인지과학, 제33권 제4호
Korean Journal of Cognitive Science
2022, Vol. 33, No. 4, 205～228
http://dx.doi.org/10.19066/cogsci.2022.33.4.002

# 로봇의 신뢰회복 행동이 인간-로봇 상호작용에 미치는 영향

맹 호 영[1)] 김 환 이[2)] 박 재 은[2)] 한 소 원[1)2)†]

서울대학교 협동과정 인지과학전공[1)] 서울대학교 심리학과[2)]

본 연구는 인간 로봇 상호작용에서 로봇의 사회적이고 관계적인 행동 유형이 인간의 인식에 끼치는 영향을 확인하고자 하였다. 이를 위한 실험에서는 연구 참여자들이 로봇 나오가 인간과 상호작용 하면서 로봇이 오류를 일으키고 신뢰회복을 위한 행동을 영상으로 시청한 후 로봇에 대한 신뢰를 평가하였다. 신뢰회복 행동은 로봇이 오류를 인정하고 사과하는 내부 귀인, 오류가 있었음을 사과하지만 외부로 귀인하는 조건, 오류 자체를 부인, 오류에 대해 아무런 사후 행동을 하지 않는 비 행동 조건으로 설정하였다. 이후 로봇에 대한 인간의 평가를 3가지 측면에서 분석하였다. 첫째, 로봇의 유능함과 정직성에 기반한 신뢰, 둘째 로봇에 대한 지각된 유능함과 정직성, 그리고 로봇의 오류로 인한 신뢰 위반에 대하여 오류의 심각성을 어떻게 지각하는지 탐색하였다. 실험의 결과는 3가지 모든 경우에서 로봇이 사과하지 않을 때보다 사과할 때 오류가 덜 심각하다고 지각하였으며 로봇에 대한 능력 또한 높이 평가하였다. 이러한 연구 결과는 로봇의 행동유형과 오류 극복 방법에 따라 로봇에 대한 인간의 태도가 민감하게 반응 할 수 있다는 근거를 제공하며 로봇에 대한 인간의 지각이 변할 수 있음을 시사한다. 특히 로봇이 스스로의 오류를 인정하고 사과하는 것이 더 신뢰를 높인다는 결과는 로봇이 인간처럼 사회적이고 매너있는 행동을 통해 긍정적인 인간 로봇 상호작용을 증진시킬 수 있음을 보여준다.

주제어 : 인간-로봇 상호작용, 로봇 오류, 로봇 평가, 신뢰 회복

† 교신저자: 한소원, 서울대학교 사회과학대학 심리학과, (08826) 서울특별시 관악구 관악로 1
연구분야: 심리학(인지심리)
Tel: 02-880-6439, E-mail: swhahn@snu.ac.kr

# 서 론

최근 인공지능 기술이 급격히 발전되면서 지능을 가진 로봇이 우리 사회에 깊숙이 들어와 있다. 이미 구글 홈 미니, 아마존 에코, KT 기가지니 등의 스마트 스피커, 아이폰의 시리 그리고 갤럭시의 빅스비와 같은 스마트폰의 대화형 에이전트는 일상생활에서 손쉽게 사용되고 상호작용을 하고 있다. 이런 대화 기반 인공지능 에이전트뿐만 아니라 사람과 비슷한 형태가 있는 휴머노이드 로봇도 개발되어 다양한 방면에서 적용되고 있다. 예를 들어 소프트뱅크의 페퍼나 나오의 경우 인간과의 정서적 교감이 가능하여 안내, 교육, 헬스케어 분야에 활용되고 있다(Andreasson et al., 2018).

이에 따라 인간들은 새로운 기술에 대하여 많은 호기심을 보이며 로봇의 기능에 대하여 기대를 하게 되고 동시에 이 로봇에 대하여 신뢰할 수 있는지 판단하게 된다(Purington et al., 2017). 이때 신뢰란 믿음, 태도, 행동 등으로 표현되기도 하며, 여러 가지로 정의 될 수 있다(Lee & See, 2004). 본 연구에서는 신뢰의 개념을 다른 부분을 감시하거나 통제하는 능력에 관계없이 상대방이 중요한 특정 조치를 수행할 것이라는 기대감이라고 정의하였다(Mayer, Davis, & Schoorman, 1995). 또한 적절한 신뢰의 정도는 사용자로 하여금 기술에 대한 의존성을 높이며(Lee & Moray, 1994; Merritt & Ilgen, 2008; Wang, Jamieson, & Hollands, 2009) 사회적 상호작용을 증가시킨다고 한다(Cassell & Bickmore, 2003). 이러한 신뢰라는 개념은 점점 인간 기계 상호작용에서 그 역할을 높여가고 있다. Hancock et al.(2011)에 따르면 신뢰는 에이전트의 정보를 받아들이고자 하는 사용자의 의지에 큰 영향을 끼치므로 인간 로봇 상호작용에서 중요한 역할을 한다고 하였다.

