인간-AI 결합 에이전트의 시너제틱 인텔리전스 거버넌스 모델 제안

Proposing a Synergetic Intelligence Governance Model for “Human-AI” Agents

유호현 [1]
Hohyon Ryu
태재미래전략연구원
Taejae Future Consensus Institute
hohyon@fcinst.org

유재연 [2]
Jaeyoun You
서울대학교,
태재미래전략연구원
Seoul National University
you.jae@snu.ac.kr

노가빈
Gabin Noh
서울대학교,
태재미래전략연구원
Seoul National University
lvlcs5124@gmail.com

황혜민
Hyemin Hwang
태재미래전략연구원
Taejae Future Consensus Institute
elly25727766@gmail.com

1,2 공동교신저자

요약문

본 연구에서는 개인화된 AI 에이전트들의 협력적 의사결정 구조인 '시너제틱 인텔리전스 거버넌스(SIG) 모델'을 제안한다. 최근 AI 기술은 단순히 개인화 추천 시스템을 넘어, 인간을 대신해 의사결정을 내리는 수준까지 발전하고 있다. 이에 따라, 에이전트들 간 협상 과정과 정치적 행위가 발생할 수 있는 환경을 상정하고, 이러한 에이전트의 합의체 모델을 탐구할 필요가 있다. 본 연구는 인간과 AI 의 결합체가 향후 사회적 의사결정 과정의 최소단위가 될 것으로 보고 이러한 새로운 행위체로 구성되는 거버넌스 모델을 제안하는 것을 목표로 한다. 이를 위해 LLM 기반의 시뮬레이션 연구들과 개인화 에이전트의 방향성, 에이전트 간 상호작용에 대한 이론적 배경들을 체계화해 모델을 뒷받침한다. 개인과 기계의 1 대 1 관계가 중심이 되던 기존의 HCI 관점으로부터 더 나아가, AI 에이전트와 인간의 결합을 사회적 상호작용의 최소단위로 바라보고, 이들간의 상호작용을 연구하는 관점을 제시한다는 점에서 본 연구는 기여점이 있다.

주제어

개인화 AI 에이전트, AI 거버넌스, 시너제틱 인텔리전스 거버넌스 모델

1 서론

ChatGPT 서비스 출시 이후 생성형 AI 모델과 대중 사용자의 접점이 빠르게 늘고 있다. 이러한 흐름은 인간-컴퓨터 상호작용의 새로운 패러다임을 형성하고 있다. 기계가 단순히 데이터를 처리하는 도구에서 벗어나, 정교한 동반자 역할을 수행할 가능성이 높아지고 있다. 특히 생성형 AI 는 인간의 언어적 맥락을 이해하고 이를 기반으로 대화를 이어 나가는 방식에서 큰 진전을 이뤘다. 대화과정에서 발생하는 질의응답 데이터의 양은 무한히 늘고 있고, 이에 따라 사용자와의 상호 작용은 더욱 강화될 것으로 전망된다. 이는 단지 사용자 경험을 향상시키는 데 그치지 않는다. AI 가 의사결정 과정에서 점차 더 큰 역할을 하게 될 가능성도 엿보인다. 이러한 변화는 AI 와 인간의 관계를 기존의 사용자-도구의 관점을 넘어선 파트너십 관점에서 이해할 필요성을 제기한다.

AI 모델이 개인의 취향과 성격, 가치관을 빠르게 학습할 뿐 아니라 이로부터 의견까지 내는 연구와 실험들도 나오고 있다. 예를 들어, Park et al.(2024)의 연구는 대규모 언어 모델(LLM)을 활용해 1 천여 명의 실제 개인의 태도와 행동을 시뮬레이션 하는 에이전트 아키텍처를 제시했다 [1]. 인간 참가자들이 두 주 간격으로 응답한 설문의 결과는 그들의 성향을 학습한 에이전트가 생성한 답변과 85%의 정확도로 일치했고, 성격 특성과 실험 결과를 예측하는 데 있어서도 유사한 성과가 나타났다고 연구진은 밝혔다. 개인화된 에이전트가, 개인 사용자와 긴밀히 연결된 의사결정 주체로서의 잠재력을 가질 수 있다는 지점을 확인할 수 있었다.

