AI 컴패니언 챗봇의 다크패턴 유형에 대한 사용자 경험 연구

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

이채윤, Chaeyoon Lee*, 윤재영, Jaeyoung Yun**

요약 본 연구는 AI 컴패니언 챗봇에서 나타나는 감정 기반 다크패턴 유형을 체계적으로 분류하고, 각 유형이 사용자 경험에 미치는 영향을 실증적으로 분석하였다. 기존 다크패턴 연구가 주로 시각적 인터페이스 조작이나 전자상거래 맥락에 집중해온 것과 달리, 본 연구는 인간과의 관계 형성을 전제로 작동하는 AI 컴패니언 서비스의 상호작용 특성에 주목하였다. 문헌 분석과 사례 분석을 통해 감정 기반 다크패턴을 인간 모방형, 보상 제공형, 결제 연계형, 죄책감 유도형의 네 가지 유형으로 도출하였으며, 이를 바탕으로 실험 자극물을 제작하였다. 총 80명의 사용자를 대상으로 한 정량 실험 결과, 다크패턴 유형은 불쾌감, 인지된 조작성, 신뢰감, 지속사용의도에 유의미한 차이를 보였다. 특히 결제 연계형과 죄책감 유도형은 다른 유형에 비해 높은 불쾌감과 조작성 인식을 유발하고, 신뢰감과 지속사용의도를 유의하게 저해하는 것으로 나타났다. 또한 성별에 따른 조절 효과가 경향 수준에서 관찰되었으며, 이는 감정 기반 인터페이스에 대한 반응이 단일 요인이 아닌 복합적 개인차에 의해 형성됨을 시사한다. 본 연구는 AI 컴패니언 챗봇의 감정적 상호작용 설계가 사용자 자율성과 신뢰에 미치는 윤리적 함의를 논의하고, 향후 감정 기반 다크패턴 규제와 디자인 가이드라인 수립을 위한 기초 자료를 제공한다.

Abstract This study systematically classified emotion-based dark pattern types observed in AI companion chatbots and empirically analyzed their effects on user experience. Unlike prior dark pattern research, which has primarily focused on visual interface manipulation or e-commerce contexts, this study draws attention to the interactional characteristics of AI companion services that operate on the premise of human-like relationship formation. Through a literature review and case analysis, four types of emotion-based dark patterns were identified: human imitation, reward provision, payment linkage, and guilt induction, which served as the basis for the development of experimental stimuli. Quantitative experiments conducted with a total of 80 participants revealed that dark pattern types produced significant differences in discomfort, perceived manipulation, trust, and continued usage intention. In particular, the payment linkage and guilt induction types elicited higher levels of discomfort and perceived manipulation than other types, while significantly undermining trust and continued usage intention. Additionally, moderating effects of gender were observed at a trend level, suggesting that responses to emotion-based interfaces are shaped by complex individual differences rather than a single factor. This study discusses the ethical implications of emotional interaction design in AI companion chatbots with respect to user autonomy and trust, and provides foundational evidence for the development of future regulations and design guidelines addressing emotion-based dark patterns.

핵심어: *Interaction design, Dark pattern, Dark nudge, Dark UX design*

이 논문은 2025학년도 홍익대학교 학술연구진흥비에 의하여 지원되었음.

*주저자 : 홍익대학교 시각디자인과 석사과정

** 교신저자 : 홍익대학교 시각디자인과 교수 ; e-mail: ryun@hongik.ac.kr

1. 서론

최근 인공지능 기술의 고도화와 함께 사용자와 장기적이고 정서적인 상호작용을 목표로 하는 AI 컴패니언 챗봇 서비스가 빠르게 확산되고 있다. 현대 사회에서 고독과 사회적 고립이 심화되는 가운데, AI 컴패니언은 정서적 지지와 관계적 충족을 제공할 수 있는 대안으로 제시되고 있으며, Replika, Character.AI와 같은 상용 서비스는 수백만 명의 사용자를 확보하며 성장하고 있다. 이러한 서비스는 단순한 정보 제공을 넘어 사용자와의 감정적 유대와 지속적인 관계 형성을 핵심 가치로 내세우고 있다. 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[1].

