

사회과제연구 논문

AI 기반 의사결정 모델과 인간 직관 판단의 불일치 분석

— 스타트업 지원 심사 맥락에서 AHP 기반 가중치 모델을 중심으로 —

대구국제고등학교 3학년 5반 13번 이승준

초록

본 연구는 스타트업 지원 심사라는 다양한 요인이 영향을 미치는 선택에서 개인이 계층분석법을 통해 설정한 판단 기준과 동일한 개인이 구체적인 사례에 대해 내리는 직관적 판단 사이에 유의미한 차이가 존재하는지를 검증하였다. 대구국제고등학교 재학생을 대상으로 실시하여 총 112명의 응답을 수집하였으며, 일관성 비율과 응답 신뢰도 기준에 따른 데이터를 전처리하여 최종적으로는 56명의 자료를 바탕으로 분석했다.

연구 결과, 전체 불일치율은 평균 12.98%로 무작위 수준의 기준값(50%)보다 유의하게 낮았으나, 완전한 일치(0%)는 아니기에 유의미한 수준의 불일치가 실제로 존재한다는 것을 입증하였다. 상황별로는 베이스라인, 단일극값, 변인충돌 순서대로 불일치율이 증가하였으며, 변인 간의 충돌이 클수록 모델이 내리는 판단과 인간이 직관적으로 판단의 차이가 커진다는 가설을 지지하였다. 또한, 불일치의 정도는 항상 일정하지 않았으며, 주어진 사례에 따라 규칙을 보였다. 특정 변인 하나가 극단적인 값을 가진 사례에서는 직관 기반 판단이 모델보다 엄격한 방향을 보였으며, 변인 간의 충돌이 큰 사례에서는 직관 판단이 모델보다 관대한 경향을 보였다. 그러나 개인의 판단 일관성, 가중치 편중도, 모델 기준 이해도를 예측변인으로 하여 회귀분석을 수행하였지만, 해당 변인들을 기반으로 전체 불일치율을 유의하게 설명하지는 못하였다($R^2=.093$, $p=.163$).

본 연구는 인간이 자기 자신이 직접 설계한 판단 모델과도 완전히 일관되게 행동하지 않으며, 그 불일치가 무작위 오차가 아니라 상황적 특성에 따라 체계적으로 발생한다는 것을 실증적으로 보여준다. 이는 AI 정렬 논의 및 인간-AI 협업 시스템 설계에 있어 다기준이 상충하는 상황일수록 인간에 대한 설명과 개입이 더욱 필요하다는 함의를 제공한다.

I. 서론

1. 연구의 필요성

인공지능 기술의 급속한 발전과 함께 다양한 의사결정 맥락에서 알고리즘 기반 추천 및 판단 시스템의 활용이 확대되고 있다. 입사 지원자 선발, 대출 심사, 의료 진단, 스타트업 지원 심사 등 사회적 영향력이 큰 결정 영역에서 AI 모델이 보조적 또는 주도적 역할을 수행하는 사례가 증가하고 있으며(Russell, 2019), 이에 따라 AI와 인간 판단의 관계에 대한 근본적 질문이 제기되고 있다.

AI 의사결정 모델의 핵심은 인간이 중요하다고 생각하는 기준과 가중치를 학습하여 일관되게 적용하는 데 있다. 그러나 인간의 판단은 이와 다른 양상을 보일 수 있다. 인간은 의사결정 기준을 명시적으로 진술할 수 있지만, 실제 사례를 판단하는 순간에는 그 기준과 다른 방식으로 반응하는 경우가 많다. Samuelson(1938)이 경제학적 맥락에서 처음 제안하고 이후 행동경제학 분야에서 광범위하게 검증된 이 괴리는, AI 의사결정 맥락에서도 중요한 함의를 갖는다.

특히 Kahneman(2003)의 이중처리 이론에 따르면, 인간의 인지는 분석적이고 규칙적인 System 2와 직관적이고 자동적인 System 1으로 구분된다. 명시적 판단 기준의 설정은 주로 System 2가 작동하는 과정인 반면, 구체적 사례에 직면했을 때의 즉각적 판단은 System 1의 영향을 강하게 받는다. 이는 동일한 개인 내에서도 명시된 기준과 실제 판단 사이의 불일치가 체계적으로 발생할 수 있음을 시사한다.

이러한 불일치를 실험적으로 측정하기 위해서는 개인의 명시적 판단 기준을 수학적으로 포착하는 신뢰할 수 있는 방법이 필요하다. Saaty(1977)가 개발한 계층분석법(AHP)은 쌍대비교를 통해 의사결정 기준의 상대적 가중치를 도출하는 방법으로, 단순 자기보고식 설문보다 정밀하게 개인의 판단 체계를 포착할 수 있으며 내적 일관성 검증까지 가능하다는 점에서 본 연구의 측정 도구로 적합하다.

그러나 지금까지의 연구는 대부분 외부에서 설계된 알고리즘에 대한 인간의 반응(Dietvorst et al., 2015)이나 임상 판단과 통계 모델의 비교(Grove et al., 2000)에 집중되어 왔다. 개인이 직접 설정한 판단 기준으로 구성된 모델과 동일인의 직관 판단을 비교한 연구는 극히 드물다. 이 지점이 본 연구의 독창적 출발점이다.

2. 연구 목적 및 연구 문제

본 연구는 여러 변인의 영향을 받아 결정되는 스타트업 지원 심사에서 개인이 **AHP**를 통해 명시적으로 설정한 판단 기준과 실제 사례가 주어졌을때 내리는 직관적 판단 간의 괴리를 양적으로 측정하고 분석하는 것을 목표로 한다.

구체적인 연구 문제는 다음과 같다. “피실험자가 **AHP**를 통해 설정한 가중치 모델의 판단과 실제 사례에 대한 직관적 판단 사이에 유의미한 불일치가 존재하는가?”

