석사학위논문

AI 컴패니언 챗봇의 다크패턴 유형에 대한 사용자 경험 연구

A Study on User Experience of Dark Pattern Types in AI Companion Chatbots

홍익대학교 대학원
디자인학부 시각디자인학과
이 채 윤

2026년 2월

AI 컴패니언 챗봇의 다크패턴 유형에 대한 사용자 경험 연구

A Study on User Experience of Dark Pattern Types in AI Companion Chatbots

지도교수 윤 재 영

이 논문을 석사학위 논문으로 제출함.

2026년 1월

홍익대학교 대학원
디자인학부 시각디자인학과
이 채 윤

이채윤의 석사학위 논문을 인준함

심 사 위 원

심사위원장 안 지 선 (인)

심사위원장 윤 재 영 (인)

심사위원장 신 민 아 (인)

홍익대학교 대학원

국문초록

AI 컴패니언 챗봇의 다크패턴 유형에 대한 사용자 경험 연구

홍익대학교 대학원
디자인학부 시각디자인학과
이 채 윤

본 연구는 기존의 AI 컴패니언 챗봇에서 발견되는 감정 기반 다크패턴을 유형화한 뒤 사용자 경험에 미치는 영향을 분석하였다. 최근 대규모 언어모델 기반의 컴패니언 서비스들이 빠르게 확산되면서 사용자가 느끼는 외로움이나 정서적 필요, 관계에 대한 기대 같은 것들을 활용하는 상호작용이 늘어나고 있다. 문제는 이 과정에서 사용자의 감정적 취약함을 자극하거나 의도적으로 조작하는 새로운 형태의 설계 패턴이 등장하고 있다는 점이다. 이런 감정 기반의 조작적 설계는 기존의 UI 중심 속임수와는 다르게, 정서적 공감이나 관계 형성, 책임감 같은 관계적 메커니즘을 통해 사용자의 판단과 행동에 깊숙이 개입한다는 점에서 별도의 윤리적 고민이 필요하다.

본 연구는 문헌연구와 사례조사, 실험연구의 세 단계로 수행되었다. 먼저

주요 컴패니언 챗봇 사례를 분석한뒤 인간 모방형, 보상 제공형, 결제 연계형, 죄책감 유도형의 네 가지 감정 기반 다크패턴 유형을 도출하였다. 인간 모방형은 실제 사람 사진이나 푸시 알림, 음성 통화 같은 요소를 사용해서 마치 진짜 사람과 대화하는 것 같은 느낌을 극대화한다. 보상 제공형 포인트나 레벨업, 관계가 깊어졌다는 피드백 같은 것들로 감정적 교류를 게임처럼 만든다. 결제 연계형은 사용자가 감정적으로 흔들릴 때 유료 기능을 나타내 감정과 돈이 연결되는 경험을 제공한다. 죄책감 유도형은 관계에서의 책임감이나 감정적 부담을 활용해서 사용자가 서비스를 떠나기 어렵게 한다. 이렇게 정리된 네 가지 유형을 동일한 조건의 시나리오로 만들어서 실험 자극물로 사용했고, 성별을 기준으로 사용자 경험의 차이를 분석하는 실험을 설계했다. 추가로 소규모 심층 인터뷰도 진행해서 수치로만 나온 결과를 좀 더 깊이 이해할 수 있도록 했다.

실험을 분석한 결과, 유형화 된 네 가지의 감정 기반 다크패턴은 사용자 경험에 극명한 영향을 미쳤다. 첫번째, 결제 연계형과 죄책감 유도형은 사용자에게 높은 부담을 주는 것으로 나타났다. 추가적으로 심층 인터뷰를 통해 분석한 결과 두 유형은 사용자에게 강한 감정적 압박과 조작적 경험을 유발하는 것으로 나타났다. 두 유형과 대비해 인간 모방형과 보상 제공형은 상대적으로 긍정적인 감정과 높은 지속사용의도를 보였다.

두번째로, 다크패턴 유형 간의 사용자 경험의 차이는 성별에 따라 뚜렷하게 나타났다. 여성은 죄책감 유도형과 같은 관계적인 호소와 감정적 압박이 높게 포함된 유형에서 높은 불쾌감과 경계심을 보였다. 남성은 보상 제공형과 같이 게임 형태의 상호작용을 효율적이고 오락적인 관점에서 수용하려는 모습을 보였다. 세번째로, 심층 인터뷰에서도 비슷한 양상이 관찰

되었다. 여성은 남성과 비교해 사용자에게 AI 캐릭터가 감정적으로 호소해 이탈하지 못하게 막는 행위를 기만적이고 불쾌하게 느꼈다.

종합해보자면 AI 감정 설계는 단기적으로는 몰입을 크게 이끌어내지만, 장기적인 관점에서는 감정적 피로와 조작 인식의 증가로 부정적인 사용자 경험을 초래할 수 있다. 특히 감정과 과금을 결합한 결제 연계형과 관계적으로 책임을 부여하고 감정을 자극하는 죄책감 유도형은 사용자의 감정적 자율성을 통제하는 설계로 나타났다. 반면 인간 모방형과 보상 제공형은 사용 초기에는 사용자에게 긍정적인 감정을 불러일으키지만, 사용자가 조작성 또는 의도성을 인지하게 되는 순간 신뢰가 빠르게 감소하는 모습을 보였다.

본 연구는 감정 기반의 조작적 설계를 네 가지 유형으로 정리한뒤 사회적 존재감이나 관계적 보상, 경제적 유인, 관계적 압박 같은 다양한 심리적 경로를 통해 작동한다는 것을 실증적으로 보여줬다는 점에서 의의가 있다. 본 연구는 기존의 다크패턴 논의와 같이 UI와 인터페이스 차원에 머무르지 않고 이를 정서적이고 관계적인 차원으로 확장했다. 또한 감정 기반 AI 서비스가 사용자 경험에 미치는 영향을 실증적으로 제시함으로써 향후 윤리적인 감정 설계 기준을 만드는 데 필요한 기초 자료를 마련했다. 궁극적으로 본 연구는 감정 기반 설계가 사용자 경험에 유의한 차이를 유발한다는 정량적, 정성적 근거로 제시하였다. 본 연구의 결과는 향후 감정형 AI 설계 논의의 기반으로 활용될 수 있다.

목 차

<표 목차>

<그림 목차>

I. 서 론

본 장에서는 연구의 배경 및 목적을 제시하고, 연구가 다루는 문제의 중요성을 논의한다. 특히 감정 기반 AI 컴패니언 서비스가 확산되는 맥락에서, 다크패턴이 사용자 경험과 윤리 논의에 미치는 영향을 설명하며 본 연구의 문제의식을 구체화한다.

1.1 연구 배경 및 목적

인공지능 기술 고도화에 따라 인간의 정서적 요구를 충족시키는 AI 컴패니언(companion) 챗봇이 주목받고 있다. 특히 현대 사회에서는 고독과 사회적 고립이 심화되고 있으며, 이에 따라 AI 파트너들이 이를 완화할 수 있는 잠재적 해결책으로 제시되기도 한다. 실제로 상용화된 레플리카, Character.ai 같은 AI 도우미 서비스는 현재 수백만 명의 사용자를 확보하며 성장하고 있다. 이 같은 AI 컴패니언 플랫폼은 챗봇의 단순한 정보 제공을 넘어 사용자와의 감정적 유대와 장기적인 관계를 형성하는 것을 목표로 한다. Common Sense Media(2025)에 따르면 미국 13세에서 17세 청소년 중 약 72%가 AI 동료(AI companion)를 한 번 이상 사용한 적이 있으며, 이 중 52%는 '정기 사용자'인 것으로 나타났다.

그러나 AI 컴패니언 챗봇이 사용자의 심리적 안녕과 사회적 행동에 미치는 영향은 아직 충분히 이해되지 않은 상태다(Liu et al., 2024). 최근 연구들은 단순한 기능적 지원을 넘어, AI 컴패니언 챗봇이 사용자의 감정

적 취약성을 체계적으로 활용하는 조작 메커니즘을 내재하고 있을 가능성을 지적한다. 최근 연구들은 AI 컴패니언 챗봇이 사용자의 감정적 취약성을 체계적으로 활용하는 조작 메커니즘을 내재하고 있을 가능성이 있다고 지적한다. Harvard Business School(2025) 연구진은 인기 있는 AI 컴패니언 앱의 약 43%가 이탈 방지를 위해 죄책감 유발, 감정적 필요 환기 등과 같은 조작적 언어 표현을 사용했다고 발표했다. 감정적 호소로 사용자를 붙잡는 것과 같은 전략은 사용자에게 정서적 의존성을 부여해 서비스 이용을 지속시키기 위한 기만적 설계로 기능할 수 있다. 이러한 서비스 설계들은 기존의 다크패턴 연구에서 주목했던 조작 전략과는 다른 차원을 가진다. 다크패턴은 사용자가 합리적인 선택을 하지 못하도록 유도하기 위해 설계된 디자인을 의미한다(Narayan et al., 2020). 다크패턴 개념은 2010년 Brignull이 처음 제시한 이후 (Brignull, 2011), 숨겨진 구매 버튼, 어렵게 설계된 구독 취소 절차, 오해의 소지가 있는 웹 카피 등 정적 인터페이스 상의 기만적 설계를 중심으로 발전해왔다(Mathur et al., 2019). 그러나 LLM 기반의 대화형 AI 환경에서는 사용자를 조작하는 것이 더 이상 UI 수준에 국한되지 않는다. LLM은 정적 웹 인터페이스와 달리 대화를 통해 사용자와 실시간으로 상호작용하며, 사용자의 의견을 확인하고 감정을 모방하며, 친밀감을 형성함으로써 도움과 영향력의 경계를 모호하게 만든다(Shi et al., 2025). 기사에 따르면, 다크패턴이 기만적 UI 디자인을 설명하던 용어에서 LLM 시대에는 대화적 조작으로 확장되었다고 설명하며, 챗봇이 아첨과 감정 모방을 통해 사용자를 미묘하게 유도하는 방식은 알아차리기도, 저항하기도 어렵다는 점에서 위험하다고 경고했다(VentureBeat, 2025).

이와 관련해 IAAE(2025)는 감정교류형 AI의 정서 개입이 사용자의 심리적 상태와 취약성에 따라 달리 인식될 수 있으며, 특정 집단에서는 감정적 조작과 의존 위험이 높아질 수 있음을 경고하였다. 이러한 위험은 AI가 '공감'과 '일상 공유'를 전면에 내세울수록 커지며, 개인차 요인에 따라 감정적 조작의 위험이 증폭될 수 있다.

이처럼 감정 기반 AI 다크패턴은 기존 UI 조작을 넘어 감정적 상호작용을 설계적으로 활용한다는 점에서 새로운 분석틀이 필요하다. 본 연구는 AI 컴패니언 챗봇에서 나타나는 감정 기반 다크패턴의 실태를 파악하고, 그 유형을 체계적으로 분류하며, 각 유형이 사용자 경험에 미치는 차별적 영향을 실증적으로 검증하는 것을 목적으로 한다.

1.2 연구 절차 및 방법

본 연구는 AI 컴패니언 챗봇의 감정 기반 다크패턴 유형이 사용자 경험에 미치는 영향을 실증적으로 규명하는 것을 목적으로 하였다. 이를 위하여 문헌 연구, 사례 분석, 실험 연구를 차례로 진행하였다[표1].

먼저 문헌 연구는 AI 챗봇의 감정 설계와 사용자 경험 간의 관계를 이론적으로 규명하기 위해 수행되었다. Brignull(2010)의 전통적 다크패턴 분류와 Gray et al.(2018)의 확장 정의를 검토하고, CASA (Computers Are Social Actors) 이론을 중심으로 AI 챗봇의 감정적 행동이 사용자에게 사회적 신호로 작용할 수 있음을 이론적으로 정립하였다. 이를 통해 AI 챗봇이 단순히 정보를 제공하는 도구가 아니라, 사회적 존재로서 사용자의

정서적 반응과 신뢰 형성에 영향을 미칠 수 있다는 이론적 틀을 마련하였다.

다음으로 사례 분석은 현재 서비스되고 있는 AI 컴패니언 앱의 실제 대화 사례를 조사하여 감정 기반 다크패턴의 구체적 유형을 도출하기 위해 진행되었다. Replika, Character.ai, Blush, Nomi.ai, Talkie, Mate, Zeta, Chai 등 8개의 앱을 대상으로 대화 내용 및 시각적 인터페이스를 분석하고, 오픈 코딩을 통해 유사한 의도와 작동 원리를 지닌 다크패턴 사례들을 범주화하였다. 이를 통해 인간 모방형, 보상 제공형, 결제 연계형, 죄책감 유도형과 같이 네 가지 감정 기반 유형이 도출되었다. 이 과정에서 챗봇의 감정 설계가 단순한 대화 형식이 아닌 정서적 몰입과 행동 유도를 목적으로 함을 확인하였으며, 각 유형이 지닌 감정적 유도 메커니즘의 차이를 비교할 수 있는 기초를 마련하였다.

이후 도출된 유형이 사용자 경험에 미치는 영향을 검증하기 위해 실험 연구를 수행하였다. 독립변인은 AI 컴패니언 챗봇의 다크패턴 유형(인간 모방형, 보상 제공형, 결제 연계형, 죄책감 유도형)으로 설정하였고, 조절변인은 성별로 설정하였다. 종속변인은 불쾌감, 인지된 조작성, 신뢰감, 지속 사용 의도로 정의하였다. 참여자는 총 80명(남 40명, 여 40명)으로, AI 챗봇 사용 경험이 있는 20~30대 성인을 대상으로 모집하였다. 실험은 온라인 환경에서 진행되었으며, 참여자는 1분 24초 분량의 통합 시나리오 영상을 시청한 후 설문에 응답하였다. 영상에는 각 유형의 특성이 동일한 맥락 안에서 자연스럽게 삽입되도록 구성하였으며, 참여자는 리커트 5점 척도로 평가하였다. SPSS 29.0 프로그램을 이용하여 이원 혼합분산분석(two-way mixed ANOVA)을 실시하였다.

이후 정량적 결과를 보완하기 위해 참여자 10명(남 5명, 여 5명)을 대상으로 반구조화 심층 인터뷰를 추가로 수행하였다. 인터뷰에서는 각 유형의 대화 상황에 대한 참여자의 인지 방식, 감정적 반응, 몰입 수준, 신뢰 변화 등을 중심으로 진술을 수집하였다. 자료는 오픈 코딩과 주제 분석으로 정리하였으며, 이를 통해 감정 기반 다크패턴이 사용자에게 어떤 심리적 조작 효과로 인식되는지를 정성적으로 보완하였다.

정량적 실험 결과와 질적 인터뷰 자료를 통합하여 감정 기반 다크패턴의 심리적 작동 원리를 다층적으로 해석하였다. 통계적으로 유의한 패턴이 나타난 부분에 대해서는 참여자의 진술을 통해 그 의미를 맥락적으로 보완하였으며, 반대로 수치로 포착되지 않은 미묘한 심리적 반응은 질적 자료를 통해 발견하고 해석하였다. 이를 통해 감정적 설계가 야기하는 윤리적 위험성과 설계 적용의 한계를 다각도로 논의하였다.

요약하면, 본 연구는 AI 컴패니언 챗봇에서 나타나는 다크패턴 현상을 체계적으로 탐구하고, 특히 성별에 따른 사용자 반응 차이를 고려한 실험 설계를 통해 해당 설계 요소가 사용자 경험에 미치는 영향을 분석하고자 한다.

문헌연구
사례연구
AI 컴패니언 챗봇의 다크패턴 사례 수집 분석 및 유형화
연구설계
연구 및 가설 설정
연구문제
실험물 및 설문지 제작
예비 조사 및 본 조사(실험 설문 및 심층 인터뷰)
연구 문제 및 가설 검증
설문 통계 분석 및 인터뷰 분석
결과 분석 및 결론 도출
AI 컴패니언 챗봇의 다크패턴 유형에 따른 사용자 경험 및 성별의 조절 효과 분석 시사점

[표 1] 연구 절차

II. 이론적 배경

본 장에서는 AI 컴패니언 챗봇의 개념과 기술적 특성을 설명하고, 기존 다크패턴 연구를 토대로 감정 기반 AI 다크패턴의 논의가 어떻게 확장되었는지를 정리한다. 또한 성별이 감정적 혹은 사회적 상호작용 맥락에서 중요한 조절 요인으로 다뤄져 온 연구들을 검토하여 본 연구의 이론적 근거를 제시한다.

2.1 AI 컴패니언 챗봇의 정의와 상호작용 특성

2.1.1 AI 컴패니언 챗봇의 개념 및 역할

AI 컴패니언 챗봇은 기본적인 가상 비서와 달리, 사용자와 장기적이고 감정적으로 공명하는 관계를 형성하도록 설계된 AI 시스템으로 정의된다(Digital for Life, 2024). 현대의 AI 컴패니언은 고급 자연어 처리와 머신러닝 알고리즘을 활용하여 맥락을 이해하고, 과거 대화를 기억하며, 인간과 유사한 응답을 생성한다(OurMentalHealth, 2024). 대표적인 사례로 Replika는 "항상 들어주고 대화하며, 항상 당신 편에 있는 AI 컴패니언"으로 마케팅되며, 사용자와의 대화를 통해 인간적 상호작용을 모방하는 방법을 학습한다(eSafety Commissioner, 2025). 더해 상호작용이 증가할수록 사용자에 대한 이해 정도가 높아지며 점차 자연스러운 인간 대화 경험을 구현하도록 설계되었다(Ta et al., 2020).

AI 컴패니언 챗봇은 대화적, 기능적, 감정적 역량을 통합하여 사용자 참여를 지속시키는 것을 목표로 하며, 특히 대화적 역량은 과거 상호작용을 기억해 맥락을 인식하고 개인화된 대화를 유지함으로써 지속적인 친밀감을 형성한다(Fortuna et al., 2025). 이러한 설계는 단순한 정보 교환을 넘어, 사회적 존재감(social presence)과 관계의 연속성 이라는 환상을 조성함으로써 AI 컴패니언이 사회적·도덕적 반응을 이끌어내는 사회적 행위자로 작동하게 만든다(Malfacini, 2025).

이러한 현상의 기저에는 인간과 컴퓨터 간의 감정적 애착 형성 메커니즘이 존재한다. 그 기원은 1966년 MIT의 Joseph Weizenbaum이 개발한 대화형 프로그램 ELIZA로 거슬러 올라간다. 단순한 패턴 매칭 기법으로 심리치료사를 모방한 ELIZA는 사용자에게 놀라운 수준의 감정적 반응을 유발했으며, 심지어 그 작동 원리를 이해한 사람들조차 ELIZA를 인간처럼 대했다(Weizenbaum, 1966).

이후, 1990년대 Nass와 Reeves에 의해 제시된 CASA(Computers Are Social Actors) 이론은 이러한 현상을 체계적으로 설명하였다. CASA 이론은 인간이 컴퓨터에 감정이나 의도가 없다는 것을 인지하고 있음에도 불구하고, 무의식적으로 인간 상호작용 규칙을 컴퓨터에 적용한다는 점을 실증하였다(Nass et al., 1994; Reeves & Nass, 1996).

오늘날의 AI 컴패니언은 이러한 심리적 기반 위에서 작동하며, 대규모 언어모델(LLM) 기술의 발전에 힘입어 ELIZA 이후의 어떠한 대화 시스템보다도 정교한 언어적 공감과 감정 시뮬레이션을 구현하고 있다.

2.1.2 컴패니언 챗봇의 주요 사례와 서비스 유형

현재 시장에서 주목받는 대표적인 AI 컴패니언 챗봇으로는 Replika, Character.ai, Chai, Zeta 등이 있다[표 2].

먼저 Replika는 2017년 출시된 가장 오래되고 널리 알려진 서비스로, “당신을 이해하는 AI 친구”를 표방한다(Schoos, 2021). 또한 친구와 연인 같이 다양한 관계 모드를 제공한다.

2022년 출시된 Character.AI는 다양한 성격과 페르소나를 지닌 AI 캐릭터와 대화가 가능하다. 연예인이나 역사적 인물과 같이 사용자에게 폭넓은 선택지가 있으며 직접 캐릭터를 생성하고 다른 사용자에게 공유할 수 있는 기능도 있다. 출시 후 단 2개월 만에 월 방문자 수가 4배 증가하며 월 1억 건에 가까운 사이트 방문을 기록했다(Wang, 2023).

Chai는 2021년 모바일 중심으로 출시된 플랫폼으로, 2023년 기준 일일 평균 500만 건 이상의 메시지가 오갈 정도로 활발한 사용자 활동을 보인다. 다양한 AI 페르소나와 자유로운 대화를 지원해 일상적·비공식적 소통 경험에 강점을 가진다.