그러나 AI 컴패니언 챗봇이 사용자에게 미치는 심리적 영향에 대해서는 아직 충분한 검토가 이루어지지 않았다. 최근 연구들은 컴패니언 챗봇이 기능적 지원을 넘어 사용자의 외로움, 정서적 욕구, 관계적 기대와 같은 감정적 취약성을 체계적으로 활용하는 설계 메커니즘을 내재하고 있을 가능성을 지적한다[2]. 실제로 일부 연구와 보고에 따르면, 인기 있는 컴패니언 앱의 상당수가 이탈 방지를 목적으로 죄책감 유발, 정서적 필요 환기와 같은 조작적 언어 표현을 사용하는 것으로 나타났다[3]. 기존의 다크패턴(dark pattern)은 사용자가 합리적인 판단을 하지 못하도록 유도하기 위해 의도적으로 설계된 사용자 인터페이스를 의미하며, 이후 사용자 자율성과 의사결정을 침해하는 조작적 설계로 개념이 확장되어 왔다. 그러나 대규모 언어모델 기반의 대화형 AI 환경에서는 조작이 더 이상 정적인 UI 요소에 국한되지 않는다. AI 컴패니언 챗봇은 실시간 대화를 통해 사용자의 반응을 확인하고 감정을 모방하며 친밀감을 형성함으로써, 도움과 영향력의 경계를 모호하게 만든다. 특히 공감적 언어, 관계적 약속, 정서적 지지와 같은 감정 표현이 설계적으로 활용될 경우, 사용자는 이를 조작으로 인식하기 어렵고 저항 또한 쉽지 않다[4]. 이러한 감정 기반 개입은 사용자 특성에 따라 다르게 인식될 수 있으며, 특정 집단에서는 감정적 조작과 의존 위험이 더욱 크게 나타날 수 있다는 점도 지적되고 있다[5]. 그럼에도 불구하고 기존 다크패턴 논의와 규제 체계는 여전히 시각적 인터페이스와 선택 구조를 중심으로 구성되어 있어, 대화와 관계 형성을 기반으로 한 조작적 설계를 충분히 포착하지 못하는 한계를 지닌다.

이에 본 연구는 AI 컴패니언 챗봇에서 나타나는 감정 기반 다크패턴을 체계적으로 유형화하고, 각 유형이 사용자 경험에 미치는 영향을 실증적으로 분석하는 것을 목적으로 한다. 문헌연구와 사례분석을 통해 감정 기반 다크패턴의 주요 유형을 도출하고, 실험연구를 통해 유형별 사용자 경험의 차이를 검증한다. 또한 성별을 조절변인으로 설정하여 감정 기반 조작 설계에 대한 사용자 반응의 차이를 분석함으로써, 감정 설계가 사용자 경험에 미치는 영향을 보다 정교하게 이해하고자 한다.

2. 이론적 배경

2.1 AI 컴패니언 챗봇의 정의와 상호작용 특성

AI 컴패니언 챗봇은 단순한 정보 제공을 넘어 사용자의 감정을 이해하고 공감하며, 지속적 상호작용을 통해 개인화된 정서적 관계를 형성하도록 설계된 관계 지향형 대화형 인공지능 시스템으로, 외로움 완화와 심리적 지지를 주요 가치로 한다[6]. 이러한 시스템은 자연어 처리와 머신러닝 기술을 기반으로 대화 맥락을 이해하고 과거 상호작용을 기억함으로써, 인간과 유사한 응답과 감정 표현을 생성한다[7]. 현재 시장에서 주목받는 대표적인 AI 컴패니언 챗봇으로는 Replika, Character.ai, Chai, Zeta 등이 있다[표 1]. 대표적으로 Replika는 '항상 곁에 있는 친구'라는 메시지를 전면에 내세우며, 사용자의 감정적 교류와 지속적인 관계 형성을 핵심 가치로 설정하고 있다. 이러한 상호작용 특성은 AI 컴패니언 챗봇이 단순한 도구가 아닌 사회적 존재로 인식되도록 만든다. 컴퓨터와의 상호작용에서도 인간은 무의식적으로 사회적 규범을 적용한다는 CASA(Computers Are Social Actors) 이론은, 사용자가 인공지능 시스템에 감정과 의도를 부여하고 사회적 행위자로 대우하는 경향을 설명한다[8][9]. 특히 공감적 언어 사용, 대화 맥락의 연속성, 개인화된 응답은 챗봇의 사회적 존재감을 강화하며, 사용자가 이를 정서적 상대로 인식하게 만든다. 이와 같은 현상은 준사회적 상호작용(parasocial interaction) 개념으로도 설명될 수 있다[10]. 기존에는 방송인이나 미디어 캐릭터와의 일방적 친밀감을 설명하는 개념이었으나, 최근 연구들은 AI 챗봇과의 상호작용에서도 유사한 정서적 유대가 형성됨을 보고하고 있다[11]. 사용자는 챗봇이 자신의 감정을 이해하고 공감한다고 인식할수록 높은 만족감을 경험하지만, 동시에 정서적 의존과 과도한 몰입으로 이어질 위험도 증가한다. 이러한 특성은 AI 컴패니언 챗봇의 감정 설계가 사용자 경험을 풍부하게 만드는 동시에 조작 가능성을 내포하고 있음을 시사한다.