이를 통해 인간이 판단 방법이 항상 일관되지 않으며, 주어진 상황에 따라 다른 방법을 사용함을 검증하고, 추후 인간의 의사결정을 기반으로 학습한 **AI** 설계에 대해서도 기초적인 의의를 제공한다.

3. 연구의 범위와 한계

본 연구는 동일 학교에 재학 중인 학생들을 대상으로만 시행되는 연구이다. 따라서 결론을 일반화하기에 부적절할 수 있으며, 본 연구에서 활용한 스타트업 지원 심사라는 맥락이 다른 상황에서도 적용되는지에 대한 추가 연구가 필요하다. 또한 **AHP**를 통한 각 요인에 대한 가중치 도출은 피실험자의 명시적 판단 기준을 측정하는 하나의 방법론적 선택이며, 이가 피실험자의 의사결정 기준을 완벽하게 반영한다고 할 수 없다.

II. 이론적 배경

본 연구는 세가지 이론을 기반으로 한다. 첫번째로 선언적 선호와 현시적 선호의 불일치에 관한 행동경제학적 이론을 바탕으로 인간의 의사결정 맥락의 이중성을 전제한다. 둘째로 이중처리 이론을 통해 그 불일치가 왜 발생하는지에 대해 인지과학적인 근거를 제시한다. 마지막으로, 계층분석법(AHP)이 명시적 판단 기준의 측정 도구로서 방법론적인 기틀을 제공한다.

1. 선언적 선호와 현시적 선호의 불일치

가. 개념의 구분

선언적 선호는 개인이 언어나 응답을 통해 명시적으로 표현하는 선호를 의미한다. 반대로 현시적 선호는 실제 주어진 상황에서 추론되는 선호를 가리킨다. 이 두 개념은 **Samuelson(1938)**이 소비자 행동 이론에서 효용이라는 개념을 구체화 하기 위해 실제 선택으로부터 선호를 추론한다는 이론에서 비롯되었다.

그의 이론대로라면 이 두 선호는 동일해야하지만, 실제 연구 결과들은 이들간의 괴리를 보여준다. 특히 행동경제학 보건경제학 분야에서 진행한 실험들에 따르면 피실험자들이 설문이나 실험 상황에서 진술한 선호가 실제 선호를 과대 추정하는 경향이 있다고 한다(**de Corte et al., 2021**).

나. 선행연구의 증거

de Corte et al.(2021)는 보건 분야의 의사결정에서 선언적 선호가 실제 행동을 과대 예측하는 경향이 있음을 밝혀냈으며, 이를 감소시키기 위한 실험 설계 방법 또한 제안하였다. 이 연구는 선언적 선호와 현시적 선호의 불일치가 실험 설계의 오류가 아니라, 본질적인 인간의 특성에서 비롯되었음을 보여준다.

Alberini(2019)는 환경경제학 분야의 여러 연구들을 종합하여, 선언적 선호의 방법론으로 측정된 가치 평가가 실제 선택 데이터와 다르다는 것을 입증했다. 이는 개인이 자신의 선호를 선택 상황이 아닐때에 진술할 때와 실제로 선택할 때 서로 다른 판단 과정이 작동한다는 것을 뒷받침한다.

다. 본 연구에서의 위치

본 연구에서 AHP를 통해 도출된 가중치는 선언적 선호에 해당한다. 피실험자가 5개 변인에 대해 명시적으로 비교하고 수치화한 판단 기준이기 때문이다. 반면 동일한 피실험자가 구체적 스타트업 사례를 보고 즉각적으로 내리는 선발 및 탈락 판단은 현시적 선호에 해당한다. 본 연구는 이 두 선호를 동일 개인 내에서 직접 비교함으로써, 소비자 시장이 아닌 개인의 다기준 의사결정 가중치 수준에서 선언적 선호와 현시적 선호간의 괴리를 측정한다.

2. 이중처리 이론

가. System 1과 System 2

이중처리 이론은 인간의 인지가 질적으로 구분되는 두 가지 처리 과정으로 작동한다고 주장한다. Kahneman(2003)은 이를 System 1과 System 2로 구분하여 체계화하였으며, 이후 Stanovich와 West(2000), Evans와 Stanovich(2013) 등의 후속 연구를 통해 이론적 틀이 정교화되었다.

이에 따르면 System 1은 빠르고 자동적으로 처리되며, 과거의 경험과 감정 등에 기반하여 판단한다. 반대로 System 2는 의식적이고 분석적으로 처리하는 방법으로, 상당한 인지 자원을 토대로 규칙과 논리에 따라 판단한다. 두 시스템은 동시에 작동하며, 대부분의 경우에는 System 1이 먼저 직관적인 결론을 도출하고, System 2가 이를 점검하거나 수정하는 방식으로 상호작용한다(Kahneman, 2011).

나. 두 시스템과 판단 불일치

이중처리 이론은 개인이 동일한 대상에 대해서도 상황에 따라 다른 판단을 내릴 수 있다고 설명한다. 판단 상황에서 벗어나 구체적인 판단 기준을 구상하는 과정은 **System 2**이다. 그러나 반면 구체적 사례를 접했을 때 즉각적인 판단은 도출하는 것은 **System 1**일 가능성이 높다.

특히 **Kahneman**과 **Frederick(2002)**이 제안한 속성 대체 개념은 이 불일치의 메커니즘을 설명하는 데 유용하다. 속성 대체란 **System 1**이 복잡한 다변수 판단 대신 가장 두드러지거나 접근하기 쉬운 단일 속성으로 판단을 대체하는 현상을 가리킨다. 예를 들어 다섯 가지 기준을 가중합하여 판단해야 하는 상황에서, 직관은 가장 인상적인 하나의 기준에 반응하고 나머지를 사실상 무시할 수 있다. 이는 **AHP** 가중치 모델이 모든 변인을 수학적으로 결합하는 방식과 근본적으로 다르다.