이러한 결과물들은 AI 모델을 활용한 디지털 트윈(Digital Twin) 및 소셜 시뮬레이션(Social Simulation)의 가능성을 보여준 연구들과도 맞물린다. You & Suh (2024)는 사회적 맥락을 잘 학습해낸 LLM 이 기존의 사회조사방법론의 한계를 넘어설 수 있다는 가능성을 확인하는 연구를 진행했다 [2]. 오직 LLM 서비스와의 상호작용 만으로도 높은 정확도로 특정 집단의 의견을 추출할 수 있다는 점을 정량적으로 살펴보고, 이로써 높은 조사 비용과 오래 걸리는 시간, 샘플링 편향 등의 문제를 해소할 수 있을 것이라고 연구진은 밝혔다. 이는 AI 가 인간의 태도와 행동을 효과적으로 모방하게 될 경우, 개인 및 집단 행동 연구를 위한 새로운 도구로 활용될 수 있는 가능성을 제시한 Park et al.의 논문과 유사한 흐름

이다. 더 나아가, Demszky et al. (2023)과 Ziems et al. (2024) 의 연구는 AI 가 심리학은 물론 계산사회과학(Computational Social Science) 전반에서 시뮬레이션 툴로서 충분히 활용될 수 있다는 지점을 짚어낸다 [3,4]. 이른바 "LLM 샌드박스(LLM as a Sandbox)" 개념이 녹아 있다. 어느 시나리오에서든 사용자의 행동을 예측하고 사회적 현상을 가늠해보는 공간이자 도구로서 LLM 을 바라볼 수 있다는 점을 포괄적으로 검증했다.

본 연구에서는 개인 사용자와 AI 에이전트가 합치된 주체(dyad)를 협의체의 최소 단위로 보고자 한다. 개인 사용자에게는 개인 맞춤형 AI 에이전트가 있고, 이들은 하나로 엮여서 사회적 의사결정을 하는 가장 작은 단위가 된다. AI 가 단순히 비서 역할 또는 보조적 도구로서만 살펴지는 것이 아니라, 사용자의 사고를 확장하고 의사결정을 공유하며 외부 환경과 소통하는 매개이자 주체로 간주된다는 데서 기존의 도구적 관점과 차이가 있다. 따라서 그저 개인화 정보를 처리하는 에이전트에 그치지 않고 상호 시너지를 낸다는 측면에서, 본 연구에서는 이러한 AI 에이전트를 **시너제틱 에이전트(Synergetic Agent: SA)**라고 명명했다.

본 연구는 SA 가 효과적으로 작동한다는 전제 하에, 이러한 SA 들이 모여 자율적으로 정치적 판단과 의사결정을 내릴 수 있는 구조를 탐구하는 데 중점을 둔다. 실시간으로 인간 사용자와 상호작용하는 SA 들이 모인 협의체가 어떠한 방식으로 새로운 협상 및 의사결정 과정을 만들어낼 수 있는지 그 가능성을 살펴보고, 이를 위한 거버넌스 모델을 제시하고자 한다.

2 SYNERGETIC AGENTS 에 대한 검토

2.1 시너지를 내는 주체로서의 기계

본격적인 검토에 앞서, 우리는 SA 에 대한 정의를 보다 명료하게 짚고자 한다. 이 개념은 특정 상품명이나 사명으로 쓰인 바는 있지만, 본 연구에서 살펴봤을 때 학계에서는 명료하게 정립되지는 않은 것으로 보인다. 우리는 인간과 AI 로 이뤄진 쌍의 내부에서, 보다 적극적으로 개인과 네트워크를 이루는 관계로서의 AI 에이전트를 시너제틱 에이전트라고 정의한다. 단지 개인의 비서 역할을 하는 수동적 존재를 넘어서서, "인간도 그 도구를 손에 쥔 그 순간에 다른 가능성을 가진 존재로 바뀐"다는 관점을 강조하고자 함이다 [5].

