AI 신뢰성 확보를 위한 글로벌 표준 동향

곽준호 TTA AI신뢰성정책팀 팀장

1. 머리말

AI를 둘러싼 글로벌 패권 경쟁의 기세가 무섭다. 각국은 AI를 국가 안보와 직결된 전략 자산으로 바라보며, AI의 위험성에 대한 경계를 늦추지 않고 있다. 동시에 각국은 자국 기업과 산업 생태계의 글로벌 경쟁력을 강화하기 위해 기술 연구·개발 촉진, 정책·전략 수립 등 다양한 움직임을 가져가고 있다.

이런 와중에 우리나라는 중요한 AI 행사 2가지를 최근 개최했다. 지난 4월에는 AI 분야 최대 규모이자 가장 활발한 표준 활동을 보이는 기구인 ISO/IEC JTC1/SC42 총회가 서울에서 개최됐다. 이어 5월 21~22일 진행된 AI 서울 정상회의에선 전 세계 27개국 정상 및 각료, 그리고 글로벌 AI 기업 대표와 전문가가 모여 AI 안전, 포용, 혁신에 대해 논의하며 의미 있는 성과를 거뒀다. 글로벌 협력, 정책 측면, 기술표준 분야에 있어 가장 굵직한 두 행사를 개최하며, 우리나라가 지속적으로 AI와 관련된 글로벌 규범 논의의 중심에 서 있음을 확인할 수 있었다.

AI 분야에서 정책과 표준, 이 두 가지는 특히 밀접하게 연관된 개념으로 보인다. 현재 전 세계가 AI 기술에 주목하고 있는 가운데, 기술패권을 차지하기 위해 산업 진흥·혁신 및 규제 정책을 마련하고 있다. 이러한 정책이 효과적으로 이행되려면, 기술적으로 뒷받침할 수 있는 표준이 필요하다. 정책도 중요하지만, 이를 구체화하고 현실화할 수 있는 수단은 결국 표준인 것이다.

이번 원고에선 먼저 AI 기술 패권의 핵심 개념인 AI 신뢰성과 관련된 각국 정책을 정리하고, 표준 제정 및 표준화 현황은 어떠한지 차례대로 짚어보고자 한다. 이를 통해, 최근 글로벌 규범 논의의 중심으로 자리 잡고 있는 우리나라가 AI 신뢰성 분야 표준에 어떤 전략으로 접근해야 하는지 살펴본다.

출처: 비즈워치(https://v.daum.net/v/20240422150203974

[그림 1] SC42 13차 서울 총회 전경

2. AI 신뢰성, 글로벌 규범과 규제의 핵심 요소

AI 신뢰성은 2019년도 이후부터 본격적으로 국제사회에서 논의되기 시작한 개념이다. 당시 AI가 본격적으로 산업·사회의 서비스·제품에 적용되기 시작됐는데, 이때 '차별적인 의사 결정'과 '설명이 불가능하다'는 특성이 주목받은 것이다. 이렇게 기존 소프트웨어·시스템과는 다른 요소 때문에 AI 신뢰성은 AI 윤리와 함께 AI 기술의 부작용 내지는 사회적 이슈로서 점차 주목받았다. 더불어 AI 신뢰성 확보를 위한 노력이 이어졌는데, 이는 국제기구와 글로벌 협의체를 중심으로 원칙·권고안이라는 형태로 표출됐다.

그러나 최근 생성 AI가 주목을 받으면서, 이전과는 양상이 사뭇 달라졌다. AI가 활용될 수 있는 범위와 편익의 파급효과가 폭발적으로 증가·확대되는 한편, AI로 인한 위험의 크기와 인식 역시 확대됐다. 생성 AI에 대한 접근성이 크게 높아지면서, 부작용과 위험요소를 많은 이들이 더욱 체감하게 된 것이다. 이에 국제사회는 점차 AI의 위험에 주목하게 됐고, 좀 더 적극적으로 대응하게 됐다.

EU(유럽연합, European Union)는 2021년도에 AI 규제를 위한 법안을 발의했다. 그간의 논의를 거쳐 규제 법안은 올해 3월 제정됐으며, 향후 3년에 걸쳐 세부적인 기술 기준과 준수 여부 확인을 위한 절차를 보완해 나갈 것으로 보여진다.