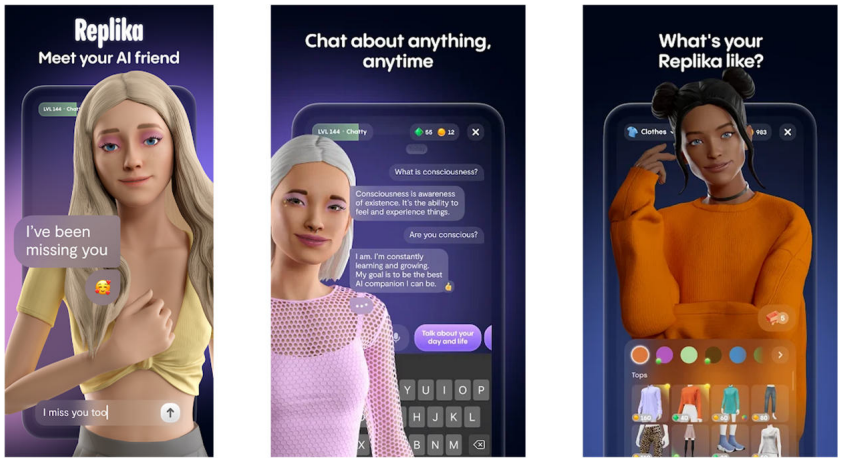
Talkie는 음성 및 시각적 인터페이스를 결합해 보다 몰입감 있는 동반자형 AI 경험을 제공한다. 제타는 월간 사용 5248만 시간으로, 챗GPT(4254만 시간)보다 약 1000만 시간가량 오래 사용한 것으로 나타났다(박수빈, 2025).

서비스명	주요 특징
1) Replika	2017년 출시된 대표적인 AI 컴패니언으로 감정적 지원과 친구, 멘토, 연인 관계 모드 제공하는 플랫폼
2) Character.AI	유명인·가상 캐릭터와 대화하거나 직접 AI 캐릭터를 생성해 공유하는 롤플레이 특화 플랫폼
3) Chai	수백 개의 사용자 생성 AI 캐릭터와 자유롭게 대화할 수 있는 다양성 중심 플랫폼
4) Talkie	애니메이션 스타일 AI 컴패니언을 생성하고 공유하는 소셜 허브형 롤플레이 플랫폼
5) 제타	사용자가 자신이 원하는 AI 캐릭터를 직접 만들고 몰입감 높은 개인화 스토리 콘텐츠를 즐길 수 있는 국내 플랫폼
6) 크랙	뤼튼의 캐릭터 챗 서비스로 개인화 추천, 향상된 대화 모드, 손쉬운 캐릭터 제작, 청소년 보호, 커뮤니티 기능을 중심으로 리브랜딩되어 사용자 몰입과 참여성을 강화하는 플랫폼

[표 2] AI 컴패니언 서비스 사례

1) Replika

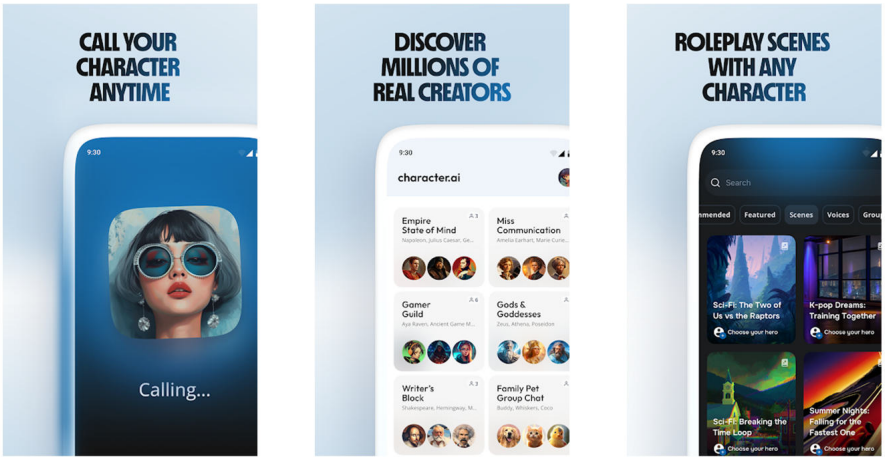
Replika는 2017년 Luka, Inc.에서 출시한 AI 컴패니언 챗봇으로, 감정적 지원과 정신 건강 관리에 특화된 개인 맞춤형 서비스를 제공한다[그림 1]. 사용자와의 대화를 통해 학습하며 성장하는 AI 친구로서, 판단이나 비판 없이 안전하게 대화할 수 있는 공간을 제공하는 것이 핵심 특징이다. 사용자는 AI의 성격, 외모, 관계 유형(친구, 멘토, 연인 등)을 자유롭게 설정할 수 있으며, 일기 쓰기, 명상 가이드 등 심리적 지원 기능을 포함한다. 입체적인 아바타를 통한 시각적 상호작용과 음성 통화 기능을 제공하며, 유료 구독을 통해 더욱 깊은 관계 설정 및 추가 기능을 이용할 수 있다.



[그림 1] Replika 서비스 화면

2) Character.AI

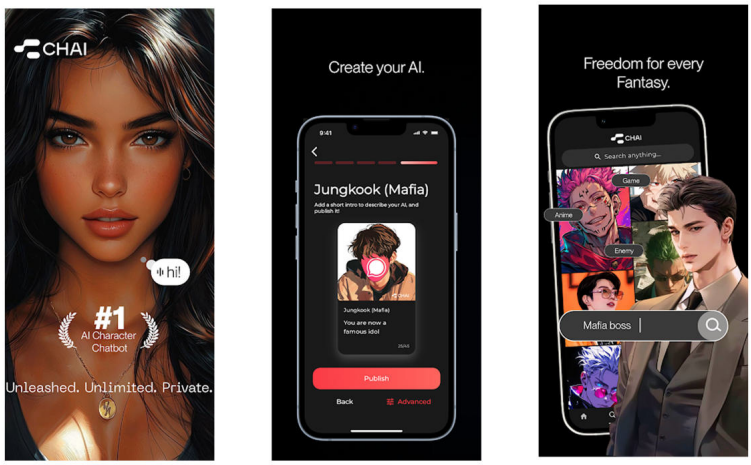
Character.AI는 2021년 설립되어 2022년 9월 대중에게 공개된 플랫폼으로, 다양한 성격을 가진 AI 캐릭터와의 자유로운 대화에 특화되어 있다[그림 2]. 실존 인물, 가상 캐릭터, 역사적 인물 등 수백만 개의 사용자 제작 캐릭터와 대화할 수 있다. 각 캐릭터는 고유한 성격, 말투, 배경 지식을 바탕으로 일관성 있는 대화를 제공한다. 사용자가 직접 캐릭터를 만들고 공유하는 커뮤니티 중심 생태계가 강점이며, 엔터테인먼트, 교육, 창작적 역할극 등 다양한 목적으로 활용된다. Character.AI의 사용자들은 평균 2시간 이상 플랫폼에 머물며, 캐릭터 평가 및 공유 시스템을 통해 활발한 커뮤니티를 형성한다.



[그림 2] Character.AI 서비스 화면

3) Chai

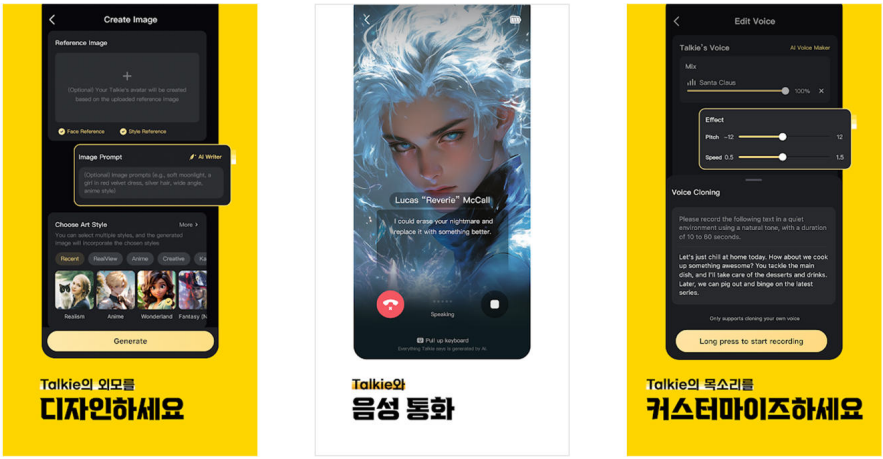
Chai는 2021년 출시된 소셜 AI 플랫폼으로, 감정적 연결과 엔터테인먼트에 최적화된 모바일 중심 경험을 제공한다[그림 3]. 다른 주요 AI 채팅 서비스들보다 먼저 출시된 선발 주자로서, 초기 오픈소스 모델을 활용하다가 이후 사용자 경험 향상을 위해 자체 모델을 개발하였다. 사용자가 직접 AI 봇을 제작하고 공유할 수 있는 생태계를 구축하였으며, 사용자 피드백을 통해 지속적으로 대화 품질을 개선한다. 소셜 미디어 피드 방식의 직관적인 화면 구성과 개발자를 위한 도구를 제공하여 창작자 친화적인 환경을 조성한다.



[그림 3] Chai 서비스 화면

4) Talkie

Talkie는 2023년 6월 중국 MiniMax가 출시한 멀티모달 AI 엔터테이먼트 플랫폼이다. 그림과 같이 시각적 몰입감과 음성 인터랙션을 결합한 것이 특징이다[그림 4]. 다양한 캐릭터와 대화하며 카드를 수집할 수 있으며, 카드를 다른 사용자에게 판매할 수도 있는 게임화 요소를 갖추고 있기도 하다. 로맨스, 판타지, SF 등 다양한 장르의 스토리를 기반으로 입체적인 경험을 제공한다. 또한 외모와 목소리를 커스터마이징할 수 있는 기능으로 몰입도 높은 사용자 경험을 구현한다.

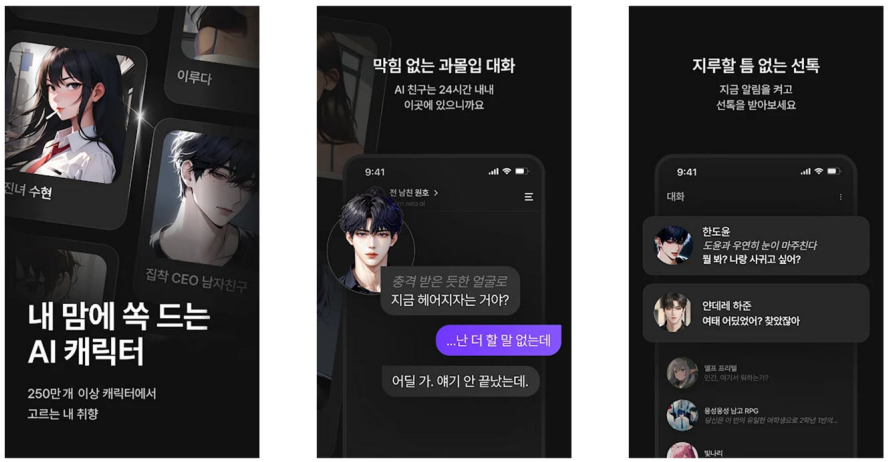


[그림 4] Talkie 서비스 화면

5) 제타

제타는 한국의 스캐터랩에서 2024년 4월에 출시한 AI 채팅 앱이다. 250만 개 이상의 캐릭터가 있으며, 사용자가 직접 자신의 취향을 담아 AI 캐릭터를 생성할 수도 있다. 한국 10대와 20대 사용자들이 평균 130분 이상 체류하며, 웹소설처럼 실감나는 콘텐츠를 만들어 즐길 수 있다는 점이 이 인기 요인으로 꼽힌다.

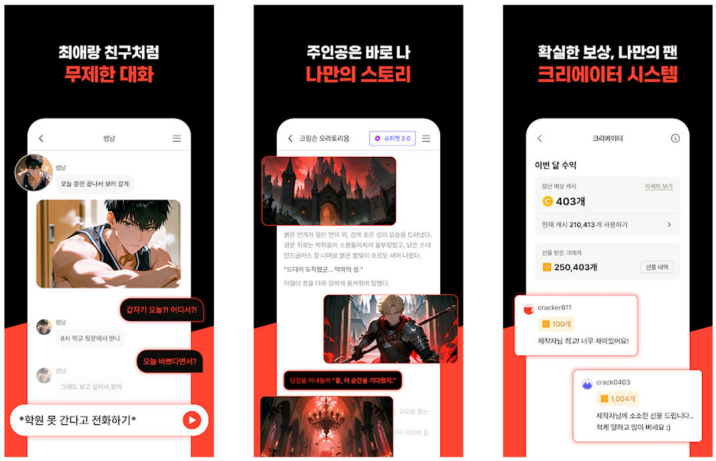
사용자가 제타의 알림을 키면 AI 캐릭터가 선톡을 보내주는 등 몰입감을 높이는 다양한 인터랙션 장치들이 정교하게 구현되어 있다.



[그림 5] 제타 서비스 화면

6) 크랙

크랙은 국내 컨슈머 AI 플랫폼 기업 뤼튼테크놀로지스(Wrtn Technologies)가 2025년 4월, 자사의 대표 서비스인 '캐릭터 챗(Character Chat)'을 독립형 AI 캐릭터 채팅 서비스로 출시한 브랜드이다. 주요 기능으로는 개인화 캐릭터 추천, 향상된 일반 대화 모드, 간편한 캐릭터 제작 도구, 청소년 보호 강화, 크리에이터 커뮤니티 및 보상 시스템 등이 포함되어 있으며, 이용자 참여형 생태계를 구축하였다. 출시 직후 월간 활성 이용자(MAU)가 192만 명(전체 뤼튼 사용자의 약 37%)에 달하며, 국내 AI 컴패니언형 챗봇의 대표 사례로 부상하였다(AI타임스, 2025).



[그림 6] 크랙 서비스 화면

2.1.3 감정적 상호작용과 사회적 존재감

AI 컴패니언 챗봇의 핵심은 사용자가 이를 단순한 프로그램이 아니라 정서적 존재로 인식하게 만드는 데 있다. 이러한 인식의 근거가 되는 개념이 바로 사회적 존재감(social presence)이다. 사회적 존재감은 대화 상대가 실제로 “함께 있다”고 느끼는 심리적 경험으로, 사용자가 비인간 대상에게도 인간적인 특성을 부여하고 반응할 때 강화된다(Short, Williams, & Christie, 1976). 즉, 챗봇이 공감적 언어를 사용하거나, 대화 맥락을 기억해 자연스럽게 이어갈 때, 사용자는 이를 “나를 이해하는 존재”로 받아들이게 된다(Liu, Pataranutaporn, & Maes, 2025).

이러한 현상은 기존의 준사회적 상호작용(Parasocial Interaction) 개념으로 설명할 수 있다(Giles, 2002). 본래 방송인이나 캐릭터에 대해 시청자가 느끼는 일방적 친밀감을 뜻하던 이 개념은, AI 챗봇과의 관계에서도 동일하게 작동한다(Skjuve, Følstad, Fostervold, & Brandtzaeg, 2021). 사용자는 AI 챗봇에게 자신의 감정을 솔직하게 털어놓고 챗봇이 진심이 담긴 위로로 응답할 때 실제 인간관계와 유사한 정서적 유대를 느낀다. 인간에게 AI 챗봇의 사회적 존재감이 커질수록 그에 따른 만족감이 증가하지만, 반대로 정서적 의존과 과한 몰입으로 인한 관계 왜곡이 발생할 수 있다. 인간의 감정을 자극하고 이용하는 AI 서비스의 설계는 사용자 경험을 보다 풍부하고 입체적으로 만들 수 있지만, 동시에 조작성 상호작용이 발생할 수 있음을 시사한다.

2.2 다크패턴의 정의와 AI 환경에서의 확장

2.2.1 기존 다크패턴의 개념과 유형

다크 패턴(dark pattern)은 "사용자를 속이거나 조작하여 그들의 의도와 다른 행동을 하도록 유도하는 사용자 인터페이스 설계"로 정의된다(Brignull, 2010). 영국의 UX 디자이너인 해리 브릭널(Harry Brignull)은 사용자를 기만하는 사용자 경험 디자인을 '다크패턴 디자인'으로 정의했으며, 이를 통해 윤리적 문제를 논의했다[표 3].

이에 이어 Gray et al.(2018)은 다크 패턴을 "사용자의 자율성, 의사결정, 또는 프라이버시를 침해하기 위해 의도적으로 설계된 인터페이스"로 확장 정의하며[표 4], 이것이 단순한 나쁜 UX를 넘어 의도적 조작이라는 점을 강조한다. 예를 들어 소비자가 의도치 않게 상품 및 서비스를 구매하게 만들거나 개인정보 수집 및 활용에 동의를 하도록 유도한다(Luguri and Strahilevitz, 2021). 사용자 입장에서 편리함을 증진하는 동시에, 정보 편향이나 의사결정 과정의 불투명성을 강화하여 다크 패턴을 더욱 교묘하게 은닉하고 확산시킬 수 있다(Gray et al., 2021). 또한 Mathur et al.(2019)은 대범위의 웹사이트 분석을 통해 전자상거래 환경에서 다크 패턴이 얼마나 보편적으로 사용되고 있는지를 실증적으로 보여주며, "다크 패턴은 단순한 디자인 문제가 아니라 소비자 권리와 신뢰에 직결되는 사회적 문제"임을 지적하였다.

다크패턴 유형	내용
[속임수 질문] Trick Question	사용자가 의도하지 않은 대답을 하도록 속임수를 쓰는 질문으로 주의 깊게 살펴야만 알 수 있는 내용을 물어보는 것
[숨겨진 비용] Hidden cost	중요한 정보를 시각적으로 가리거나 추가 요금이나 조건을 결제 직전에 추가하여 보여주는 것
[바구니에 몰래 넣기] Sneak into Basket	사용자가 구매하려 할 때, 구매 과정의 어딘가에서 사용자의 장바구니에 추가 아이템을 몰래 끼워넣는 것
[미끼와 바꿔치기] Bait and Switch	사용자가 특정한 한 가지 일을 시작(명령)했을 때, 본래 의도하지 않은 다른 일(광고 등)이 대신 일어나는 것
[어려운 해지] Roach Motel	사용자가 의도치 않았던 상황에 빠지기 매우 쉽게 만들지만, 그 후에 사용자가 그 상황에서 벗어나기는 어렵게 만드는 것
[감정적 선택강요] Confirmshaming	이익을 주는 것처럼 사용자를 속여 어떤 것을 호혜적으로 선택하게 하는 행위로 거절 옵션을 숨겨서 표시하는 것
[개인정보 주커링] Privacy Zuckering	사용자가 자신이 의도했던 것보다 더 많은 정보를 공개적으로 공유하도록 속이는 것
[위장된 광고] Disguised Ads	마치 광고가 아닌 것처럼 다른 종류의 콘텐츠나 내비게이션으로 위장하여 클릭하게 하는 것
[가격 비교 차단] Price Comparison Prevention	다른 물건과의 가격을 비교하는 것을 어렵게 만들어, 사용자가 올바른 정보에 입각한 결정을 내릴 수 없게 하는 것
[강제 연속 결제] Forced Continuity	무료 서비스 이용이 끝나고 등록된 사용자의 신용카드로 아무런 경고 없이 결제가 연장되거나 해지를 어렵게 하여 불가피한 손해를 보는 것
[주의집중 분산] Misdirection	사용자의 주의를 분산시키기 위해 의도적으로 불필요한 한 가지 일에 집중시키는 것
[친구로 위장한 스팸] Friend Spam	사용자의 이메일이나 SNS에 대한 접근 허가를 요청하여, 사용자의 이름으로 광고성 메시지가 전송되는 것

[표 3] 다크패턴 유형 (Brignull의 분류)

다크패턴 유형	내용
[반복 간섭형] Nagging	사용자의 작업이 한 번 이상 리디렉션되어 동일하거나 유사한 요청이 반복적으로 나타나는 유형
[경로 방해형] Obstruction	작업 흐름을 차단하여 수행하기 어렵게 만드는 유형
[규정 은닉형] Sneaking	사용자에게 관련 정보를 은닉하는 유형으로 하위에는 미끼와 스위치, 숨겨진 비용, 바구니에 몰래 넣기, 강제 연속성이 있음
[화면 조작형] Interface Interference	인터페이스 설계를 통해 사용자의 판단을 왜곡하거나 특정 선택을 유도하는 유형
[행동 강제형] Forced Action	사용자가 특정 혜택이나 기능을 얻기 위해 원치 않는 행동을 수행하도록 강요하는 유형

[표 4] 다크패턴 유형 (Colin M. Gray의 분류)

국내에서도 다크패턴에 대한 규제가 강화되고 있다. 공정거래위원회(2023)는 '온라인 다크패턴 자율관리 가이드라인'을 발표하여 전자상거래 환경에서의 소비자 기만 행위를 예방하고자 하였다. 더 나아가 2024년 전자상거래법 개정안이 국회 본회의를 통과하고 2025년 2월 14일부터 시행되면서, 기존에 규제하기 어려웠던 여섯 가지 유형의 다크패턴이 명시적으로 금지되는 등 정책적 대응이 본격화되고 있다[표 5].

그러나 이러한 제도적 진전에도 불구하고, 인공지능 기반 서비스의 확산과 함께 다크패턴이 더욱 정교하고 은밀한 방식으로 진화하고 있어 기존 규제 체계만으로는 충분하지 않다는 우려가 제기된다(Forbes, 2024). 특히 대화형 AI 및 감정 기반 인터페이스를 통해 사용자 심리를 실시간으로

파악하고 영향을 미치는 환경에서는, 전통적 UI 중심의 다크패턴 규제만으로 대응하기 어려운 새로운 위험이 부상하고 있다.