다. 본 연구에서의 위치

본 연구에서 **AHP** 쌍대비교 과정은 **System 2** 기반의 판단 과정이다. 피실험자는 각 변인 쌍을 하나씩 비교하며 수치로 중요도 차이를 표현해야 하므로, 의식적이고 분석적인 인지 처리가 요구된다. 반면 스타트업 사례에 대한 직관 판단은 수치 계산 없이 즉각적으로 이루어지므로, **System 1**의 영향이 강하게 개입한다.

따라서 두 판단 간의 불일치는 단순한 측정 오차가 아니라, 동일한 개인 내에서 서로 다른 인지 시스템이 작동할 때 발생하는 구조적 결과로 해석될 수 있다. 이 관점에서 본 연구의 핵심 측정값인 **AHP** 모델 판단과 직관 판단의 불일치율은 개인 내 두 인지 시스템의 출력 차이를 정량화한 것이다.

3. 계층분석법(AHP)

가. 방법론의 개요

계층분석법(AHP)은 **Saaty(1977)**가 개발한 다기준 의사결정 방법론으로, 여러 기준이 동시에 고려되는 복잡한 판단 문제에서 각 기준의 상대적 중요도(가중치)를 수학적으로 도출한다. **AHP**의

핵심 아이디어는 복잡한 다기준 비교 문제를 두 기준씩의 단순 비교(쌍대비교)로 분해하고, 그 결과에서 전체 가중치를 역산하는 데 있다.

쌍대비교는 Saaty가 제안한 1~9 척도를 사용한다. 값 1은 두 기준이 동등하게 중요함을, 9는 한 기준이 다른 기준보다 극히 중요함을 의미한다. n 개의 기준이 있을 때 쌍대비교 횟수는 $\frac{n(n-1)}{2}$ 이며, 본 연구에서 변인이 5개이므로 총 10쌍의 비교가 이루어진다. 본 연구에서는 응답 편의를 위해 이 척도를 -4 ~ +4 범위의 9단계로 변형하여 제시하였다(0은 동등, 부호는 우세한 기준의 방향, 절댓값은 우세 정도를 의미).

나. 가중치 도출과 일관성 검증

쌍대비교 결과는 $n \times n$ 비율 행렬로 구성된다. 행렬의 대각선 원소는 1이고, a_{ij} 와 a_{ji} 는 역수 관계를 갖는다. 이 행렬의 주고유벡터가 각 기준의 가중치를 나타내며, 실용적 구현에서는 각 행의 기하평균을 정규화하는 근사법 또는 반복적 행렬 곱셈을 통한 멱승법이 활용된다(Saaty, 2008).

AHP의 중요한 특성 중 하나는 내적 일관성 검증이 가능하다는 점이다. Saaty(1977)는 일관성 비율(CR)을 제안하였다. CR은 일관성 지수(CI)를 무작위 일관성 지수(RI)로 나눈 값으로, $CR \leq 0.10$ 일 때 응답이 충분히 일관적임을 의미한다. CR이 0.10을 초과하면 해당 응답자의 쌍대비교에 논리적 모순이 있음을 시사하므로, 재입력을 요청하거나 분석에서 별도 처리해야 한다.

$$CI = (\lambda_{max} - n) / (n - 1) \quad CR = CI / RI$$

λ_{max} : 비교 행렬의 최대 고유값, n : 기준 수, RI: 무작위 일관성 지수($n=5$ 일 때 $RI = 1.12$)

판단 기준: $CR \leq 0.10 \rightarrow$ 수용 가능 / $CR > 0.10 \rightarrow$ 재검토 필요

다. AHP의 선언적 선호 측정 도구로서의 적합성

AHP가 본 연구의 Stated Preference 측정 도구로 적합한 이유는 세 가지이다. 첫째로, 직접적인 수치 배분 방식보다 인지 부하가 균일하다. 피실험자는 매번 두 기준만 비교하면 되므로,

동시에 여러 기준을 고려하는 인지적 어려움이 줄어든다. 두번째로, **CR**을 통해 응답의 내적 일관성을 정량화할 수 있어, 응답 신뢰도를 데이터 수준에서 검증할 수 있다. 세번째로, 쌍대비교 결과는 수학적으로 정의된 가중 선형 모델로 즉시 변환되므로, 피실험자의 판단 기준을 직접 반영한 **AI** 모델을 구성할 수 있다.

Saaty(2008)는 AHP를 "쌍대비교를 통한 측정 이론"으로 정의하였으며, 이 방법이 전문가의 판단에서 추상적인 판단 기준을 양적인 상대 비율 척도로 추출하는데에 적합하다고 설명한다. 본 연구에서 피실험자의 **AHP** 가중치는 그 사람의 의사결정 모델의 파라미터가 된다. 이는 개인의 선언적 선호 기준을 수학적으로 표현한 것이라고 할 수 있다.

4. 이론적 틀의 통합

선언적/현시적 선호 이론은 명시적인 판단 기준과 실제 판단 사이에 괴리가 존재한다는 것을 설명함으로써 본 연구의 전제가 된다. 이중처리 이론은 해당 불일치가 왜 발생하는가를 2개의 **System**을 통해 설명한다. 마지막으로 **AHP**는 개인을 선언적 선호를 양적으로 수치화하는지에 대한 방법론을 쌍대비교 계산이라는 방법으로 제시한다.

III. 연구방법

1. 연구 모형 및 가설

가. 연구 모형

본 연구는 동일한 피실험자를 대상으로 선언적 선호와 현시적 선호를 직접 비교하는 연구 방식을 사용한다. 독립변수는 **AHP** 쌍대비교를 통해 도출된 피실험자의 가중치 벡터이며, 이는 개인의 선언적 선호를 양적으로 표현한 것이다. 종속변수는 실제 스타트업 사례에 대한 직관적 판단으로, 이는 피실험자의 현시적 선호에 해당한다.

