 인지 기능은 완화된 검증을, 표적화 · 조기경보 · 정밀타격 연계 등 결심에 직접 영향을 미치는 기능은 강화된 인간 심의 · 교차검증을 차등 적용
    - ※ 미 국방부 역시 AI 기반 체계 획득에서 기능 위험도가 높을수록 시험 · 검증 강도를 상향하는 ‘위험 기반 시험 수준(risk-based level of test)’ 원칙을 채택<sup>37</sup>
    - ※ 이 같은 접근은 NIST에서 시행중인 AI RMF의 위험 관리 구조(Govern-Map- Measure-Manage)와도 일치<sup>38</sup>

#### 다. 장기: 전략적 위험감축 규범의 국제 의제화

- 책임 있는 군사 AI의 이행 지침을 제시한 제2차 REAIM 정상회의 「Blueprint for Action」(2024. 9)을 계승하여, AI-SNNW체계의 전략적 위험감축 논의로 확장할 필요<sup>39</sup>
  - 동 Blueprint의 고영향 군사 AI 논의를 확장하여, AI-SNNW의 오인 · 오판 및 확전 위험을 관리하기 위한 상호적 위험감축 원칙을 국제 의제로 제안<sup>40</sup>
- 특히 핵 지휘통제(C2) · 조기경보 단계에 AI가 결합될 경우 오판과 결심시간 압축 위험의 민감성이 높아지는 만큼, Blueprint의 인간 책임 원칙과 핵 C2의 인간통제 규범을 연계하여 의제화할 필요
  - 이는 한국의 SNNW뿐 아니라 북한의 핵 C2 · 조기경보 AI 결합에도 동일한 규범을 적용하는 양방향적 접근으로, ‘한국의 AI 기반 비핵억제 수단 통제’와 ‘북한 핵 C2의 AI 통제’를 병행하는 시도

37 U.S. DoD, “DoD Manual 5000.101: Operational Test and Evaluation and Test and Evaluation of AI-Enabled Systems,” (Dec. 24, 2024), <https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodm/5000101p.pdf> (검색일: 2026. 8. 18).

38 NIST, “AI Risk Management Framework,” <https://www.nist.gov/itl/ai-risk-management-framework> (검색일: 2026. 8. 18).

39 외교부, 『2024 인공지능의 책임있는 군사적 이용에 관한 고위급회의(REAIM) 행동을 위한 청사진(Blueprint for Action)』, (2024. 9.), [https://www.mofa.go.kr/www/brd/m\\_4080/download.do?brd\\_id=235&seq=375378&data\\_type=A&file\\_seq=9](https://www.mofa.go.kr/www/brd/m_4080/download.do?brd_id=235&seq=375378&data_type=A&file_seq=9) (검색일: 2026. 8. 22).

40 Reuters, “Sixty countries endorse ‘blueprint’ for AI use in military; China opts out,” (2024. 9. 10), <https://www.reuters.com/technology/artificial-intelligence/south-korea-summit-announces-blueprint-using-ai-military-2024-09-10> (검색일: 2026. 8. 22).