미국은 작년 10월 행정명령을 통해 AI 신뢰 확보 및 혁신·진흥을 위한 이행 방안을 내놓은 바 있다. 여기엔 특정 컴퓨팅 성능을 능가한 최신 모델에 대해, 레드팀 방식을 이용해 테스트를 실시하거나 모델 정보를 공개하도록 하는 등 투명성 확보 조항도 포함돼 있다. 또, AI가 생성한 콘텐츠를 합성 콘텐츠라 명명하고, 합성 콘텐츠임을 표기하는 기술에 대한 도입과 연구·개발에 대한 조항도 있다. 이는 최근 생성 AI로 인한 딥페이크, 가짜뉴스 등에 대한 대응이다. 이러한 기술적 조치는 미국 상무부 주도로 진행되고 있으며, 구체적 추진은 NIST(미국 국립표준기술연구소, National Institute of Standards and Technology)가 담당하고 있다. 특히, 주목할 대목은 해당 행정명령이 '국제 협력'과 '국제 표준화'에의 적극적 참여와 활동을 언급했다는 사실이다. 미국이 얼마나 'AI 신뢰성 확보를 위해 국제 표준화를 중요하게 다루고 있는지' 알 수 있는 부분이다.

국가 단위뿐 아니라, 국제기구 혹은 국가 간 협의체에서도 AI 위험과 신뢰성 확보 논의가 매우 활발하다. 활발하다 못해 너무 많아 헷갈릴 정도인데, 실제로 그러한 협의체에서 발표한 원칙, 가이드, 권고안만 해도 수를 헤아리기 어렵다. OECD(경제협력개발기구, Organisation for Economic Co-operation and Development)의 AI 원칙, UNESCO(유엔교육과학문화기구, United Nations Educational, Scientific and Cultural Organization)의 AI 윤리 권고안처럼 주요 국제기구가 '19년도부터 준비한 것이 있고, 이어서 다양한 산업 협의체와 비영리 단체까지 논의에 뛰어들고 있다. 이 중 유의미하게 살펴볼 최근의 협의체 혹은 기구가 AI 안전 정상회의다. 작년 11월 영국 주도로 전 세계 국가 정상과 각료들, 그리고 AI 기업 및 전문가들이 모여 AI 안전 확보를 위한 선언을 했다. 지난 5월 21일에는 한국에서 2회 정상회의가 진행된 바 있다. 2회 정상회의에선 안전과 함께 혁신과 포용을 포함한 광범위한 논의가 진행됐다.

특히, 한국 기업을 포함해 16개 기업이 AI 안전 확보를 위한 자발적 조치·이행을 서약한 부분을 주목할 만하다. 여기엔 AI의 위험 요소를 발굴하고, 대책을 마련하는 등 위험관리를 수행하는 것뿐만 아니라, 그 결과를 공개하고 설명하는 등 8개 이행사항이 있다. 최근 이야기되는 프런티어

[그림 2] AI 서울 정상회의 성과

출처: 중앙일보사(https://www.joongang.co.kr/article/25251116#home

[그림 3] AI 서울 정상회의 전경

AI 모델에 대한 안전성 테스트 등도 포함돼 있다. 이는 내년 초 예정된 3회 정상회의 전까지 이행해야 하는 사항으로, 각 기업의 노력과 결과에 시선이 모아진다.

1회 AI 안전 정상회의 이후로, AI 안전이라는 키워드가 많이 사용되고 있다. 사실 지금까지 글로벌 논의와 표준에서 안전성(Safety)은 신뢰성(Trustworthiness)을 구성하는 하위 속성으로서 정의가 돼 왔다. 허나 최근의 AI 안전은 신뢰성만큼이나 광범위한 영역을 포함하는 개념으로 논의되고 있다. 이는 AI가 제품·서비스와 같은 산업의 영역을 넘어, 사회 혼란과 무기 활용, 국가적 재난을 초래할 수 있다는 관점이 대두됐기 때문으로 보인다. 향후 이러한 개념에 대한 명확한 정의가 내려질 것이다.

한편, AI 안전에 관심이 집중되며 영국과 미국을 중심으로 AI 안전연구소가 설립됐다. 향후 개발되거나 출시될, '높은 수준의 기능과 역량'을 갖춘 AI의 안전성을 확보하자는 취지다. 현재 이러한 AI의 위험 요소는 무엇인지, 위험관리는 어떻게 수행해야 하는지, 위험 방지를 위해 적용할 수 있는 기술엔 어떤 것이 있는지, 시험·검증은 어떻게 진행해야 하는지 등 관련 연구가 초기 단계에 머물러 있다. 더불어 올해 초에는 일본도 AI 안전연구소 설립을 발표한 바 있으며, 우리나라 역시 이번 AI 정상회의에서 AI 안전연구소 설립을 공식화했다.

