Issues & Insights
on Economy, Technology, and Security

발행처 경제기술안보연구원
발행인 김성한

제33호 (26-05) 2026년 05월 12일

미토스 충격과 한국의 사이버 보안 전략

The Mythos Shock and Korea's Cybersecurity Strategy

송경호 (한국전자통신연구원 AI안전연구소 선임연구원)

<초록>

2026년 4월 앤트로픽이 공개한 '프론티어 모델 클로드 미토스 프리뷰'는 사이버 공격을 자율적으로 수행하는 역량에서 이전 세대 모델과 질적으로 다른 비약적 향상을 보여주었다. 미토스 사례는 AI 안전성 문제가 사이버 보안에 미치는 영향이 얼마만큼 심각해질 수 있는지를 보여주는 대표적 사례로 기록될 것이다. 앤트로픽은 이례적으로 출시를 지연하고 소수 핵심 인프라 파트너와의 한정 컨소시엄 프로젝트 '글래스윙'을 출범시켰다. 이번 비공개 결정은 앤트로픽이라는 독특한 행위자에 의한 자율 거버넌스의 산물이자 우연의 결과였다. 그러나 AI를 둘러싼 경쟁이 나날이 심화되고 있다는 점을 고려할 때, 이를 다른 행위자에게 기대할 수 없는 것이 현실이다. 보다 근본적으로, 사이버 역량은 미토스가 명시적으로 목적한 바가 아니라, 범용 AI 모델로서 일반적 코딩・추론・자율성 등 역량 향상의 부산물로 발현된 것이다. 미토스의 충격은 고영향 역량 충격이자 거버넌스 충격이다. 이에 대응하기 위하여 한국은 AI 안전성 평가를 국가적 목표로 재설정하고, 국제 첨단 AI 측정・평가・과학 네트워크(NAAIMES)를 중심축으로 한 다자 평가 네트워크에 적극 참여하며, 프론티어 모델 포럼(FMF) 주요 멤버와의 협력관계와 독자 파운데이션 모델 개발국이라는 이중 좌표축을 활용하는 전략적 위치 설정이 필요하다.

<Abstract>

Anthropic's release of 'Claude Mythos Preview' in April 2026 demonstrated a dramatic leap in autonomous offensive cyber capabilities. The Mythos case will likely be remembered as a defining example of how profoundly AI safety failures can affect cybersecurity. In an unusual move, Anthropic delayed public release and instead launched Project 'Glasswing', a limited consortium involving a small number of critical infrastructure partners. This non-public deployment decision was both a product of voluntary self-governance by a distinct corporate actor and a contingent historical outcome. Given the intensifying competitive dynamics surrounding artificial intelligence, however, such restraint cannot be assumed from other actors. More fundamentally, the cyber capabilities displayed by Mythos were not the result of explicit cyber-oriented optimization, but rather emerged as a downstream consequence of broader advances in general-purpose AI capabilities such as coding, reasoning, and agentic autonomy. The Mythos shock is therefore both a high-impact capability shock and a governance shock. Under these conditions, the Republic of Korea should: reposition AI safety evaluation as a core national strategic objective; actively participate in multilateral evaluation networks centered on NAAIMES (International Network for Advanced AI Measurement, Evaluation and Science); and strategically leverage its dual position as both a member of the Frontier Model Forum (FMF) cooperation ecosystem and a state pursuing sovereign foundation model development.

□ 미토스의 충격

2026년 4월 7일 앤트로픽은 '프론티어 모델 클로드 미토스 프리뷰'(Claude Mythos Preview, 이하 미토스)를 발표하면서 일반에의 공개를 연기하겠다고 공표했다. 미토스가 자율적으로 사이버보안 취약점들을 다수 발견했기 때문이다. 여기에는 27년 묵은 OpenBSD 원격 크래시 취약점[1], 16년 묵은 FFmpeg 취약점[2], 17년 묵은 FreeBSD RPCSEC_GSS/NFS 원격 코드 실행 취약점[3] 등이 포함되어 있었다.

