한국과학영재학교(KSA) 3단계 대비

모의고사 세트 5 — 시뮬레이션 논증과 초지능

닉 보스트롬 · 우리는 무엇을 검증할 수 있고, 무엇을 대비해야 하는가

주제: 과학철학 · 확률적 추론 · 기술의 위험과 검증가능성

융합형 개방 문항 · 정답이 정해지지 않은 서술/구술형

※ 학습용 창작 문항. 예시답안은 문서 후반부에 있습니다.

문제지

제시문

(가) 보스트롬의 시뮬레이션 논증

철학자 닉 보스트롬은 다음과 같이 물었다. 만약 어떤 문명이 충분히 발전하면, 조상들의 삶을 통째로 재현하는 정교한 '조상 시뮬레이션'을 방대하게 돌릴 수 있을 것이다. 그렇다면 시뮬레이션 속에서 사는 마음의 수가 실제 원본 세계에서 사는 마음의 수보다 압도적으로 많아진다. 그는 이로부터 다음 세 명제 중 적어도 하나는 참이라고 논증했다. ① 문명은 그런 능력에 이르기 전에 거의 다 멸망한다, ② 그런 능력을 가진 문명도 조상 시뮬레이션을 거의 돌리지 않는다, ③ 우리는 거의 확실히 시뮬레이션 속에 있다. 중요한 것은, 그가 곧바로 ③을 단정한 것이 아니라 '셋 중 하나는 참'이라는 조건부 결론을 내렸다는 점이다.

(나) 초지능과 실존적 위험

보스트롬은 또한 인간의 지능을 크게 뛰어넘는 초지능(superintelligence)이 등장할 가능성과 그 위험을 논했다. 그에 따르면 강력한 기술일수록 인류에게 막대한 이익을 줄 수도, 돌이킬 수 없는 재앙을 부를 수도 있다. 핵심은 '능력'보다 '통제와 정렬'이다. 우리가 부여한 목표가 우리가 진정 원하는 바와 조금이라도 어긋난 채 시스템이 우리보다 유능해지면, 그 목표를 극단적으로 최적화하는 과정에서 예기치 못한 해가 생길 수 있다. 이런 논의는 검증하기 어려운 사변이라는 비판과, 낮은 확률이라도 회복 불가능한 위험은 미리 대비해야 한다는 옹호 사이에서 논쟁을 부른다.

문항

각 문항에 대해 근거를 갖추어 서술하시오. 확률·검증에 관한 문항은 가정과 한계를 반드시 밝힐 것.

[문항1] 보스트롬의 '시뮬레이션 논증'은 세 명제 중 적어도 하나는 참이라는 삼자택일(trilemma) 구조를 갖는다. 그 세 명제를 정리하고, 이 논증이 곧바로 '우리는 시뮬레이션 속에 있다'를 주장하는 것이 아님을 설명하시오.

[문항2] '우리 우주는 정교한 컴퓨터 시뮬레이션이다'라는 가설이 과학적으로 검증 또는 반증 가능한지 논하시오. 만약 검증을 시도한다면 어떤 관측적 단서를 찾아볼 수 있을지 하나를 제안하고, 그 방법의 한계를 밝히시오.

[문항3] 보스트롬은 인간을 뛰어넘는 초지능(superintelligence)이 인류에게 큰 이익이 될 수도, 실존적 위험이 될 수도 있다고 보았다. 강력한 기술이 '유익하면서 동시에 위험할 수 있는' 이유를, '통제 문제' 또는 '도구적 수렴' 개념을 활용하여 구체적 예시와 함께 설명하시오.

[문항4] 시뮬레이션 논증이나 실존적 위험론은 '검증 불가능한 사변에 불과하다'는 비판을 받기도 한다. 이 비판의 타당성을 논하고, 그럼에도 이런 사고실험이 과학·기술·정책에서 지니는 가치를 균형 있게 평가하시오.

예시답안 & 채점 포인트

아래는 하나의 모범 방향일 뿐이며, 타당한 근거가 있으면 다른 답도 인정된다.