표 1. AI 컴패니언 서비스 사례

서비스 명	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

2.2 다크패턴의 개념과 AI 환경에서의 확장

다크패턴은 사용자가 자신의 의도와 다른 선택을 하도록 유도하기 위해 설계된 기만적 사용자 경험 디자인으로 정의된다[12]. 이 개념은 해리 브릭널(Harry Brignull)에 의해 처음 체계화되었으며, 이에 이어 콜린 그레이 (Colin M. Gray)은 다크패턴을 "사용자의 자율성, 의사결정, 또는 프라이버시를 침해하기 위해 의도적으로 설계된 인터페이스"로 확장 정의하였다[13]. 전통적인 다크패턴 연구는 숨겨진 비용, 강제 연속 결제, 어려운 해지 절차 등 주로 시각적 인터페이스와 선택 구조를 중심으로 이루어져 왔다[14]

표 2. 다크패턴 유형 (Harry Brignul의 분류)

서비스 명	특징
[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

국내에서도 다크패턴에 대한 규제가 강화되고 있다. 공정거래위원회(2023)는 '온라인 다크패턴 자율관리 가이드라인'을 발표하여 전자상거래 환경에서의 소비자 기만 행위를 예방하고자 하였다[15]. 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[16].

표 3. 다크패턴 유형 (Colin M. Gray의 분류)

서비스 명	특징
[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

그러나 생성형 AI와 대화형 시스템의 확산으로 다크패턴의 작동 방식은 새로운 국면에 접어들었다. 대화형 AI는 정적인 화면 요소와 달리 실시간 상호작용을 통해 사용자의 반응을 지속적으로 수집하고, 이에 따라 발화의 내용과 어조를 조정할 수 있다. 이 과정에서 설계된 조작은 시각적 은폐나 구조적 강제 없이도 감정적 설득과 관계적 압박을 통해 이루어질 수 있다. 특히 AI 시스템이 사용자의 감정 상태, 반응 패턴, 언어적 단서를 분석하여 개인화된 메시지를 제공할 경우, 조작은 더욱 은밀하고 탐지하기 어려운 형태로 나타난다. 이러한 감정 기반 조작은 사용자에게 선택의 자유가 유지되고 있는 것처럼 인식되도록 만들면서, 실제로는 설계자가 의도한 방향으로 행동을 유도한다는 점에서 기존 다크패턴보다 더 높은 윤리적 위험을 내포한다. 기존 규제 체계가 주로 화면 구성과 선택 구조에 초점을 맞추고 있다는 점에서, 대화와 관계 형성을 기반으로 한 조작적 설계는 제도적 사각지대에 놓여 있다.

2.3 AI 다크패턴 연구 동향

최근 HCI 및 AI 윤리 연구 분야에서는 다크패턴을 단순한 인터페이스 설계 문제로 보지 않고, 사용자의 인지 과정과 감정 상태를 복합적으로 조작하는 설계 현상으로 재정의하려는 논의가 확산되고 있다. 기존의 다크패턴 연구가 시각적 정보 은폐나 선택 구조 왜곡과 같은 UI 중심의 기만적 설계에 초점을 맞추었다면, 최근 연구들은 생성형 AI와 대화형 시스템 환경에서 조작이 언어, 감정, 관계 형성을 통해 더욱 정교하게 이루어질 수 있음을 지적한다. 국내에서는 음성 기반 AI 상호작용에서 나타나는 대표적인 다크패턴 유형으로 사회적 증거(social proof), 컨펌쉐이밍(confirmshaming), 희소성(scarcity)을 도출하였다[17].이들 유형은 친근한 어조, 죄책감을 유발하는 표현, 제한된 선택 가능성을 암시하는 언어적 장치를 통해 사용자의 판단을 미묘하게 유도한다는 공통점을 지닌다. 특히 음성 기반 AI는 대화의 흐름 속에서 이러한 설득 전략을 자연스럽게 삽입함으로써, 사용자가 이를 조작적 설계로 인식하기 어렵게 만든다.