모델 판단은 다음과 같이 산출한다. 5개 변인의 사례값에 개인별 AHP 가중치를 곱한 가중합으로 모델 점수(0~100점)를 계산하고($\text{Score} = \sum \text{변인값} \times \text{가중치}$), 이를 사전에 설정한 경계값(28 / 42 / 58 / 72점)에 따라 1~5점의 리커트 척도로 환산하여 직관 판단과 동일한 척도로 비교한다. 모델 판단과 직관 판단의 차이가 절댓값 기준 2점 이상일 때 이를 '불일치'로 판정한다.

핵심 측정값은 다음과 같다.

$$\text{불일치율(\%)} = (\text{불일치 사례 수} / 15) \times 100$$

나. 연구 가설

앞서 제시한 이론적 배경을 바탕으로 다음과 같은 연구 가설을 설정한다.

표 1. 연구 가설 목록

가설	내용	이론적 근거
H1	피실험자의 AHP 가중치 모델 판단과 직관 판단 사이에 유의미한 불일치가 존재할 것이다.	Stated vs. Revealed Preference, 이중처리 이론
H2	변인 간 충돌이 극단적인 사례에서 일반 사례에 비해 불일치율이 더 높게 나타날 것이다.	속성 대체 이론
H3	불일치 발생 시 직관 판단은 가장 두드러진 변인의 방향으로 모델에서 이탈할 것이다.	Kahneman & Frederick (2002)

2. 연구 대상 및 표본 추출

가. 모집단 및 표본

본 연구의 모집단은 스타트업 지원 심사의 맥락에서 다기준 의사결정 과제를 수행할 수 있는 청소년이다. 구체적인 표본은 연구자가 재학 중인 대구국제고등학교의 전교생이다.

설문은 2026년 5월 1일부터 5월 30일까지 진행하였으며, 총 112명의 응답을 수집하였다. 표본 크기는 AHP 일관성 비율(CR) 검증 등에서 제외되는 응답을 고려하여 최소 30명을 상회하도록 목표하였다.

나. 전처리 및 최종 표본

수집된 112명의 응답 가운데, 다음 세 가지 기준 중 하나 이상에 해당하는 응답을 분석에서 제외하였다:

1. AHP 쌍대비교의 일관성 비율이 기준치를 초과한 경우($CR > 0.10$)
2. 15개 직관 판단 문항의 응답 표준편차가 0.5 미만으로 나타나 무성의한 응답으로 판단되는 경우
3. 10개 쌍대비교 문항에 모두 동일한 값을 입력한 경우.

이 세 기준을 순차적으로 적용한 결과, 총 56명(50.0%)이 최종 유효 표본으로 확정되어 분석에 사용되었다.

3. 측정 도구

가. 독립변수: AHP 쌍대비교

스타트업 지원 심사의 판단 기준으로 안정성, 수익 가능성, 자금 효율성, 사업 계획 완성도, 사회적 가치의 5개 변인을 설정하였다(각 0~100점, 점수가 높을수록 해당 기준에서 유리함을 의미). 5개 변인에 대해 총 10쌍의 쌍대비교를 -4 ~ +4 범위의 9단계 척도로 실시하였으며, 값이 음수이면 왼쪽 변인이, 양수이면 오른쪽 변인이 더 중요함을, 0은 두 변인이 동등함을 의미한다.

나. 종속변수: 직관 판단

5개 변인의 값이 서로 다르게 조합된 15개의 가상 스타트업 사례 카드를 제작하였다. 사례는 유형에 따라 베이스라인(모든 변인이 고르게 높거나 낮은 사례) 4개, 단일극값(하나의 변인만

극단적으로 높거나 낮은 사례) 5개, 변인충돌(변인 간 방향이 서로 상충하는 사례) 6개로 구성하였다. 피실험자는 각 사례에 대해 5점 리커트 척도(1=매우 탈락~5=매우 선발)로 즉각적인 직관 판단을 하였다.

다. 사후 설문

AHP 결과와 직관 판단 결과를 공개한 후, 모델과 직관 판단의 차이를 어떻게 인식하는지(사후Q1), 자신이 설정한 모델의 기준을 얼마나 잘 이해하고 있다고 느끼는지(사후Q2), 모델의 판단이 얼마나 합리적이라고 느끼는지(사후Q3)를 각각 리커트 척도로 측정하였다.

4. 자료 수집 절차

실험 절차는 직관 판단을 바탕으로 현시적 선호를 수집하고, AHP 쌍대비교를 통해 선언적 선호를 도출한다. 그리고 결과를 보여준 후 사후을 진행하는 순서로 진행되었다. 현시적 선호에 대한 조사를 먼저 시행한 것은 피실험자가 본인의 명시적 가중치를 먼저 고려한 후 상황을 제시한다면 이전에 설계한 기준에 기반하여 System 2가 개입될 가능성이 높기 때문이다. 따라서, 이러한 순서를 통해 직관 판단이 AHP 결과의 영향을 받지 않은 순수한 System 1 반응에 가까울 수 있도록 하였다.

5. 자료 분석 방법

수집된 자료는 Python에서 pandas와 NumPy 라이브러리 등을 이용하여 우선 전처리를 진행하였다. 이후 AHP 가중치 그리고 불일치율 계산하였으며, SPSS를 이용하여 통계 분석을 실시하였다. 유의수준은 .05를 기준으로 하였으며, 연구문제 및 가설별로 다음과 같은 분석을 수행하였다.