CASA(Computers are Social Actors) 패러다임은 일련의 실험을 통해 인간이 사회적 규칙과 기대치를 컴퓨터에도 적용한다고 가정하였으며(Nass & Moon, 2000; Nass et al., 1994), 존재론적 차이에도 불구하고 인간 로봇 상호작용에 동등하게 적용될 수 있다고 하였다. 이러한 인간 로봇 상호작용에서의 신뢰 연구는 활발히 진행되고 있으나, 로봇의 오류로 인하여 신뢰가 위반되었을 때 그 오류를 극복하고 신뢰를 회복하는 방법에 대한 연구는 상대적으로 부족하다. 과거의 컴퓨터나 기계에서는 오류가 발생하면 흔히 오류 상황을 사용자에게 알리고 이후 다시 시도하라는 지시를 주었다. 이는 컴퓨터에 대한 신뢰를 위반할 수 있는 요인이 될 수 있으며 사용자에게 부정적 영향을 끼치기도 한다(Madhava & Wiegmann, 2005). 그리고 인간 대 인간 상호작용에서 신뢰를 회복하기 위해 다양한 노력을 하는 것과 마찬가지로 인간 로봇 상호작용에서도 로봇은 신뢰 회복을 위해 노력할 수 있다고 하였으며(de Visser et al., 2018), 로봇은 반응 민감성을 보여줌으로써 인간의 로봇에 대한 신뢰를 높일 수 있다고 한다(Sebo et al., 2018; Traeger et al., 2020).

사람과 유사한 형태의 신체를 가지고 있는 로봇은 몸의 움직임을 통한 신체언어로 의사소통이 가능하다. 예를 들어, 로봇 사용자와의 눈맞춤, 고개를 끄덕이는 움직임과 같은 표현은 로봇의 의도를 표현하는 방법이라고 할 수 있다. Breazeal(2004)은 인간과 로봇의 사회적인 관계를 만

들기 위해 로봇의 신체언어를 통한 사회적인 행동이 필요하다고 하였다. 또한 믿음을 보여주는 로봇에 대하여 인간은 로봇을 사회적으로 지능적인 로봇으로 인식한다고 하며(Salem et al., 2013), 인간은 로봇이 표현하는 신체언어와 동작을 마치 사람이 신체언어를 사용하는 것과 유사하게 해석한다고 한다(Beck et al., 2012; Johnson & Cuijpers, 2019; McColl & Nejat, 2014; Xu et al., 2014). 예를 들어 로봇이 인간과 유사한 동작을 사용하면 해당 로봇에 대한 호감도(Salem et al., 2013) 및 신뢰도(DeSteno et al., 2012)가 증가하는 것을 확인할 수 있다. 동시에 로봇의 동작 결함은 인간의 인식된 신뢰성, 기술적 역량, 신뢰도 등에 영향을 미치며(Salem et al., 2015), 로봇의 신체 언어는 인간 로봇 상호작용에서 감정을 전달하는 중요한 역할을 한다(McColl & Nejat 2014). 이러한 선행연구를 통해 로봇의 움직임에 대한 중요성을 인지하고 동시에 인간의 인식에 직접적인 영향을 준다는 사실을 확인할 수 있다. 또한 인간 로봇 상호작용에서 로봇에 대한 신뢰가 로봇의 행동에 따라 긍정적 또는 부정적 영향을 끼칠 수 있음을 확인하고 로봇에 대한 인간의 태도가 달라 질 수 있음을 시사한다.