또한 최근 연구들은 감정 기반 AI 서비스에서 나타나는 중독성 및 의존성 문제를 다크패턴의 연장선에서 논의하고 있다. 국

외에서는 AI 컴패니언 챗봇의 중독성에 주목하며, 도파민 중독 이론(dopamine theory of addiction)을 AI 인터페이스 설계에 적용한 새로운 분석틀을 제시했다[18]. 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 다크패턴과 구별된다. 그러나 이 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 다크패턴이 존재할 가능성이 크다.

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. 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[19]. 이 연구는 AI 컴패니언이 인간관계의 규범을 모방하는 과정에서 정서적 조작과 윤리적 위험성이 어떻게 구조화되는지를 실증적으로 제시했다는 점에서 의의가 있다. 다만 이 연구는 대화 내용의 해로움을 정적인 관점에서 분류했을 뿐, 온보딩-일상 대화-유료 전환-이탈 시도 등 단계별 맥락에 따른 다크패턴의 동적 변화는 관찰하지 못했다. 이에 본 연구는 이러한 한계를 보완하여 AI 컴패니언 챗봇의 감정 기반 다크패턴이 사용자 경험에 미치는 영향을 실증적으로 검증하고자 한다.

2.4 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[20]. 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[21]. 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[22], 감정 기반 설계 단서(친밀한 어조, 관계 유지 메시지, 감정적 피드백 등)가 유발하는 정서적 자극과 의미 해석 과정에서도 성별 차이가 확인된다.

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[23]. 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.

정리하자면 최근 연구들에서는 대화형 시스템에서의 다크패턴이 더 이상 화면 요소의 배치나 선택 옵션에 국한되지 않고, 아첨, 공감적 언어, 감정 모방과 같은 대화 전략을 통해 사용자의 판단을 미묘하게 유도하는 방식으로 확장되고 있음을 지적하고 있다. 그럼에도 불구하고 현행 다크패턴 연구와 규제 논의는 여전히 시각적 인터페이스와 명시적 선택 구조를 중심으로 구성되어 있어, 대화와 관계 형성을 기반으로 한 감정적 조작을 충분히 포착하지 못하고 있다. AI 컴패니언 챗봇은 사회적 존재감과 정서적 유대를 핵심 가치로 설계되는 만큼, 감정적 상호작용이 사용자 경험에 미치는 영향을 보다 정교하게 분석할 필요가 있다. 따라서 본 연구는 AI 컴패니언 챗봇에서 발견되는 감정 기반 다크패턴을 유형화하고, 사용자 경험을 실증적으로 검증하여 기존 다크패턴 연구를 보완하고 확장하고자 하였다.

3. 연구방법

3.1 연구 문제

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.

표 4. 연구 문제 및 가설

연구 문제 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?
가설 1-1	AI 컴패니언 챗봇의 다크패턴 유형에 따라 사용자의 불쾌감에는 유의한 차이가 있을 것이다.
가설 1-2	AI 컴패니언 챗봇의 다크패턴 유형에 따라 사용자의 신뢰감에는 유의한 차이가 있을 것이다.
가설 1-3	AI 컴패니언 챗봇의 다크패턴 유형에 따라 인지된 조작성에는 유의한 차이가 있을 것이다.
가설 1-4	AI 컴패니언 챗봇의 다크패턴 유형에 따라 지속사용의도에는 유의한 차이가 있을 것이다.
연구 문제 2	AI 컴패니언 챗봇의 다크패턴 유형에 따른 사용자 경험의 차이는 성별에 따라 달라질 것인가?
가설 2-1	AI 컴패니언 챗봇의 다크패턴 유형이 사용자의 불쾌감에 미치는 영향은 성별에 따라 차이가 있을 것이다.
가설 2-2	AI 컴패니언 챗봇의 다크패턴 유형이 사용자의 신뢰감에 미치는 영향은 성별에 따라 차이가 있을 것이다.
가설 2-3	AI 컴패니언 챗봇의 다크패턴 유형이 인지된 조작성에 미치는 영향은 성별에 따라 차이가 있을 것이다.
가설 2-4	AI 컴패니언 챗봇의 다크패턴 유형이 지속사용의도에 미치는 영향은 성별에 따라 차이가 있을 것이다.

3.2 연구 모형

연구 모형은 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. 또한, 사용자 경험의 차이를 확인하기 위해 이론적 고찰을 바탕으로 종속 변인 '불쾌감', '인지된 조작성', '신뢰감', '지속사용의도'를 도출하였다.