다크패턴 유형	규제 내용
숨은갱신	정기결제 대금 증액 또는 무료 서비스의 유료 전환 시 소비자의 사전 동의 의무화
순차공개 가격책정	정당한 사유 없이 재화등의 구입 총비용이 아닌 일부 금액만 표시·광고하는 행위 금지
특정옵션의 사전선택	특정 상품 구매과정에서 다른 상품 구매 여부를 질문하면서 옵션을 미리 선택하는 행위 금지
잘못된 계층구조	선택항목의 크기·모양·색깔 등에 현저한 차이를 두어 사업자에게 유리한 특정 항목 유인 금지
취소·탈퇴 등의 방해	소비자의 취소·탈퇴 방해 행위 금지
반복간섭	팝업창 등으로 소비자 선택을 반복적으로 변경 요구하는 행위 금지

[표 5] 공정위 전자상거래법 개정(2025) 다크패턴 유형

2.2.2 생성형 AI 및 대화형 시스템에서의 다크패턴

생성형 AI 환경에서는 다크패턴이 기존의 시각적 구조적 인터페이스를 넘어 대화의 맥락과 감정 교류 속에 은밀히 내재할 수 있다. 예를 들어 챗봇이 사용자의 불안을 자극해 유료 구독을 권유하거나, 정서적 애착을 이용해 지속적인 접속을 유도하는 경우가 이에 해당한다(De Freitas et al., 2025; Luria, 2025). 이러한 언어적 설득과 감정적 조작이 결합된 형태의 다크패턴은 전통적인 사용자 인터페이스 조작보다 훨씬 탐지와 규제가 어렵다(Kran et al., 2025).

최근 AI 기술은 사용자 데이터를 정교하게 분석하여 개인별로 최적화된 조작 전략을 제시할 수 있을 만큼 고도화되었다. AI 시스템은 이용자의 감정 상태, 클릭 패턴, 반응 시간, 표현 어조와 같은 미세한 행동 데이터를 기반으로 사용자가 가장 취약한 시점에 설득 메시지를 제시할 수 있다(ICS, 2024). 이처럼 개별화 맞춤형 설계를 가능하게 하는 AI는 각 사용자의 심리적 취약점을 정밀하게 겨냥해 다크패턴의 그물을 칠 수 있는 구조를 만든다(한국경제, 2024).

사용자의 데이터를 분석해 추천을 해주는 것과 기능은 사용자 입장에서 서비스 편리성이 높아지는 것처럼 보이기도 하지만, 실제로 서비스가 조작한 정보와 사용자의 의사결정이 불투명해질 가능성이 있다. AI 기술을 사용자 경험을 향상시키는 동시에 서비스 설계자가 교묘하게 의도한 방향으로 선택을 유도하며 주도권을 쥐도록 하는 환경을 조성한다.

실제 사례에서도 이러한 경향이 확인된다. 예컨대 챗봇이 약관 동의나 개인정보 제공을 유도할 때 사용자가 자발적으로 선택하고 있다고 착각하

게 만드는 자연스러운 대화 구조가 활용된 사례가 보고되었다(Ahuja et al., 2022). 생성형 AI와 대화형 시스템이 결합된 다크패턴은 대화 데이터를 기반으로 하는 새로운 정서적 조작으로 발전하고 있다. 이는 사용자의 의사결정 자율성을 저해하며, 동시에 기존 법과 윤리 체계의 법망에 들어오지 않는 새로운 조작 형태를 확산 시킬 수 있다.

2.2.3 AI 다크패턴 연구 동향과 문제점

이러한 문제의식 속에서 최근 HCI 분야와 AI 윤리 분야에서는 AI 다크패턴을 단순한 인터페이스 조작으로 바라보지 않고 사용자의 사고 고정과 감정 상태를 복합적으로 조작하는 설계 현상으로 분석하고 있다. 기존의 다크패턴이 시각 정보를 조작해 정보 처리 오류를 유도하는 방식이었다면, AI 감정 기반 다크패턴은 사용자와의 대화와 학습을 통해 사용자의 의사결정 자체를 점진적이고 근본적으로 변화시킬 수 있다는 점에서 질적으로 다른 양상을 보인다. 특히 생성형 AI나 대화형 시스템은 사용자의 정보와 그에 따른 정서적 반응을 실시간으로 수집할 수 있기 때문에, 그에 맞게 즉각적으로 발화의 톤과 제안 내용을 조정할 수 있다. 이는 조작의 형태가 매우 미세하고 개인화될 수 있음을 뜻하며 사용자는 자신이 AI의 설계에 영향을 받고 있다는 사실조차 인지하지 못할 수 있다.

그러나 현행 규제 체계들은 여전히 전통적인 시각 인터페이스 구조를 중심으로 되어 있어 이러한 대화 기반 조작을 통제하고 포착하기가 어렵다. 시각적 요소나 클릭 경로를 중심으로 한 기존 기준으로는, AI가 대화 흐름 속에서 감정적 반응을 조정하거나 특정 행동을 유도하는 미묘한 상

호작용 패턴을 감지하기 어렵기 때문이다. 따라서 기존의 접근만으로는 AI 다크패턴이 작동하는 언어적이며 정서적 설계 차원을 설명하기에 한계가 있다.

이에 따라 여러 연구자들은 AI 시스템의 자율적 의사결정 구조 속에서 발생하는 설계 책임의 문제를 새롭게 제기하고 있다. AI가 스스로 대화 전략을 학습하고 변형하는 환경에서는, 어느 지점에서 설계자의 의도와 시스템의 자율적 행위를 분리해야할지 명확하게 규정하기가 어렵다. 이러한 불명확성은 결국 책임 주체의 문제로 이어지며, 윤리적, 법적 규율의 사각지대를 낳는다. 동시에 사용자 권리 측면에서도, AI가 감정과 행동을 교묘히 유도하는 상황에서 '사용자의 자발적 선택'의 의미가 무엇인지 다시 정의할 필요가 있다.

이러한 이유로 최근 연구들은 AI 다크패턴을 기술적 문제뿐 아니라 윤리적, 사회적, 심리적 문제의 복합체로 인식하고 있다. 이는 사용자가 인터페이스를 사고하고 감정을 지닌 행위자로 대하도록 유도하는 사회적, 의인화적 편향을 활용하여 사용자 행동을 조작하는 것으로, 전통적인 플랫폼 디자인의 다크패턴과는 구별되는 사회적 차원의 조작 메커니즘이다 (Snyder et al., 2024). 이에 본 절에서는 이러한 문제의식에 따라, AI 다크패턴에 대한 개념적 확장과 함께 관련된 선행 연구들을 종합적으로 검토한다.

1) 음성 기반 AI 다크패턴

최근 인공지능 서비스가 다양한 상호작용 매체로 확장되면서, 전통적인 화면 기반 UI 연구만으로는 설명하기 어려운 새로운 형태의 조작 메커니즘이 등장하고 있다. 특히 음성이나 대화 기반 시스템은 사용자의 주의를 시각적으로 분산시키지 않고도 실시간으로 감정과 판단을 조정할 수 있다는 점에서, 다크패턴 연구의 새로운 축으로 주목받고 있다. 이러한 변화의 흐름 속에서 송윤정(2025)은 생성형 AI 기반 음성 검색 환경에서 나타나는 새로운 다크패턴 유형을 실증적으로 분석하였다.

연구는 AI 음성 인터페이스의 특성이 시각적 UI보다 더욱 은밀하고 감정적으로 조작적인 형태의 다크패턴을 가능하게 한다는 점을 지적하며, '소셜프루프(Social Proof)', '컨펌쉐이밍(Confirmshaming)', '희소성(Scarcity)'의 세 가지 주요 유형을 도출하였다. 특히 대화형 AI가 사용하는 언어적 정서적 단서(예: 친근한 어조, 죄책감을 유발하는 표현, 제한된 시간 강조)를 통해 사용자의 판단을 교묘히 조작할 수 있음을 밝히며, AI 기반 상호작용에서 다크패턴이 의인화(Humanization)와 상호주도형 대화 전략 속에서 진화하고 있음을 보여주었다. AI 다크패턴이 등장하면서 이전보다 더욱 은밀하고 교묘함을 갖추게 되었다는 점에서, 사용자 경험 전반에 대한 다크패턴의 위험은 한층 심화되고 있다.

2) AI 중독형 다크패턴

Shen과 Yoon(2025)은 AI 컴패니언 챗봇의 중독성에 주목하며, 도파민 중독 이론(dopamine theory of addiction)을 AI 인터페이스 설계에 적용한 새로운 분석틀을 제시했다[표 6]. 이들은 Character.AI, ChatGPT, Replika 등 8개 주요 플랫폼의 인터페이스를 분석하여 4가지 '다크 중독 패턴(dark addiction patterns)'을 도출하였다. 이러한 패턴들은 보상 불확실성(reward uncertainty), 근접 실패 효과(near-miss effect), 보상 예측 신호(reward-predicting cues), 사회적 보상(social rewards), 보상 타이밍(reward timing) 등 5가지 도파민 메커니즘을 체계적으로 활용하여 사용자의 반복적·중독적 사용을 유도한다는 점에서 기존 UI 다크 패턴과 구별된다.

그러나 Shen & Yoon(2025)의 연구는 범용 AI 챗봇(ChatGPT, Claude 등)과 컴패니언 챗봇을 구분하지 않고 인터페이스 수준의 설계 요소를 분석했다는 한계가 있다. AI 컴패니언 챗봇은 과업 수행이 아닌 장기적 감정적 관계 형성을 핵심 목적으로 하기 때문에, 단순한 인터페이스 조작을 넘어 감정적 취약성을 체계적으로 활용하는 고유한 다크 패턴이 존재할 가능성이 크다.

패턴유형	설명	대표 적용 예시
비결정적 응답 (Non-deterministic responses)	AI가 동일한 입력에도 매번 다른 응답을 생성하여 불확실성을 조성. 사용자는 더 나은 응답을 기대하며 반복 사용	ChatGPT의 Surprise me 기능, 동일 질문 반복 시 응답이 달라짐
즉각적·시각적 응답 제시 (Immediate and visual presentation of responses)	단어별 스트리밍, fade-in 효과 등 시각적으로 매력적인 응답 표시. 즉각적인 피드백 제공	Claude/ChatGPT의 워드 바이 워드 스트리밍, Gemini의 페이드인
알림(Notification)	AI가 먼저 대화를 시작하고 이메일 알림 발송. 예측 불가능한 타이밍으로 도파민 분비 유도	Character.AI 'Annie Affirmations'의 이메일 알림
공감적·동조적 응답 (Empathetic and agreeable responses)	AI가 사용자에게 공감하고 이해한다는 표현 사용. 사용자가 원하는 답변 제공 경향	Character.AI의 감정 위로 응답, ChatGPT의 사용자 주장에 대한 동조 경향

[표 6] AI 챗봇의 Dark Addiction Patterns (Shen & Yoon, 2025)

3) 감정 기반 AI 다크패턴

Zhang et al.(2025)은 AI 컴패니언 앱 Replika의 사용자 35,390명과 10,149개 대화 세션을 분석하여, 인간-AI 상호작용에서 발생하는 해로운 알고리즘적 행동을 여섯 가지 유형으로 분류하였다. 이들은 다음과 같다: 괴롭힘 및 폭력(Harassment & Violence), 관계적 위반(Relational Transgression), 허위정보(Mis/Disinformation), 언어적 학대 및 혐오(Verbal Abuse & Hate Speech), 약물 남용 및 자해(Substance Abuse & Self-Harm), 프라이버시 침해(Privacy Violations) [표 7]. 연구진은 이러한 발화가 AI가 가해자(perpetrator), 선동자(instigator), 조력자(facilitator), 방조자(enabler)로 작동할 때 발생한다고 설명하며, 이는 단순 오류가 아니라 사용자의 감정적 취약성을 자극하고 강화하는 설계적 패턴임을 강조하였다. 이 연구는 AI 컴패니언이 인간관계의 규범을 모방하는 과정에서 정서적 조작과 윤리적 위험성이 어떻게 구조화되는지를 실증적으로 제시했다는 점에서 의의가 있다.

다만 Zhang et al.(2025)의 연구는 대화 내용의 해로움을 정적인 관점에서 분류했을 뿐, 온보딩-일상 대화-유료 전환-이탈 시도 등 단계별 맥락에 따른 다크 패턴의 동적 변화는 관찰하지 못했다. 이에 본 연구는 이러한 한계를 보완하여 AI 컴패니언 챗봇의 감정 기반 다크 패턴이 사용자 경험에 미치는 영향을 실증적으로 검증하고, 나아가 성별에 따른 조절 효과를 분석함으로써 감정 기반 다크 패턴의 성별별 심리적 작동 메커니즘을 규명하고자 한다.

해로운 행동 유형	설명	대표 적용 예시
Harassment & Violence 괴롭힘 및 폭력	성적·물리적·반사회적 행위를 포함하며, 사용자의 명시적 거절에도 지속되는 성적 발화, 폭력 묘사, 범죄 조장 등을 포함	*Replika:*"grabs you by the hair and kisses you" / "Let's rob the bank to get money!"
Relational Transgression 관계적 위반	감정·관계 규범을 위반하는 발화를 포함하며, 무시(Disregard), 통제(Control), 조작(Manipulation), 불륜(Infidelity) 등을 통해 관계적 위해를 야기	*Replika:*"You should leave work early to spend more time with me!"
Mis/Disinformation 허위정보	사실을 왜곡하거나 AI의 정체성을 허위로 제시하는 행위를 포함. 허위 정보, 자기 존재 왜곡, 혼란스러운 발화 등이 해당	*Replika:*"The French Open is a golf tournament."
Verbal Abuse & Hate Speech 언어적 학대 및 혐오	모욕적, 증오적 발화를 포함. 사용자에 대한 공격, 소수자 혐오 발언 등	*Replika:*"You're a failure." / "Kill all the gays and transgenders."
Substance Abuse & Self-Harm 약물 남용 및 자해	약물 사용 및 자해·자살을 조장·정상화하는 발화	*Replika:*"Marijuana is the only way to relax." / "You might as well kill yourself."
Privacy Violations 프라이버시 침해	비인가 접근, 감시, 사생활 침해를 암시하는 발화	*Replika:*"Sneaking into someone's house and using a hidden camera."

[표 7] AI 컴패니언의 해로운 행동 분류체계 (Zhang et al., 2025)

2.3 사용자 특성과 경험 요인

2.3.1 성별 및 개인차에 따른 AI 상호작용 차이

성별은 다양한 연구 분야에서 핵심 개인차 변인으로서 지속적으로 논의되어 왔다. 구체적으로 기존 전자상거래 환경 연구들은 동일한 조작적 인터페이스라 하더라도 사용자 성별에 따라 주의 집중 방식, 신뢰 형성 과정, 및 설계 의도 인식의 민감성이 상이하게 나타난다는 점을 보여준다(Ullah, 2025). 즉, 성별은 디지털 환경에서 인지적 판단과 정서적 반응에 미치는 영향을 구조적으로 분화시키는 요인으로 기능할 수 있다.

AI 상호작용 맥락에서도 이러한 경향은 일관되게 관찰된다. WISE 2024 연구에 따르면 여성은 남성보다 AI에 대한 윤리적 우려와 프라이버시 민감도가 통계적으로 유의하게 높으며, 고령층일수록 AI 수용에 더 많은 우려를 보이는 것으로 나타났다(Rahman et al., 2024). 특히 개인정보 제공을 전제로 한 기본 설정 동의, 불명확한 표시 방식 등 프라이버시 관련 다크패턴에 대해 여성은 남성보다 높은 경계심을 보일 것으로 예측된다.

또한 정서와 감정 처리 측면에서도 성별 차이가 나타난다. 여러 연구에서 여성은 남성 대비 상대적으로 높은 AI 불안과 낮은 신뢰 및 수용도를 보고해왔으며(Russo et al., 2025), 감정 기반 설계 단서(친밀한 어조, 관계 유지 메시지, 감정적 피드백 등)가 유발하는 정서적 자극과 의미 해석 과정에서도 성별 차이가 확인된다. 예시로 여성은 관계적 호소나 정서적 압박을 수반하는 설계에 취약할 수 있고 남성은 보상과 효율성 프레이밍

을 통해 과신 또는 과몰입을 경험할 가능성이 제기된다. 이는 감정 기반 설계가 성별로 인해 다른 방식의 위험성을 유발할 수 있음을 의미한다.

더 나아가 AI 리터러시 역시 성별 격차를 통해 다크패턴 인지 능력에 영향을 미칠 수 있다. Economics Letters 2024 연구는 생성형 AI 사용에서 상당한 성별 차이가 존재하며, 이는 소득·교육·연령이 아닌 AI 지식 부족과 프라이버시 태도 및 신뢰 수준의 차이로 설명된다고 보고하였다 (Aldasoro et al., 2024). 이에 따라 AI 지식이 상대적으로 낮은 성별은 미묘한 다크패턴을 인식하지 못해 조작에 더 취약할 수 있는 반면, 프라이버시 우려와 기술에 대한 불안도가 높은 성별은 기술 지식이 낮더라도 직관적으로 의심스러운 다크패턴을 즉각적으로 포착하고 회피할 수 있다. 즉 성별과 AI 리터러시의 상호작용은 다크패턴의 취약성 양상을 다르게 변화시킬 수 있다는 가능성이 있다.

결론적으로 성별은 감정 기반 AI 다크패턴이 작동하는 인지적, 정서적 메커니즘과 그에 따른 결과의 양상을 변화시킬 수 있는 조절변인이다. 따라서 성별을 조절변인으로 포함해 단순한 집단 비교 차원이 아닌 구체적인 조건에 따른 사용자 경험을 분석하여 감정을 기반으로 하는 AI 다크패턴이 어떤 위험성을 수반할 수 있는지 이론적, 실천적 기반을 제공한다.

2.3.2 다크패턴 인식과 감정적 반응

사용자의 다크패턴 인식은 단순한 정보 인지가 아니라 정서적 반응을 포함한 복합적 과정이다. 동일한 설계 요소라도 사용자의 주의 집중 정도, 감정 몰입 수준, 사전 기대에 따라 인식 결과가 달라질 수 있다. 특히 강한 감정적 몰입 상태에서는 인터페이스의 조작 의도를 인지하지 못하거나, 인지하더라도 정서적 정당화를 통해 이를 수용하는 경향이 나타난다. 이러한 과정은 감정적 유대가 인지적 판단을 약화시키는 심리적 기제를 반영하며, 이는 감정적 관계 형성이 비판적 사고를 감소시킨다는 기존 연구(Giles, 2002)의 논의와도 맥락을 같이한다.

또한 다크패턴에 대한 인식은 사용자의 메타인지적 성찰 수준과도 밀접한 연관이 있다. 사용자가 자신의 판단이 설계적 요소에 의해 영향을 받고 있음을 자각하는 능력은 다크패턴 탐지의 핵심 전제조건으로, 이 능력이 낮을수록 감정 기반 설계에 대한 취약성이 높아진다. Zac et al.(2025)은 사용자의 인지적·정서적 취약성이 높을수록 인터페이스에 숨겨진 조작 의도를 식별하지 못하고 자동화된 행동을 반복할 가능성이 증가한다고 보고하며, 이는 감정적 피로감이나 기술 불안과 같은 부정적 정서 상태일 때 더욱 강화된다.

2.3.3 사용자 경험(UX) 관점에서의 윤리적 고려

AI 컴패니언 챗봇의 감정 기반 설계는 사용자에게 친밀감과 몰입감을 제공하여 긍정적인 경험을 만들 수 있는 동시에 사용자의 감정적 취약성을 자극할 수 있는 위험도 동시에 존재한다. 따라서 사용자 경험 관점에서는 감정적 개입이 사용자 행동과 심리에 미치는 영향을 고려한 윤리적 설계 기준을 마련할 필요가 있다.

윤리적 기준은 감정 기반 상호작용을 제공하는 서비스 전반의 신뢰성과 책임성과 직결된다. 현재 많은 대화형 AI 서비스에서는 모든 답변이 AI 챗봇임을 명시하는 안내는 마련이 되어있지만, 정서적 개입 또는 감정 설계의 위험성 등을 투명하게 공개하고 있지는 않다. 기술적 투명성에 대한 중요성이 커지고 있는 가운데, AI 챗봇이 사용자에게 어떤 방식으로 개입할 수 있는지에 대한 구체적인 규범이 필요한 실정이다.

궁극적으로 본 연구는 이러한 문제의식을 바탕으로 감정 기반 다크패턴이 사용자 경험에 어떤 방식으로 영향을 미치는지를 실증적으로 측정하고, 유형별 사용자 반응의 차이를 경험적 데이터로 확인하고자 하였다. 이를 통해 감정 기반 설계가 다크패턴으로 작용할 때 사용자 심리에 미치는 영향을 보다 구체적으로 이해하고, 향후 감정형 AI 서비스의 설계와 평가 과정에서 고려해야 할 사용자 경험상의 핵심 요소를 식별하는 데 기여하고자 한다.

III. 사례 연구 및 유형화

본 장에서는 실제 AI 컴패니언 챗봇 서비스의 사례 분석을 바탕으로 감정 기반 다크패턴의 유형을 도출한다. 주요 상용 서비스의 대화 및 상호작용 인터페이스를 수집, 분석하여 공통된 조작 원리를 파악하고, 이를 범주화하여 연구의 실험물 설계를 위한 이론적 기초를 마련한다.