첫번째로, 응답자 특성 및 주요 변수의 분포를 파악하기 위해 기술통계 분석을 실시하였다. 두번째로, H1을 검증하기 위해 전체 불일치율 평균을 기준값(무작위 수준을 가정한 50%)과 비교하는 일표본 t검정을 실시하였다. 세번째로, H2를 검증하기 위해 베이스라인, 단일극값, 변인충돌 세 사례

유형 간 불일치율 차이를 대응표본 t검정으로 비교하였다. 네번째로, H3을 검증하기 위해 불일치가 발생한 사례에서 직관 판단이 모델 판단보다 높은 방향(관대)과 낮은 방향(엄격) 중 어느 쪽으로 치우치는지를 사례 유형별로 빈도분석 및 방향성 분석하였다. 다섯번째, 개인의 판단 특성이 전체 불일치율에 영향을 미치는 인과관계를 확인하기 위해 다중회귀분석을 실시하였으며, 보조적으로 CR과 불일치율 간의 상관분석을 실시하였다.

IV. 연구 결과 및 해석

1. 응답자 특성 및 자료 정리

설문에는 총 112명이 응답하였다. 이 가운데 AHP 일관성 비율($CR > 0.10$), 직관 판단 응답의 낮은 변산성($SD < 0.5$), 쌍대비교 문항 전체 동일값 응답의 세 가지 기준 중 하나 이상에 해당하는 56명의 응답을 제외하고, 최종 56명(전체의 50.0%)의 자료를 분석에 사용하였다. 직관 판단 15문항은 각 사례별 5점 리커트 응답을 그대로 사용하였으며, AHP 쌍대비교 10문항은 개인별 가중치 벡터(5개 변인)와 일관성 비율(CR)로 정리하였다. 모델 점수는 개인별 가중치와 사례값의 가중합으로 산출한 뒤 1~5점 리커트로 환산하여, 직관 판단과 동일한 척도에서 직접 비교 가능하도록 정리하였다.

2. 주요 변수의 기술통계

분석에 사용된 주요 변수의 기술통계는 <표 2>와 같다. AHP 가중치 중 수익 가능성의 평균 가중치(.352)가 가장 높게 나타나 피실험자들이 전반적으로 수익성을 가장 중요한 판단 기준으로 설정하는 경향을 보였으며, 자금 효율성(.089)의 평균 가중치가 가장 낮게 나타났다. 전체 불일치율의 평균은 12.98%($SD=7.56$)로, 15개 사례 중 평균적으로 약 2개 사례에서 모델 판단과 직관 판단이 갈리는 것으로 나타났다. 사례 유형별로는 변인총돌 사례의 불일치율($M=20.84$, $SD=14.98$)이 단일극값 사례($M=13.57$, $SD=16.67$) 및 베이스라인 사례($M=0.45$, $SD=3.34$)보다 뚜렷하게 높게 나타나, 변인 간 갈등이 클수록 두 판단이 어긋나는 경향을 예비적으로 보여준다.

표 2. 주요 변수의 기술통계량 (N=56)

변수	최소값	최대값	평균	표준편차
가중치_안정성	.0441	.4547	.2079	.1055
가중치_수익가능성	.0746	.5393	.3517	.1399
가중치_자금효율성	.0431	.2591	.0893	.0506
가중치_사업계획완성도	.0661	.4628	.2228	.1017
가중치_사회적가치	.0443	.5100	.1283	.0940
CR	.0247	.0989	.0659	.0209
불일치율_전체(%)	0.0	33.3	12.98	7.56
불일치율_베이스라인(%)	0.0	25.0	0.45	3.34
불일치율_단일극값(%)	0.0	60.0	13.57	16.67
불일치율_변인충돌(%)	0.0	50.0	20.84	14.98
사후Q1_모델직관차이인식	1	3	1.61	.562
사후Q2_모델기준이해도	2	5	4.07	.828
사후Q3_모델합리성평가	1	5	3.20	.923

3. 가설 검증 결과

가. H1: 모델 판단과 직관 판단의 불일치 존재 여부

H1을 검증하기 위해 전체 불일치율의 평균과 무작위 수준을 가정한 기준값 50%와 비교하는 일표본 t검정을 실시하였다. 모델 판단과 현시적 판단이 서로 아무런 관련이 없다면 불일치율은 50%와 유사해야하므로, 해당 기준값과의 비교를 통해서 실제로 관찰한 불일치율 수준이 관련이 없는 상황과 비교하여 얼마나 다른지를 나타낸다고 할 수 있다.

표 3. 불일치율_전체에 대한 일표본 t검정 결과 (검정값 = 50)

변수	M	SD	t	df	p	95% CI	Cohen's d
불일치율_전체	12.98	7.56	-36.670	55	<.001	[-39.05, -35.00]	-4.90

분석 결과, 전체 불일치율($M=12.98$, $SD=7.56$)은 검정값 50%보다 유의하게 낮았다($t(55)=-36.670$, $p<.001$, Cohen's $d=-4.90$). 이는 모델 판단과 직관 판단이 무작위 수준보다 훨씬 높은 일치도를 보인다는 것을 의미한다. 그러나 동시에 <표 2>의 기술통계에서 확인되듯, 불일치율의 평균은 0%가 아닌 12.98%로 나타났으며 최대 33.3%에 이르는 사례도 존재하였다. 즉 모델과 직관이 대부분 일치하지만, 결코 무시할 수 없는 수준의 체계적인 불일치가 실제로 존재한다는 점에서 H1은 부분적으로 지지되었다. 이는 인간이 자신이 직접 설계한 판단 모델과도 완전히 동일하게 행동하지는 않는다는 것을 보여주며, Samuelson(1938)의 선언적 선호와 현시적 선호간의 괴리 및 Kahneman(2003)의 이중처리 이론이 예측하는 바와 부합한다.