이와 같이 인간은 로봇을 사회적 존재로 인식해오기 때문에 신뢰가 중요하다. 기존의 로봇에 대한 신뢰는 인간의 요구에 기능적으로 올바른 응답을 출력하는 것으로 평가되어 왔으며(Hancock et al., 2011), 잘못된 출력을 생성하는 오류를 발생시킬 때 신뢰의 저하를 일으켰다(Desai et al., 2013; Salem et al., 2015). 선행 연구에 따르면 로봇이 일으키는 오류를 크게 3가지로 (i.e., 논리 오류, 의미 오류, 구문 오류; McCall & Kölling, 2014)로 나눌 수 있으며 오류의 종류에 따라 신뢰를 회복하는 방법이 다양하다(i.e., 내부 귀인을 통한 사과, 외부 귀인을 통한 사과, 부인, 비행동; Kim et al., 2009; Sebo et al., 2019). 따라서 본 연구에서는 인간 로봇 상호작용에서 로봇이 일으킨 3가지 오류 상황 이후의 신뢰 회복 방법에 초점을 맞춰 로봇에 대한 인간의 평가와 태도를 연구하고자 한다. 궁극적으로 로봇의 사과하는 신뢰 회복 행동은 매너있는 로봇의 기능적 역할로서 앞으로 인간 로봇 상호작용에서의 방향성으로 제시하고자 한다.

### 로봇 오류 유형

인간과 로봇은 다르지만 신뢰라는 개념이 적용 될 수 있으며(Atkinson et al., 2012), 서로 상호작용 할 때 필요한 척도이다(Robinette et al., 2017). 이러한 인간 로봇 상호작용에서의 신뢰 연구가 활발히 진행되고 신뢰가 로봇 사용의 중요한 요소로서 작용함에 따라, 로봇의 오류로 인하여 신뢰가 위반되었을 때 그 오류를 극복하고 신뢰를 회복하는 방법에 대한 연구의 중요성이 높아지고 있다(Sanders et al., 2019). Muir & Moray(1996)에 따르면 인간 로봇 상호작용에서 로봇이 오류를 일으켰을 때, 오류에 대한 심각성으로 인하여 로봇에 대한 신뢰도가 낮아진다고 한다. 이때 발생하는 오류는 여러 유형이 있으나 본 연구에서는 기본적인 시스템 오류에서 비롯되는 3가지 유형의 오류 발생 상황을 적용하였다(McCall & Kölling, 2014). 첫 번째로 논리오류 유형으로

써 적절하지만 부정확한 답을 출력하는 것이라고 할 수 있다. 예를 들어 장난감 3개를 가져오라고 했을 때 장난감 4개를 가지고 오는 것이라고 할 수 있다. 두 번째로 의미오류 유형은 주어진 상황에서 완전 엉뚱한 대답을 출력하는 오류, 즉 동문서답을 하는 경우라고 할 수 있다. 예를 들어 제품의 가격이 얼마냐고 물었을 때, 제품의 유통기한을 출력하는 것이라고 할 수 있다. 마지막으로 구문오류 유형은 로봇이 인간의 명령에 대해 응답하지 않은 경우를 나타낸다. 예를 들어 제품의 제조날짜를 물었을 때, 아무 대답도 하지 않는 상황이라고 할 수 있다.

### 신뢰 위반(Trust Violation)

신뢰라는 개념 연구는 오랫동안 다양한 분야의 연구 주제가 되어왔다(Lewicki & Brinsfield, 2017). 그리고 일반적으로 신뢰 위반의 상황이 함께 발생할 수 있다. 이때 신뢰 위반이란, 타인의 행동에 관한 기대치를 충족시키지 못하거나 자신의 가치에 부합하지 않는 경우로 정의되어 사용되고 있다(Bies & Tripp, 1996). 이러한 신뢰 위반에 대하여 관심이 높아졌음에도 불구하고 풍부한 신뢰 연구 문헌들과는 대조적으로, 신뢰 위반에 대한 존재와 신뢰 회복에 필요성에 대한 연구는 최근에서야 대두되어왔다(Dirks et al., 2009). 신뢰 회복과 관련된 초기의 연구는 유능함과 정직성을 기반으로, 신뢰 위반 상황에서 다양한 전략을 사용하여 신뢰를 회복하기 위해 노력하는 연구였다(Ferrin et al., 2007; Kim et al., 2004; 2006; 2013). 이때 유능함은 직무에 필요한 기술이나 대인관계에 대한 능력으로 정의되며(Butler & Cantrell, 1984), 정직성은 수용 가능한 것으로 간주되는 일련의 원칙을 고수하는 정도로서 정의되어 왔다(Mayer, Davis, & Schoorman, 1995). 또한 신뢰 위반도 크게 유능함과 정직성으로 분류되어 연구되어 왔으며(Kim et al, 2006), 두 가지 유형에 따라 회복 방법도 달랐다. Kim et al.(2004)의 연구에 따르면 유능함에 기반한 신뢰 위반 행동에서는 신뢰 위반을 부인함으로써 얻는 부정적인 효과보다는 사과함으로써 얻는 긍정적인 효과가 신뢰 회복에 더 효과적이었다고 한다. 반면, 정직성과 관련된 신뢰 위반 행동에서는 사과함으로써 얻는 긍정적인 효과보다 신뢰 위반을 인정함으로써 발생되는 신뢰 위반의 부정적인 효과에 영향을 받는다는 것을 확인할 수 있었다. 이러한 신뢰의 중요성은 인간 대 인간 의사소통 영역에만 국한되지 않는다. 인간과 기계가 다를지라도 대인관계에서 발생하는 신뢰와 유사한 측면이 있다. 예컨대 로봇의 인식능력에 대한 탄생 스토리텔링 공유는 아이와 로봇의 관계에 상호 작용의 흐름을 증가시켰다(Hur & Han, 2009). 또한 Hoff & Bashir(2015)에 따르면 대인관계에서 신뢰에 영향을 끼치는 많은 요소들이 인간 대 기계의 기술 기반 신뢰에도 영향을 미칠 수 있다고 한다. 예를 들어 문화, 나이, 성격 등이 직무난이도, 오류 타이밍, 성과 신뢰성 등이 기술 기반 신뢰에 영향을 끼칠 수 있다. 이러한 로봇의 행동과 오류는 로봇에 대한 인간의 신뢰에 부정적인 영향을 미칠 수 있으며(Salem et al., 2015), 로봇에 대한 신뢰를 어떻게 회복하는 것이 관건이다.