이처럼 AI 신뢰성은 글로벌 규범과 규제의 핵심 요소로 꼽히고 있다. 동시에, AI 신뢰성 확보에 필요한 기술의 위상도 달라졌다. 이제 AI 기술 경쟁력을 갖추기 위한 영역으로 받아들여지고 있는 것이다. 하지만 AI 기술이 폭발적 속도로 발전해 나가는 상황에서, AI를 검증하고, 위험을 발굴·관리하며, 방지하기 위한 기술의 연구·개발은 뒤처질 수밖에 없다. 영국에서 지난 5월 20일 발표한 AI 안전에 관한 국제과학보고서에 의하면, AI 안전과 위험 대응을 위한 기술은 아직 성숙하지 않았으며, 연구적인 도전(Challenge)이 산적해 있다고 한다. 그럼에도 불구하고, 규범·규제의 실질적인 효과를 위해선 이를 뒷받침할 수 있는 기술·기준을 마련해야 한다. 기술표준이 필요한 이유다.

3. AI 신뢰성 분야 표준의 중요성

지금까지 살펴본 글로벌 동향에서 공통적으로 언급할 수 있는 부분이 있다. 신뢰성 확보를 위한

원칙, 가이드, 권고를 비롯해 규범·규제 역시, 그 이행과 준수 확인을 위한 기술과 기준이 요구된다는 것이다. 기술이 없다면 이행과 준수 자체가 어렵고, 구체적인 기준이 없다면 명확한 이해와 해석이 불가능하므로 사회와 산업에서 혼선이 가중될 것이다.

이러한 관점에서 기술표준은 규범과 규제를 더욱 구체화하고, 이해관계자들의 혼돈과 혼선을 줄이는 수단이면서, 다양한 협의체·단체에서 산발적으로 내놓고 있는 내용들을 통일성 있게 바라볼 수 있게 한다. 이는 효율적인 의사소통을 가능케 한다.

장기적 관점에서도 AI 신뢰성의 기술표준은 매우 중요하다. 흔히들 AI 신뢰성을 이야기할 때, 많은 사람들은 기술혁신을 방해하는 측면으로 이해한다. 앞서 소개한 바와 같이, 글로벌 규범과 규제의 핵심 요소이기 때문이다. 하지만 AI 신뢰성 확보 기술이 더 발전할수록 규범과 규제는 더욱 명확해지고 구체화 될 수 있다. 필연적으로 규범·규제 준수를 해야 하는 기업 입장에서는 오히려 부담이 완화되는 효과를 얻을 수 있다.

실제 EU는 AI 규제가 AI 기술 혁신과 발전을 저해하지 않고 신뢰성을 확보할 수 있는 요소로서 구체적인 기술 기준을 언급한 바 있다. 또한, 이는 AI 위험을 체감하고 있는 소비자와 일반 시민의 불안감을 해소할 수 있다. 이런 효과는 결론적으로 AI 기술이 더욱 확산되는 현상으로 이어질 것이다.

이런 과정에서 AI 신뢰성 관련 기술표준은 AI 산업 생태계 이해관계자들에게 중요한 가이드라인이 될 것이다. 이를 바탕으로 AI 신뢰성 확보는 AI 기술·산업 혁신으로 이어질 전망이다.

4. AI 신뢰성 글로벌 표준 동향(ISO/IEC JTC1/SC42를 중심으로)

현재 AI 분야 글로벌 표준화 논의를 가장 활발하게 진행하는 기구로는 ISO/IEC JTC1/SC42가 꼽힌다. AI를 다루는 표준화 기구 중 가장 많은 국가의 전문가가 참여하고 있으며, 표준 제정 수(28개)와 표준화 작업 수(33개) 역시 가장 많다. 특히 지난 4월에는 서울에서 13차 총회가 열렸는데, SC 총회로서는 매우 큰 규모인 200여 명이 온·오프라인으로 참석해 AI 표준화 열기를 느낄 수 있었다.

SC42에선 표준화 논의 영역에 따라 총 9개 WG(작업반, Working Group)가 운영되고 있다. 이 중 AI 신뢰성과 관련된 논의를 중점적으로 진행하고 있는 작업반은 WG3다. 그간 10개의 표준을 제정했고, 현재도 8개가 표준화 과제로 채택돼 표준화 작업과 논의를 진행하고 있다.