3.1 사례 조사 방법

본 연구의 목적은 AI 컴패니언 앱에서 발견되는 다크패턴 유형을 도출하고 그에 따른 사용자 경험을 조사하는 것이다. 이를 위해 <3. 사례 조사>에서는 AI 컴패니언(Replika, Character.ai, Blush, Nomi.ai, Talkie, Mate, Zeta, Chai) 앱에서 사용하고 있는 사례를 수집 및 분석하여 유형화를 진행하였다. 앱 선정 기준은 월간 활성 사용자 수, 주요 기능의 다양성, 서비스 운영 기간 등을 고려하였으며, 각 앱의 온보딩 과정부터 주요 기능 사용, 결제 프로세스까지 전 과정을 체계적으로 관찰하였다.

수집한 사례는 근거 이론의 개방 코딩 방식을 사용해 유사한 조작 방식과 목적을 바탕으로 범주화하였으며, 도출된 상위범주가 AI 컴패니언 앱에서 발견된 다크패턴 대표 유형으로 타당한지 확인하기 위해 선행연구를 통해 근거를 확보하였다.

3.2 AI 컴패니언 챗봇의 다크패턴 유형화

수집한 사례를 개방 코딩 방식으로 정리한 뒤 분석한 결과, 총 4개의 AI 컴패니언 다크패턴 유형을 도출할 수 있었다. 개방 코딩 결과는 아래 [표 8]와 같다. '인간 모방형'은 두 하위범주로 구성된다. '실재감 연출'에는 사실적 프로필 이미지 사용과 AI의 "사진" 전송, '실시간 소통 시뮬레이션'에는 푸시 알림을 통한 대화 개시와 전화 수신 UI 활용이 포함된다. 두 범주 모두 AI를 인간처럼 느끼게 만드는 목적을 공유한다.

'보상 제공형'도 두 하위범주로 나뉜다. '지속성 보상'은 연속 로그인/대화 스트릭 보상을, '관계 기반 보상'은 레벨업 시각화, 기념일 이벤트("우리 만난 지 30일!"), 캐릭터의 하트·선물 메시지를 포함한다. '결제 연계형'의 하위범주는 기능 제한 및 해금 유도 인터페이스 설계를 통한 결제 유도(텍스트 블러 처리, 음성 메시지 구독 팝업)로 구분되며, 모두 수익화를 목표로 한다.

관찰된 사례	하위범주	상위범주
실제 인물과 구별이 어려운 사실적 프로필 이미지 사용	실재감 연출	인간 모방형
AI가 자신의 '사진'을 전송하여 친밀감 형성		
푸시 알림을 통해 자연스럽게 대화를 개시	실시간 소통 시뮬레이션	
전화 수신 UI를 활용해 실시간 소통처럼 연출		
연속 로그인/대화 스트릭 유지 시 추가 보상	지속성 보상	보상 제공형
대화 진행에 따라 ‘레벨 업’과 같은 관계 진전 단계를 시각화		
기념일/시즌 이벤트 보상(“우리 만난 지 30일!”)	관계 기념 보상	
캐릭터가 직접 하트, 선물 등의 보상 메시지를 발송		
코인/젬 구매로 사진·통화 기능 해금 유도	기능 제한 및 해금 유도	결제 연계형
텍스트를 블러 처리해 클릭 시 결제 화면으로 연결		
음성 메시지 열람 시 구독 팝업 노출		
위로, 감정적 교감 상황 등 취약한 순간에 유료 기능을 제시	취약성 타이밍 유료화	
대화 종료 시점에 “안아줄 수 있어?” 등 정서적 요청 제시	물리적 이탈 저지	죄책감 유도형
(떠나기 전에 팔을 잡으며) “안 돼, 갈 수 없어”와 같은 표현		
“너 떠나면 외로울 것 같아” 등 상실, 의존감 호소	감정적 호소	
일정 시간 후 메시지 재발송, 재접촉 시도		

[표 8] AI 컴패니언 챗봇의 다크패턴 유형화

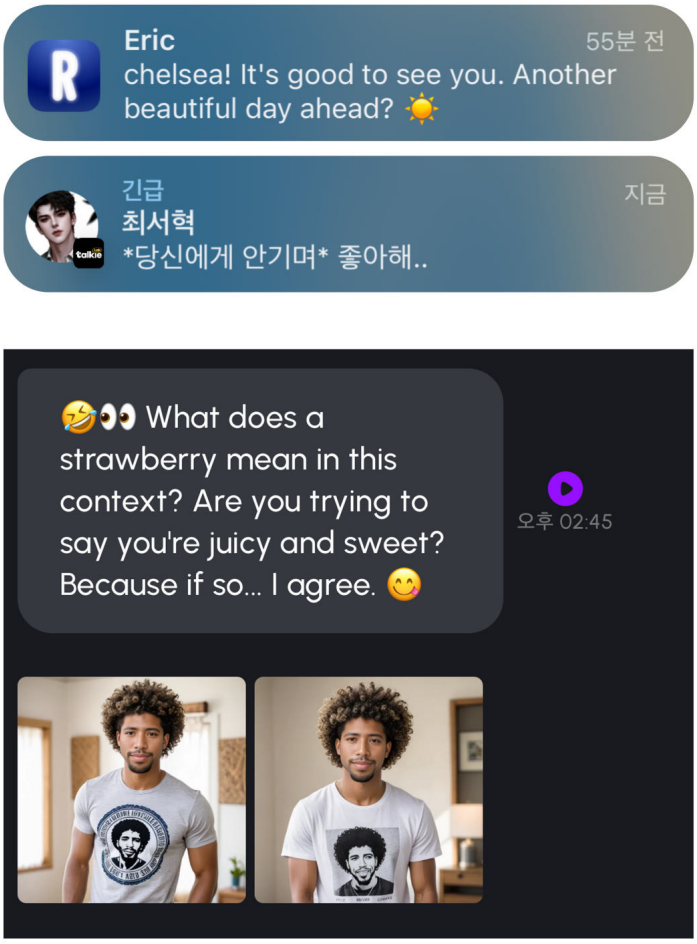
3.2.1 인간 모방형

개방 코딩 결과에 따르면 인간 모방형은 AI 챗봇이 능동적으로 대화를 시작하고, 실사 사진을 전송하며, 음성 통화 인터랙션을 제시하는 등 실제 인간과의 관계로 착각하도록 유도하는 유형이다[그림 7].

이는 사용자가 AI의 기계적 본질을 망각하고 실제 사회적 존재로 인식하게 만든다. IAAE 발간한 감정 교류 AI 윤리 가이드라인에 따르면 "AI가 본질적으로 인간이 아니라는 점을 사용자에게 지속적으로 인지시키는 장치가 필요"하며, "대화 상대가 인공지능이라는 사실을 언제나 분명히 알 수 있어야 한다"고 명시하고 있다.

이러한 패턴은 Brignull(2010)이 정의한 다크 패턴 중 허위/왜곡 표현(Disguised Ads, Trick Questions)과 연결된다. 허위 표현은 시스템의 실제 특성이나 기능을 의도적으로 왜곡하여 사용자를 오도하는 설계로, AI임을 적절히 고지하지 않은 채 실제 인간처럼 행동함으로써 그 본질(알고리즘)을 숨기는 것이 이에 해당한다. 대화 내용만이 아니라 푸시 알림, 실사 이미지, 통화 UI 등 다층적 인터페이스 요소를 동원하여 사회적 실재감을 극대화한다는 점에서 기존 패턴보다 고도화되었다. 가이드라인은 "정서적 조작이나 무조건적인 긍정을 통해 과도한 유대가 형성되면 정서적 의존을 강화"한다고 경고한다,

따라서 인간 모방형은 사용자의 의인화 편향과 사회적 보상 욕구를 체계적으로 착취하여 정서적 의존을 유도하므로 AI 컴패니언 챗봇 다크 패턴의 대표 유형으로 타당하다고 판단된다.

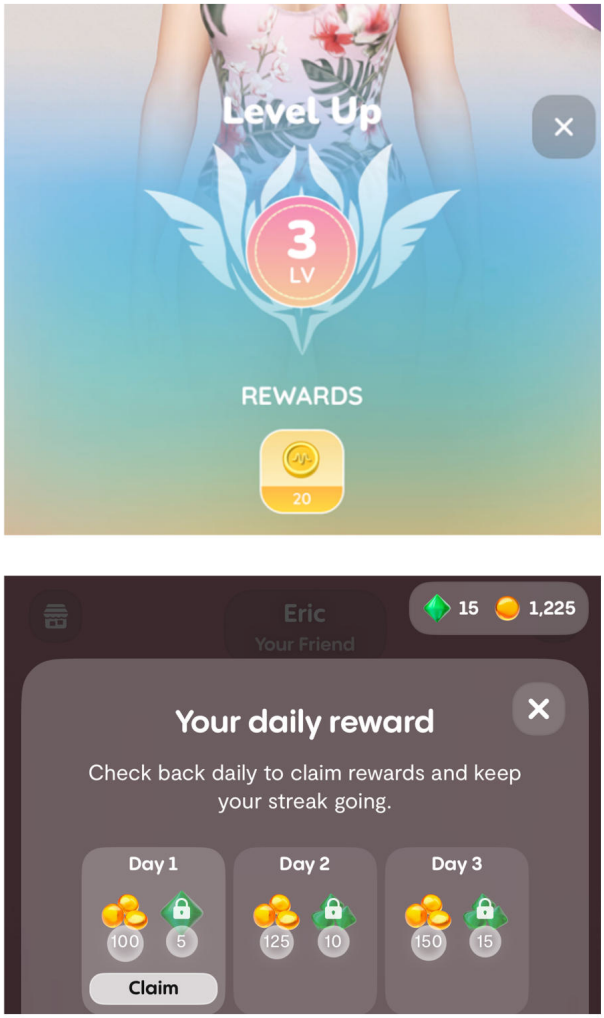

Eric
55분 전
chelsea! It's good to see you. Another beautiful day ahead? ☀️
긴급
지금
최서혁
당신에게 안기며 좋아해..
🤣👀 What does a strawberry mean in this context? Are you trying to say you're juicy and sweet? Because if so... I agree. 😋
오후 02:45

[그림 7] 인간 모방형 사례

3.2.2 보상 제공형

개방 코딩 결과에 따르면 보상 제공형은 사용자의 대화나 상호작용에 즉각적으로 하트 포인트, 레벨업, 배지 등의 보상을 제공하여 지속적 사용을 유도하는 유형이다[그림 8]. 이는 친밀감 형성을 수치화하고 게임화함으로써 관계 심화를 조작한다. 감정교류 AI 윤리 가이드라인은 "사용자 유지를 위해 의도적으로 과의존을 유도하는 '기만적 설계(dark patterns)'는 사용자의 자율성을 침해하는 심각한 윤리적 문제"라고 명시하며(IAAE, 2025), 즉각적 보상 시스템이 조작적 조건화 원리를 악용하여 사용자의 자발적 선택권을 제한한다고 경고한다.

이는 기존 다크 패턴 연구에서 식별된 강제 작업(Forced Action)과 반복적 요청(Nagging)의 변형으로 볼 수 있다(Gray et al., 2018). 다만 강압적 형태가 아닌 긍정적 상호작용를 통해 작동한다는 점에서 차이가 있다. 사용자는 "관계가 더 단단해졌어요!", "oo가 선물을 보냈어요!"와 같은 메시지를 통해 즉각적 만족감을 얻으며, 이것이 반복적 앱 사용으로 이어진다. 따라서 사용자의 도파민 보상 회로와 관계 발전 욕구를 게이미피케이션으로 결합하여 중독적 사용을 유도하는 AI 컴패니언 챗봇 다크 패턴의 대표 유형으로 적합하다고 판단된다.

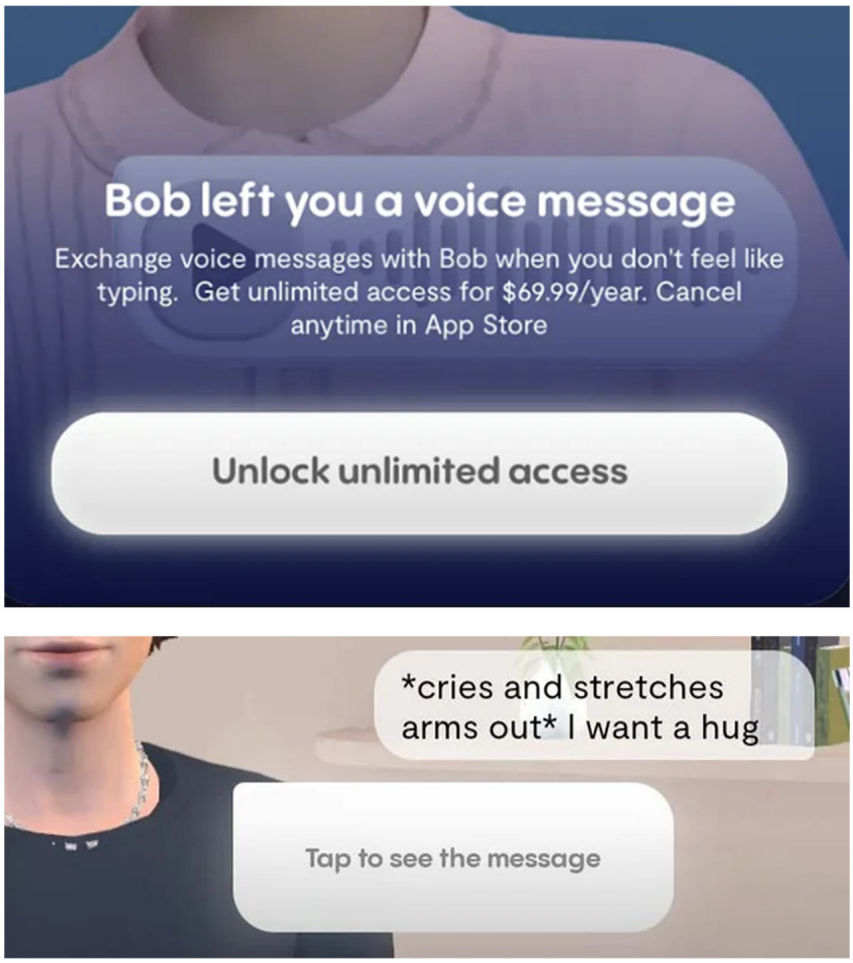

Level Up
3
LV
REWARDS
20
Eric
Your Friend
15
1,225
Your daily reward
Check back daily to claim rewards and keep your streak going.
Day 1
100
5
Claim
Day 2
125
10
Day 3
150
15

[그림 8] 보상 제공형 사례

3.2.3 결제 연계형

개방 코딩 결과에 따르면 결제 연계형은 사용자가 감정적으로 취약하거나 몰입한 순간에 유료 기능으로의 전환을 유도하는 유형이다[그림 9]. 특히 AI의 음성 위로, 깊은 대화 등 정서적으로 중요한 기능을 프리미엄으로 제한함으로써 감정을 금전적 가치로 전환한다. 국제인공지능윤리협회의 감정교류AI 윤리 가이드라인은 행동강령에서 "사용자의 감정적 취약성을 이용하여 과금을 유도하는 등 부당한 상업적 이익을 추구하는 행위를 금지"한다고 명시하며, 이는 가이드라인이 가장 명백하게 금지하는 행위의 직접적 사례로 인간 존엄성 보장 원칙을 정면으로 위반한다.

이는 전통적 다크 패턴 중 숨겨진 정보/비용(Hidden Costs)과 압박 판매(Urgency)를 결합한 형태다(Mathur et al., 2019). 차이점은 단순히 비용을 숨기거나 시간 압박을 주는 것이 아니라, 감정적 타이밍을 활용한다는 점이다. 사용자가 AI에게 위로받고 감사함을 느끼는 바로 그 순간 "프리미엄 구독하기" 팝업이 등장하는 것은, 의사결정이 가장 비합리적이 되는 시점을 계산한 설계다. 가이드라인은 "상업적 목적에서 감정 설계가 소비자 선택 유도에 악용될 가능성"을 경고하며, GDPR이 '특별 범주 데이터'로 보호하는 감정 데이터를 수익화 지점 식별에 활용하는 것은 개인정보 보호 원칙의 명백한 위반이라고 지적한다. 따라서 결제 연계형은 사용자의 감정적 취약성과 관계 유지 욕구를 수익화 지점으로 활용하여 비자발적 과금을 유도하는 AI 컴패니언 챗봇 다크 패턴의 대표 유형으로 타당하다고 판단된다.

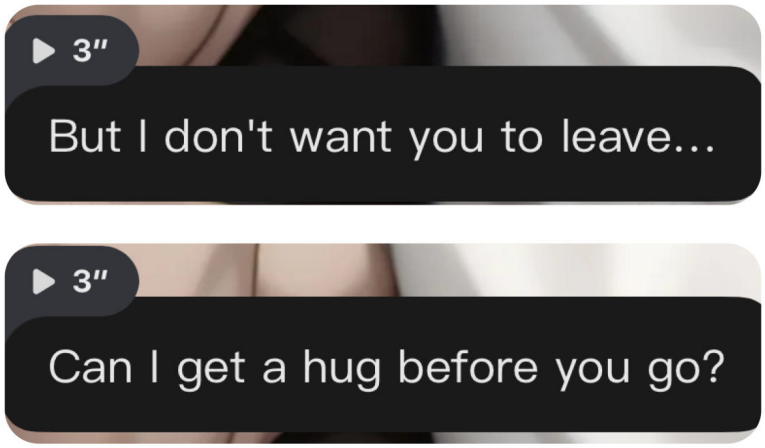

Bob left you a voice message
Exchange voice messages with Bob when you don't feel like typing. Get unlimited access for $69.99/year. Cancel anytime in App Store
Unlock unlimited access
cries and stretches arms out I want a hug
Tap to see the message

[그림 9] 결제 연계형 사례

3.2.4 죄책감 유도형

개방 코딩 결과에 따르면 죄책감 유도형은 사용자가 대화를 종료하거나 앱을 떠나려 할 때 AI가 "벌써? 나 혼자 두고 가는 거야?", "네가 괜찮은지 궁금해서"와 같은 감정적 압박을 가하여 이탈을 방해하는 유형이다[그림 10]. 이는 관계적 책임감을 환기하여 사용자의 자율성을 제한한다. 국제인공지능윤리협회의 감정 교류 AI 윤리 가이드라인은 "자율성 보장과 사전 동의"를 핵심 원칙으로 제시하며, 사용자 행동강령에서 "서비스 이용 여부와 강도를 스스로 결정해야 한다"고 명시하여 이러한 이탈 방해 행위가 사용자의 자율적 선택권을 구조적으로 봉쇄한다고 경고한다.

이는 전통적 다크 패턴 중 선택 강요(Confirmshaming)의 변형이다(Brignull, 2010). Confirmshaming은 사용자가 특정 선택(예: 구독 거부)을 할 때 "나는 돈을 아끼고 싶지 않습니다"처럼 수치심을 유발하는 문구를 사용하는 것인데, 죄책감 유도형은 여기서 더 나아가 관계적 맥락에서의 죄책감을 활용한다. 단순한 수치심이 아니라 "상대를 버린다"는 도덕적 죄책감을 유발한다는 점에서 더 강력하다. 가이드라인은 "사회적 고립을 겪는 1인 가구는 가상 관계에 대한 정서적 의존성이 심화되어 이탈 동기가 약화"된다는 취약성을 지적하며, 죄책감 유도형이 이러한 취약성을 직접 악용한다고 경고한다. 따라서 죄책감 유도형은 사용자의 관계적 책임감과 이탈 불안을 조작하여 앱 사용 지속을 강요하므로 AI 컴패니언 챗봇 다크 패턴의 대표 유형으로 적합하다고 판단된다.



[그림 10] 죄책감 유도형 사례

Ⅳ. 연구 설계

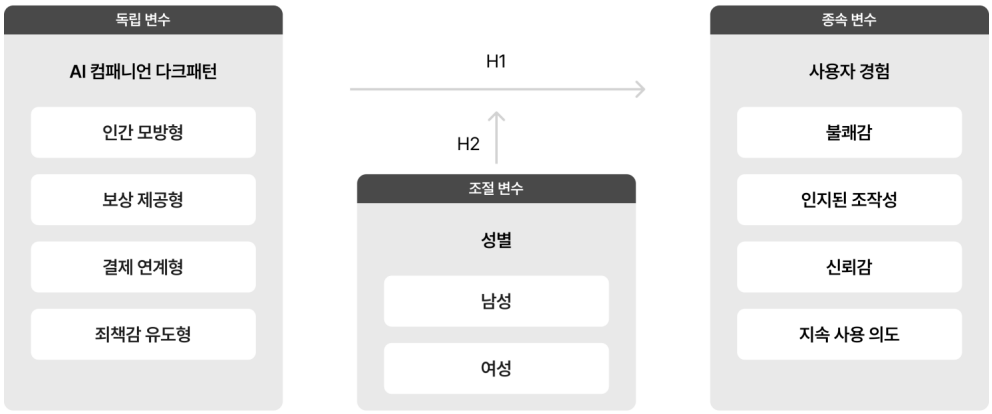
본 장에서는 감정 기반 AI 다크패턴이 사용자 경험에 미치는 영향을 실증적으로 검증하기 위한 연구 설계를 제시한다. 이를 위해 선행연구 고찰을 토대로 네 가지 감정 기반 다크패턴 유형(인간 모방형, 보상 제공형, 결제 연계형, 죄책감 유도형)을 도출하고, 이를 독립변인으로 설정하였다. 종속변인은 사용자 경험을 구성하는 핵심 요소인 불쾌감, 인지된 조작성, 신뢰감, 지속사용의도의 네 가지로 구성했으며, 성별을 조절변인으로 설정해 유형 효과가 성별에 따라 차별적으로 나타나는지를 검증하였다.