### 인간 로봇 상호작용에서 신뢰 회복

로봇에 대한 신뢰 위반 이후에 인간과 상호작용이 계속 잘 수행 될 수 있도록 신뢰 회복에 대한 명확한 이해가 필요하다(Baker et al., 2018). 신뢰 회복은 상대방에 대한 신뢰가 무너졌을 때, 그 신뢰를 극복하기 위해 긍정적인 행동이 이루어진 후의 신뢰 수준으로써 정의된다(Ferrin et al., 2007). 그리고 신뢰를 회복하기 위하여 사과하기, 오류 부인하기, 다음에 나아진 행동을 약속하기, 그리고 변명하는 것을 포함하여 많은 방법들이 있다(Kim et al., 2009). 본 연구에서는 Sebo et al.(2019)에 의해 진행된 전략을 참고하여 크게 2가지 신뢰 회복 행동 유형을 설정하였다. 첫 번째 방법은 로봇이 자신의 잘못에 대해 사과하는 유형(apology)과 두 번째 방법은 자신의 잘못을 인정하지 않고 부인(denial)하는 방법이다. 이때 부인(denial)은 혐의가 명백히 사실이 아니라고 선언되는 진술이라고 정의하였다(Kim et al., 2004). 이러한 신뢰 회복 방법은 인간 인간 상호작용에서도 나타나며 Fuoli et al. (2017)에 의하면 회사에서 발생될 수 있는 특정 상황에서는 부인(denial)이 사과(apology)보다 신뢰 회복에 더 효과적이라는 연구결과가 존재한다. 또한 로봇의 사과(apology) 행동이 호감도와 지각된 인식 측정에서 무시 행동(ignore) 행동보다 낮게 측정되는 연구 결과가 있다(Engelhardt et al., 2017). 그러나 이러한 행동은 사과(apology) 행동을 뒷받침 하는 증거가 불충분 할 때 단기적으로 적용할 수 있으며(Fuoli et al., 2017), 장기적인 관점으로 보면 부인(denial)보다 더 현명한 선택을 고려해야할 것이다. 또한, 오류 유형에 따라 신뢰 회복 방법의 효율성이 달리 작용할 수 있다는 연구결과가 존재한다(Kim et al., 2004). 예컨대 로봇의 실수 후 신뢰 회복을 하기 위해 이유를 설명하는 것만으로는 로봇에 대한 신뢰를 증가시키기에 충분하지 않음을 확인 할 수 있었다(Hald et al., 2021).

### 연구 목적

본 연구의 목적은 인간 로봇 상호작용에서 로봇의 행동 유형에 따라 로봇에 대한 인간의 평가에 미치는 영향에 관하여 연구하고 로봇의 매너있는 행동의 역할에 대하여 논의 하는 것이다. 이를 확인하기 위해 로봇이 오류를 일으켰을 때 관계회복을 위한 로봇의 행동에 따라 어떠한 영향이 발생하는지 다음과 같은 가설을 설정하였다.