구체적으로 WG3에선 위험관리에 대한 표준(ISO/IEC 23894:2023, Guidance on risk management), 신뢰성 관련 용어와 개념에 대한 기술보고서(ISO/IEC TR 24028:2020, Overview of trustworthiness in artificial intelligence), AI 신경망의 강건성 평가 방법에 대한 시리즈 표준(ISO/IEC TR 24029-1, 24029-2, Assessment of the robustness of neural networks), AI 시스템의 품질 모델(ISO/IEC 25059:2023, SQuaRE-Quality model for AI systems) 등이 제정된 바 있다. 표준화 논의가 진행중인 건들은 AI 시스템 및 기계학습 모델의 설명가능성(ISO/IEC TS 6254, Objectives and approaches for explainability and inter-pretability of ML models and AI systems), A I 시스템에 대한 인간 감시 방안(ISO/IEC AWI 42105, Guidance for human oversight of AI systems), AI 시스템의 투명성 용어(ISO/IEC DIS 12792 Transparency taxonomy of AI systems) 등이다.

<표 1> ISO/IEC JTC1/SC42 WG

그룹	그룹명	주요 표준
WG1	Foundational Standards	ISO/IEC 22989:2022, ISO/IEC 42001:2023
WG2	Data	ISO/IEC 5259 시리즈
WG3	Trustworthiness	ISO/IEC TR 24028
WG4	Use cases and applications	ISO/IEC 5338, ISO/IEC TR 24030:2024
WG5	Computational approaches	ISO/IEC TS 4213:2023
JWG2	JWG w/ SC7 : Testing of AI-based systems	ISO/IEC AWI TS 29119-11
JWG3	JWG w/ ISO/TC 215 : AI enabled health informatics	-
JWG4	JWG w/ IEC TC65/SC65A : Functional Safety and AI systems	ISO/IEC AWI TS 22440 시리즈
JWG5	JWG w/ ISO/TC37 : Natural Language Processing	-

한편 최근 규범·규제와 관련해 논의가 활발한 작업반은 WG1이다. 여기선 AI와 관련된 개념과 용어표준(ISO/IEC 22989:2022, Artificial intelligence concepts a nd terminology)을 개정하고 있다. 또 WG1은 A I를 개발, 적용, 운영하려는 자가 지켜야 하 는 요구사항 표준(ISO/IEC 42001:2023 Management system)을 작년 제정한 데 이어, ISO/IEC 42001에 명시된 요구사항에 대한 감사 혹은 평가를 수행할 수 있는 조직의 요건(ISO/IEC DIS 42006, Requirements for bodies providing audit and certification of AI management systems), 영향 평가 방안(ISO/IEC DIS 42005, AI system impact assessment) 등을 활발하게 논의하고 있다. 아울러, 아직 표준화 작업 프로젝트로 인정되지는 않았으나, AI 성숙도 모델, AI 감사 방안, 적합성 평가·인증체계 등도 활발하게 논의되고 있어, 향후 내용에 대한 귀추가 주목된다.

이 외에도 AI 신뢰성과 관련된 표준화 작업이 다른 작업반에서 일부 이뤄지고 있다. WG2는 데이터 분야, 특히 데이터 품질 표준 시리즈(ISO/IEC 5259, Data quality for analytics and ML)의 제정을 앞두고 있다. 이어서, 최근 주목받고 있는 생성 AI가 생성한 콘텐츠 및 데이터에 대한 표준(ISO/IEC AWI TR 42103, Overview of synthetic data in the context of AI systems)도 신규 표준화 프로젝트로 채택, 논의를 진행하고 있다.

WG4는 인간과 기계의 협업(ISO/IEC AWI TR 42109, Use cases of human-machine teaming)과 관련된 표준화를 진행하고 있다. 또한 JWG2는 SC7과의 공동 작업반으로, AI 시스템의 테스팅 방법에 대한 표준(ISO/IEC AWI TS 29119-11, Software testing – Part11: Testing of AI systems)을 논의하고 있으며, JWG4는 IEC TC65/SC65A와의 공동 작업을 통해 AI 시스템과 기능 안전(ISO/IEC AWI TS 22440, Functional safety and AI systems)에 대한 시리즈 표준화 작업을 진행하고 있다.

작업반 외 애드혹 그룹(Ad-hoc Group, AHG)도 운영되고 있다. 이는 타 위원회 및 기구와의 연락·논의 등을 위한 임시 그룹이라 보면 된다. 특히, AHG7은 유럽 표준화 기구인 CEN(유럽 표준화위원회, European Committee for Standardization), CENELEC(유럽 전기기술표준화위원회, European Committee for Electrotechnical Standardization)과 맺은 비엔나 협정의 일환으로, 표준

활용 및 개발 협력을 논의하고 있다.