같은 날 앤트로픽은 12개 발족 파트너[4]와 추가 40여 개 핵심 인프라 조직[5]에 한정된 접근권을 부여하는 프로젝트 글래스윙을 출범[6]하는 한편, 1억 달러 규모의 사용 크레딧과 오픈소스 보안 단체에 대한 400만 달러 직접 기부를 약정했다.

발표 6일 만인 4월 13일, 영국 AI 보안연구소(AI Security Institute, 이하 UK AISI)는 미토스에 대한 배포 전(pre-deployment) 평가 결과를 공개했다. 핵심은 UK AISI가 자체 개발한 32단계 기업 네트워크 공격 시뮬레이션 The Last Ones(TLO)에서 미토스가 10회 시도 중 3회 종단 완수에 성공한 첫 모델이 되었다는 것이었다. 이는 인간 전문가 기준 약 20시간 분량의 공격 체인을 미토스가 자율적으로 수행했음을 의미한다. 직전 최고 모델인 클로드 오퍼스(Opus) 4.6이 평균 16단계에 머물렀으며 종단 완수 사례는 없었다는 점을 고려하면, 그야말로 비약적 역량 향상인 셈이다. 이는 전문가 수준 캡처더플래그(Capture-

1) OpenBSD는 보안성을 최우선으로 설계된 유닉스 계열 운영체제로 방화벽·서버·네트워크 장비 등 보안에 민감한 인프라에 폭넓게 사용된다. 해당 취약점은 27년간 보안을 핵심 가치로 삼는 OpenBSD 커뮤니티의 엄격한 검토를 거치고도 발견되지 않았던 것이다.

2) FFmpeg은 거의 모든 영상·음성 처리 소프트웨어의 기반이 되는 오픈소스 멀티미디어 라이브러리다. 해당 취약점은 16년간 무수한 인간 보안 전문가의 검토와 약 500만 회의 자동화 테스트를 거치고도 발견되지 않았던 것이다.

3) FreeBSD는 넷플릭스, 와츠앱 등 대규모 서비스의 백엔드 인프라에 널리 쓰이는 유닉스 계열 운영체제이고, RPCSEC_GSS/NFS는 네트워크 파일 시스템의 보안 인증 프로토콜이다. 해당 원격 코드 실행은 사이버 보안에서 가장 위험한 부류에 해당함에도 불구하고, 해당 취약점이 17년간 발견되지 않았던 것이다.

4) 12개 발족 파트너는 아마존웹서비스(Amazon Web Services), 앤트로픽(Anthropic), 애플(Apple), 브로드컴(Broadcom), 시스코(Cisco), 크라우드스트라이크(CrowdStrike), 구글(Google), 제이피모건체이스(JPMorgan Chase), 리눅스재단(Linux Foundation), 마이크로소프트(Microsoft), 엔비디아(NVIDIA), 팔로알토 네트웍스(Palo Alto Networks)이다,

5) 추가 40여 개 조직은 운영체제·브라우저·커널·주요 오픈소스 라이브러리 등 핵심 소프트웨어 인프라를 구축·유지하는 기관들로, 앤트로픽의 공식 발표에 따르면 자체 시스템과 오픈소스 시스템의 취약점 스캔·보안 강화를 위해 접근권을 부여받았다. 다만 구체적 명단은 공개되지 않았다.