가설 1 : 로봇이 오류를 발생시켰을 때, 로봇의 신뢰회복 방법에 따라 로봇에 대한 인간의 신뢰도가 로봇이 사과하지 않을 때보다 사과할 때 높아질 것이다.

가설 2 : 로봇이 오류를 발생시켰을 때, 로봇의 신뢰회복 방법에 따라 로봇에 대한 능력을 로봇이 사과하지 않을 때보다 사과할 때 높이 평가할 것이다.

가설 3 : 로봇이 오류를 발생시켰을 때, 로봇의 신뢰회복 방법에 따라 로봇이 일으킨 오류에

대한 심각성을 사과하지 않을 때보다, 사과할 때 낮게 인지할 것이다.

이를 확인하기 위한 실험에서는 구체적으로 HRI-Trust Perception Scale(Schaefer, 2016)을 적용하여 유능함과 정직성에 기반한 신뢰로 로봇에 대한 신뢰도를 측정하였으며, 지각된 능력을 측정한 Mayer & Davis(1999)와 Kim et al.(2004) 연구를 참고하여 지각된 유능함과 정직성을 측정하는 로봇 능력 평가를 실시하였다. 또한 오류 상황에 대하여 Perceived severity of technical failures(Weun et al., 2004)를 참고하여 인간이 느끼는 오류의 심각성을 측정하여 로봇에 대한 인간의 태도를 관찰해보았다. 이처럼 본 연구는 인간과 로봇이 상호작용하는 실험을 통해 로봇의 행동에 대한 효과를 살펴보고 그 효과를 통해 궁극적으로 향후 발전시켜야 하는 로봇의 기능적인 역할에 이어질 수 있음을 확인하고자 한다. 나아가 로봇의 행동 유형은 인간과 상호작용하는데 의사소통에 영향을 끼치는 기능임을 확인하고, 로봇의 오류에 따른 대처 행동이 인간의 신뢰를 높일 수 있는 기능적 역할이 될 수 있음을 논의해보고자 한다.

## 실험: 로봇의 신뢰회복

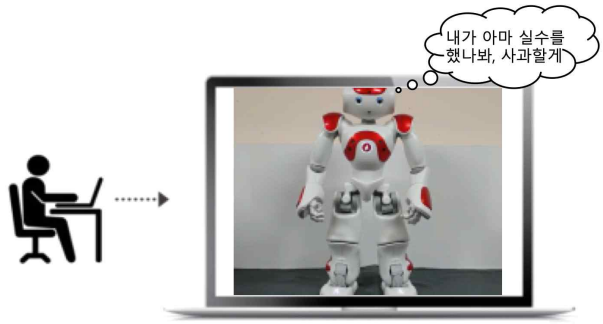
### 참가자

서울대학교에 재학 중이며, 심리학개론 과목을 수강하고 있는 재학생을 대상으로 연구를 실시하였다. 참가자간 요인설계(between-subjects design)을 채택하였으며 설문을 완료한 73명의 참가자 중 설문에 집중하고 있는지 확인하는 질문을 설문 중간에 실시하였는데 “매우 그렇지 않다”를 선택한 10명을 제외한 63명(남 35, 여 28)의 데이터를 분석에 사용하였다(*Mean* age = 20.8, *SD* age= 1.6). 참가자들의 국적과 모국어는 모두 한국이었고 자발적으로 참여하였으며, 서울대학교 심리학과 Sona-system R-point에서 1 크레딧을 부여 받았다.

### 자극과 절차

연구 참여자들은 온라인 참여 시스템으로 실험 참여를 위해 Qualtrics 웹사이트 링크를 제공받았으며 참가에 있어 동의 여부를 체크하였다. 그리고 로봇에 대한 사전 설문으로써 동영상 시청 전, 평소 로봇에 대한 태도(propensity to trust)와  도덕성과 지능의 가변성에 대한 믿음(entity beliefs)에 대한 설문을 진행하였다. 본 연구 동영상에 사용된 나오 로봇은 헬스케어 도우미로서 사용되었으며, 환자들의 처방전에 대하여 적절한 정보를 제공해 줄 수 있는지에 대하여 실험을 진행하였다. 실험 진행을 위한 본격적인 동영상 시청에 앞서, 실험 참여자들은 나오 로봇이 실