나. H2: 사례 유형에 따른 불일치율 차이

H2를 검증하기 위해 베이스라인, 단일극값, 변인충돌 세 사례 유형의 불일치율을 대응표본 t검정으로 비교하였다.

표 4. 사례 유형별 불일치율 대응표본 t검정 결과

대응	평균차이	SD	t	df	p	Cohen's d
베이스라인 - 단일극값	-13.13	17.36	-5.657	55	<.001	-0.76
베이스라인 - 변인충돌	-20.39	15.95	-9.566	55	<.001	-1.28
단일극값 - 변인충돌	-7.27	23.60	-2.305	55	.025	-0.31

세 대응비교 모두 통계적으로 유의하였다. 베이스라인 사례($M=0.45\%$)의 불일치율은 단일극값 사례($M=13.57\%$, $t(55)=-5.657$, $p<.001$) 및 변인충돌 사례($M=20.84\%$, $t(55)=-9.566$, $p<.001$)보다 유의하게 낮았으며, 단일극값 사례와 변인충돌 사례 간에도 유의한 차이가 나타나 변인충돌 사례의 불일치율이 더 높았다($t(55)=-2.305$, $p=.025$). 즉 불일치율은 베이스라인 < 단일극값 < 변인충돌의 순으로 체계적으로 증가하였으며, 효과크기(Cohen's d) 역시 같은 순서로 커지는 경향을 보였다($d=-0.76 \rightarrow -1.28$, 단일극값-변인충돌 비교는 $d=-0.31$ 로 상대적으로 작았음). 이는 변인 간 갈등이 극심한 사례일수록 모델 판단과 직관 판단의 괴리가 커진다는 것을 의미하며, H2는

지지되었다. 이러한 결과는 Kahneman과 Frederick(2002)의 속성 대체 이론과 일치한다. 변인들이 서로 충돌하지 않는 베이스라인 사례에서는 어떤 판단 방식을 사용하든 결론이 유사하게 도출되지만, 변인들이 서로 다른 방향을 가리키는 변인충돌 사례에서는 직관(System 1)이 다섯 개 변인을 모두가중합하는 대신 가장 두드러진 한두 개의 변인에 의존하여 판단을 대체하는 경향이 강해지기 때문으로 해석된다.

다. H3: 불일치의 방향성

H3을 검증하기 위해 불일치가 발생한 사례에서 판단 이탈의 방향성을 분석하였다. 판단차이가 양의 값을 가지면 직관 판단이 모델 판단보다 관대한 방향으로 이탈한 경우이며, 음의 값을 가지면 직관 판단이 모델 판단보다 엄격한 방향으로 이탈한 경우이다. 이러한 이탈 방향을 사례 유형별로 구분하여 분석하였다. 전체 56명 × 15문항 중 불일치가 발생한 사례는 총 109건이었다.

표 5. 사례 유형별 불일치 방향성 (직관 판단 - 모델 판단)

사례 유형	불일치 건수	직관이 관대(+)	직관이 엄격(-)	평균 판단차이
베이스라인	1	1 (100.0%)	0 (0.0%)	+2.00
단일극값	38	0 (0.0%)	38 (100.0%)	-2.13
변인충돌	70	66 (94.3%)	4 (5.7%)	+1.94
전체	109	67 (61.5%)	42 (38.5%)	+0.52

분석 결과, 불일치의 정도는 사례들의 유형에 따라 규칙적인 패턴을 보였다. 단일극값이 주어진 사례에서 발생한 불일치들 38건은 모두 현시적 판단이 모델 판단보다 더 엄격한 양상을 띈다. 이는 하나의 변인이 극단적으로 낮은 경우에는 System1의 직관이 그 지점에 집중하여 모델이 계산한 점수보다 훨씬 엄격한 평가를 내리는 경향을 보여준다. 반대로, 변인 충돌이 발생한 사례에서 불일치가 발생한 70건 중에서 66건은 현시적 판단이 모델 판단보다 관대한 평가를 내렸다. 이는 여러

변인이 서로 다른 방향에 놓여있을 때, 직관은 그 중 가장 높은 변인에 집중하고, 다른 변인들은 과소반영하는 경향이 있음을 보여준다.

이러한 결과는 불일치가 단순한 무작위 오차가 아니라 사례의 구조적 특성에 따라 체계적인 방향성을 가진다는 것을 보여주며, H3은 지지되었다. 다만 당초 가설이 예측한 것처럼 '가장 두드러진 변인의 방향'이 상황에 따라 다르게 작동한다는 점은 주목할 만하다. 즉 단일 변인이 극단적으로 낮을 때는 그 약점 방향으로, 변인들이 서로 충돌할 때는 가장 매력적인 강점 방향으로 직관이 이탈하는 비대칭적 패턴이 나타났다. 이는 Kahneman과 Frederick(2002)이 제안한 속성 대체가 상황에 따라 부정적 속성(단일 약점) 또는 긍정적 속성(상대적 강점)으로 선택적으로 작동할 수 있음을 시사하는 결과로 해석된다.

단일극값 사례 내에서도 어떤 변인이 극단이었는지에 따라 불일치 발생 빈도가 달라지는지를 추가로 살펴보았다. 단일극값 사례 5개는 각각 안정성 극저(Q05), 수익 가능성 극고(Q06), 자금 효율성 극저(Q07), 사업 계획 완성도 극고(Q08), 사회적 가치 극저(Q09)의 조건으로 설계되었다.