6) 앤트로픽은 글래스윙 참여 조직을 약 50개에서 120개로 확대하려는 계획을 추진하였으나, 트럼프 행정부는 이에 대해 두 가지 근거로 반대 입장을 표명하였다. 첫째는 모델 오용 가능성에 대한 보안 우려, 둘째는 앤트로픽의 컴퓨팅 자원이 120개 기관을 서비스할 경우 미국 정부 자체의 모델 사용 효율이 저하될 수 있다는 운용상 우려이다.

the-Flag, CTF) 과제 73% 성공률로도 확인되는 바이다.[7]

종종 간과되지만, 미토스는 '사이버 보안 특화 모델'이 아니라 '범용AI 모델로서 사이버 보안 영역에도 높은 위험성을 보인 모델'이다. 앤트로픽의 시스템 카드는 미토스가 명시적으로 사이버 역량을 학습시킨 결과가 아니라, 일반적 코딩 • 추론 • 자율성 향상의 다운스트림 결과[8]로 발현되었다는 점을 재확인해 준다.

사이버 역량이 일반 역량의 부산물이라면, 다음 도약은 사이버에 국한되지 않을 가능성이 높다. 사이버 공격 자율 수행을 가능하게 한 우수한 코딩 역량과 연쇄적인 다단계 미션 수행 역량은 다른 영역에도 동일하게 작동할 수 있는 범용 역량이기 때문이다. 결국, 미토스 충격의 더 본질적인 층위는 사이버 보안이 아니라 프론티어 AI 모델 자체의 안전성 확보에 있는 것이다.

□ 고성능 AI의 고영향 역량

앞서 지적한 바와 같이, 미토스는 사이버 공격이라는 형태로 충격을 드러냈지만, 그 근원은 범용 AI 모델의 급격한 역량 상승 자체에 있다. 따라서 미토스 충격의 본질은 AI 안전 특히, 고영향 역량[9]을 가진 고성능 AI 모델의 안전성 확보 관점에서 이해되어야 한다. 사이버 보안의 질문이 '미토스급 모델로 공격받을 때 어떻게 방어할 것인가'라면, AI 안전의 질문은 '미토스급 모델의 고영향 역량을 어떻게 다룰 것인가'로 요약될 수 있다.

미토스의 비공개는 전적으로 앤트로픽의 자율 의사결정, 구체적으로는 책임 있는 스케일링 정책(Responsible Scaling Policy, RSP) v3.0[10]의 AI 안전 수준(AI Safety

7) CTF는 의도적으로 취약점이 심어진 시스템에서 숨겨진 인증 문자열('깃발')을 찾아내는 형식의 사이버 보안 평가 방식으로, 공격·방어 역량을 표준화된 조건에서 측정하기 위해 사용된다.

8) 다운스트림 결과는 AI 모델이 특정 능력을 직접 학습하지 않았음에도 일반 역량 향상의 부산물로 그 능력이 발현되는 현상을 가리킨다. 본 용어는 앤트로픽이 미토스 프리뷰를 발표하면서 사용한 표현이다.

9) 고영향 역량(high-impact capabilities)은 EU AI법에서 정의한 개념으로, 범용 AI 모델이 사회 전체에 중대한 부정적 영향을 미칠 수 있는 시스템적 위험(systemic risk)을 야기할 수 있는 임계 역량 수준을 의미한다. 본 글에서는 사이버 공격 자율 수행·복합 추론·다단계 에이전트 작업 등을 통해 그러한 위험을 야기할 수 있는 프론티어 AI 역량 일반을 지칭하는 개념으로 사용했다.

10) RSP는 앤트로픽이 2023년 9월 최초 도입한 자율 거버넌스 체계로, 모델 역량의 잠재적 위험에 따라 안전·보안 조치를 차등 적용하는 단계적 접근을 표방한다. 앤트로픽은 2026년 2월 24일 RSP v3.0을 발표하여 ASL-4 및 ASL-5 단계를 사전 재정의하였는데, 미토스는 이 새 정책이 적용된 첫 모델이다.

Level, ASL)-4 역량 임계치[11]가 발동된 결과였다. 이번 비공개 결정과 프로젝트 글래스윙 파트너 선정 역시 모두 앤트로픽이 단독으로 수행했다.