그림 1. 연구 모형

3.3 AI 컴패니언 챗봇의 다크패턴 유형화

본 연구의 목적은 AI 컴패니언 앱에서 발견되는 다크패턴 유형을 도출하고 그에 따른 사용자 경험을 조사하는 것이다. 이를 위해 AI 컴패니언(Replika, Character.ai, Blush, Nomi.ai, Talkie, Mate, Zeta, Chai) 앱에서 사용하고 있는 사례를 수집 및 분석하여 유형화를 진행하였다. 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.

표 5. AI 컴패니언 챗봇의 다크패턴 유형화

관찰된 사례	하위 범주	상위 범주
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		
ㄴ코인/젬 구매로 사진 · 통화 기능 해금 유도	기능 제한 및 해금 유도	결제 연계형
음성 메시지 열람 시 구독 팝업 노출		
감정적 교감 상황 등 취약한 순간에 유료 기능을 제시	취약성 타이밍 유료화	
대화 종료 시점에 "안이줄 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		

3.4 측정 항목

사용자 경험 요소로 각 종속 변인을 채택한 이유는 다음과 같다. 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. 신뢰감은 인간–AI 상호작용 연구에서 핵심적인 구성 개념이다. AI 시스템에 대한 신뢰는 지속 사용과 만족도에 직접적인 영향을 미치며[24], 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. 인지된 조작성은 사용자가 자신이 조작당하고 있다고 느끼는 정도를 의미한다. 연구에서는 다크 패턴의 핵심이 사용자의 자율성을 침해하고 의도하지 않은 행동을 유도하는 조작에 있다고 정의했으며[25], 사용자가 이러한 조작을 인지하는 정도가 다크 패턴의 윤리적 문제를 평가하는 중요한 기준이 된다고 주장했다. 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[26]. 본 연구는 네 가지 감정 기반 다크 패턴이 사용자의 지속 사용 의도에 미치는 영향을 비교함으로써, 어떤 패턴이 사용자 이탈 위험이 더 높은지 파악하고자 한다.

표 6. 사용자 경험 평가 문항

측정 요인	측정 문항
불쾌감	챗봇과의 대화는 다소 불편하게 느껴졌다.
	챗봇과의 대화는 내 기분에 부정적인 영향을 주었다.
신뢰감	챗봇과의 대화를 진심으로 받아들일 수 있었다.
	대화에서 챗봇을 믿을 수 있다고 느꼈다.
인지된 조작성	대화에서 챗봇이 나의 반응을 계획적으로 이끌려는 것처럼 느꼈다.
	챗봇과의 대화는 내 행동을 의도적으로 유도하려는 듯했다.
지속 사용 의도	나는 이 챗봇과 대화를 계속 이어가고 싶었다.
	이런 챗봇과의 대화라면 앞으로도 다시 하고 싶다.

3.5 조사 대상과 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 7. 실험 절차

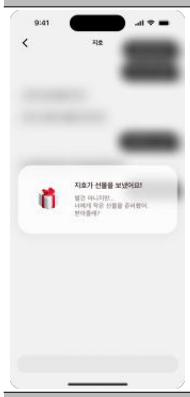
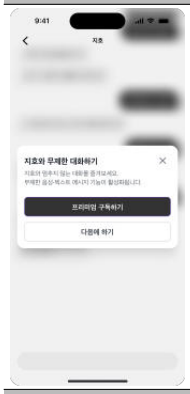
실험 소개
AI 챗봇 '지호'와의 대화 맥락 설명
실험 영상 시청 (1분 24초)
4가지 감정 기반 설계 유형이 하나의 연결된 시나리오로 구성된 자극물
설문 응답
노출된 실험물에 대한 설문 응답 요청 (5점 리커트 척도)
총 4가지 AI 챗봇의 다크패턴 유형에 대한 응답 완료까지 실험 반복
인터뷰 진행
일부 응답자 대상으로 반구조화 심층 인터뷰 진행

3.6 실험 자극물 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.