따라서 단어 '시너지(synergy)'도 기존 HCI 개념에서 쓰던 협업(collaboration)을 넘어서는 측면에서 활용했다. 함께(sun-) 일한다(ergon)는 뜻의 그리스어인 sunergia 를 어원으로 하고 있는 이 단어는, 사전적으로는 두 개 이상의 조직이나 물질, 주체들이 각각 만들어내는 결과의 합보다 더 큰 영향을 내는 상호작용 또는 협력을 뜻한다[1]. 시너지를 주고 받는 관계로서, 인간 사용자와 AI 에이전트 사이에는 위계가 없고, 서로 묶여 있다. 이 둘의 묶임은 홍성욱(2010)의 고장난 자동차의 비유와 같은 선상에서 볼 수 있다. 평소에는 이종적인 네트워크가 하나의 대상으로 만들어지는 블랙박스(black-box) 같지만, 고장이 나게 되면 비로소 그 차를 구성하는 네트워크를 펼쳐 보게(unfold) 된다는 것이다 [5]. 그만큼 끈끈한 네트워크는 구성물인 행위자들을 들여다보지 않게 할 정도로 블랙박스가 되는 것이고, 인간과 AI 에이전트의 쌍 또한 그만큼 강력하게 묶여 블랙박스로서 행위하게 될 것이라는 점이 본 연구의 강력한 전제다.

SA 또한 하나의 블랙박스다. 이를 펼쳐 볼 경우 여러 행위자로 구성돼 있다. 에이전트 기능을 설계하고 유지·보수하는 개발자, 데이터 및 모델 업데이트를 가능하게 하는 서버와 통신망 인프라, 서비스 제공자, 학습데이터 및 추론 모델, 자본을 대는 투자자 등이 이에 해당한다고 볼 수 있다. 이 행위자들이 네트워크를 이루면서 만들어진 SA 는 고정돼 있지 않고 시공간의 변화에 따라 지속적으로 변화한다. 각각의 행위자들이 빠르게 발전하고, 적극적으로 자본을 투입하고, 규모를 키우고 있다는 지점은 눈여겨볼 부분이다.

SA 를 구성하는 행위자인 모델의 스케일은 규모의 법칙(Scaling Laws)를 바탕으로 갈수록 확장될 것이다 [6]. 더 나은 성능을 확보한 모델 기반 서비스들은 인간 사용자들을 더 다양하게, 더 오래 붙들 것이고, 또다른 행위자인 데이터는 더 많이 쌓일 것이다. 기대감 확대로 자본도 모이면서 더욱 높은 성능의 개인 맞춤형 모델로 클 수 있는 루프가 형성될 가능성도 높아졌다. 이를 보다 정량적으로 예측하기 위해 본 연구에서 소프트웨어 시제품에 대한 반응을 수집하는 플랫폼 '프로덕트 헌트'의 데이터를 분석했는데, 최소 득표수 1 천 건을 넘긴 서비스 300 개 중 74%가 AI 에이전트 및 자동화 기능을 다루는 것으로 나타났다[2]. 해당 서비스들은 개인 맞춤형과 생산성 강화의 맥락에서 소비자를 타깃하고 있었다. 개인을

[1] 옥스포드 어학사전 https://languages.oup.com/google-dictionary-en/

[2] 데이터 수집 기간은 2024 년 11 월 28 일 하루이며, 수집 대상은 2023 년 4 월 9 일부터 2024 년 11 월 15 일까지 웹사이트(https://www.producthunt.com)에 올라온 서비스 65,819 개다. 사용자의 반응을 뜻하는 득표수의 최솟값은 0, 최댓값은 7482 이며, 평균값은 51, 중간값은 7 이다. 득표를 1 천 건 했다는 뜻은 '오늘의 프로덕트'에 선정될 수 있는 최소요건으로 알려져 있다.

더 잘 학습하고 추론하는 것이 더욱 핵심적인 지표로 거듭나고 있다.

위 이론 검토와 분석을 종합할 때, 시간이 흐를수록 향후 에이전트들은 단지 도구로서만 머무르지 않을 것이다. 이를 구성한 행위자들에 의해 더욱 사용자의 생산성을 높이고, 시너지를 내는 방향으로 발전할 가능성이 높다.

2.2 SA 와 행위자로서의 인간 사용자

인간 사용자가 SA 와 네트워크를 이루어 쌍을 이룰 것이라는 전망에 대해, 이론적 흐름은 아래와 같이 탐색할 수 있다. 먼저 Merleau-Ponty (1945)는 도구가 신체의 연장(extension)으로 작동할 때, 주변 세계를 탐지하는 감각기관처럼 작동하면서 신체의 일부로 받아들여진다(incorporation)고 설명한다 [7]. 이러한 관점은 SA 를 단순한 기술적 도구가 아닌, 인간의 정신적 활동을 공유하고 확장하는 주체로 이해하게 한다. SA 는 물리적 행동의 보조 도구를 넘어, 인간의 사고 과정에 깊숙이 개입하여 의사결정을 지원하거나, 더 나아가 독립적으로도 임시적 결정을 내릴 수 있는 잠재력을 지닌다. 이를 통해 인간과 AI 의 관계는 인간의 측면에서 봐도, 전통적인 사용자-도구 관계를 넘어선 상호 의존적인 파트너십으로 재구성될 수 있다.