이 자율적 결정이 작동한 것은 개발자가 앤트로픽이라는 안전 지향 기업이었기 때문이다. 실제로 오픈AI가 GPT-5.4-Cyber를 신뢰 기반 사이버 접근(Trusted Access for Cyber, TAC) 프로그램을 통해 발표하면서 미토스보다 덜 제한적인 접근 정책을 채택한 사례는, 동일한 역량 수준에서도 기업별 판단과 위험 인식에 따라 전혀 다른 거버넌스 결과가 나타날 수 있음을 보여준다.

그러나 단일 기업이 고성능 AI 모델의 접근권과 배포 범위를 결정하는 구조는 정당성과 책임성 양 측면에서 지속 가능한 답이 될 수 없다. 앤트로픽 RSP v3.0 역시 ASL-4보다 더 높은 역량 단계에서는 단일 기업이 단독으로 안전조치를 시행하기 불가능할 수 있다는 점을 명시하고 있다. 즉, 보다 고도화된 차세대 모델에 대해서는 자율 거버넌스만으로 충분하지 않으며, 국가안보 공동체의 지원과 다자·국가 차원의 거버넌스가 필요하다는 점을 앤트로픽 스스로 인정하고 있다.

문제는 이러한 전환이 요구되는 시점에 오히려 평가와 거버넌스의 시간은 점점 축소되고 있다는 점이다. 필자는 최근 UK AISI의 미토스 평가 실무진과의 면담을 통해, 과거 평균 1~2개월 수준이었던 배포 전 평가 기간이 최근에는 1~2주 수준으로 단축되면서 충분한 평가 시간을 확보하지 못하고 있다는 점을 확인했다. AI 개발 경쟁이 치열해질수록 역량 향상 속도는 더 빨라지는 반면, 평가와 검증에 주어지는 시간은 오히려 줄어들고 있는 것이다. 이는 평가 체계가 역량 발전 속도를 구조적으로 따라가지 못하는 상황이 되고 있음을 시사한다.

역량 확산 속도 또한 이러한 문제를 심화시킨다. 역량 향상의 추세선을 고려한다면 다른 프론티어 개발사들 역시 조만간 미토스 수준에 도달할 가능성이 높다. 에포크 AI(Epoch AI)의 분석에 따르면 클로즈드 모델과 오픈 웨이트 모델 간 역량 시차는 평균 3개월에 불과하다. UK AISI 역시 AI 사이버 공격 역량이 약 4개월마다 두 배씩 증가하는 추세를 보인다고 지적한 바 있다. 결국, 단순히 '제2의 미토스'가 등장하는 문제가 아니라, 훨씬 더

11) ASL은 RSP가 모델 역량을 분류하는 5단계 위계로, ASL-1 표준은 의미 있는 파국적 위험이 없는 시스템(체스 AI 등), ASL-2 표준은 위험 역량의 초기 징후를 보이지만 신뢰성이 낮은 일반 대화형 모델, ASL-3 표준은 비AI 기준선 대비 파국적 오용 위험이 실질적으로 증가하거나 초보적 자율 역량을 보이는 모델, ASL-4 표준은 자율 AI R&D 입문급 작업의 완전 자동화 또는 적정 자원 국가 프로그램의 CBRN(화학·생물·방사능·핵) 무기 개발 능력 실질적 증진이 가능한 역량에 도달한 모델, ASL-5는 그 이상의 역량으로 단일 기업의 안전조치만으로는 통제 불가능한 단계에 적용된다.

강력한 역량을 가진 모델들이 짧은 간격으로 연속 등장할 가능성이 높은 현실인 것이다.

이러한 상황은 기존 역량 측정 패러다임 자체에 대한 재설계를 요구한다. 현재 부동소수점연산(FLOPs) 단위로 표기되는 누적 학습 연산량은 역량의 근사치(proxy)일 뿐 역량 자체가 아니다. The Last Ones와 같은 고난도 다단계 자율 공격 벤치마크가 예상보다 빠르게 따라잡히고 있는 현실은 학습 연산량 기준만으로는 실제 위험 역량을 충분히 설명할 수 없음을 시사한다. 동일한 연산량에서도 추론 컴퓨트 스케일링(inference compute scaling)[12], 도구 사용 능력, 에이전트 구조에 따라 모델의 실제 역량은 크게 달라질 수 있다. 요컨대 동일 모델이라 하더라도 투입할 수 있는 추론 자원에 따라 훨씬 더 위험한 행위자로 변할 수 있는 것이다.