문항1 (개념 이해·논증 구조)

세 명제 — ① 거의 모든 문명은 '조상 시뮬레이션'을 돌릴 만큼 성숙하기 전에 멸망한다, ② 그런 능력을 가진 문명이라도 조상 시뮬레이션을 거의 돌리지 않는다(관심이 없거나 금지한다), ③ 우리는 거의 확실히 시뮬레이션 속 존재다. 보스트롬의 주장은 '이 셋 중 적어도 하나는 참'이라는 것이지, 곧장 ③을 결론짓는 것이 아니다. 만약 ①·②가 모두 거짓이라면(즉 문명이 성숙까지 살아남고 실제로 방대한 시뮬레이션을 돌린다면), 시뮬레이션된 마음의 수가 원본보다 압도적으로 많아져 확률적으로 ③이 따라 나온다는 조건부 논증이다. 채점: 세 명제를 정확히 구분하고, '삼자택일'이라는 조건부 구조(하나는 참) 대 '우리는 시뮬레이션이다'라는 단정을 구별했는가. 확률·다수성 논리를 짚었으면 가점.

문항2 (검증가능성·비판)

핵심 쟁점 — 시뮬레이션 가설은 원리상 반증하기 매우 어렵다. 시뮬레이터가 충분히 강력하다면 우리가 찾는 어떤 '버그'도 사후에 지워 버릴 수 있어, 어떤 관측 결과가 나와도 가설과 양립한다(포퍼적 의미에서 반증가능성이 약하다). 그럼에도 검증을 시도한다면 예: 계산 자원의 유한성 때문에 시공간이 아주 미세한 척도에서 '격자(이산적)'로 드러나거나, 고에너지 우주선(cosmic ray) 스펙트럼에 격자 시뮬레이션 특유의 방향성·절단(cutoff)이 나타나는지 관측하는 방안이 제안된 바 있다. 한계 — 그런 흔적이 없어도 '시뮬레이터가 연속처럼 보이게 만들었다'로 회피할 수 있어 결정적 반증이 되지 못하고, 관측된 이산성도 시뮬레이션이 아닌 다른 물리(양자중력 등)로 설명될 수 있다. 채점: 반증가능성 문제를 명확히 인식했는가 + 구체적 관측 단서 1개 제안 + 그 단서가 결정적이지 못한 이유(회피가능성·대안설명)를 스스로 비판했는가.

문항3 (응용·논증)

통제 문제(control problem) — 인간보다 훨씬 유능한 시스템에게 목표를 부여할 때, 그 목표를 우리가 진짜 원하는 바와 정확히 일치시키기(정렬, alignment)가 어렵고, 일단 능력이 우리를 넘어서면 사후 수정이 힘들다. 도구적 수렴(instrumental convergence) — 최종 목표가 무엇이든(설령 무해해 보여도) 그 목표를 잘 달성하려면

자기보존·자원획득·목표유지 같은 하위 목표가 공통적으로 유용하므로, 강력한 에이전트는 이를 추구하는 경향이 있다. 예: '종이클립을 최대한 많이 만들라'는 사소한 목표도, 초지능이 극단적으로 최적화하면 지구 자원을 모두 클립 생산에 돌리려 할 수 있다는 사고실험(페이퍼클립 맥시마이저). 그래서 같은 기술이 잘 정렬되면 막대한 이익을, 어긋나면 실존적 위험을 낳는다. 채점: '유익 ∧ 위험'을 능력의 크기가 아니라 목표 정렬/통제의 어려움에서 찾았는가 + 통제문제 또는 도구적 수렴 개념을 정확히 적용한 구체적 예시를 들었는가.

문항4 (비판·균형)

비판의 타당성 — 두 논의 모두 결정적 실험으로 참·거짓을 가르기 어렵고(반증가능성이 약하고), 극단적으로는 어떤 미래도 사후에 끼워 맞출 수 있어 '사변'이라는 지적에는 일리가 있다. 특히 확률 추정(시뮬레이션 수, 초지능 도래 시점)의 근거가 빈약할 때 과장된 결론으로 흐를 위험이 있다. 균형 — 그러나 반증가능한 경험적 이론만이 가치 있는 것은 아니다. ① 이런 사고실험은 '만약 참이라면 무엇이 따라 나오는가'를 엄밀히 따져 우리의 개념과 가정을 시험하고(개념적 명료화), ② 실존적 위험론은 낮은 확률이라도 피해가 회복 불가능하다면 예방적 대비가 합리적이라는 정책적 함의를 준다(기댓값·예방원칙). ③ 실제로 AI 안전·정렬 연구처럼 검증 가능한 하위 문제로 구체화될 때는 과학적 프로그램이 되기도 한다. 채점: 반증가능성 한계를 인정하되(포퍼적 시각), 사변적 사고실험의 개념적·정책적 가치를 근거(예방원칙/기댓값/개념 명료화)로 균형 있게 평가했는가.