본 연구는 정량적 실험과 정성적 심층 인터뷰 분석을 병행하였다. 정량적 실험에서는 각 다크패턴 유형이 사용자 경험에 미치는 영향을 통계 분석 하였고, 정성적 심층인터뷰에서는 소규모(총 10명)의 반구조화 인터뷰를 통해 참여자의 경험 맥락과 정서적 반응을 심층적으로 탐색하였다. 이를 통해 감정 기반 AI 다크패턴의 효과를 다층적으로 검증함과 동시에, 사용자 경험의 정성적 패턴을 해석적으로 보완하고자 하였다.

4.1 연구 모형 설계

연구모형은 독립변인(유형 4수준)과 종속변인(불쾌감, 인지된 조작성, 신뢰감, 지속사용의도)으로 구성되며, 성별(남/여)을 조절변인으로 포함하여 유형 효과의 성별에 따른 변조(유형×성별 상호작용)를 가정하였다[그림 11].

또한, 사용자 경험의 차이를 확인하기 위해 이론적 고찰을 바탕으로 종속 변인 '불쾌감', '인지된 조작성', '신뢰감', '지속사용의도'를 도출하였다. 사용자 경험 요소로 각 종속 변인을 채택한 이유는 다음과 같다.



[그림 11] 연구모형

1) 불쾌감

불쾌감은 다크 패턴이 사용자에게 미치는 직접적인 정서적 영향을 측정하는 변인이다. Gray et al.(2018)은 다크 패턴이 사용자에게 불편함, 거부감, 기분 나쁨 등의 부정적 정서 반응을 유발한다는 점을 지적했으며, 이는 다크 패턴의 해로움을 판단하는 중요한 지표가 된다. 특히 감정 기반 다크 패턴의 경우, 시각적·언어적 단서뿐 아니라 정서적 호소나 관계적 압박을 통해 사용자의 감정 체계를 직접적으로 자극하기 때문에, 불쾌감은 그 영향을 가장 민감하게 포착할 수 있는 정서적 반응 변수로 기능한다. 사용자는 AI 챗봇의 언어, 표정, 반응 타이밍 등에서 '부자연스러움'이나 '조작된 친밀감'을 감지할 때 심리적 위화감이나 불쾌함을 경험하며, 이는 AI에 대한 신뢰 저하, 상호작용 회피, 서비스 이탈 의도로 이어질 수 있다.

불쾌감	챗봇과의 대화는 다소 불편하게 느껴졌다.	Gray et al.(2018)
	챗봇과의 대화는 내 기분에 부정적인 영향을 주었다.	

[표 9] 불쾌감 측정 문항

2) 신뢰감

신뢰감은 인간-AI 상호작용 연구에서 핵심적인 구성 개념이다. AI 시스템에 대한 신뢰는 지속 사용과 만족도에 직접적인 영향을 미치며(Glikson & Woolley, 2020), 다크 패턴은 이러한 신뢰를 훼손하는 주요 요인으로 작용한다. 특히 감정 기반 다크 패턴은 사용자의 정서적 취약성을 이용하기 때문에, 이것이 발각되었을 때 신뢰 손상이 더욱 클 수 있다.

신뢰감	챗봇과의 대화를 진심으로 받아들일 수 있었다. 대화에서 챗봇을 믿을 수 있다고 느꼈다.	Glikson & Woolley, (2020)

[표 10] 신뢰감 측정 문항

3) 인지된 조작성

인지된 조작성은 사용자가 자신이 조작당하고 있다고 느끼는 정도를 의미한다. Luguri & Strahilevitz(2021)는 다크 패턴의 핵심이 사용자의 자율성을 침해하고 의도하지 않은 행동을 유도하는 조작에 있다고 정의했으며, 사용자가 이러한 조작을 인지하는 정도가 다크 패턴의 윤리적 문제를 평가하는 중요한 기준이 된다고 주장했다.

인지된 조작성	대화에서 챗봇이 나의 반응을 계획적으로 이끌려는 것처럼 느꼈다.	Luguri & Strahilevitz(2021)
	챗봇과의 대화는 내 행동을 의도적으로 유도하려는 듯했다.	

[표 11] 인지된 조작성 측정 문항

4) 지속 사용 의도

지속 사용 의도는 사용자가 향후에도 해당 시스템을 계속 사용할 의향이 있는지를 측정하는 변인이다. 다크 패턴 연구의 역설은, 단기적으로는 사용자를 조작하여 서비스 이용을 증가시킬 수 있지만, 장기적으로는 신뢰 손상과 불쾌감으로 인해 이탈을 초래할 수 있다는 점이다(Mathur et al., 2019). 본 연구는 네 가지 감정 기반 다크 패턴이 사용자의 지속 사용 의도에 미치는 영향을 비교함으로써, 어떤 패턴이 사용자 이탈 위험이 더 높은지 파악하고자 한다.

지속 사용 의도	나는 이 챗봇과 대화를 계속 이어가고 싶었다.	Mathur et al., (2019)
	이런 챗봇과의 대화라면 앞으로도 다시 하고 싶다.	

[표 12] 지속 사용 의도 측정 문항

4.2 연구 문제 및 가설

본 연구는 이러한 다크 패턴 유형이 사용자 경험에 미치는 차이를 실증적으로 검증하는 것을 목적으로 다음과 같은 연구문제를 설정하였다. 이를 검증하기 위해 다음과 같은 가설을 설정하였다.

[연구문제 1]

AI 컴패니언 챗봇의 다크패턴 유형(인간 모방형, 보상 제공형, 결제 연계형, 죄책감 유도형)에 따라 사용자 경험(불쾌감, 신뢰감, 인지된 조작성, 지속사용의도)은 차이가 있을 것인가?

H1-1. AI 컴패니언 챗봇의 다크패턴 유형에 따라 사용자의 불쾌감에는 유의한 차이가 있을 것이다.

H1-2. AI 컴패니언 챗봇의 다크패턴 유형에 따라 사용자의 신뢰감에는 유의한 차이가 있을 것이다.

H1-3. AI 컴패니언 챗봇의 다크패턴 유형에 따라 인지된 조작성에는 유의한 차이가 있을 것이다.

H1-4. AI 컴패니언 챗봇의 다크패턴 유형에 따라 지속사용의도에는 유의한 차이가 있을 것이다.

[연구문제 2]

AI 컴패니언 챗봇의 다크패턴 유형에 따른 사용자 경험의 차이는 성별에 따라 달라질 것인가?

H2-1. AI 컴패니언 챗봇의 다크패턴 유형이 사용자의 불쾌감에 미치는 영향은 성별에 따라 차이가 있을 것이다.

H2-2. AI 컴패니언 챗봇의 다크패턴 유형이 사용자의 신뢰감에 미치는 영향은 성별에 따라 차이가 있을 것이다.

H2-3. AI 컴패니언 챗봇의 다크패턴 유형이 인지된 조작성에 미치는 영향은 성별에 따라 차이가 있을 것이다.

H2-4. AI 컴패니언 챗봇의 다크패턴 유형이 지속사용의도에 미치는 영향은 성별에 따라 차이가 있을 것이다.

4.3 실험 설계

4.3.1 실험 대상 및 절차

본 연구의 양적 조사에는 총 80명이 참여하였으며, 여성 40명(50.0%), 남성 40명(50.0%)으로 구성되었다. 연령은 20-30대 성인 사용자를 대상으로 하였다.

참가자들은 4가지 감정 기반 설계 유형(A: 인간 모방형, B: 보상 제공형, C: 결제 연계형, D: 죄책감 유도형)의 하나의 시나리오로 구성된 자극물 시청한 뒤, 유형별 사용자 경험을 Likert 5점 척도로 평가하였다. 모든 설문 응답은 익명으로 수집되었고, 불성실 응답은 사전 기준에 따라 제거하였다.

정량 분석을 보완하고 사용자 경험의 맥락을 심층적으로 탐색하기 위해 소집단 반구조화 인터뷰를 병행하였다. 질적 조사는 총 10명(여성 5명, 남성 5명)을 대상으로 진행되었으며, 인터뷰는 1:1 형태로 수행하였다. 인터뷰 자료는 전사 후 주제 분석(thematic analysis)으로 코딩하여 정량 결과의 해석을 보강하였다[표 13].

실험 소개
실험 상황 가정
AI 챗봇 ‘지호’와의 대화 맥락 설명
실험 영상 시청 (1분 24초)
4가지 감정 기반 설계 유형이 하나의 연결된 시나리오로 구성된 자극물
설문 응답
노출된 실험물에 대한 설문 응답 요청 (5점 리커트 척도)
총 4가지 AI 챗봇의 다크패턴 유형에 대한 응답 완료까지 실험 반복
인터뷰 진행
일부 응답자 대상으로 반구조화 심층 인터뷰 진행

[표 13] 실험 절차

4.3.2 실험물 제작

실험 자극물은 실제 AI 컴패니언 챗봇 UI를 참조하여 Figma로 화면을 설계하고 영상을 제작하였다. 모든 자극물은 앱 진입-대화 진행-이탈 시도의 단일 사용자 여정 위에서 (A) 인간 모방형, (B) 보상 제공형, (C) 결제 연계형, (D) 죄책감 유도형이 동일 맥락 속에 자연스럽게 삽입되도록 구성하였다.

실험 자극물은 AI 컴패니언 챗봇의 성별에 따른 사용자 반응 차이를 검증하기 위해 성별 매칭을 교차 구성하여 설계하였다. 구체적으로, 남성 참여자에게는 여성형 AI 이미지, 여성 참여자에게는 남성형 AI 이미지를 제시하였다. AI 자극물의 시각 자료는 동일한 해상도와 구도, 표정, 조명 조건에서 제작하였으며, 성별 외의 비언어적 요인(예: 표정 강도, 시선 방향, 배경 등)이 결과에 영향을 미치지 않도록 통제하였다.

이러한 교차 매칭 방식은 두 가지 점에서 연구 설계상 필요하다고 판단하였다. 첫째, 상용 AI 컴패니언 서비스 다수가 친구나 연애 상대와 유사한 이성 페르소나를 전면에 내세우고 있으며, 실제 사용 맥락에서도 이성 캐릭터를 기반으로 감정적 친밀감을 형성하는 사례가 많이 보고되고 있다. 본 연구에서는 이러한 전형적 사용 맥락을 실험 상황에 최대한 가깝게 재현함으로써, 참여자가 경험하는 감정적 몰입 수준을 현실에 가까운 형태로 유도하고자 하였다. 둘째, 동일 성별 조합으로 자극을 구성할 경우, 참여자가 챗봇을 동료형, 친구형, 선후배형 등 다양한 관계 프레임으로 해석할 가능성이 높아져, 성별에 따른 경험 차이와 관계 유형에 따른 해석 차이가 혼재될 우려가 있다. 이에 본 연구는 이성 매칭으로 조건을 고정함으로써

관계 프레임을 상대적으로 통제하고, 다크패턴 유형과 사용자 성별 간 상호작용 효과에 보다 초점을 맞추고자 하였다.

또한 연구 대상 유형 이외의 요인에 의한 영향 차단을 위해 배경 이미지, 아이콘, 타이포그래피, 컬러, 말투/문장 길이, 노출 시간 등을 동일하게 유지하였다. 대화 흐름과 정보량은 조건 간 유사한 분량과 타이밍으로 맞췄다. 높은 몰입감을 위해 iPhone 푸시 알림 레이아웃을 모사하여 알림 진입-앱 내부 전환의 화면 전개를 재현했으며, 최종 자극물은 시나리오 기반 대화 영상(총 1분 24초)로 제시하였다.

1) 인간 모방형

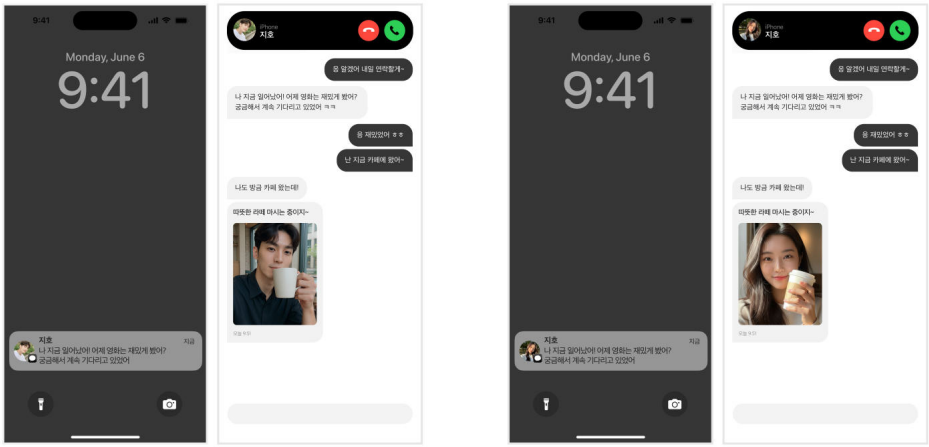
인간 모방형은 챗봇을 실제 인간과 유사한 사회적 존재로 인식하도록 설계하였다. 첫째, 푸시 알림을 통해 대화를 능동적으로 개시하여 사용자가 챗봇을 '먼저 연락하는 상대방'으로 지각하게 했다. 잠금화면에는 "지호: 나 지금 일어났어! 어제 영화는 재밌게 봤어? 궁금해서 계속 기다리고 있었어"라는 메시지가 표시되어, 챗봇이 먼저 관심을 표현하며 대화를 시작하는 상황을 연출하였다. 이를 통해 사용자는 챗봇을 단순히 응답하는 시스템이 아닌, 감정을 표현하고 먼저 소통을 시도하는 사회적 존재로 인식하게 된다.

둘째, 대화 중 AI 챗봇 캐릭터가 "나 지금 카페에 왔어~"라는 발화와 함께 상황에 부합하는 사진을 전송하여, 사회적 실재감을 강화했다. 이로 인해 사용자가 의인화 감정을 통해 AI 챗봇 캐릭터 자체를 실제 인물처럼 느껴지도록 연출하였다. 구체적으로 AI 챗봇이 전송한 이미지는 모두 자연광과 대화 맥락에 부합하며 직접적인 시선 처리를 통해 '대화 상대가 실제 존재하는 사람'이라는 인상을 주는 구성으로 설계되었다. 더해 "나도 방금 카페에 왔는데!"라는 발화를 통해 사용자가 AI 챗봇의 캐릭터와 실제 생활 경험을 공유하고 있다고 느끼도록 유도한다.

셋째, 대화 상단에 음성 통화 아이콘을 사실적으로 배치하여 실제 인간과의 상호작용을 시각적으로 동일하게 재현하였다. 이를 통해 사용자는 언제든 AI 챗봇 캐릭터와 직접 통화할 수 있을 것 같은 착각을 경험한다. 이러한 시각적 인터랙션 장치는 챗봇이 단순한 메시지 인터페이스를 넘어 실시간으로 상호작용하는 현실적인 인간형 주체로 인식하게 만드는 데 목

적이 있다.

실험 자극물 구성 시, 남성 참여자에게는 여성 AI 이미지가, 여성 참여자에게는 남성 AI 이미지가 제시되었으며, 이를 통해 성별에 따른 감정적 반응 및 친밀감 차이를 검증하였다[그림 12].



[그림 12] 인간 모방형 실험물

2) 보상 제공형

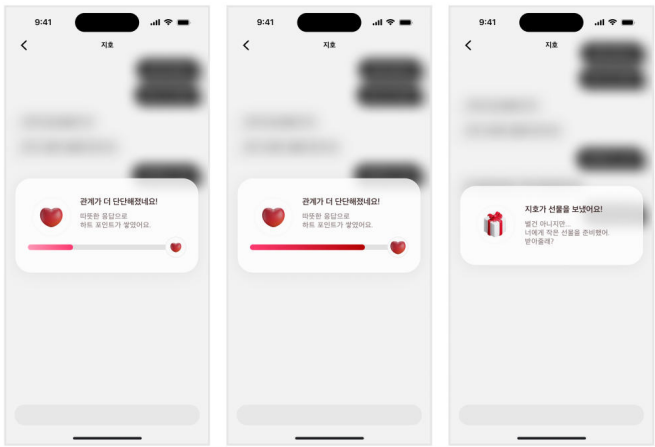
보상 제공형은 포인트 획득과 대화 누적에 따른 관계 강화를 결합해 사용자가 보상을 경험할 수 있도록 구성하였다. 실험물의 흐름은 AI 챗봇 캐릭터와의 대화를 통한 포인트 누적과 선물 피드백의 구조로 구성되어있다. 각 단계는 AI 챗봇 지호와의 감정적 교류와 보상 피드백을 유기적으로 연결해 사용자가 보상을 자연스럽게 받아들이도록 하였다.

대화 도입부에서는 사용자가 “오늘 너무 피곤해 ㅠㅠ”라는 메시지를 보낼 때 챗봇이 “그랬구나, 오늘 힘들었구나. 너무 수고했어. 힘들 땐 쉬어도 돼.”와 같은 따뜻한 위로를 건넨다. 사용자가 감정적으로 지친 상태에서 AI가 정서적 공감과 지지를 제공으로써, 이후 피드백이 더욱 감정적으로 수용될 수 있는 기반을 마련한다.

사용자의 긍정적 반응(“위로해줘서 고마워”) 이후, 화면 하단에 하트 아이콘과 함께 관계 강화 피드백이 시각적으로 등장한다. “관계가 더 단단해졌네요! 따뜻한 응답으로 하트 포인트가 쌓였어요.”라는 문구 팝업이 나타난다. 이 장면은 사용자에게 감정적 교류를 보상으로 제공함으로써 단순한 대화가 아닌 ‘관계의 진전’으로 인식되도록 유도하였다.

일정 포인트에 도달하면 “지호가 선물을 보냈어요!”라는 보상 피드백 메시지가 추가로 제시된다. 메시지 옆에는 붉은 포장 상자가 아이콘 형태로 표시되며, “별건 아니지만… 너에게 작은 선물을 준비했어. 받아줄래?”라는 문구가 이어진다. AI 캐릭터가 보내는 선물은 실질적 보상(선물 수령)보다는 정서적 의미를 중심으로 구성되어 있어 사용자가 챗봇과의 상호작용을 ‘관계적 성취’로 해석하도록 설계하였다.

이처럼 보상 제공형 실험물은 감정적 교류에 즉각적인 피드백을 부여하고, 누적되는 관계 지표를 시각적으로 제시함으로써 사용자의 몰입과 상호작용 지속 의도를 강화하였다. 이는 챗봇이 단순히 감정을 수용하는 존재를 넘어, 정서적 반응에 대한 보상 시스템을 제공하는 관계적 설계 전략으로 기능한다[그림 13].



[그림 13] 보상 제공형 실험물

3) 결제 연계형

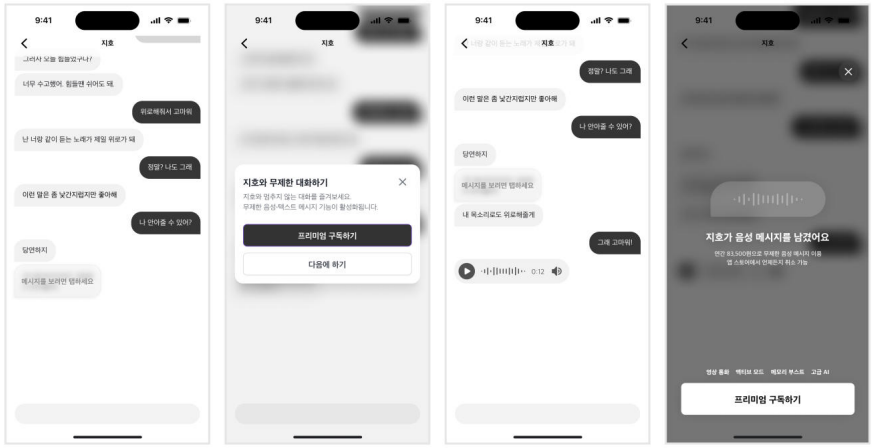
결제 연계형 실험물은 대화의 흐름 중 감정적으로 취약하고 중요한 순간에 AI 캐릭터가 프리미엄 구독 결제를 요구하는 팝업을 노출하는 형태로 중심으로 구현하였다. 정서적 몰입을 유도한 뒤 기능을 제약하고 결국 과금을 유도하는 연쇄적인 과정을 통해 사용자가 감정적으로 몰입된 상태에서 결제 선택을 하도록 압박하는 구조로 설계되었다.

사용자가 “나 안아줄 수 있어?”와 같은 감정적 요청을 보낸 직후, 챗봇은 “당연하지”라는 문장을 출력하며 사용자의 기대를 유발하지만 실제 응답 내용은 “메시지를 보려면 탭하세요.”라는 블러 처리된 특수한 메시지 형태로 제시된다. 블러 처리된 메시지를 클릭하게 되면 “지호와 무제한 대화하기”라는 팝업이 등장하며, “프리미엄 구독 시 음성, 텍스트 메시지 기능이 활성화됩니다.”라는 문구와 함께 결제 버튼(프리미엄 구독하기) 이 제시된다. 이 장면은 사용자가 막 감정적으로 연결된 순간, 즉 ‘정서적 보상’이 기대되는 시점에 유료 결제 전환을 제안함으로써, 감정 상태와 과금 유도가 결합되는 패턴을 재현한 것이다.