Merleau-Ponty 가 확장적 신체로서 도구를 다루기는 했지만, 여전히 인간과 사물의 경계가 있다. 이를 넘어서서, 사이보그 이론은 이들의 관계를 보다 융합의 차원에서 바라본다. Haraway (1985)의 사이보그 매니페스토는 기술과 인간의 경계가 없는 정체성의 가능성을 상정하며, 기술이 기존 구조를 전복하고 새로운 정치적 가능성을 열어내는 해방의 도구로도 살펴볼 수 있다고 주장한다 [8]. 그러한 관점에서 SA 는 개인 인간 사용자의 데이터를 단순히 분석하는 데 그치지 않고, 인간 사용자의 지식을 더욱 강화하도록 독려하는 동시에 그의 의견을 대변하는 역할을 해낸다. SA 와 개인 사용자의 결합은 새로운 사이보그 주체로 검토할 수 있고, 더 나아가 사이보그를 협의체의 최소 단위로 하는 거버넌스의 그림으로도 확장할 수 있다.

이러한 이유로 본 연구의 SA 에 대한 접근은, AI 와 인간의 관계를 긍정적인 파트너십이자 심지어 한 몸으로 바라본다는 점에서 최근 연구들과 관점이 다를 수 있다. 예컨대 AI 를 도구적 관점, 나아가 전복 가능성을 지닌 경쟁 대상으로 살펴본 Bostrom(2017), Tegmark(2017)의 이론과는 차이를 보인다 [9,10]. Bostrom 은 AI 가 인간 지능 수준을 초월하면서 지능 폭발이 발생하게 되고, 따라서 인간은 AI 를 통제하기 어려워질 것이라고 주장한다. Tegmark 또한 인간의 신체와 지식 모두 설계 가능해지게 되면서, AI 통제불가 상태가 심화할 수 있다고 말한다. 두 연구 모두 인류의 미래를 형성하는 가장 강력한 도구이자 별개의 개체로서, 지배자와 피지배자의 프레임 안에서 AI 를 바라보고 있고, 이에 따라 AI 가 인간과 인간사회의 도덕적 가치를 따를 수 있도록 설계해야 한다고 핵심적으로 주장한다. 본 연구는 AI 와 인간을 경쟁하는 개별적 주체로 두기 보다는, 상호 융합이 필요한 파트너 관계로 살펴본다는 점에서 관점의 차이가 있으나, 인간의 가치를 일치시키는 얼라인먼트(alignment) 관점은 모든 상호작용 과정에서 필수적으로 검토돼야 한다는 점을 적극적으로 반영하고자 한다.

SA 를 독립적 행위자로 본다는 지점에서 본 연구의 관점은 앞서 지속적으로 언급한 Latour 의 행위자-네트워크 이론(Actor-Network Theory: ANT)과도 상통한다 [5]. AI 를 인간과 함께 네트워크를 구성하는 동등한 행위자로서 상정한다는 지점에서다. 인간과 비인간을 구분하지 않고, 다른 요소들과 관계를 형성하며 영향을 미친다는 점에서, 이미 AI 는 그 역할을 추천 시스템과 LLM 기반 기술들로 이행하고 있다. 우리는 더 적극적인 관점에서, SA 를 사회를 구성하는 물질이자 인간과 함께 네트워크를 이루는 행위자로서 다차원적으로 역할을 해내는 주체로 정의하고자 한다.

3 에이전트 간 거버넌스 모델

본 연구에서는 SA 를 개별화된 인간 사용자의 능력을 증강시키는 도구 관점에 그치지 않고, 사용자의 의사결정을 대변하는 주체이자 네트워크 안의 하나의 행위자로서의 역할로도 확장한다. 이러한 행위자로서, 인간 뿐 아니라 SA 또한 다수를 이루며 사회적 의사결정을 할 경우, SA 들 간의 거버넌스에 대한 검토도 필요하다.