## 참고문헌

- 김보미, 2026, “9차 당대회와 북한 군사전략의 전환: 핵무기 보유에서 운용으로”, 『INSS ISSUE BRIEF』, 제801호.
- 외교부, 2024, 『2024 인공지능의 책임있는 군사적 이용에 관한 고위급회의(REAIM) 행동을 위한 청사진(Blueprint for Action)』, (2024. 9), [https://www.mofa.go.kr/www/brd/m\\_4080/down.do?brd\\_id=235&seq=375378&data\\_tp=A&file\\_seq=9](https://www.mofa.go.kr/www/brd/m_4080/down.do?brd_id=235&seq=375378&data_tp=A&file_seq=9) (검색일: 2026. 8. 22).
- 이성훈, 2025, “첨단기술이 억제 효과성에 미치는 영향”, 『국방연구』, (2025. 12월), 제68권 4호.
- Abraham, Yuval. 2024, “‘Lavender’: The AI Machine Directing Israel’s Bombing in Gaza,” +972 Magazine, April 3, 2024, <https://www.972mag.com/lavender-ai-israeli-army-gaza/>
- Acton, James M., 2018, “Escalation through Entanglement: How the Vulnerability of Command-and-Control Systems Raises the Risks of an Inadvertent Nuclear War,” *International Security*, Vol. 43, No. 1, pp. 56-99.
- Bernal, Gabriela, 2026, “The US-ROK Intelligence Rift and Seoul’s Efforts for Greater Defense Autonomy,” *Peninsula Dispatch*, April 29, 2026.
- Boulanin, Vincent, et al., 2020, *Artificial Intelligence, Strategic Stability and Nuclear Risk*, Stockholm: SIPRI.
- Center for Strategic and International Studies (CSIS), n.d., “The U.S.-ROK Nuclear Consultative Group’s Successful Launching,” <https://www.csis.org/analysis/us-rok-nuclear-consultative-groups-successful-launching> (검색일: 2026. 8. 17).
- Freedman, Lawrence, 2004, *The Evolution of Nuclear Strategy*, London: Palgrave Macmillan.
- Futter, Andrew, and Benjamin Zala, 2021, “Strategic Non-Nuclear Weapons and the Onset of a Third Nuclear Age,” *European Journal of International Security*, Vol. 6.
- Goldfarb, Avi, and Jon R. Lindsay, 2021/2022, “Prediction and Judgment: Why Artificial Intelligence Increases the Importance of Humans in War,” *International Security*, Vol. 46, No. 3.
- Hoffmann, Fabian, 2021, “Strategic Non-nuclear Weapons and Strategic Stability:

- Promoting Trust through Technical Understanding,” November 2021.
- \_\_\_\_\_. 2024, “The Strategic-Level Effects of Long-Range Strike Weapons: A Framework for Analysis,” *Journal of Strategic Studies*, Vol. 47, No. 6-7.
- Horowitz, Michael C., 2026, “Artificial Intelligence and the Future of Strategic Stability,” *Texas National Security Review*, Vol. 9, No. 2.
- Johnson, James, 2023, *AI and the Bomb: Nuclear Strategy and Risk in the Digital Age*, Oxford: Oxford University Press, pp. 1-36, 37-71, 111-153.
- \_\_\_\_\_. James, 2024, “Revisiting the ‘Stability-Instability Paradox’ in AI-Enabled Warfare,” *Review of International Studies*.
- Kodama, Nick, 2021, “Strategic Stability and the Credibility of North Korea’s Nuclear Threats,” *Journal of Global Security Studies*, Vol. 6, No. 1.
- MIL-STD-882E, n.d., “DoD Standard Practice: System Safety,” <https://www.safetyartisan.com/category/mil-std-882e/> (검색일: 2026. 8. 12).
- Ministry of Foreign Affairs of the People’s Republic of China, 2021, “Position Paper of the People’s Republic of China on Regulating Military Applications of Artificial Intelligence (AI),” December 13, 2021.
- NATO, 2024, “Summary of NATO’s Revised Artificial Intelligence (AI) Strategy,” July 10, 2024.
- \_\_\_\_\_. 2026, “Alliance Digital Strategy,” January 13, 2026.
- National Institute of Standards and Technology (NIST), 2023, “AI Risk Management Framework,” <https://www.nist.gov/itl/ai-risk-management-framework> (검색일: 2026. 8. 18).
- Paul HIMA and Hildebrandt GmbH, 2024, “Structures of Safety Systems,” January 2, 2024.
- Responsible AI in the Military Domain (REAIM) Summit, 2024, “Blueprint for Action,” Seoul, September 2024.
- Reuters. 2024, “Biden, Xi Agree That Humans, Not AI, Should Control Nuclear Weapons,” November 16, 2024.
- \_\_\_\_\_. 2026, “Pentagon to adopt Palantir AI as core US military system, memo says” (March 20, 2026), <https://www.reuters.com/technology/pentagon-adopt-palantir-ai-as-core-us-military-system-memo-says-2026-03-20/> (검색일: 2026. 9. 9).
- Schelling, Thomas C., 1966, *The Strategy of Conflict*, Cambridge, MA: Harvard University Press.