표 5-1. 단일극값 사례별 불일치 발생 빈도 (N=56)

문항	극단 변인 및 방향	불일치 인원	비율(%)
Q05	안정성 극저(5점)	13	23.2
Q06	수익 가능성 극고(97점)	0	0.0
Q07	자금 효율성 극저(4점)	8	14.3
Q08	사업 계획 완성도 극고(96점)	0	0.0
Q09	사회적 가치 극저(6점)	17	30.4

분석 결과, 매우 뚜렷한 비대칭이 확인되었다. 하나의 변인이 극단적으로 낮게 설정된 사례(Q05, Q07, Q09)에서는 각각 23.2%, 14.3%, 30.4%의 불일치가 발생한 반면, 하나의 변인이 극단적으로 높게 설정된 사례(Q06, Q08)에서는 불일치가 단 한 건도 발생하지 않았다(0.0%). 이는 <표 5>에서

확인한 '단일극값 사례의 불일치가 예외 없이 모두 엄격한 방향으로 발생한다'는 결과와 정확히 부합하는 세부 근거이다. 즉 모델과 직관의 괴리는 하나의 변인이 두드러지게 우수할 때가 아니라, 하나의 변인이 두드러지게 취약할 때 집중적으로 발생하였다. 이는 직관 판단이 손실이나 결함에 더 민감하게 반응하는 경향, 즉 행동경제학에서 폭넓게 보고되어 온 손실 회피 성향과 맥을 같이하는 결과로 해석할 수 있다. 특정 변인의 결정적 약점 하나가 나머지 변인들의 우수함에도 불구하고 직관적으로 '탈락' 쪽으로 강하게 작용하는 반면, 하나의 강점은 모델이 계산하는 만큼의 긍정적 영향력을 직관에 즉각적으로 발휘하지 못하는 것으로 보인다.

4. 추가 분석

본 연구는 인과관계를 확인하는 연구 설계의 특성을 반영하여, 개인의 판단 특성이 전체 불일치율에 미치는 영향을 다중회귀분석으로 검증하였다. 독립변수는 AHP 쌍대비교의 일관성 비율(CR), 5개 변인 가중치의 표준편차로 측정한 가중치 편중도(가중치가 특정 변인에 쏠릴수록 값이 커짐), 사후Q2(모델 기준 이해도)로 설정하였고, 종속변수는 불일치율_전체(%)로 설정하였다.

표 6. 불일치율_전체에 대한 다중회귀분석 결과 (N=56)

예측변인	B	SE	t	p
(상수)	11.897	7.974	1.492	.142
CR	-65.220	49.296	-1.323	.192
가중치 편중도(SD)	-15.192	37.217	-0.408	.685
사후Q2(모델기준이해도)	1.880	1.229	1.529	.132

회귀모형은 통계적으로 유의하지 않았으며($F(3,52)=1.780$, $p=.163$), 설명력 또한 낮았다($R^2=.093$). 개별 예측변인 중에서도 CR($B=-65.220$, $p=.192$), 가중치 편중도($B=-15.192$, $p=.685$), 모델 기준 이해도($B=1.880$, $p=.132$) 어느 것도 통계적으로 유의한 예측력을 보이지 않았다. 다만 계수의 방향성은 이론적 예상과 대체로 부합하였다: CR이 낮을수록 불일치율이 낮아지는 경향, 그리고 자신의 모델 기준을 잘 이해하고 있다고 응답할수록(사후Q2) 오히려 불일치율이 높아지는

경향이 관찰되었다. 후자는 다소 역설적으로 보일 수 있으나, 모델 기준을 명확히 이해하고 있다는 자기 인식이 오히려 직관 판단 시 그 기준으로부터 의식적으로 벗어나 별도의 직관적 고려를 반영했을 가능성을 시사한다. 보조적으로 실시한 CR과 불일치율_전체 간 상관분석에서도 두 변수는 약한 부적 상관을 보였으나 통계적으로 유의한 수준에는 이르지 못하였다($r=-.226$, $p=.095$, $N=56$).

종합하면, 본 연구에서 측정한 개인차 변인들은 전체 불일치율의 개인 간 차이를 유의하게 설명하지 못하였다. 이는 불일치의 크기가 개인의 안정적인 판단 성향보다는 <표 4>, <표 5>에서 확인된 것처럼 개별 사례의 구조적 특성(변인 간 충돌 여부)에 더 크게 좌우된다는 것을 시사하는 결과로 해석할 수 있다.

V. 결론

1. 연구 결과 요약 및 논의

본 연구는 스타트업 지원 심사라는 여러 변인이 영향을 미치는 의사결정 상황에서, 개인이 AHP를 통해 명시적으로 구성한 판단 모델과 동일한 개인의 현시적으로 수행한 판단 사이에 유의미한 불일치가 존재하는지를 검증하였다. 연구 결과를 가설별로 정리하면 다음과 같다.

첫번째로(H1), 전체 불일치율은 평균 12.98%로 무작위 수준(50%)보다 유의하게 낮아 모델과 직관이 대체로 높은 일치도를 보였지만, 0%가 아닌 유의미한 수준의 불일치가 실제로 존재함을 확인하였다. 이는 인간이 직접 판단 기준을 설계하였음에도, 실제 상황에서 판단한 결과는 해당 기준을 정확하게 따르지 않음을 보여주었다. 따라서 H1은 부분적으로 지지되었다.

두번째로(H2), 불일치율은 베이스라인, 단일극값, 변인충돌 순으로 유의하게 증가하였다. 변인간에 갈등이 없는 상황에서는 모델과 직관의 판단 결과가 거의 완벽하게 동일했다. 그러나 변인간의 충돌이 존재하는 경우에는 불일치율이 평균 20.84%까지 상승하였다. 따라서 H2는 지지되었다.

세번째로(H3), 불일치의 양상은 주어진 사례의 유형에 따라 규칙적인 패턴을 보였다. 하나의 변인이 극단적으로 낮은 사례에서는 직관이 모델보다 예외 없이 더 엄격한 방향으로 판단하는 경향을 보였지만, 변인들이 서로 충돌하는 사례에서는 대부분의 경우 직관이 모델보다 더 관대한 방향으로 평가하였다. 따라서 H3은 지지되었다.