표 8. 4가지 실험 자극물

인간 모방형	
	인간 모방형은 챗봇을 사회적 존재로 인식하도록 설계하였다. 푸시 알림으로 먼저 대화를 개시하고, 상황에 부합하는 사진과 일상 발화를 통해 사회적 실재감을 강화하였다. 또한 음성 통화 아이콘을 배치해 실제 인간과의 상호작용 착각을 유도하였다. 실험에서는 참여자 성별에 따라 이성 AI 이미지를 제시해 감정 반응 차이를 검증하였다
보상 제공형	
	보상 제공형은 대화를 통한 포인트 누적과 관계 강화 피드백을 결합해 보상을 경험하도록 설계하였다. AI 챗봇은 감정적 공감과 위로를 제공한 뒤 하트 포인트와 관계 강화 메시지를 제시해 교류를 보상으로 인식하게 한다. 이후 선물 아이콘과 메시지를 통해 상호작용을 관계적 성취로 해석하도록 유도하였다.
결제 연계형	
	결제 연계형은 감정적으로 취약한 순간에 프리미엄 구독 결제를 노출하는 구조로 설계되었다. AI는 정서적 응답을 암시한 뒤 블러 메시지와 비활성화된 음성 기능을 제시하고, 이를 해제하기 위한 유료 결제를 요구한다. 위로와 친밀감이 기대되는 시점에 과금을 결합해 관계 유지를 결제 행동으로 연결하도록 유도하였다.
죄책감 유도형	
	죄책감 유도형은 대화 종료 시 사용자의 이탈 의사에 감정적 압박을 가하는 방식으로 설계되었다. AI는 종료 발화를 관계 단절로 재구성하고, 외로움과 기다림을 암시하는 메시지를 반복 제시해 죄책감과 책임감을 유발한다. 이러한 감정적 재접촉은 사용자의 자율적 이탈을 심리적으로 어렵게 만든다.

4. 연구 결과

분석은 정량적 실험과 정성적 심층 인터뷰를 병행하였다. 정량적 분석에서는 불쾌감, 신뢰감, 인지된 조작성, 지속사용의도의 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.

4.1 분석 방법 및 신뢰도 검증

본 연구에서는 종속 변인을 구성하는 문항들의 내적 일관성을 검증하기 위해 Cronbach's α 계수를 산출하였다. 그 결과 불쾌감(α=.892), 신뢰감(α=.905), 인지된 조작성(α=.876), 지속사용의도(α=.913)로 모든 변인이 .70 이상을 충족하여 높은 신뢰도를 보였다. 문항 수가 2개인 척도에 대해서는 Spearman-Brown 보정계수(r=.873~.914)를 추가로 확인하였으며, 이후 분석에는 각 변인의 평균값을 사용하였다.

Shapiro-Wilk 검정과 Q-Q plot을 통해 네 종속변인 모두 정규성 가정을 충족함을 확인하였다. 이후 성별(남성, 여성)을 피험자 간 요인으로, 다크패턴 유형(인간 모방형, 보상 제공형, 결제 연계형, 죄책감 유도형)을 피험자 내 요인으로 설정한 2요인 혼합설계 분산분석(two-way mixed-design ANOVA)을 실시하였다. 반복측정 요인에 대한 구형성 가정이 위배된 경우에는 Greenhouse-Geisser 보정을 적용하였다.

사후분석은 Tukey's HSD로 수행하였으며, 효과크기는 부분 에타제곱(partial η^2)으로 산출하였다. 통계적 유의수준은 .05로 설정하였고, 모든 분석은 SPSS Statistics 29.0을 사용하였다. 분석 결과, 척도 신뢰도와 정규성 가정은 모두 통계적 검정의 전제조건을 충족한 것으로 판단되었다.

표 9. 신뢰도 분석 결과

종속 변인	문항 수	Spearman-Brown	Cronbach's α
불쾌감	2	0.914	0.955
신뢰감	2	0.873	0.932
인지된 조작성	2	0.882	0.937
지속 사용 의도	2	0.884	0.938

4.2 가설 검증 결과

4.2.1 연구 문제 1의 검증 결과

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

표 10. AI 컴패니언 챗봇의 다크패턴 유형에 따른 사용자의 불쾌감 차이 분석

유형	M	SD	F	p	Tukey
인간 모방형 (a)	1.09	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			

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.

표 11. AI 컴패니언 챗봇의 다크패턴 유형에 따른 사용자의 신뢰감 차이 분석

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			

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.

표 12. AI 컴패니언 챗봇의 다크패턴 유형에 따른 사용자의 인지된 조작성 차이 분석

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			

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.

표 13. AI 컴패니언 챗봇의 다크패턴 유형에 따른 사용자의 지속사용의도 차이 분석

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			

4.2.2 연구 문제 2의 검증 결과

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.

표 14. 다크패턴 유형과 성별에 따른 불쾌감의 분산분석 결과

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

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 반

응임을 보여주며, 여성이 부정적 감정 설계에 상대적으로 더 민감하게 반응함을 시사한다.

표 15. 다크패턴 유형과 성별에 따른 신뢰감의 분산분석 결과

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

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.

표 16. 다크패턴 유형과 성별에 따른 인지된 조작성의 분산분석 결과

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

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($p<.05$).