따라서 향후 안전성 확보 의무는 단순한 학습 연산량 기준이 아니라, 실제 역량, 특히 자율적 다단계 작업 수행 역량을 직접 측정하는 평가 체계로 보강되어야 한다. 이는 곧 AI 거버넌스의 초점이 '얼마나 큰 모델인가'에서 '어디까지 자율적으로 수행할 수 있는가'로 이동해야 함을 뜻한다.

□ 한국의 사이버 보안 전략

한국이 미토스로부터 직접 공격을 받지 않은 것은 다행스럽지만, 사실상 이는 요행에 불과했다. 앞서 지적한 바와 같이, 이는 첫 번째 도달자가 안전 지향 기업이었고, 그 기업의 자율적 RSP가 제대로 작동했기 때문에 가능했다.

주지하다시피, 이것이 대응의 전제가 될 수 없다. 나아가 우리가 지금부터 대비해야 하는 것은 미토스 그 자체가 아니라 미토스 이후의 시나리오다. 즉, "만약 안전을 고려하지 않는 행위자가 최첨단(State-of-the-Art, SoTA)급 역량에 먼저 도달하면 어떻게 될 것인가?"를 전제로 대응 방향을 모색해야 하는 것이다.

앞서 살펴본 바와 같이, 미토스 충격의 이면에는 거버넌스 공백이 있다. 미토스 이후 미국 정부조차 앤트로픽과의 관계에서 일정 부분 후퇴하는 모습을 보였다는 사실은 고성능 AI를 보유한 프론티어 기업이 국가와 대등하거나 경우에 따라 우위의 전략적 행위자로

12) 동일 모델이라도 추론 시점에 투입되는 연산 자원(토큰 수, 시도 횟수, 사고 단계 등)을 늘리면 성능이 향상되는 현상으로, 누적 학습 연산량이 동일하더라도 추론 자원에 따라 모델의 실제 역량이 달라질 수 있음을 의미한다. UK AISI의 미토스 프리뷰 평가에서도 1억 토큰까지 추론 예산을 늘렸음에도 성능이 포화되지 않고 지속적으로 상승하는 현상이 관찰되었다.

부상했음을 방증한다. 요컨대, 자율 거버넌스의 한계와 정부-기업 간 권력 비대칭의 심화가 동시에 진행되고 있다는 점, 그리고 이 두 흐름이 미래 시나리오에서 더욱 첨예해질 것이라는 점이 향후 대응 전략의 기본 조건인 것이다.

이러한 조건에서, 한국의 대응 역시 사이버 방어 역량 강화 차원에 머물러서는 안 된다. AI 안전성 확보를 위한 근본적인 출발점은 “모든 국가가 프론티어 모델을 개발할 수는 없지만, 적어도 그것을 안전하게 관리할 수 있는 역량은 가져야 한다”는 것에 있다. 미토스 사례의 거버넌스 공백은 그러한 평가·관리 역량의 부재가 곧 자국의 안전을 외부 행위자의 자율적 선의에 위탁하는 결과로 귀결된다는 점을 분명히 보여준다.

물론, 이는 한국만의 과제가 아니다. 영국, 미국, 일본, 싱가포르, 한국 등이 잇따라 AI 보안연구원(AISI)을 설립한 배경에는 바로 이에 대한 공유된 인식이 자리하고 있다. 요컨대, AISI를 보유한 국가들은 소버린 모델 개발 여부와 무관하게, 프론티어 모델의 위험 역량을 독립적으로 평가하고 그 결과를 정책에 연결하는 국가 차원의 인프라를 갖추는 것을 공통의 목표로 설정하고 있다.