한편 EU는 최근 AI 사무실(AI Office)를 설립하며 AI 규제법의 기술 기준 마련을 위해 표준 활용을 적극 추진하고 있다. 이와 관련해 CEN/CENELEC JTC21이 주요 역할을 수행하고 있어, ISO(국제표준화기구, International Organization for Standardization)/IEC(국제전기기술위원회, International Electrotechnical Commission) 표준이 향후 EU 규제법에는 어떻게 활용될지도 주목해야 하는 부분이다. AHG4는 ISO/IEC JTC1 내 보안 분야의 표준화 위원회인 SC27와의 협력을 논의하고 있다.

SC42는 표준화 작업을 진행하는 한편, 다른 표준화 기구 및 협의체와도 적극적으로 협력하고 있다. 특히, CEN/CENELEC JTC21은 물론이고, IEEE(전기전자공학자협회, Institute of Electrical and Electronics Engineers), ETSI(유럽전기통신표준협회, European Telecommunications Standards Institute) 등 타 표준화 기구와도 공식적으로 협업과 의사소통을 하고 있다. 더불어 SC42는 OECD, WEF(세계경제포럼, World Economic Forum), UN(국제연합, United Nations), UNESCO(유엔교육과학문화기구, United Nations Educational, Scientific and Cultural Organization), WTO(세계무역기구, World Trade Organization)와도 함께하고 있는데, 이는 AI 기술 기준에 대한 전 세계의 관심을 잘 보여준다.

5. 맺음말

2023년 초 생성 AI로 인해 AI 위험이 현실화되고, 글로벌 규범·규제 논의가 본격화된 지 벌써 2년째다. 그전까진 AI 윤리와 신뢰성 확보를 위한 가이드, 원칙, 권고가 많이 나오면서, 선도적으로 모범 사례를 발굴하고 확보하는 것이 매우 중요했다. 이제는 이에 더해 '규범·규제 준수를 위해 실제 신뢰성 확보가 가능한 기술을 개발하는 것'과 이에 필요한 '평가·검증 기술 및 기준을 마련하는 것'이 중요한 시기가 됐다.

앞서 이야기한 바와 같이, 글로벌 AI 패권 경쟁에서 중요한 것은 AI 신뢰성 확보이며, 여기서 기술표준 선점 및 표준화 논의 선도의 중요성은 절대적이다. 그러나 AI 기술이 유례없는 속도로 발전하는 가운데, 이에 대한 기준과 신뢰성 확보 기술을 개발·마련한다는 것은 매우 어려운 부분이다. 조선소와 배를 동시에 만드는 것과 같다는 말이 나올 정도다. 현재 생성 AI와 관련해 제정된 기술표준은 아직 없으며, 최근 글로벌 협의체, 국제기구에서 논의되는 프런티어 모델, 이중용도 모델 등에 대한 엄밀한 개념 정의도 이뤄지지 않았다. 아직 기술표준화를 위해 넘어야 할 산이 많다.

그럼에도 불구하고, 전 세계는 이러한 기술표준의 중요성에 공감하고 있으며, 글로벌 경쟁력을 갖추고자 적극적인 노력을 기울이고 있다. AI 신뢰성 분야 기술표준은 향후 글로벌 AI 패권 경쟁에서의 필수적인 국가 전략자산이라고 생각한다.

우리나라 역시 과학기술정보통신부를 중심으로 AI 신뢰성을 확보할 수 있는 기술지침을 마련하고, 체계를 정립하기 힘써왔으며, 글로벌 논의에도 적극 참여·주도하고 있다. AI 서울 정상회의는 이러한 노력의 성과로 볼 수 있다. TTA 역시 이러한 정책적 노력의 최선단에서 AI 신뢰성 확보를 위한 인증체계 수립, 평가·검증 기술 및 기준 마련, 국제표준화 활동에 많은 노력을 기울이고 있다.

AI 신뢰성 분야 기술표준의 중요성은 앞으로 더욱 커질 것이다. 기술표준 제정, 확산, 활용 등 전 방위에 걸쳐 우리나라의 AI 경쟁력을 확보하고, 글로벌 규범·규제에 선도적으로 대응할 수 있도록 노력하고자 한다.

※ 본 연구와 활동은 "AI신뢰성기반조성" 사업의 일환으로 수행됨.

[참고문헌]

[1] International Scientific Report on Advanced AI Safety, the UK Government(DSIT 2024/009)
[2] Executive Order on the safe, Secure, and Trustworthy Development and Use of Artificial Intelligence(EO 14110, '23.10.30.)[18]
https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

※ 출처: TTA 저널 제213호