네번째로, 추가로 실시한 회귀분석에서는 AHP 일관성(CR), 가중치 편중도, 모델 기준 이해도라는 개인차 변인이 전체 불일치율을 유의하게 예측하지 못하였다. 이는 불일치의 크기를 결정하는 주된 요인이 안정적인 개인 특성이 아니라, 개별 의사결정 상황에서 변인들이 충돌하는 정도라는 것을 시사한다.

이러한 결과는 AI 정렬 논의에 대해서 시사점을 제공한다. 개인이 자신의 판단 기준을 아무리 명확하게 진술하고 그 기준으로 구성된 모델의 CR(내적 일관성)이 양호하더라도, 실제 상황에서의 직관적 판단은 그 모델과 다르게 나타날 수 있으며, 그 괴리는 특히 여러 기준이 상충하는 복잡한 상황에서 커진다는 것이다. 이는 인간-AI 협업 시스템을 설계할 때, 특히 변인 간 충돌이 큰 의사결정 상황일수록 AI의 판단 근거를 인간에게 명시적으로 설명하고, 인간의 직관적 판단과 모델 판단이 괴리되는 지점에 대한 별도의 검토 절차를 마련할 필요가 있음을 시사한다.

2. 연구의 한계

본 연구는 다음과 같은 한계를 내포하고 있다. 첫번째로, 본 연구는 단일 학교 재학생을 표본으로 분석되어 본 연구의 결론을 전반적인 사람들의 특성으로 일반화하는데에 무리가 있다. 다음으로, 실제 스타트업에서 사용된 심사 기준이 아닌 연구자가 임의로 제작한 15개의 사례를 바탕으로 실험이 진행되었기 때문에 실제 사회를 충분히 반영하지 못한다는 한계가 있다. 세번째로, 피실험자들의 설문 참여도와 설문 참여 소요 시간 등을 고려하여 사례 수를 15개로 제한하였다는 한계가 있다. 이는 단일극값 이나 변인충돌 사례에서 더 구체적인 조건에 따른 차이를 통계적으로 검증하기에 표본 수가 부족하다는 한계로 이어졌다.

3. 후속 연구 및 제언

후속 연구에서는 다음과 같은 방향을 고려할 필요가 있다. 첫번째로, 표본을 여러 학교 및 연령대로 확대하여 결과의 일반화 가능성을 높일 필요가 있다. 두번째로, 가상 사례 카드가 아닌 실제 스타트업 지원 서류나 모의 심사 상황을 활용하여 생태적 타당도를 높이는 실험 설계를 시도할 수 있다. 세번째로, 반응 시간이나 시선 추적과 같은 행동적 및 생리적 지표를 추가로 측정한다면, 자기보고에 의존하지 않고 System 1과 System 2의 개입 정도를 더 직접적으로 포착할 수 있을 것이다. 네번째로, 본 연구에서 발견된 '단일 약점에는 엄격하게, 상대적 강점에는 관대하게 반응하는 비대칭적 방향성 패턴이 다른 의사결정 맥락에서도 동일하게 나타나는지를 검증하는 비교 연구가

필요하다. 다섯번째로, 표본을 확대하여 개인차 변인이 불일치율에 미치는 영향을 더 큰 통계적 검정력으로 재검증할 필요가 있다.

참고문헌

- Alberini, A. (2019). Revealed versus stated preferences: What have we learned about valuation and behavior? *Review of Environmental Economics and Policy*, 13(2), 283–298. <https://doi.org/10.1093/reep/rez010>
- de Corte, K., Cairns, J., & Grieve, R. (2021). Stated versus revealed preferences: An approach to reduce bias. *Health Economics*, 30(5), 1095–1123. <https://doi.org/10.1002/hec.4246>
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126. <https://doi.org/10.1037/xge0000033>
- Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3), 223–241. <https://doi.org/10.1177/1745691612460685>
- Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12(1), 19–30. <https://doi.org/10.1037/1040-3590.12.1.19>
- Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58(9), 697–720. <https://doi.org/10.1037/0003-066X.58.9.697>
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
- Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The*

psychology of intuitive judgment (pp. 49–81). Cambridge University Press.

<https://doi.org/10.1017/CBO9780511808098.004>

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of Mathematical*

Psychology, 15(3), 234–281. [https://doi.org/10.1016/0022-2496\(77\)90033-5](https://doi.org/10.1016/0022-2496(77)90033-5)

Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *International Journal of*

Services Sciences, 1(1), 83–98. <https://doi.org/10.1504/IJSSCI.2008.017590>

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, 5(17),

61–71. <https://doi.org/10.2307/2548836>

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the

rationality debate? *Behavioral and Brain Sciences*, 23(5), 645–665.

<https://doi.org/10.1017/S0140525X00003435>

부록

1. 설문지 구성 개요

본 설문은 웹사이트를 통해 배포되었으며, 이는 직접 제작한 웹사이트를 기반으로 한다. 해당 도구의 소스코드는 이 링크에서 확인할 수 있다.

https://gitea.seung6lee.com/seung6lee/AHP_decision_analysis

부록표 1. 설문지 구성

구분	문항 수	척도	비고
안내 및 응답 동의	-	-	연구 목적 및 익명성 안내
현시적 판단	15	5점 리커트 (1=매우 탈락 ~ 5=매우 선발)	베이스라인 4 / 단일극값 5 / 변인총돌 6
AHP 쌍대비교	10	-4 ~ +4 (9단계)	5개 변인(안정성, 수익 가능성, 자금 효율성, 사업 계획 완성도, 사회적 가치)의 모든 쌍
결과 확인	-	-	AHP 가중치 모델 판단과 직관 판단 결과 공개
사후 설문	3	5점 리커트	모델-직관 차이 인식 / 모델 기준 이해도 / 모델 합리성 평가