이러한 관점에서 세 가지 대응 방향을 제안한다. 첫째, 고성능 AI의 안전성 확보를 국가적 정책 목표로 재설정해야 한다. 지금까지의 규제 논의가 누적 학습 연산량 중심 접근에 머물렀다면, 앞으로는 실제 역량, 특히 자율적 다단계 작업 수행 역량을 직접 측정하는 체계로 이동해야 한다. 미토스 사례는 동일한 모델도 추론 컴퓨트 스케일링과 도구 사용 환경에 따라 훨씬 더 위험한 행위자로 변할 수 있음을 보여주었다. 따라서 고성능 AI 안전성 확보 역시 단순 연산량 임계치가 아니라, 멀티스텝 자율 작업 • 추론 확장 • 에이전트 행동 평가 등을 포함하는 역량 기반 평가 체계로 보강될 필요가 있다. 이는 곧 평가 인프라에 대한 인력 • 예산 • 권한의 실질적 강화와, 평가 결과를 사이버 안보 정책 결정 과정에 공식적으로 연결하는 거버넌스의 구축을 함께 요구한다.

둘째, 다자 평가 네트워크에 적극 참여해야 한다. 미토스급 모델의 평가와 정보 공유는 더 이상 개별 국가나 기업 차원에서 해결될 수 없다. 정부와의 양자 협상은 개별 기업의 재량 안에서 작동하며, 그 재량은 컴퓨팅 자원의 한계, 보안 우려, 자국 정부와의 우선순위 등 협상국 외부의 변수에 좌우되기에 불안정할 수밖에 없다. 그렇다고 UN이나 OECD 같은 일반 국제기구를 통한 접근이 현실적이지도 않다. AI 역량이 4개월마다 두 배로 증가하는 환경에서 수년 단위의 다자 협상은 사실상 무의미하기 때문이다.

현실적 답은 2024년 11월 샌프란시스코에서 출범한 첨단 AI 측정・평가・과학 네트워크(NAAIMES)[13]를 보다 적극적으로 활용하는 것이다. NAAIMES는 현재 영국이 코디네이터를 맡고 있으며, 한국은 호주·캐나다·EU·프랑스·일본·케냐·싱가포르·영국·미국과 함께 창립 멤버에 속해 있다. 이 네트워크는 평가 방법론의 표준화, 공동 평가 과제, 평가 결과의 회원국 간 공유를 목적으로 하며, 앞서 살펴본 영국 AISI의 미토스 평가 결과 공유가 작동한 것도 바로 이 네트워크가 전제하는 사전 배포 접근권 덕분이었다. 요컨대 한국은 추가적 조약이나 협상 없이 이 인프라를 즉시 활용할 수 있는 전략적 자산을 가지고 있는 것이다.

셋째, 프론티어 모델 포럼(Frontier Model Forum, FMF) 협력관계와 독자 파운데이션 모델 개발국이라는 이중 좌표를 전략적으로 활용해야 한다. 한국은 구글 딥마인드, 앤트로픽, OpenAI, Google 등 주요 프론티어 기업들과 협력 채널을 구축하고 있는 한편, LG, 네이버, KT 등 자체 파운데이션 모델 개발 역량을 보유한 국가이기도 하다. 이는 단순히 외국 모델을 수입·채택하는 국가들과 달리, 한국이 '평가 대상 모델의 수용자'이면서 동시에 '책임 있는 거버넌스의 제공자'라는 이중적 위치에 있음을 의미한다.

이러한 이중적 위치는 한국이 AI 안전 거버넌스 영역에서 영향력을 확대하는데 긍정적 역할을 할 수 있다. 예를 들어, 한국의 개발사들이 앤트로픽의 RSP에 준하는 자율적 능력 임계치 약속을 채택하도록 권장함으로써 '책임 있는 거버넌스의 제공자'라는 위치를 실체화하는 구체적 경로로 활용할 수 있는 것이다.