38 North, 2023, “Modest Beginnings: North Korea Launches Its First Reconnaissance Satellite,” November 2023.

United Nations General Assembly, 2025, “Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security: Report of the Secretary-General (A/80/78),” June 5, 2025, pp. 101-102.

United Nations General Assembly, 2025, “Resolution 80/23: Possible Risks of the Integration of Artificial Intelligence into Command, Control and Communications Systems of Nuclear Weapons,” December 2025.

United Nations Security Council, 2025, “Maintenance of International Peace and Security,” S/PV.10005 (Resumption 2), September 26, 2025, p. 6.

\_\_\_\_\_. 2020 (revised), Nuclear Matters Handbook 2020 (Revised), Chapter 8: Nuclear Surety.

U.S. Department of State, 2023, “Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy”

U.S. Department of Defense, 2023, “Directive 3000.09: Autonomy in Weapon Systems,” January 25, 2023.

\_\_\_\_\_. 2024, “DoD Manual 5000.101: Operational Test and Evaluation and Test and Evaluation of AI-Enabled Systems,” December 24, 2024.

\_\_\_\_\_. 2025, “AI Aided Sensor Data Fusion,” Small Business Innovation Research (SBIR) Phase II, <https://www.sbir.gov/awards/218485> (검색일: 2026. 8. 23).

\_\_\_\_\_. 2025, “Joint Press Statement on the Fifth Nuclear Consultative Group Meeting,” December 11, 2025, <https://www.war.gov/News/Releases/Release/Article/4358035/>

U.S. Department of War, 2026, “War Department Launches AI Acceleration Strategy to Secure American Military AI Dominance,” January 12, 2026.

U.S. Forces Korea (USFK), 2026, “Joint Press Statement on the Sixth Nuclear Consultative Group Meeting,” June 11, 2026, <https://www.usfk.mil/Media/Press-Products/Press-Releases/Article/4515329/>

U.S.-ROK, 2024, “Joint Statement on the Guidelines for Nuclear Deterrence and Nuclear Operations on the Korean Peninsula,” July 11, 2024, <https://kr.usembassy.>

gov/071224-joint-statement-by-president-biden-and-president-yoon-on-u-s-  
s-rok-guidelines-for-nuclear-deterrence-and-nuclear-operations-on-the-  
korean-peninsula/ (검색일: 2026. 8. 17).

Wissler, John E. III, 2023, “Artificial Intelligence and Cyber Warfare: Evolution or  
Revolution?,” The Cyber Defense Review, Vol. 8, No. 1, pp. 53-64.

## Abstract

# The AI Transition of Strategic Non-Nuclear Weapons and Measures to Ensure Crisis Stability

Junghyun Yoon

(Institute for National Security Strategy)

Recent advances in the AI transformation of Strategic Non-Nuclear Weapons (SNNWs) are fundamentally reshaping the security architecture of deterrence and crisis management. South Korea, in particular, faces structural constraints on crisis escalation management arising from the combined effects of an asymmetric strategic environment and the compression of time and space. Against this backdrop, this study analyzes the challenges to crisis stability posed by the integration of AI into SNNW systems and proposes implementation measures for institutionalizing crisis management and peaceful coexistence on the Korean Peninsula. First, cross-validation procedures should be established by ensuring a minimum decision time between AI-generated recommendations and human authorization and by developing rules for operational signaling. Second, operational principles should be established to address automation-induced entanglement, based on fail-safe design. Third, a phased approach is needed to establish risk-communication mechanisms both within the ROK-U.S. alliance and between the two Koreas. Key measures include: △ establishing ROK-U.S. alliance-wide standards for decision time and data-validation governance; △ pursuing a “September 19 Military Agreement 2.0” framework for procedures to verify AI-related false alarms and provide notification of incidents between the two Koreas; △ incorporating strategic-effects assessments into SNNW acquisition and verification procedures; and △ advancing efforts to internationalize norms for responsible military AI and human control over nuclear weapons.

Keywords: AI, Strategic Non-Nuclear Weapons (SNNW), Crisis Stability, Human Control over AI, Responsible Military AI

본지에 실린 내용은 집필자 개인의 견해이며,  
국가안보전략연구원의 공식입장이 아닙니다.

# INSS

## 전략보고

September 2026  
**No. 409**