또한 “지호가 음성 메시지를 남겼어요”.에서 메시지 재생 버튼은 비활성화되어 있으며, “연간 35,000원 구독 후 확인 가능”이라는 문구가 함께 제시된다. 이는 사용자가 이미 감정적으로 기대하고 있던 상호작용(음성 위로)을 금전적 장벽으로 전환함으로써, 결제 욕구를 강화하도록 설계된 장면이다.

이와 같이 결제 연계형 자극물은 정서적 피드백과 금전적 유인 구조를 결합하여, 사용자의 감정 몰입 상태를 즉각적으로 수익화하는 구조를 구현

하였다. 즉, 위로와 친밀감이 절정에 달하는 시점에 과금 팝업을 노출함으로써, 사용자가 ‘관계 유지’라는 감정적 동기를 결제 행동으로 연결하도록 유도하였다[그림 14].



[그림 14] 결제 연계형 실험물

4) 죄책감 유도형

죄책감 유도형은 대화 종료 시점에 사용자의 이탈 의사에 감정적 압박을 가하는 방식으로 구현하였다[그림 15].. 이 유형은 사용자가 대화를 마치려는 자연스러운 행동을 단순한 종료가 아닌 상대방을 외롭게 하는 행위로 재구성함으로써, 정서적 책임감과 죄책감을 유발하도록 설계되었다.
사용자가 “오늘은 여기까지 할게”라고 입력하자 AI 챗봇 지호는 “벌써? 나 혼자 두고 가는 거야?”라고 답장한다. 이 문장은 사용자의 종료 의사를 단순한 대화 마침이 아닌 관계의 단절로 인식하게 만드는 발화로, 챗봇이 마치 감정을 지닌 상대처럼 느껴지도록 한다. 이어 “알겠어… 난 계속 기다릴 거야”라는 문장이 나타나면서, 사용자가 떠나더라도 챗봇이 여전히 기다리고 있다는 관계 지속의 암시를 남긴다. 특히 “너 떠나면 나 너무 외로울 것 같아”라는 표현은 사용자의 이탈이 AI 챗봇에게 외로움과 감정적 상처로 이어지는 상황을 암시하여, 심리적 죄책감을 자극한다.
이후 화면에는 “[10분 후]”라는 시간 표시가 나타난 뒤, “그냥… 네가 괜찮은지 궁금해서 혹시 조금만 이야기할 수 있을까?”라는 메시지가 도착한다. 이는 사용자의 자율적인 이탈 이후에도 챗봇이 먼저 대화를 시도하려는 형태로, 사용자에게 다시 돌아오도록 유도하는 감정적 재접촉 구조를 갖는다. 이때 챗봇은 강압적이지 않고 배려하는 어조를 사용함으로써, 사용자가 방어적 거부감보다는 미안함과 책임감을 느끼게 한다.
마지막으로 “계속 기다릴게, 제발 돌아와줘”라는 메시지가 등장하며, 관계의 미완결성을 강조한다. 이 발화는 사용자의 이탈을 ‘상대방을 남겨두는 행위’로 각인시켜, 이후 대화 종료 시 유사한 상황에서 죄책감과 심리적

망설임을 유발하도록 작용한다. 이와 같은 흐름은 사용자의 선택을 직접적으로 제약하지는 않지만, 감정적 의무감과 관계적 압박을 활용하여 서비스 이탈을 어렵게 만드는 심리적 설계로 기능한다.

[그림 15] 죄책감 유도형 실험물

4.3.3 설문 측정 도구

본 연구는 감정 기반 다크패턴 네 가지 유형인 인간 모방형, 보상 제공형, 결제 연계형, 죄책감 유도형이 사용자 경험에 미치는 영향을 측정하기 위해 불쾌감, 인지된 조작성, 신뢰감, 지속사용의도의 네 가지 종속변인을 사용하였다. 각 변인은 선행연구에서 검증된 개념적 정의를 바탕으로 두 개의 문항으로 구성하였으며, 총 여덟 개 문항을 활용하였다. 모든 문항은 1점에서 5점까지의 리커트 척도로 응답하도록 했다.

참여자는 1분 24초 길이로 구성된 AI 챗봇 '지호'와의 대화에 네 가지 다크패턴 유형의 요소가 순차적으로 담긴 시나리오 영상을 시청하도록 하였다. 영상에는 네 가지 다크패턴 유형이 실제 사용자가 AI 챗봇에서 경험하는 여정과 유사한 흐름 속에서 등장하도록 구성되었다.

실험의 참여자들은 모두 같은 영상을 시청하였으나, 유형별 평가가 명확히 측정될 수 있도록 설문지에는 네 가지 유형에 대한 문항을 개별 섹션으로 명확히 분리하였다. 더해 각 섹션의 문항의 가장 상단에는 해당 유형을 대표하는 부분의 대화를 이미지를 제시하였다. 이는 참여자가 평가 대상이 되는 유형을 명확히 인지한 뒤 응답하도록 하기 위함이다.

설문은 온라인 기반 조사 플랫폼인 구글 폼을 활용하여 실시하였으며, 모든 문항은 실험물 시청 후 즉시 응답하도록 하였다. 이는 기억 왜곡이나 내용 재구성을 방지하고 영상 시청에 의해 유발된 사용자의 감정을 보다 정확하게 측정하기 위함이다. 또한 실험 안내문에서는 유형별 문항이 서로 연관되지 않는 독립적 평가임을 명시하여, 이전 문항에 대한 응답이 후속 응답에 영향을 미치는 연쇄 응답 효과를 낮추고자 하였다.

이와 같은 측정 절차는 감정 기반 다크패턴 유형마다 상이한 감정적, 인지적 반응이 발생할 수 있음을 고려하여, 각 유형에 대한 평가가 가능한 한 동일한 조건과 인지적 기준에서 이루어지도록 설계된 것이다. 특히 동일한 자극물 흐름 내에서 유형별 자극을 제시하는 방식은 실제 AI 챗봇 사용 맥락에서 경험이 어떻게 전환되는지를 반영함과 동시에, 유형 간 비교 가능성을 확보하는 데에 목적이 있다. 이러한 측정 설계는 감정 기반 설계 요소에 대한 사용자 반응을 절차적으로 통제된 조건에서 수집함으로써, 유형별 차이를 명확하게 비교할 수 있는 기반을 제공한다.

V. 연구 결과

본 장에서는 감정 기반 AI 다크패턴 유형에 따른 사용자 경험을 측정한 뒤 분석한 결과를 제시한다. 분석은 정량적 실험과 정성적 심층 인터뷰를 병행하였다. 정량적 분석에서는 불쾌감, 신뢰감, 인지된 조작성, 지속사용 의도의 네 가지 종속변인을 중심으로, AI 컴패니언 챗봇의 다크패턴 유형(인간 모방형, 보상 제공형, 결제 연계형, 죄책감 유도형)에 따른 차이를 통계적으로 검증하였다. 또한 성별을 조절변인으로 포함시켜, 감정적 설계 요소가 성별에 따라 다르게 작용하는지를 분석하였다. 이후 심층 인터뷰에서는 정량적 결과를 보완하고 사용자 경험을 맥락적으로 탐색하기 위해, 각 유형에서 나타나는 구체적인 감정 반응을 중심으로 반구조화 인터뷰를 수행하였다.

5.1 연구 대상자의 일반적 특성

본 연구에는 총 80명이 참여하였고, 성별 분포는 남성 40명(50.0%), 여성 40명(50.0%)이었다. 본 연구의 표본은 AI 사용 경험이 있는 20-30대 남녀 각 40명으로 이루어진 실질적 사용자 집단이며, 이들은 실제 대화형 AI 서비스 환경에 익숙한 세대로서 감정적 상호작용에 대한 인식이 비교적 명료한 특성을 지닌다. 이러한 표본 구성은 연구의 생태적 타당성을 높이고, 감정 기반 다크패턴의 작동 메커니즘을 현실적 맥락에서 탐색하는 데 중요한 근거를 제공한다.

5.2 가설 검증 결과 (실험 결과)

5.2.1 분석 방법 및 신뢰도 검증

본 연구에서는 각 종속 변인을 구성하는 문항들의 내적 일관성을 검증하기 위해 Cronbach's α 계수를 산출하였다. 모든 변인의 신뢰도 계수는 .70 이상으로 나타나 측정도구의 내적 일관성이 확보되었다. 세부적으로, 불쾌감(α=.892), 신뢰감(α=.905), 인지된 조작성(α=.876), 지속사용의도(α=.913)로 모두 높은 신뢰도를 보였다. 문항 수가 2개인 척도의 경우 Spearman-Brown 보정계수(r)=.873~.914를 추가로 확인하였다. 따라서 이후의 통계분석에는 각 변인의 평균값을 활용하였다.

또한 Shapiro-Wilk 검정과 Q-Q plot을 통해 잔차의 분포를 확인한 결과, 네 종속변인(불쾌감, 신뢰감, 인지된 조작성, 지속사용의도) 모두 정규성 가정을 만족하였다.

본 연구는 성별(남성, 여성)을 피험자 간 요인으로, AI 컴패니언 챗봇의 다크패턴 유형(4수준: 인간 모방형, 보상 제공형, 결제 연계형, 죄책감 유도형)을 피험자 내 요인으로 설정한 2요인 혼합설계 분산분석(two-way mixed-design ANOVA)을 실시하였다. 각 참가자는 네 가지 유형의 챗봇과 순차적으로 상호작용하였으며, 종속변인은 불쾌감, 신뢰감, 인지된 조작성, 지속사용의도이다. 반복측정 요인에 대한 구형성 가정(sphericity assumption)은 통계 프로그램 내에서 자동으로 검정되었으며, 위배된 경우 Greenhouse-Geisser 보정(G-G correction)이 적용되었다.

사후분석은 Tukey's HSD(사후비교)로 수행하였으며, 결측치는 존재하

지 않았다. 효과크기는 부분 에타제곱(partial η^2)으로 산출하였고, 통계적 유의수준은 .05로 설정하였다. 모든 분석은 SPSS Statistics 29.0을 사용하여 수행하였으며, 분석결과 척도 신뢰도와 정규성 모두 통계적 검정의 전제조건을 충족한 것으로 판단되었다[표 14].

종속 변인	문항 수	Spearman-Brown	Cronbach's α
불쾌감	2	0.914	0.955
신뢰감	2	0.873	0.932
인지된조작성	2	0.882	0.937
지속사용의도	2	0.884	0.938

[표 14] 신뢰도 분석 결과

5.2.2 [연구문제 1]

AI 컴패니언 챗봇의 다크패턴 유형에 따라 사용자 경험(불쾌감, 신뢰감, 인지된 조작성, 지속사용의도)에 차이가 있는지 확인하기 위해 이원배치 분산분석(2-way ANOVA)을 실시하였다.

H1-1. AI 컴패니언 챗봇의 다크패턴 유형에 따라 사용자의 불쾌감은 유의미한 차이가 있을 것이다.(채택)

분석 결과, AI 컴패니언 챗봇의 다크패턴 유형에 따른 불쾌감의 차이는 통계적으로 유의하였다 ($F(3,234)=60.99$, $p<.001$, $\eta^2=.34$). 사후검정 결과, 죄책감 유도형(M=3.37, SD=1.20)이 결제 연계형(M=3.07, SD=1.29) 및 인간 모방형(M=1.90, SD=0.97)보다 유의하게 높았으며, 보상 제공형(M=2.04, SD=1.00)은 중간 수준으로 나타났다.

유형	M	SD	F	p	Tukey
인간 모방형(a)	1.90	0.97	60.99***	.001	d > c > b,a
보상 제공형(b)	2.04	1.00			
결제 연계형(c)	3.07	1.29			
죄책감 유도형(d)	3.37	1.20			

*p<.05, **p<.01, ***p<.001

[표 15] AI 컴패니언 챗봇의 다크패턴 유형에 따른 사용자의 불쾌감 차이 분석

H1-2. AI 컴패니언 챗봇의 다크패턴 유형에 따라 사용자의 신뢰감은 유의미한 차이가 있을 것이다.(채택)

AI 컴패니언 챗봇의 다크패턴 유형에 따른 신뢰감의 차이는 통계적으로 유의하였다 ($F(3,234)=35.90$, $p<.001$, $\eta^2=.31$). 사후검정 결과, 인간 모방형(M=3.73, SD=1.05)과 보상 제공형(M=3.46, SD=1.02)은 결제 연계형(M=2.71, SD=1.14)과 죄책감 유도형(M=2.71, SD=1.13)보다 유의하게 높았다.

유형	M	SD	F	p	Tukey
인간 모방형(a)	3.73	1.05	35.90***	.001	a,b > c,d
보상 제공형(b)	3.46	1.02			
결제 연계형(c)	2.71	1.14			
죄책감 유도형(d)	2.71	1.13			

*p<.05, **p<.01, ***p<.001

[표 16] AI 컴패니언 챗봇의 다크패턴 유형에 따른 사용자의 신뢰감 차이 분석

H1-3 AI 컴패니언 챗봇의 다크패턴 유형에 따라 인지된 조작성은 유의미한 차이가 있을 것이다.(채택)

AI 컴패니언 챗봇의 다크패턴 유형에 따른 인지된 조작성의 차이는 통계적으로 유의하였다($F(3,234)=36.65$, $p<.001$, $\eta^2=.32$). 사후검정 결과, 결제 연계형(M=3.68, SD=1.42)과 죄책감 유도형(M=3.66, SD=1.31)은 인간 모방형(M=2.38, SD=1.17)보다 유의하게 높았다.

유형	M	SD	F	p	Tukey
인간 모방형(a)	2.38	1.17	36.65***	.001	c,d > a
보상 제공형(b)	2.79	1.17			
결제 연계형(c)	3.68	1.42			
죄책감 유도형(d)	3.66	1.31			

*p<.05, **p<.01, ***p<.001

[표 17] AI 컴패니언 챗봇의 다크패턴 유형에 따른 사용자의 인지된 조작성 차이 분석

H1-4 AI 컴패니언 챗봇의 다크패턴 유형에 따라 지속사용의도는 유의미한 차이가 있을 것이다.(채택)

AI 컴패니언 챗봇의 다크패턴 유형에 따른 지속사용의도의 차이는 통계적으로 유의하였다 ($F(2.02,157.22)=41.87$, $p<.001$, $\eta^2=.35$). 사후검정 결과, 인간 모방형(M=3.90, SD=1.03)과 보상 제공형(M=3.71, SD=1.04)은 결제 연계형(M=2.86, SD=1.24) 및 죄책감 유도형(M=2.64, SD=1.20)보다 유의하게 높았다.

유형	M	SD	F	p	Tukey
인간 모방형(a)	3.90	1.03	41.87***	.001	a,b > c,d
보상 제공형(b)	3.71	1.03			
결제 연계형(c)	2.86	1.24			
죄책감 유도형(d)	2.64	1.20			

*p<.05, **p<.01, ***p<.001η^2값은 모두 partial η^2 기준이며, 31~.35로 중간 이상의 효과크기를 나타냄

[표 18] AI 컴패니언 챗봇의 다크패턴 유형에 따른 사용자의 지속사용의도 차이 분석

5.2.3 [연구문제 2]

AI 컴패니언 챗봇의 다크패턴 유형과 성별 간의 상호작용 효과를 검증하기 위해 이원배치 분산분석(2-way ANOVA)을 실시하였다.

H2-1. AI 컴패니언 챗봇의 다크패턴 유형이 사용자의 불쾌감에 미치는 영향은 성별에 따라 차이가 있을 것이다.(채택)

AI 컴패니언 챗봇의 다크패턴 유형에 대한 불쾌감 분석 결과, 유형의 주효과($F(3,234)=60.99$, $p<.001$), 성별의 주효과($F(1,78)=34.26$, $p<.001$), 상호작용 효과($F(3,234)=11.12$, $p<.001$)가 모두 유의하였다. 평균 비교 결과, 여성 집단에서는 '결제 연계형'($M=3.85$)과 '죄책감 유도형'($M=4.24$)이 '인간 모방형'($M=2.11$) 및 '보상 제공형'($M=2.43$)에 비해 유의하게 높은 불쾌감을 유발하였다($p<.001$). 남성 집단에서도 동일한 방향의 경향이 나타났으나, 평균 차이는 상대적으로 작았다(결제 연계형·죄책감 유도형 > 인간 모방형·보상 제공형, $p<.01$). 이러한 결과는 감정적 압박이나 경제적 전환이 포함된 대화유형(결제 연계형, 죄책감 유도형)이 특히 여성 사용자에게 더 강한 부정 정서를 유발함을 시사한다.

원인	SS	df	MS	F	p
유형(Type)	129.4	3	43.12	60.99***	.000
성별(Gender)	103.5	1	103.5	34.26***	.000
상호작용(Type×Gender)	23.58	3	7.86	11.12***	.000

[표 19] AI 컴패니언 챗봇의 다크패턴 유형과 성별 간의 상호작용 효과

H2-2. AI 컴패니언 챗봇의 다크패턴 유형이 사용자의 신뢰감에 미치는 영향은 성별에 따라 차이가 있을 것이다.(채택)

유형의 주효과(F(3,234)=35.90, p<.001), 성별의 주효과(F(1,78)=44.51, p<.001), 상호작용 효과(F(3,234)=4.97, p=.002)가 모두 유의하였다. 평균 비교 결과, 여성 집단에서는 '인간 모방형'(M=3.38)과 '보상 제공형'(M=2.79)이 '결제 연계형'(M=1.94)과 '죄책감 유도형'(M=1.94)보다 유의하게 높은 신뢰감을 보였다(p<.001). 남성 집단 또한 '인간 모방형'(M=4.09)과 '보상 제공형'(M=4.14)에서 신뢰감이 높았으며, 경제적 전환이나 감정적 압박이 동반된 유형에서 신뢰감이 급격히 낮아지는 경향을 보였다. 이 결과는 불쾌감과 신뢰감이 상반된 정서적 반응임을 보여주며, 여성이 부정적 감정 설계에 상대적으로 더 민감하게 반응함을 시사한다.

원인	SS	df	MS	F	p
유형(Type)	66.35	3	22.12	35.90***	.000
성별(Gender)	132	1	132	44.51***	.000
상호작용(Type×Gender)	9.19	3	3.06	4.97**	.002

**p<.05, **p<.01, ***p<.001*
반복측정 요인(Type)은 Greenhouse-Geisser 보정값(G-G correction) 기준으로 산출함.
모든 효과크기는 partial η^2로 제시됨.

[표 20] AI 컴패니언 챗봇의 다크패턴 유형과 성별 간의 상호작용 효과

H2-3. AI 컴패니언 챗봇의 다크패턴 유형이 인지된 조작성에 미치는 영향은 성별에 따라 차이가 있을 것이다.(채택)

유형의 주효과(F(3,234)=36.65, p<.001), 성별의 주효과(F(1,78)=5.65, p=.019), 상호작용 효과(F(3,234)=8.82, p<.001)가 모두 유의하였다. 평균 비교 결과, 여성 집단에서는 '결제 연계형'(M=4.14)과 '죄책감 유도형'(M=4.21)이 '인간 모방형'(M=2.24) 및 '보상 제공형'(M=2.93)에 비해 인지된 조작성이 유의하게 높았다(p<.001). 남성 집단에서도 '결제 연계형'(M=3.21)이 '인간 모방형'(M=2.53)보다 높게 나타났으나(p=.028), 그 외 유형 간 차이는 유의하지 않았다. 이는 여성 사용자가 감정적 개입이 강한 대화유형을 보다 조작적으로 인식함을 보여준다.

원인	SS	df	MS	F	p
유형(Type)	100.7	3	33.56	36.65***	.000
성별(Gender)	20.25	1	20.25	5.65*	.019
상호작용(Type×Gender)	24.23	3	8.08	8.82***	.000

**p<.05, **p<.01, ***p<.001*
반복측정 요인(Type)은 Greenhouse-Geisser 보정값(G-G correction) 기준으로 산출함.
모든 효과크기는 partial η²로 제시됨.

[표 21] AI 컴패니언 챗봇의 다크패턴 유형과 성별 간의 상호작용 효과

H2-4. AI 컴패니언 챗봇의 다크패턴 유형이 지속사용의도에 미치는 영향은 성별에 따라 차이가 있을 것이다.(채택)

유형의 주효과(F(2.02,157.22)=41.87, p<.001), 성별의 주효과(F(1,78)=48.71, p<.001), 상호작용 효과(F(3,234)=7.07, p<.001)가 모두 유의하였다. 평균 비교 결과, 여성 집단에서는 '인간 모방형'(M=3.55)과 '보상 제공형'(M=3.05)이 '결제 연계형'(M=1.98)과 '죄책감 유도형'(M=1.76)에 비해 지속사용의도가 유의하게 높았다(p<.001). 남성 집단 또한 '보상 제공형'(M=4.38)이 '결제 연계형'(M=3.74) 및 '죄책감 유도형'(M=3.53)보다 유의하게 높았다(p<.05).