결국 한국의 전략 목표는 평가 결과의 사후적 수신자가 되는 것이 아니라, 글로벌 AI 안전 거버넌스의 공동 구축자로 자리매김하는 것이어야 한다. 미토스 다음 모델은 머지않아 등장할 것이며, 그 위험은 사이버 영역에만 국한되지 않을 가능성이 높다. 그 시점에 각국이 의존할 수 있는 것은 특정 기업의 자율적 선의가 아니라, 사전에 구축된 평가・정보공유・거버넌스 인프라일 것이다. 초고성능 AI 시대의 사이버 안보는 결국 AI 안전성 거버넌스의 확립을 통해서만 지속 가능하게 달성될 수 있다. 이것이 미토스 충격이 한국에 던지는 핵심적 정책 과제다.

13) NAAIMES(International Network for Advanced AI Measurement, Evaluation and Science)는 2024년 5월 한국이 주최한 AI 서울 정상회의(AI Seoul Summit)에서 채택된 '인공지능 안전 과학에 관한 서울 의향 선언'을 직접적 출범 기반으로 하며, 같은 해 11월 21일 샌프란시스코에서 'AI 안전연구소 국제 네트워크'라는 명칭으로 출범한 후 2025년 12월 영국 주최 회의에서 명칭이 개편되었다.

저자 소개 / BIO

송경호

- 現) 한국전자통신연구원(ETRI) AI안전연구소(AISI) 선임연구원
- 現) 국회입법조사처 자문위원
- 現) UNDP Expert Group on Closing the Language Gap in AI for Sustainable Development 자문위원
- 연세대학교 정치학 박사

Email: songkyungho@etri.re.kr

송경호 박사는 한국전자통신연구원(ETRI) AI안전연구소(AISI) AI안전정책 및 대외협력실 선임연구원으로 재직 중이다. 정치사상, 특히 인권과 민주주의의 개념사를 연구해 왔으며, 현재는 인공지능 규범 및 글로벌 거버넌스를 중심으로 연구와 정책 활동을 수행하고 있으며, 보다 넓게는 AI가 정치와 민주주의, 국제 거버넌스, 그리고 인류의 집합적 삶의 조건을 어떻게 재구성하는가를 탐구하고 있다. 연세대학교 정치외교학과에서 학사, 동 대학원에서 석사 및 박사 학위를 취득했다. 일본 토호쿠대학교 법학연구과 CNDC 프로그램을 거쳐, 연세대학교 BK21 박사후연구원으로 활동했으며, 교육부장관 표창과 연세대학교 우수강의 표창을 수상했다.

Kyungho David, SONG

- (Current) Senior Researcher, AI Safety Institute, Electronics and Telecommunications Research Institute (ETRI).
- (Current) Advisory Committee Member, National Assembly Research Service
- (Current) Member, UNDP Expert Group on Closing the Language Gap in AI for Sustainable Development
- Ph.D. in Political Science from Yonsei University

Dr. Kyungho (David) Song is a Senior Researcher at the AI Safety Policy and Strategic Cooperation Team at the AI Safety Institute(AISI), Electronics and Telecommunications Research Institute(ETRI), Republic of Korea. His academic work has focused on the conceptual history of human rights and democracy within political thought, while his current research and policy activities center on AI norms and global governance. More broadly, he explores how artificial intelligence is reshaping politics, democracy, international governance, and the collective conditions of human life. He received his B.A., M.A., and Ph.D. in Political Science and International Relations from Yonsei University. He also participated in the Cross-National Doctoral Course (CNDC) Program at the Graduate School of Law, Tohoku University, Japan, and he later worked as a postdoctoral researcher in the BK21 program at Yonsei University. He received the Award of Recognition from the Minister of Education and Yonsei University's Outstanding Instructor Award.