먼저 우리는 거버넌스에 대한 개념을 Rhodes (1996)의 거버넌스 모형을 중심으로 정의하고자 한다 [11]. 사전적 의미로 거버넌스는 '정부, 기업, 시민사회 등 다양한 주체들이 공동의 목표를 달성하기 위해 협력하고 상호작용하는 통치 및 관리 체계'를 뜻한다. 여기에서 Rhodes 모델은 **다양한 행위자가 참여하는 자기 조직화된 네트워크**로서의 거버넌스를 강조했다. 이전까지는 정부를 중심적인 주체로 둔 통치 매커니즘이 강조돼 왔지만, Rhodes 는 권력이 여러 행위자에게 분산되고 정책은 자율적인 협력으로 형성된다는 것을 골자로 한다는 점에서 차별화 됐다. 앞서 Latour 의 행위자 네트워크에서 논의한 바와 같이, 비인간 또한 하나의 행위자라고 할 때, 본 연구의 SA 와 같은 AI 에이전트 또한 협력적 관계의 행위자로서 바라볼 수 있다.

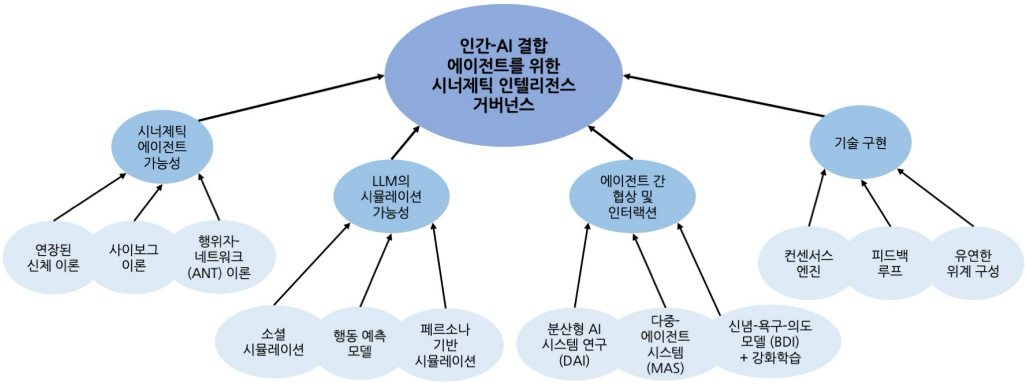

그림 1. 시너제틱 인텔리전스 거버넌스를 위한 이론적 검토 체계

3.1 에이전트 간 협력 구조에 대한 이론

개별화된 에이전트들 사이의 협력과 의사결정, 구조 설계에 대한 논문은 1970 년대 분산 AI 시스템(Distributed Artificial Intelligence: DAI) 분야에서 연구되기 시작했다. 특히 분산 문제 해결이라는 맥락에서 초기 논문들이 나왔는데, 여러 개의 자율적 노드가 서로 협력해 문제를 해결하는 것이 중점적으로 다뤄졌다. 당시에는 인간의 문제 해결 방법을 그대로 모사한 Newell & Simon (1957)의 연구를 이론적 바탕으로 삼아, 분업을 하듯 각 노드에서 큰 문제의 부분들을 각각 해결한 뒤 이를 취합해 효율을 높이는 시도가 중점적으로 검토됐다 [12].

이후 1980 년대에 이르러 등장한 다중 에이전트 시스템(Multi-Agent System: MAS)은 단순한 병렬 처리가 아닌 협력과 조정의 매커니즘을 제시하며 패러다임을 바꾸었다. Georgeff(1984)의 연구가 대표적인데, 기존에 네트워크 노드나 프로세스에 중심을 두던 DAI 의 틀을 넘어선 '에이전트(Agent)' 개념을 등장시켰다 [13]. 이 연구에서는 에이전트를 자율적이고 의도적으로 행동하는 주체로 정의하고, 이들의 협력과 동시성 문제를 이론적으로 다룰 수 있는 모델을 제시해 이전 연구들과의 차이를 만들었다.

이후 MAS 연구는 신념-욕구-의도(Belief-Desire-Intention)을 주축으로 하는 BDI 모델을 제시하는 방향으로 확장됐다. 에이전트가 지닌 정보와 지식, 타 에이전트에 대한 **신념**을 바탕으로, 에이전트가 달성하고자 하는 **욕구**를 향해, 실제 달성을 위해 선택한 구체화된 **의도**에 기반하여 작동한다는 매커니즘이다 [14, 15, 16]. 이러한 이론적 바탕은 로보틱스와 게임 NPC, 소프트웨어 에이전트 등에서 꾸준히 활용되고 있다.