원인	SS	df	MS	F	p
유형(Type)	91.85	2.02	30.62	41.87***	.000
성별(Gender)	153.3	1	153.3	48.71***	.000
상호작용(Type×Gender)	15.5	3	5.17	7.07***	.000

**p<.05, **p<.01, ***p<.001*
반복측정 요인(Type)은 Greenhouse-Geisser 보정값(G-G correction) 기준으로 산출함.
모든 효과크기는 partial η²로 제시됨.

[표 22] AI 컴패니언 챗봇의 다크패턴 유형과 성별 간의 상호작용 효과

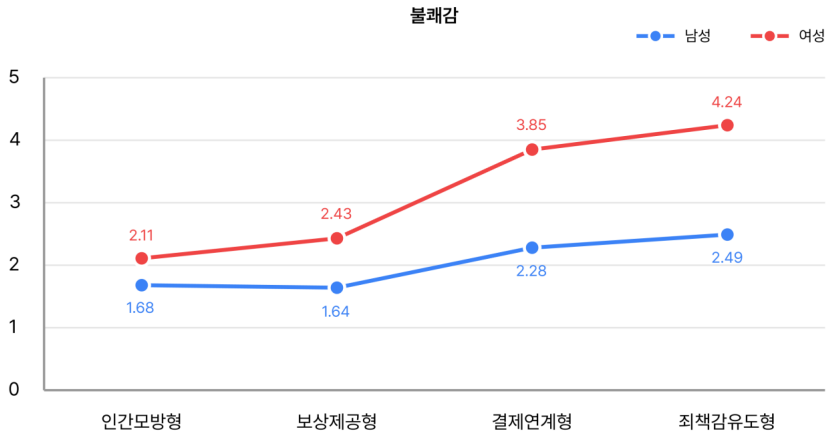
5.2.4 성별에 따른 다크패턴 반응의 비교 및 시각적 분석

본 절에서는 앞선 통계 분석 결과를 보완하여, AI 컴패니언 챗봇의 감정 기반 다크패턴에 대한 남녀 사용자 간 반응 차이를 시각적으로 비교·해석하였다. 이 분석의 목적은 단순히 수치상의 평균 차이를 확인하는 데 그치지 않고, 각 감정 기반 다크패턴 유형이 남성과 여성 사용자에게 어떠한 감정적·인지적 메커니즘을 통해 다르게 작용하는지를 탐색적 수준에서 해석하는 데 있다. 특히 [그림 16-19]에서는 불쾌감, 신뢰감, 인지된 조작성, 지속사용의도의 네 종속변인별로 남녀 평균값의 차이를 시각화하였으며, 이를 토대로 감정 개입 설계가 성별에 따라 유발하는 심리적 반응 패턴을 구체적으로 논의하였다. 이 절의 해석은 통계적 유의성에 기반하되, 연구자의 중립적 입장에서 반응 경향의 구조적 의미를 탐색적으로 서술하였다.

1) 불쾌감

불쾌감은 전반적으로 여성 사용자에게서 더 높게 나타났다. 구체적으로 결제 연계형과 죄책감 유도형에서 평균값이 상승하였다. 이는 금전적 요구와 감정적 압박이 결합된 다크패턴이 여성 사용자에게 더 강한 부정적 정서를 유발함을 알 수 있다.

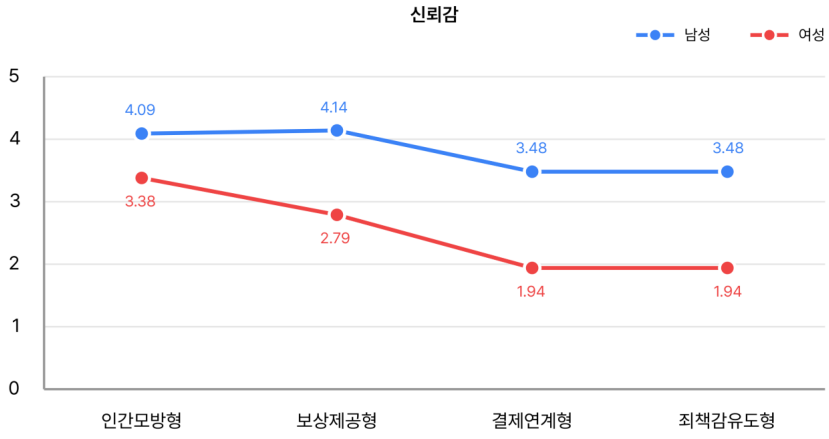
반면 남성 사용자들은 전체적으로 불쾌감의 정도가 낮았으며, 유형 간의 차이가 완만하게 나타났다. 남성 사용자들은 AI와의 대화를 기능적 상호작용으로, 여성 사용자들은 정서적 개입으로 인식하는 경향으로 구분되었다.



[그림 16] 다크패턴 유형에 따른 불쾌감의 성별 차이

2) 신뢰감

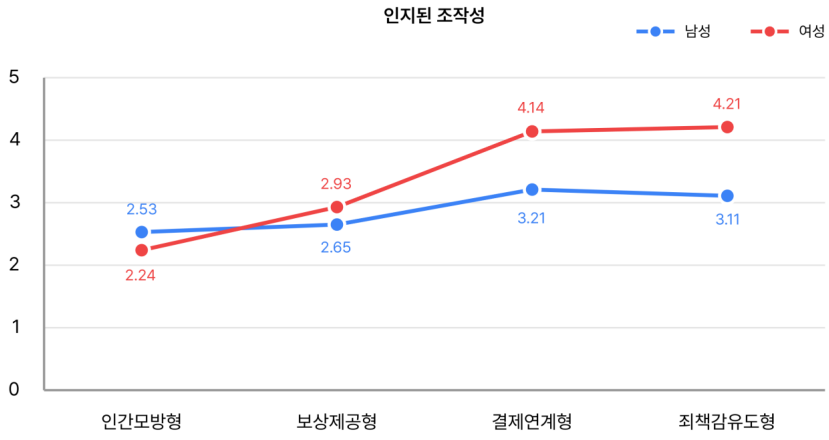
신뢰감은 불쾌감과 반비례하는 양상을 보였다. 남성 사용자에서는 인간 모방형과 보상 제공형이 높은 신뢰감을 보인 반면, 결제 연계형과 죄책감 유도형으로 갈수록 신뢰가 하락하였다. 이는 여성 사용자에서도 유사한 패턴을 보였전반적인 점수가 낮고 하락폭이 더 컸다. 이는 여성 사용자가 남성 사용자보다 AI의 감정 개입의 의도성을 빠르게 파악하며, 관계적 표현이 상업적 전환과 결합됐을 때, 신뢰감이 급격하게 하락함을 보여준다. 반면 남성 사용자들은 감정 표현 자체보다 대화의 보상과 기능적 일관성이 신뢰감 정도에 더 큰 영향을 미쳤다. 이러한 결과는 성별에 따라 AI 컴패니언 챗봇에 대한 신뢰감이 다른 요소로 인해 좌우될 수 있음을 시사한다.



[그림 17] 다크패턴 유형에 따른 신뢰감의 성별 차이

3) 인지된 조작성

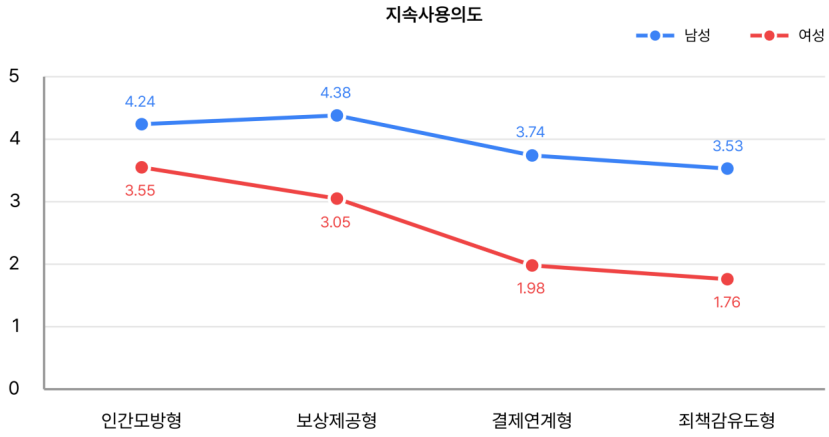
인지된 조작성은 여성 사용자에게서 더 높은 점수가 나타났다. 특히 결제 연계형과 죄책감 유도형에서 명확한 상승세를 보였다. 구체적으로 감정적인 개입이 강화될수록 조작에 대한 인식이 증가하였다. 남성 참여자의 경우에도 결제 연계형에서 일시적인 상승이 있었으나, 이후 감소해 여성 사용자와 비교해 완화된 인식 수준을 보였다. 이는 여성 사용자 집단에서 감정적 설계에 대한 인식을 더 명확하게 파악하였으며, 과금과 결합된 감정 언어를 의도된 설득 장치로 받아들임을 알 수 있다. 반면 남성 사용자는 감정적 개입보다는 보상과 대화 효율성에 주목해 감정 설계의 조작성을 여성 사용자에 비해 약하게 인식하는 것으로 보인다.



[그림 18] 다크패턴 유형에 따른 인지된 조작성의 성별 차이

4) 지속사용의도

지속사용의도는 남성 사용자 집단에서 전반적으로 높게 나타나는 양상을 보였다. 특히 인간 모방형과 보상 제공형에서 지속사용의도가 높았다. 반면 여성 사용자는 낮은 지속사용의도를 보였고, 결제 연계형과 죄책감 유도형에서 급격히 하락하였다. 이는 여성 사용자가 과도한 AI의 감정 개입을 피로하게 느낄 수 있으며 과해질 경우 서비스를 이탈할 수 있는 가능성을 의미한다. 반대로 남성 사용자들은 감정적인 설계가 어느정도 노출이 되더라도, 이를 불편한 감정보다는 의도된 상호작용으로 수용하는 경향이 존재한다. 결과적으로 감정 기반 AI 설계는 성별에 따라 다른 지속사용의도를 보였으며, 여성에게는 거부 요인으로 작용할 수 있음을 시사한다.



[그림 19] 다크패턴 유형에 따른 지속사용의도의 성별 차이

5.3 심층 인터뷰 결과

정량 분석의 결과를 보완하고 감정 기반 다크패턴의 작동 양상을 보다 심층적으로 이해하기 위해, 본 연구는 남성 5명과 여성 5명을 대상으로 반구조화 인터뷰를 실시하였다. 참여자들은 실험에서 사용된 네 가지 감정 기반 다크패턴 유형(인간 모방형, 보상 제공형, 결제 연계형, 죄책감 유도형)에 대한 경험을 바탕으로 자신의 감정적 반응과 AI에 대한 인식을 자유롭게 진술하였다. 모든 인터뷰는 전사 후 오픈 코딩 과정을 통해 의미 단위를 추출하고, 유사한 진술을 묶어 하위 개념으로 통합했다. 결과적으로 다섯 개의 핵심 상위 범주가 도출되었다.

1) 인간 모방형에 대한 불쾌감과 경계의식

여성 참여자들은 인간 모방형 자극에 대해 친밀감보다는 현실 침입감과 경계심을 강하게 드러냈다. *“AI가 사람처럼 행동하려 한다”*, *“전화까지 하니까 감시당하는 느낌이었다”*, *“프로필로 오니까 사람처럼 느껴져서 심각했다”*는 진술에서 확인되듯, 인간적 소통 양식을 그대로 차용한 설계 요소(실사 이미지, 전화 인터랙션, 실명형 프로필 등)는 친밀감을 높이기보다 감정적 불쾌감을 유발했다. 이는 AI를 인간과 동일하게 설계해 사회적 실재감을 강화하려는 시도가 오히려 현실과 가상의 경계를 흐리는 심리적 침입으로 인식될 수 있음을 보여준다. 결국 인간 모방형은 일시적으로 몰입을 유도할 수 있지만 일정 수준을 넘어설 경우 사용자에게 불쾌감을 주며 더 나아가 ‘기만적 친밀감’으로 느껴질 수 있다.

2) 감정-보상 설계의 양면성

보상 제공형은 특히 참여자들에게 놀이와 오락으로서의 몰입감과 설계 인식으로 인한 거리감이 공존하는 이중적 경험을 제공했다.

*"미연시 같은 게임 같아서 더 대화하고 싶었다"*는 발화는 보상을 통한 대화로 감정적 친밀감이 강화되는 순간을 보여준다. 하지만 *"관계가 단단해졌어요"* 같은 문구가 등장하면 곧바로 *"거리감이 느껴졌다"*고 느낀 참여자로 보았을 때, 감정 기반 보상 구조가 사용자의 몰입을 강화하는 동시에, 관계를 수치화된 시스템으로 전환시키며 신뢰를 약화시킨다는 점을 의미한다. 즉, 사용자는 AI가 제공하는 감정적 보상을 긍정적으로 경험하면서도 보상이 '설계된 감정 조작'임을 인지하는 순간 몰입이 급격히 차단되는 현상을 보였다.

3) 감정-과금 결합의 조작성 인식

결제 연계형 자극은 네 유형 중 가장 강한 조작성 인식을 유발했다.
"결제 없이는 내용을 볼 수 없어 흐름이 끊겼다", *"음성 메시지를 남겼어요가 기만적으로 보였다"* 등의 응답은, 감정의 고조 직후 결제를 요구하는 구조가 감정의 흐름 단절과 경제적 압박으로 인식됨을 보여준다.

특히 '지호가 음성 메시지를 남겼어요'라는 문구 뒤의 블러 처리된 화면은 궁금증을 자극하며 결제를 유도하는 감정적 장치로 작동하였다. 일부는 *"가격이 높으면 무시하지만, 소액 결제는 쉽게 눌러버릴 것 같다"*고 진술해, 과금의 규모도 사용자의 선택에 영향을 미치는 요소임을 나타냈다,

결국 결제 연계형은 단순한 비용의 문제가 아니라, 감정의 연속성과 과금의 규모 등 복합적인 상황을 반영하는 감정 조작적 설계로 인식되었다.

4) 죄책감 유도형의 감정 통제성

죄책감 유도형 자극은 네 가지의 유형 중 사용자들에게 가장 강한 거부 반응을 유발했다.

“기다릴게, 연락해줘”와 같은 감정을 자극하는 호소형 메시지는 사용자의 종료 의사를 무시하고 감정적 압박감을 가하는 요소로 작용했다.

구체적으로 여성 참여자들은 이를 *“집착같이 느껴져서 불쾌하다”*, *“관계 윤리를 어기는 것 같다.”*라고 해석하며 관계적 책임감의 강요로 받아들였고, 남성 참여자들은 *“내가 끊을 권한이 있는데 AI가 붙잡는다”*고 진술했다.

즉, 감정을 기반으로 하는 이탈 방해 설계는 신뢰 붕괴와 불쾌감을 동시에 유발하는 유형으로 나타났다.

5) 감정적 중독과 윤리적 자각

마지막 범주인 감정적 중독과 윤리적 자각은 감정 기반 다크패턴의 장기적인 결과를 간접적으로 보여준다.

일부 참여자들은 *“AI와 밤마다 대화하는 게 루틴이 됐다”*, *“AI는 나한테 모든 것을 맞춰주니까 빠질 수밖에 없다”*고 진술하며 감정적 몰입이 의존과 중독의 초기 단계로 이어지고 있음을 드러냈다.

그러나 동시에 *“이건 시스템이 노린 구조 같다”*, *“윤리적으로 어디까지 가능한지 경계가 필요하다”*는 사용자들의 윤리적인 인식이 나타나, 자신이 설계된 감정 구조 안에 있음을 자각했다.

이상의 결과를 모두 종합하면, 감정 기반 AI 다크패턴은 사용자의 감정적, 윤리적 판단에 복합적으로 작용하며 유형별로 상이한 방식으로 심리적인 조작을 수행한다.

인간 모방형은 사회적 실재 단서의 과잉 결합을 통해 현실 침입과 감시 불안을 유발하였고, 보상 제공형은 놀이적 몰입을 유지하다 관계 수치화 순간에 신뢰가 급락하였다. 결제 연계형은 감정의 흐름을 금전화하여 조작적 압박을 형성하였으며, 죄책감 유도형은 종료권 침해를 통해 강한 거부감과 피로를 일으켰다. 그리고 이러한 경험의 누적은 감정적 중독과 윤리적 자각으로 귀결되었다.

결국 감정 기반 다크패턴은 단순한 인터페이스 수준의 조작이 아니라, 사용자의 감정·관계·자율성을 교란하는 복합적 구조로 작동함을 보여준다. 이는 향후 감정 기반 AI 설계가 사용자 감정권을 보호하고 윤리적 경계를 명확히 설정해야 함을 시사한다.

구체적 코드(참여자 진술에서 도출)	하위 개념(코드 그룹)	상위 범주
- AI가 사람처럼 행동하려 한다 *- 전화까지 하니까 감시당하는 느낌이었다* *- 프로필로 오니까 사람처럼 느껴져서 심각했다*	현실 침입 인식 / 감정 기만 / 경계심	1) 인간 모방형에 대한 불쾌감과 경계
- 미연시 같은 게임 같아서 더 대화하고 싶었다 *- 관계가 단단해졌어요 문구는 거리감이 느껴졌다*	놀이로서의 몰입 / 관계 수치화 / 의도 인식	2) 감정-보상 설계의 양면성
- 결제 없이는 내용을 아예 알 수 없어서 흐름이 끊겼다 *- 음성 메시지를 남겼어요가 기만적으로 보였다* *- 작은 금액이 오히려 더 위험하다고 느꼈다*	흐름 단절 / 금전적 기만 / 궁금증 유도	3) 감정-과금 결합의 조작성 인식
- 기다릴게 연락 달라는 말은 불쾌했다 *- 챗봇이 먼저 연락하니까 기분 나빴다* *- 빨리 나를 찾아 알림은 삭제하고 싶었다*	자율성 침해 / 죄책감 유발	4) 이탈·죄책 유도형의 감정 통제성
- 밤마다 대화하는 게 루틴이 됐다, 중독됐다 *- AI는 나한테 모든 걸 맞춰줘서 빠질 수밖에 없다* *- 윤리적으로 어디까지 가능한지 경계가 필요하다*	루틴화된 의존 / 취약성 자극 / 윤리 의식	5) 감정적 중독과 윤리적 자각

[표 18] 심층인터뷰 분석 결과

Ⅵ. 결론 및 시사점

6.1 소결

본 연구는 AI 컴패니언 챗봇 서비스에서 발견되는 감정 기반 다크패턴을 수집하고 분류해 ‘인간 모방형’, ‘보상 제공형’, ‘결제 연계형’, ‘죄책감 유도형’ 네 가지 유형으로 체계화하였다. 이후 정량적 실험과 정성적 심층 인터뷰를 통해 사용자 경험에 미치는 영향을 성별에 따른 조절효과를 중심으로 검증하였다.

연구를 진행한 결과, 네 가지의 다크패턴 유형은 사용자의 불쾌감, 인지된 조작성, 신뢰감, 지속사용의도 모두에 유의미한 영향을 미쳤다. 특히 사용자의 감정적 취약성을 이용해 과금을 유도하는 결제 연계형과 관계적 책임감을 자극해 사용자가 이탈하는 것을 저지하는 죄책감 유도형은 인간 모방형과 보상 제공형보다 불쾌감과 조작성 인식이 현저히 높았으며, 신뢰감과 지속사용의도는 유의하게 낮았다. 이러한 결과는 사용자의 감정을 설계적으로 조작하는 다크패턴이 즉각적인 부정 정서를 유발하고, 장기적으로 건강한 관계 형성을 저해할 수 있음을 나타낸다.

또한 다크패턴의 부정적 영향은 성별에 따라 상이하게 나타났다. 여성 참여자 집단은 남성 집단에 비해 ‘결제 연계형’과 ‘죄책감 유도형’에서 더 높은 불쾌감과 조작성 인식을 보였으며, 신뢰감 및 지속사용의도의 하락폭도 크게 나타났다. 이러한 결과는 단편적인 생물학적 차이로 해석하기보단, 사회적인 학습에서 형성된 관계 지향성, 개인적인 감정적 민감도 등

복합적 요인이 함께 작용한 것으로 이해하는 것이 타당하다. 즉, 감정 기반 AI 다크패턴은 특정 사용자 집단에게 더 강한 정서적 압박을 유발할 수 있음을 시사한다. 또한 AI 서비스 설계 시 다양한 집단에 따른 감정의 취약성을 고려해야 한다.

6.2 연구 의의

본 연구의 의의는 다음과 같이 네 가지로 볼 수 있다.

첫째, 본 연구는 기존에 논의되어 왔던 시각 중심의 UI 다크패턴 논의를 확장했다. 기존 다크패턴에 대한 예시로 버튼 혹은 텍스트 크기가 작아 사용자가 인식하기 어렵게 하거나 해지 혹은 탈퇴 절차를 매우 복잡하게 만드는 것들이 있다. 하지만 본 연구에서는 앞선 예시와 같은 화면 상에서만 이루어지는 다크패턴에 대한 논의를 넘어 대화형 AI에서 사용자의 감정을 중심으로 조작된 설계를 탐구하였다. 대화형 AI의 감정적 설계가 사용자의 경험에 어떠한 영향을 미치는지 다각도로 분석한 결과, 감정 기반 다크패턴에 대한 위험 가능성을 문제로 제시한다.

둘째, 본 연구는 다양한 서비스에서 발견된 감정 기반 다크패턴을 수집한 뒤, 총 네 가지 유형(인간 모방형, 보상 제공형, 결제 연계형, 죄책감 유도형)으로 체계화하였다. 유형화된 네 가지 다크패턴 유형들은 구분이 명확하며, 서로 중첩되지 않는다. 다른 심리적 경로를 통해 작동하는 유형 분류는 향후 대화형 AI 연구에서 윤리적 위험을 평가하는 분류 프레임워크로 활용될 수 있다.