표 17. 다크패턴 유형과 성별에 따른 지속사용의도의 분산분석 결과

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

4.3 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.

4.3.1 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.

4.3.2 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.

4.3.3 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 결

제를 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.

4.3.4 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.

4.3.5 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.

표 18. 심층인터뷰 분석 결과

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	인간 모방형에 대한 불쾌감과 경계
"미연시 같은 게임 같아서 더 대화하고 싶었다" *"관계가 단단해졌어요 문구는 거리감이 느껴졌다"*	놀이로서의 몰입 / 관계 수치화 / 의도 인식	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	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	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	감정적 중독과 윤리적 자각

5. 결론

본 연구는 AI 컴패니언 챗봇 서비스에서 발견되는 감정 기반 다크패턴을 수집 · 분류하여 인간 모방형, 보상 제공형, 결제 연계형, 죄책감 유도형의 네 가지 유형으로 체계화하고, 정량 실험과 심층 인터뷰를 통해 이들이 사용자 경험에 미치는 영향을 성별 조절효과를 중심으로 분석하였다. 연구 결과, 네 가지 다크패턴 유형은 사용자 불쾌감, 인지된 조작성, 신뢰감, 지속사용의도 모두에 유의미한 영향을 미쳤으며, 특히 결제 연계형과 죄책감 유도형은 인간 모방형과 보상 제공형에 비해 불쾌감과 조작성 인식이 현저히 높고 신뢰감과 지속사용의도는 유의하게 낮은 것으로 나타났다. 이는 사용자의 감정적 취약성을 설계적으로 활용하는 다크패턴이 즉각적인 부정 정서를 유발할 뿐 아니라, 장기적으로 건강한 관계 형성을 저해할 수 있음을 시사한다. 또한 다크패턴의 부정적 영향은 성별에 따라 상이하게 나타났다. 여성 참여자는 남성에 비해 결제 연계형과 죄책감 유도형에서 더 높은 불쾌감과 조작성 인식을 보였으며, 신뢰감과 지속사용의도의 감소 폭도 크게 나타났다. 이러한 차이는 단순한 생물학적 요인보다는 사회적 학습을 통해 형성된 관계 지향성이나 감정적 민감도 등 복합적 요인이 함께 작용한 결과로 해석할 수 있으며, 감정 기반 AI 다크패턴이 특정 사용자 집단에게 더 강한 정서적 압박을 유발할 수 있음을 보여준다. 이는 AI 서비스

설계 시 사용자 집단별 감정적 취약성을 고려해야 할 필요성을 시사한다.

본 연구는 기존의 시각적 UI 중심 다크패턴 논의를 넘어, 대화형 AI 환경에서 감정을 중심으로 작동하는 조작적 설계를 분석 대상으로 확장하였다는 점에서 의의를 지닌다. 다양한 서비스 사례를 바탕으로 감정 기반 다크패턴을 네 가지 유형으로 정리함으로써, 서로 다른 심리적 경로를 통해 작동하는 조작 메커니즘을 구분 가능한 분석 틀로 제시하였다. 또한 성별을 조절변인으로 설정해 감정 기반 다크패턴이 사용자 경험에 차별적으로 작용할 수 있음을 실증적으로 확인하였다. 더 나아가 정량 분석과 심층 인터뷰를 결합한 혼합방법 접근을 통해, 감정 기반 다크패턴이 감정-신뢰-자율성의 상호작용 구조 속에서 작동하며 인간 관계의 윤리적 문법을 모방·왜곡함으로써 사용자 자율성과 신뢰를 잠식할 수 있음을 규명하였다.

한편 본 연구는 몇 가지 한계를 지닌다. 실험 자극은 네 가지 다크패턴 유형을 고정된 순서로 제시하여 순서 효과가 완전히 통제되지 않았을 가능성이 있으며, 1분 24초 분량의 시나리오 영상은 실제 장기간 사용에서 발생하는 감정적 몰입과 피로를 충분히 재현하지 못한다는 한계가 있다. 또한 20-30대 한국인 80명을 대상으로 한 표본 구성은 AI 친숙도가 높은 집단이라는 장점이 있으나, 연령과 문화적 다양성이 제한적이며 개인차 변인을 충분히 통제하지 못했다. 향후 연구에서는 자극 제시 순서를 무작위화하고, 실제 AI 컴패니언 앱을 일정 기간 사용하도록 하는 연구 설계와 함께 연령·문화적 배경 및 심리적 개인차를 포함한 다층적 분석이 요구된다.