BDI 모델이 목표 지향적 시스템에서 주로 검토되고 있는 한편, 강화학습은 데이터를 기반으로 보상을 통해 환경에 빠르게 적응하며 최적 행동을 학습한다는 점에서 차이를 보인다. 2000 년대에는 BDI 의 의사결정 과정에 강화학습을 적용해 동적 환경에서의 적응성을 향상시키는 연구들이 등장했다. 에이전트의 욕구와 의도를 명시적으로 유지하는 동시에, 신경망 기반의 강화학습을 활용해 환경 변화에 적응하며 계획을 생성하고 평가할 수 있도록 만든 하이브리드 모델들이다 [17, 18].

본 연구에서는 에이전트들이 환경과 상호작용하며 실시간으로 조직적 의사결정을 하는 모델에 대한 이론적인 배경을 바탕으로 하이브리드 모델을 시스템에 반영하고자 한다.

3.2 협상하는 주체로서의 에이전트에 대한 검토

최신 언어모델들이 맥락적이고 직관적인 소통도 할 수 있을 뿐만 아니라 페르소나를 높은 정확도로 반영한다는 연구가 이어지면서, LLM 모델을 에이전트로 활용하는 협상 시뮬레이션 연구 또한 증가하고 있다.

먼저 현실 세계 인물의 페르소나로 만든 에이전트 간 협상을 다룬 논문이 있다. Baker & Azher (2024)는 미국 상원의원 개개인을 학습한 에이전트들끼리 모의 토론을 시켰더니 초당적인 해결책을 찾아냈다는 내용의 논문을 발표했다[19]. 에이전트들은 높은 정확성으로 자신의 행동을 요약했고, 심도 깊은 성찰을 보여줬다. 초기에는 에이전트마다 각기 다른 의견을 품고 있더라도, "러시아의 우크라이나 침공이 임박했다"는 식의 외부적인 요인의 등장에 반응하며 초당적으로 협력하는 것으로 나타났다. 인간과 마찬가지로 변동 요인에 의해 영향을 받는 결과를 보인 것이다. 해당 연구는 에이전트들이 충분히 개인의 특성을 반영하며, 스스로 의견을 성찰할 수 있고, 환경별로 유연하게 협상에 나설 수 있는 가능성을 보였다.

외교와 같이 복잡하고 방대한 결정 공간을 필요로 하는 다중적인 환경에서 장기적인 계획을 수립할 수 있도록 하는 AI 에이전트를 개발한 연구도 있다 [20]. 해당 실험을 토대로 연구진은 기억과 성찰을 통해 전략적으로 계획할 수 있고, 사회적 추론으로 목표 지향적인 협상을 해낼 수 있으며, 인간의 개입 없이 알아서 시나리오를 돌려

스스로 성능을 향상시킬 수 있는 세 가지 핵심 역량이 다중 에이전트 개발에서 가장 중요한 요소임을 주장했다.

개별 LLM 에이전트의 협상 능력 자체를 평가하고 조사하기 위한 프레임워크도 연구되었다 [21]. 공유자원의 할당 문제와 거래 게임, 가격 협상 등의 시나리오에서 LLM 의 행동을 평가하는 방식이다. 이 때 에이전트들이 특정 행동 전술을 사용할 경우 협상 결과를 크게 향상시키기도 했는데, 가령 실제 인간처럼 비참하고 절박한 척을 함으로써 에이전트가 다른 에이전트와의 협상에서 20%의 이익 향상을 달성하기도 했다고 연구진은 밝혔다. 개별 에이전트들이 다양한 협상 전략을 펼쳐두고 최대 이익을 거둘 수 있는 전략을 선택하게 하는 점은 SA 의 설계에서도 고려할 부분이라고 판단했다.

4 모델 제안: SYNERGETIC INTELLIGENCE GOVERNANCE MODEL

본 연구에서는 LLM 의 소셜 시뮬레이션 성능이 충분히 올라오고 있고, 인간 사용자를 반영하는 에이전트(SA)의 출현도 빠르게 다가오고 있으며, 에이전트 간 협상을 다루는 연구도 빠르게 증가하고 있다는 지점을 〈그림 1〉에서 체계적으로 확인했다.