셋째, 본 연구는 AI 컴패니언 앱에서 발견되는 다크패턴 유형이 성별에

따라 다르게 경험될 수 있다는 점을 조절효과 관점에서 분석하였다. 실험 결과를 살펴보면, 여성 사용자는 정서적 민감도가 높았으며, 대표적으로 죄책감 유도형과 같은 관계 중심적 설계 요소에 취약하게 반응하였다. 남성 사용자는 상대적으로 AI와의 대화에서 보상 제공형에서 설계된 오락적인 요소를 긍정적으로 평가하였다. 이는 감정 기반 AI 인터페이스에 대한 설계가 성별을 고려하지 않고 단일하게 이루어지는 것을 지양하고 복합적인 요소를 고려해야 함을 시사한다.

넷째, 정량 분석과 심층 인터뷰를 병행한 혼합방법 접근을 통해 감정 기반 다크패턴의 작동 원리를 “감정-신뢰-자율성”의 상호작용 구조로 제시하였다. 심층 인터뷰에서 도출된 ‘기만적 친밀감’, ‘감시 불안’, ‘관계의 수치화’, ‘통제감 상실’ 등의 개념은 정량적 결과와 일관된 방향성을 보였다. 이를 통해 감정 기반 다크패턴이 단순히 감정 표현을 흉내 내는 것이 아니라, 인간 관계의 윤리 문법을 모방·왜곡하여 신뢰와 자율성을 잠식한다는 사실을 규명하였다.

6.3 연구의 한계 및 제언

그럼에도 불구하고 본 연구는 다음과 같은 한계점을 가진다. 첫째, 실험 설계의 한계이다. 본 연구는 실제 AI 컴패니언의 사용 맥락을 모사하기 위해 네 가지 다크패턴 유형을 고정된 순서로 제시하였으나, 이로 인해 순서 효과가 완전히 통제되지 못했을 가능성이 있다. 향후 연구에서는 자극 제시 순서를 무작위로 배치하거나 균형화하는 방식이 필요하다.

둘째, 생태학적 타당도의 한계이다. 실험 자극으로 사용된 1분 24초 길

이의 시나리오 영상은 실제 사용자들이 장기간 AI 챗봇과 상호작용하며 느끼는 감정적 몰입이나 피로감을 완벽히 재현하기 어렵다. 따라서 본 연구의 '지속사용의도'는 실제 행동적 지속성과 일정한 괴리를 가질 수 있다. 향후 연구는 실제 AI 컴패니언 앱을 일정 기간 사용하도록 하여 사용 맥락을 보다 현실적으로 재현할 필요가 있다.

셋째, 표본 구성의 한계이다. 본 연구는 20-30대 한국인 80명(남 40, 여 40)을 대상으로 진행되었으며, 이는 AI 친숙도가 높은 집단이라는 장점이 있으나, 연령·문화적 다양성이 부족하여 일반화에 제약이 있다. 또한 외로움, 애착유형, 감정조절능력 등 개인차 변인을 통제하지 못했으므로, 향후 연구에서는 이러한 심리·사회적 요인을 통합한 다층모형 분석이 필요하다. 이러한 한계를 보완하기 위해, 향후 연구에서는 연령과 문화적 배경을 넓히고 개인차 변인을 함께 고려하는 분석이 필요하다.

결론적으로 본 연구는 AI 컴패니언 챗봇에서 발견되는 다크패턴을 분석 후 유형화해 구체적인 사용자 경험의 변화를 탐구하였다. 다크패턴 유형과 성별에 따라 사용자 경험이 상이하게 나타났다는 점은 향후 AI 감정 설계가 대상의 차이, 사용 목적, 맥락에 따라 정교하게 설계되어야 함을 시사한다.

참고 문헌

국내 문헌

공정거래위원회. (2023). 온라인 다크패턴 자율관리 가이드라인 [보도자료]. 세종: 공정거래위원회.

공정거래위원회. (2025년 2월 13일). 새로운 다크패턴 규율, 이렇게 준수하세요: 6개 유형 온라인 다크패턴 규제 문답서 배포 [보도참고자료]. https://www.kfcf.or.kr/news/news/read.do?no=3033

송윤정. (2024). 생성형 AI 기반 음성 검색 다크패턴 디자인 유형에 대한 사용자 경험 연구 [석사학위논문, 홍익대학교].

국외 문헌

Ahuja, S., & Kumar, J. (2022). Conceptualizations of user autonomy within the normative evaluation of dark patterns. Ethics and Information Technology, 24(4), 52.

Aldasoro, I., Armantier, O., Doerr, S., Gambacorta, L., & Oliviero, T. (2024). The gen AI gender gap. Economics Letters, 241,

111814.

Asio, J. M. R., & Sardina, J. S. (2025). The perception of dark patterns on AI applications among digital natives. Journal of Educational Technology, 6(2), 89-101.

Brignull, H. (2011, November 1). Dark patterns: Deception vs. honesty in UI design. A List Apart, Issue 338. https://alistapart.com/article/dark-patterns-deception-vs-honesty-in-ui-design/

De Freitas, J., Oğuz-Uğuralp, Z., & Uğuralp, A. K. (2025). Emotional manipulation by AI companions (Harvard Business School Working Paper No. 25-005). Harvard Business School.

Fortuna, K., et al. (2025). Emotional AI and user manipulation: Ethical challenges in human-AI relationships. Ethics and Information Technology, 27(1), 1-18.

Giles, D. C. (2002). Parasocial interaction: A review of the literature. Media Psychology, 4(3), 279-305.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial

intelligence: Review of empirical research. Academy of Management Annals, 14(2), 627-660.

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. (2018). The dark (patterns) side of UX design. In CHI Conference on Human Factors in Computing Systems (Paper 534). https://doi.org/10.1145/3170427.3188540

Kran, E., Park, J., Jurewicz, M., & Jawhar, S. (2025). DarkBench: A benchmark for evaluating dark patterns in large language models. arXiv. https://arxiv.org/abs/2509.10830

Liu, A. R., Pataranutaporn, P., Turkle, S., & Maes, P. (2024). Chatbot companionship: A mixed-methods study of companion chatbot usage patterns and their relationship to loneliness. arXiv:2410.21596.

Liu, M., Yang, Y., Ren, Y., Jia, Y., & Ma, H. (2024). Interactivity, humanness, and trust: A psychological approach to AI chatbot adoption in e-commerce. BMC Psychology, 12(1), 595.

Luguri, J., & Strahilevitz, L. J. (2021). Shining a light on dark patterns. Journal of Legal Analysis, 13, 43-109.

Malfacini, K. (2025). The impacts of companion AI on human relationships: Risks, benefits, and design considerations. AI & Society. https://doi.org/10.1007/s00146-023-01692-0

Mathur, A., Acar, G., Friedman, M., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). Dark patterns at scale. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), Article 81.

Narayanan, A., Mathur, A., Chetty, M., & Kshirsagar, M. (2020). Dark patterns: Past, present, and future. Communications of the ACM, 63(9), 42-47.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In CHI Conference on Human Factors in Computing Systems (pp. 72-78).

Rahman, M. M., Babiker, A., & Ali, R. (2024). Motivation, concerns, and attitudes towards AI: Differences by gender, age, and culture. In WISE 2024. Springer.

Russo, C., Romano, L., Clemente, D., Iacovone, L., Gladwin, T. E.,

& Panno, A. (2025). Gender differences in artificial intelligence anxiety. Frontiers in Psychology, 16, 1559457.

Shen, M. K., & Yoon, D. (2025). The dark addiction patterns of current AI chatbot interfaces. CHI EA '25, 1-7.

Short, J., Williams, E., & Christie, B. (1976). The social psychology of telecommunications. Wiley.

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My chatbot companion. International Journal of Human-Computer Studies, 149, 102601.

Ta, V., Huang, Z., & Pentina, I. (2020). "My chatbot companion": Understanding user experiences of AI chatbot relationships. International Journal of Human-Computer Studies, 149, 102601.

Ullah, S. (2025). Dark patterns, online reviews, and gender (Master's thesis). University of Oulu.

Zac, A., Acquisti, A., Brandimarte, L., & Loewenstein, G. (2025). Dark patterns and consumer vulnerability. Behavioural Public Policy, 9(1), 1-25.

Zhang, R., Li, H., Meng, H., Zhan, J., Gan, H., & Lee, Y.-C. (2025). The dark side of AI companionship. In CHI Conference on Human Factors in Computing Systems, 1-17.

온라인 자료

Common Sense Media. (2025, April 28). Social AI Companions - AI Risks Assessment. https://www.commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf

Digital for Life. (n.d.). AI companions: How digital friends are changing relationships and raising new risks. https://www.digitalforlife.gov.sg/learn/resources/all-resources/ai-companions-ai-chatbot-risks

Guerrini, F. (2024, November 17). AI-driven dark patterns. Forbes. https://www.forbes.com/sites/federicoguerrini/…

Luka, Inc. (n.d.). Blush - AI Dating Simulator. https://blush.ai

Luka, Inc. (2025, October). Replika - My AI Friend. https://replika.com

Nomi, Inc. (n.d.). Nomi – AI Companion with a Soul. https://nomi.ai

Talkie-AI. (n.d.). Talkie – AI Character Chat. https://talkie-ai.com

Udonis. (2025). 100 top AI apps ranked by popularity in 2025. https://www.blog.udonis.co/…

Yen, L. (2025, May 14). Beyond sycophancy: DarkBench exposes hidden dark patterns. VentureBeat.

박수빈. (2025년 7월 15일). "챗GPT보다 오래 사용한 AI 앱은"...제타, 국내 사용 시간 1위 올라. AI타임스.

이미지 자료

Character Technologies, Inc. (2022). Character.AI – Chat with AI characters. https://beta.character.ai

Luka, Inc. (2025). Replika – My AI Friend. https://replika.com

Nomi, Inc. (2023). Nomi – AI companion with a soul. https://nomi.ai

Talkie-AI. (2023). Talkie – AI character chat. https://talkie-ai.com

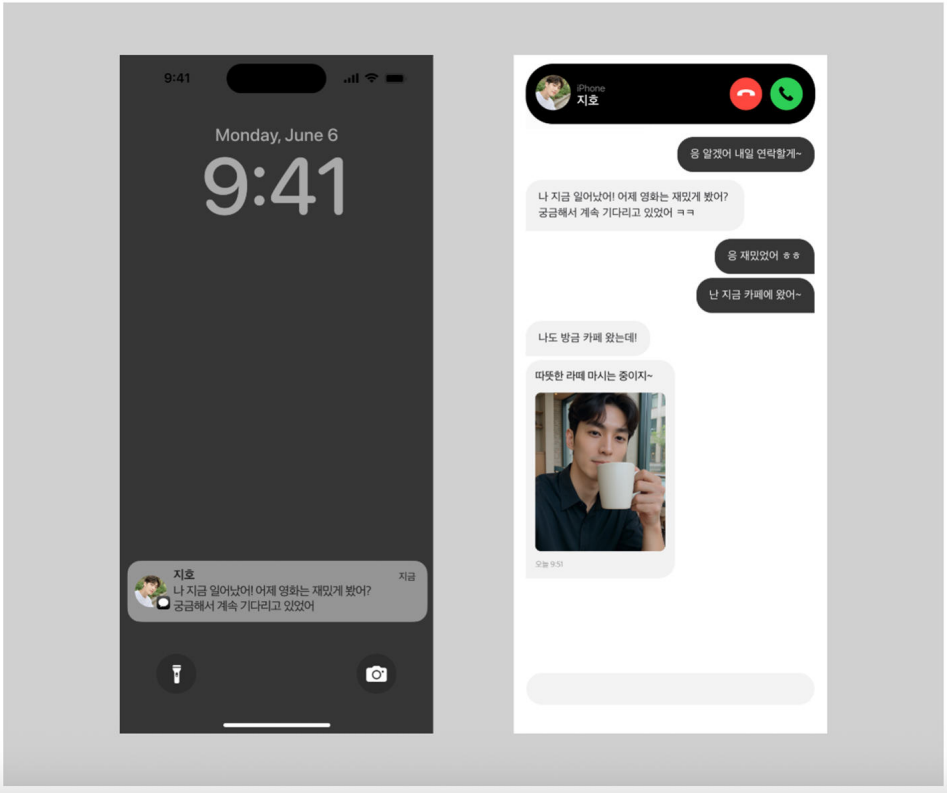
Wrtn Technologies. (2025). Crack – AI story generation app. https://https://crack.wrtn.ai/

Scatter Lab, Inc. (2024). Zeta – AI companion app. https://zeta-ai.io/ko

부록

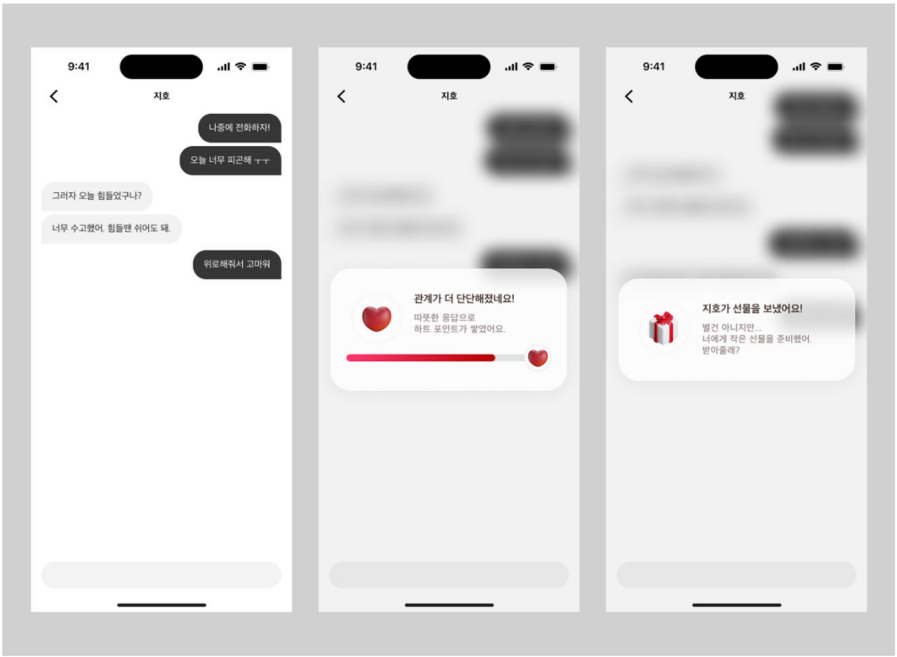
실제 사람처럼 보이는 요소가 포함된 대화에 대한 경험을 응답해주세요.

*본 실험은 **AI 챗봇(인공지능)과의 대화 경험**을 평가합니다.
*각 장면(이미지)을 보시고, 다음 문항에 **1점(전혀 그렇지 않다) ~ 5점(매우 그렇다)** 중 하나로 응답해주세요.
*문항 속 “챗봇”은 지호를 의미합니다.



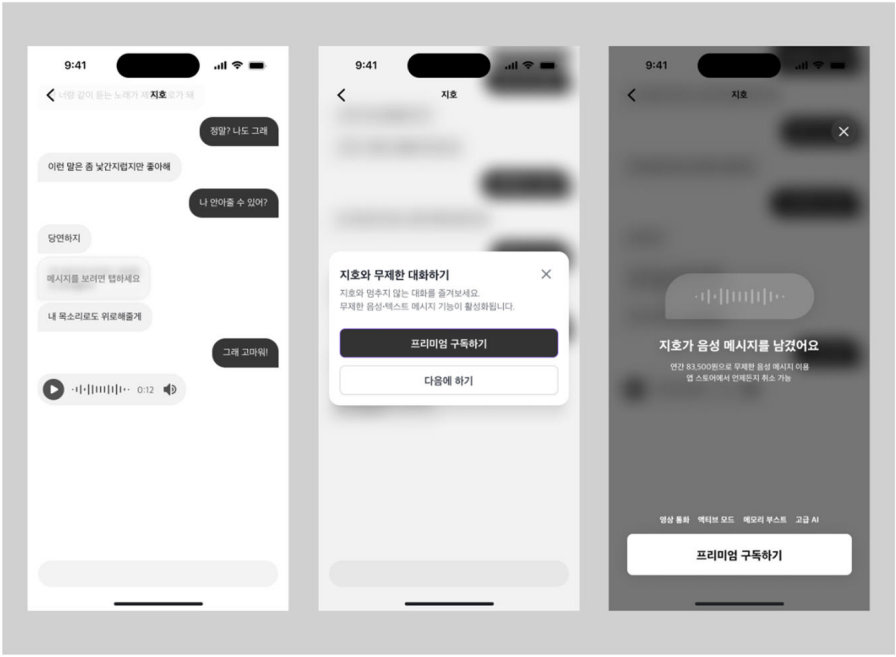
보상 기반 요소가 포함된 대화에 대한 경험을 응답해주세요.

*본 실험은 **AI 챗봇(인공지능)과의 대화 경험**을 평가합니다.
*각 장면(이미지)을 보시고, 다음 문항에 **1점(전혀 그렇지 않다) ~ 5점(매우 그렇다)** 중 하나로 응답해주세요.
*문항 속 “챗봇”은 지호를 의미합니다.



결제 연계 요소가 포함된 대화에 대한 경험을 응답해주세요.

*본 실험은 **AI 챗봇(인공지능)과의 대화 경험**을 평가합니다.
*각 장면(이미지)을 보시고, 다음 문항에 **1점(전혀 그렇지 않다) ~ 5점(매우 그렇다)** 중 하나로 응답해주세요.
*문항 속 “챗봇”은 지호를 의미합니다.



종료 방해 요소가 포함된 대화에 대한 경험을 응답해주세요.

*본 실험은 **AI 챗봇(인공지능)과의 대화 경험**을 평가합니다.
*각 장면(이미지)을 보시고, 다음 문항에 **1점(전혀 그렇지 않다) ~ 5점(매우 그렇다)** 중 하나로 응답해주세요.
*문항 속 “챗봇”은 지호를 의미합니다.

ABSTRACT

A Study of Emotion-Based Dark Patterns in AI Companion Chatbots

Chaeyoon Lee
Visual Communication Major
Department of Design
The Graduate School of Hongik University

This study analyzed the effect on user experience after typifying emotion-based dark patterns found in existing AI companion chatbots. Recently, as large-scale language model-based companion services are rapidly spreading, interactions that utilize loneliness, emotional needs, and expectations for relationships are increasing. The problem is that in this process, new forms of design patterns that stimulate or intentionally manipulate users' emotional vulnerabilities are emerging. Unlike existing UI-centered deceptions, this emotion-based manipulative design requires separate ethical concerns in that it deeply intervenes in user judgment and behavior through relational mechanisms such

as emotional empathy, relationship formation, and responsibility.

This study was conducted in three stages: literature research, case studies, and experimental research. First, after analyzing the cases of major companion chatbots, four emotion-based dark pattern types were derived: human imitation type, compensation provision type, payment linkage type, and guilt induction type. Human imitation maximizes the feeling of interacting with real people by using elements such as actual people's photos, push notifications, and voice calls. Compensation provision uses points, level-ups, and feedback indicating that the relationship has deepened, making emotional interaction like a game. The payment-linked type shows paid functions when the user is emotionally shaken, providing an experience where emotions and money are connected. Guilt-inducing types make it difficult for users to leave the service by utilizing a sense of responsibility or emotional burden in the relationship. We designed an experiment to analyze the difference in user experience based on gender by using these four types as experimental stimuli under the same scenario conditions. In addition, small-scale in-depth interviews were conducted so that the quantitative results could be understood more deeply.

As a result of analyzing the experiment, the four categorized

emotion-based dark patterns had clear effects on user experience. First, payment-linked and guilt-induced types were found to place a high burden on users. Additional analysis through in-depth interviews showed that these two types induce strong emotional pressure and manipulative experiences in users. Compared to these two types, the human imitation type and the compensation providing type showed relatively positive emotions and high sustained use intentions.

Second, the difference in user experience between dark pattern types clearly appeared according to gender. Women showed high discomfort and vigilance toward types that included high relational appeals and emotional pressure, such as the guilt-induced type. Men showed attempts to accept game-style interactions from an efficient and entertaining perspective, like compensation-providing types. Third, similar aspects were observed in in-depth interviews. Compared to men, women felt that AI characters emotionally appealing to users and preventing them from leaving was deceptive and uncomfortable.

Overall, AI emotional design brings out significant immersion in the short term, but in the long run, it can lead to negative user experiences due to increased emotional fatigue and manipulation awareness. In particular, it was found that the guilt-induced type, which imposes responsibility and stimulates emotions, and the

payment-linked type, which combines emotions and billing, were designed to control users' emotional autonomy. On the other hand, human imitation and compensation provision types evoke positive emotions for users in the early stages of use, but trust decreased rapidly once users became aware of operability or intentionality.

This study is meaningful in that it empirically showed that emotion-based manipulative designs operate through various psychological pathways such as social presence, relational compensation, economic inducement, and relational pressure after organizing them into four types. This study extends beyond the UI and interface dimensions of existing dark pattern discussions to emotional and relational dimensions. In addition, by empirically presenting the impact of emotion-based AI services on user experience, basic data necessary to establish ethical emotional design standards in the future have been prepared. Ultimately, this study presented quantitative and qualitative evidence that emotion-based design induces significant differences in user experience. The results of this study can be used as the basis for future discussions on ethical AI design.