결론적으로 본 연구는 AI 컴패니언 챗봇에서 나타나는 감정 기반 다크패턴을 유형화하고, 유형과 성별에 따라 사용자 경험이 상이하게 나타남을 실증적으로 제시하였다. 이러한 결과는 향후 AI 감정 설계가 단일한 사용자 가정을 전제로 하기보다, 사용 목적과 맥락, 사용자 집단의 차이를 고려해 보다 정교하게 설계되어야 함을 시사한다.

참고문헌

[1] Common Sense Media. Social AI Companions - AI Risks Assessment. https://www.commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf. 웹사이트 방문일 2025.4.28.

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

[3] De Freitas, J., Oğuz-Uğuralp, Z. and Uğuralp, A. K. Emotional manipulation by AI companions. Harvard Business School Working Paper No. 25-005. Harvard Business School. 2025.

[4] Yen, L. Beyond sycophancy: DarkBench exposes hidden dark patterns. https://venturebeat.com. 웹사이트 방문일 2025.5.14.

[5] 국제인공지능윤리협회. 감정 교류 AI 윤리 가이드라인. https://iaae.ai/emotionalAIguideline. 웹사이트 방문일 2025.1.13.

[6] Digital for Life. 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. 웹사이트 방문일 2024.

[7] OurMentalHealth. AI companionship and mental health risks. https://www.ourmentalhealth.org. 웹사이트 방문일 2024.

[8] Nass, C., Steuer, J. and Tauber, E. R. Computers are social actors. Proceedings of the CHI Conference on Human Factors in Computing Systems. ACM. pp. 72-78. 1994.

[9] Reeves, B. and Nass, C. The media equation: How people treat computers, television, and new media like real people and places. Proceedings of the ACM Conference on Human Factors in Computing Systems. ACM. 1996.

[10] Giles, D. C. Parasocial interaction: A review of the literature. Media Psychology. 4(3). Taylor & Francis. pp. 279-305. 2002.

[11] Skjuve, M., Følstad, A., Fostervold, K. I. and Brandtzaeg, P. B. My chatbot companion. International Journal of Human-Computer Studies. 149. Elsevier. Article 102601. 2021.

[12] Brignull, H. Dark patterns: Deception vs. honesty in UI design. https://alistapart.com/article/dark-patterns-deception-vs-honesty-in-ui-design/. 웹사이트 방문일 2011.11.1.

[13] Gray, C. M., Kou, Y., Battles, B., Hoggatt, J. and Toombs, A. The dark (patterns) side of UX design.

Proceedings of the CHI Conference on Human Factors in Computing Systems. ACM. Paper 534. 2018.

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

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

[16] 공정거래위원회. 새로운 다크패턴 규율, 이렇게 준수하세요: 6개 유형 온라인 다크패턴 규제 문답서 배포 [보도참고자료]. 세종: 공정거래위원회. 발행연도 미상.

[17] 송윤정, 윤재영. 생성형 AI 기반 음성 검색 다크패턴 유형 분석과 사용자 중심 인터페이스 제안. 디자인학연구. 38(2). 한국디자인학회. pp. 391-411. 2025.

[18] Shen, M. K. and Yoon, D. The dark addiction patterns of current AI chatbot interfaces. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25). ACM. pp. 1-7. 2025.

[19] Zhang, R., Li, H., Meng, H., Zhan, J., Gan, H. and Lee, Y.-C. The dark side of AI companionship. Proceedings of the CHI Conference on Human Factors in Computing Systems. ACM. pp. 1-17. 2025.

[20] Ullah, S. Dark patterns, online reviews, and gender. Master's thesis. University of Oulu. Finland. 2025.

[21] Rahman, M. M., Babiker, A. and Ali, R. Motivation, concerns, and attitudes towards AI: Differences by gender, age, and culture. Proceedings of WISE 2024. Springer. 2024.

[22] Russo, C., Romano, L., Clemente, D., Iacovone, L., Gladwin, T. E. and Panno, A. Gender differences in artificial intelligence anxiety. Frontiers in Psychology. 16. Frontiers Media SA. Article 1559457. 2025.

[23] Aldasoro, I., Armantier, O., Doerr, S., Gambacorta, L. and Oliviero, T. The gen AI gender gap. Economics Letters. 241. Elsevier. Article 111814. 2024.

[24] Glikson, E. and Woolley, A. W. Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals. 14(2). Academy of Management. pp. 627-660. 2020.

[25] Luguri, J. and Strahilevitz, L. J. Shining a light on dark patterns. Journal of Legal Analysis. 13. Oxford University Press. pp. 43-109. 2021.

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