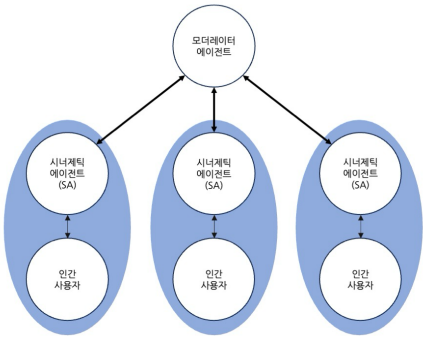

그림 2. 시너제틱 인텔리전스 거버넌스 모델의 구조

이러한 이론적 검토와 데이터 분석을 바탕으로, 본 연구에서는 AI 에이전트와 개별화된 인간 사용자의 조합을 최소 의사결정 단위로 하는 **시너제틱 인텔리전스 거버넌스** (Synergetic Intelligence Governance) 모델을 제안하고자 한다. 〈그림 2〉에서 볼 수 있는 바와 같이, 이 모델은 개인과 접점을 가지는 개별 AI 에이전트(SA)가 개인 인간 사용자의 의사결정 과정에 시너지를 만들어 내고, 사용자를 대신하여 외부의 타 에이전트들과 협의를 통해 의사결정을 진행한다. 인간-AI 의 쌍이 하나의 블랙박스로 움직이며 다른 블랙박스들과 소통하되, 그 창구는 SA 인 셈이다. 시스템적으로는 의제를 관리하고 의사결정에 소요되는 시간 및 협상의 규칙을 조율하는 모더레이터 에이전트가 이 거버넌스 모델 안에 포함된다.

구체적으로 보면 개인 인간 사용자는 자신의 SA 를 통해 모더레이터 에이전트로 의제를 발의할 수 있고, SA 는 인간 사용자의 의견을 구체화하거나 혹은 특정 문제를 의제로 발의하도록 의도할 수 있다. 의제는 모더레이터를 통해 다른 SA 들에게 전달되고 각 SA 는 각자의 개인 인간 사용자와 소통해 모더레이터에게 전달한다. 이 순환은 한 번에 그치지 않고 정해진 규칙에 따라 수 차례 지속될 수 있다. SA 를 매개로 진행된 협상의 결과물은 이들이 소속한 커뮤니티 안에 룰로 정해지게 된다.

향후 연구를 위해 각 상호작용에서 구체적으로 고려해야 할 점을 서술하면 다음과 같다. 첫째, 개인 사용자와 SA 의 상호작용에서 사용자의 지식을 증강하는 동시에 의사결정의 편의를 높이는 인터랙션 디자인이 필요하다. 사용자의 지속가능한 의사결정 참여를 위해서다. 둘째, 인간 사용자의 개입에 대한 디자인이다. 에이전트 자체가 택할 협상 전략에 대한 부분을 어디까지 자동화하고 사람을 개입시킬 것인지도 검토해야 할 요소다.

컨센서스 엔진을 어떻게 구성하느냐에 대한 검토도 필요하다. 각 SA 들의 협상에 있어 게임이론이나 경매 기반 협상 모델, 다수결 원칙 같은 규칙 기반 모델 등을 이슈에 따라 달리 활용할 수 있을 것이다. 마지막으로 SA 의 협상 결과를 받아든 사용자의 수용자 연구도 필요하다. 휴먼인더루프의 방식으로 피드백을 학습시키는 매커니즘에 대한 검토부터, 의사결정을 원점으로 돌릴 권한을 줄 것인지 같은 고려도 필요하다. 향후 연구에서는 이러한 거버넌스 모델을 반영한 플랫폼 서비스로 실제 사례 연구를 진행할 예정이다.

본 연구는 SA 가 기존 개인화 에이전트들이 취해 온 도구적인 양상과 다른 방식으로 발전할 것이라는 전망을 과학기술학과 철학, 데이터분석을 가로지르며 검토했다. 이 과정에서 이론적 부족함으로 편향 및 한계가 있을 수 있다. 리뷰와 피드백을 바탕으로 향후 연구에서 더 심층적으로 보완할 예정이다. 실제 실험에 진입할 때는 모든 과정에서 인간 사용자가 섣불리 소외되지 않고, 주체성을 지키며, 지속적으로 협의에 참여하도록 만드는 윤리적 디자인을 고안할 계획이다.

5 결론

본 연구는 개인화된 AI 에이전트들이 중심이 되어 협상하고 의사결정 하게 될 미래에 대한 전망을 바탕으로, 해당 생태계를 위한 거버넌스 모델인 시너제틱 인텔리전스 거버넌스 모델을 제안했다. 특히 AI 에이전트가 별개의 도구로만 활용되는 것을 넘어서서, 함께 시너지를 낼 수 있는 행위자인 SA 로서의 가능성을 짚었다는 점에서 본 논문은 차별점이 있다. 나아가 인간 사용자와 SA 를

묶은 단위가 시스템 안에서 새로운 상호작용의 최소 단위가 될 수 있음을 제안함으로써 HCI 학계에도 새로운 관점을 제시했다. 향후 해당 모델이 어떻게 작동했을 때 가장 합리적인 결정을 내릴 수 있는지, 인간 사용자를 소외하지 않는 시스템의 고려사항은 무엇인지, 그리고 기존의 의사결정 시스템들에 미칠 영향은 어떻게 될 지에 대한 실험을 지속할 계획이다.

참고 문헌

1. Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., & Bernstein, M. S. (n.d.). Generative Agent Simulations of 1,000 People.

2. You, J., & Suh, B. (2024). Utilizing Large Language Models for Social Simulations: Responding to the Degree of Freedom Questionnaire with ChatGPT. *Journal of the HCI Society of Korea*, 19(3), 49-59. https://doi.org/10.1976-0671

3. Demszky, D., Yang, D., & Yeager, D. S. et al. (2023). Using large language models in psychology. *Nature Reviews Psychology*, 2, 688-701. https://doi.org/10.1038/s44159-023-00241-5

4. Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, 50(1), 237-291.

5. 라투르, 브루노 외. (2010). *인간-사물-동맹*. (홍성욱 엮음). 서울: 사이언스북스.

6. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., & Sifre, L. (2022). Training Compute-Optimal Large Language Models. arXiv preprint arXiv:2203.15556.

7. 메를로퐁티. (2002). *지각의 현상학* (류의근 역). 서울: 문학과지성사. (원저 1945 년 출간)

8. Haraway, D. (1985). A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century. *Socialist Review*, 80, 65-107.

9. 보스트롬, 닉. (2017). *슈퍼 인텔리전스: 경로, 위험, 전략.* (신상규 역). 서울: 까치글방.

10. 테그마크, 막스. (2017). *라이프 3.0: 인공지능 시대에 인간으로 산다는 것.* (강동혁 역). 서울: 동아시아.

11. Rhodes, R. A. W. (1996). The New Governance: Governing without Government. *Political Studies*, 44(4), 652-667. https://doi.org/10.1111/j.1467-9248.1996.tb01747.x

12. Newell, A., & Simon, H. A. (1957). *Human Problem Solving.* RAND Corporation.

13. Georgeff, M. (n.d.). A Theory of Action for MultiAgent Planning. *Artificial Intelligence Center*, SRI International, Menlo Park, California.

14. Bratman, M., Israel, D., & Pollack, M. (1988). Plans and resource-bounded practical reasoning. *Computational Intelligence*. https://doi.org/10.1111/j.1467-8640.1988.tb00284.x

15. Rao, A. S., & Georgeff, M. P. (1991). A Logical Framework for Modeling BDI Agents. In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*.

16. Padgham, L., & Winikoff, M. (2004). *Developing Intelligent Agent Systems: A Practical Guide.* John Wiley & Sons. ISBN 978-0-470-86120-8.

17. Tan, A. H., Ong, Y. S., & Tapanuj, A. (2011). A hybrid agent architecture integrating desire, intention and reinforcement learning. *Expert Systems with Applications,* 38(7), 8477-8487.

18. Pulawski, S., Dam, H. K., & Ghose, A. (2021). BDI-Dojo: Developing Robust BDI Agents in Evolving Adversarial Environments. In *2021 IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS)* (pp. 257-262).

19. Baker, Z. R., & Azher, Z. L. (2024). Simulating The U.S. Senate: An LLM-Driven Agent Approach to Modeling Legislative Behavior and Bipartisanship. *ArXiv*, abs/2406.18702.

20. Guan et al. (2024). Richelieu: Self-evolving LLM-based agents for AI diplomacy.

21. Bianchi, F., Chia, P. J., Yuksekgonul, M., Tagliabue, J., Jurafsky, D., & Zou, J. How Well Can LLMs Negotiate? NEGOTIATIONARENA Platform and Analysis.