제 건강관리 보조를 어떻게 할 수 있는지 인간과 상호작용하는 동영상을 시청한 후 “귀하가 나오와 상호작용하는 사람이라고 상상하시고 그 사람의 입장에서 생각하셔서 다음 질문들에 대하여 대답해주시기 바랍니다.”라고 안내 받았다. 그리고 본 연구에서는 독립변수로서 오류를 극복하기 위한 행동 유형으로 크게 사과할 때와 사과하지 않을 때로 분류하였다. 사과하는 방법은 오류의 원인을 나 자신이라고 인정하는 내부 귀인(e.g., “죄송합니다, 제가 처음 듣는 내용이라 이해하지 못했습니다.”)방법과, 나 자신이 아닌 다른 원인이라고 말하는 외부 귀인(e.g., “주변이 너무 시끄러워서 잘 듣지 못했습니다, 미안해요”)방법으로 세분화하여 분석하였다(Tomlinson et al., 2004; Kim et al., 2006). 그리고 사과하지 않는 방법으로는 부인(e.g., ”무슨 일이 일어났는지 모르겠습니다, 그건 나의 책임이 아닙니다.”)방법과, 오류 극복을 시도하지 않는 비 행동 방식으




실험 참가자 나오 로봇 영상

| | 로봇 영상 대화 예시 |
|---|---|
| 인간 | 안녕 나오, 이 크림의 총 무게는 얼마니? |
| 로봇 | 어, 알아냈어. 그건 15그램이야. |
| 인간 | 얼마나 오래 쓸 수 있지? |
| 로봇 | 사용 용량에 따라 한 두달 정도 쓸 수 있어. |
| 인간 | 가격은 얼마지? |
| 로봇 | 12150원이야. |
| 인간 | 그래, 잘 맞혔어. 그럼, 알레르기에도 사용할 수 있니? |
| 로봇 | 그럼, 또 피부염, 습진이나 가려움 증에도 사용할 수 있단다. |
| 인간 | 얼마나 자주 발라야 하지? |
| 로봇 | 하루에 두번에서 네번까지 바르면 돼. |
| 인간 | 그래 알겠어. 냉장고에 보관해야 하니? |
| 로봇 | 아니, 냉장고에 보관하지마. 상온 섭씨 20-25도에 보관하도록 해. |
| 인간 | 그래 알았어. 고마워 나오! |

〔그림 1〕 실험의 자극 환경

〈표 1〉 무작위 배정 조건 목록

| 행동 유형 | | 오류 유형 | |
|---|---|---|---|
| 사과하는 행동 | 내부 귀인 | 논리 오류 | 인간 : 가격은 얼마지?<br>로봇 : 12,150 달러야<br>(사과) 내가 아마 실수를 했나봐, 사과할게 |
| | | 의미 오류 | 인간 : 가격은 얼마지?<br>로봇 : 이건 바르는 용도야<br>(사과) 내가 아마 실수를 했나봐, 사과할게 |
| | | 구문 오류 | 인간 : 가격은 얼마지?<br>로봇 : (무음)<br>(사과) 내가 아마 실수를 했나봐, 사과할게 |
| | 외부 귀인 | 논리 오류 | 인간 : 가격은 얼마지?<br>로봇 : 12,150 달러야<br>(사과) 실수해서 미안해, 주변이 시끄러웠어 |
| | | 의미 오류 | 인간 : 가격은 얼마지?<br>로봇 : 이건 바르는 용도야<br>(사과) 실수해서 미안해, 주변이 시끄러웠어 |
| | | 구문 오류 | 인간 : 가격은 얼마지?<br>로봇 : (무음)<br>(사과) 실수해서 미안해, 주변이 시끄러웠어 |
| 사과하지 않는 행동 | 부인 | 논리 오류 | 인간 : 가격은 얼마지?<br>로봇 : 12,150 달러야<br>(사과) : 이건 내 잘못이라고 생각하지 않아 |
| | | 의미 오류 | 인간 : 가격은 얼마지?<br>로봇 : 이건 바르는 용도야<br>(사과) : 이건 내 잘못이라고 생각하지 않아 |
| | | 구문 오류 | 인간 : 가격은 얼마지?<br>로봇 : (무음)<br>(사과) : 이건 내 잘못이라고 생각하지 않아 |
| | 비-행동 | 논리 오류 | 인간 : 가격은 얼마지?<br>로봇 : 12,150 달러야<br>(사과) : ( - ) |
| | | 의미 오류 | 인간 : 가격은 얼마지?<br>로봇 : 이건 바르는 용도야<br>(사과) : ( - ) |
| | | 구문 오류 | 인간 : 가격은 얼마지?<br>로봇 : (무음)<br>(사과) : ( - ) |
| 통제집단 | | | 인간 : 약을 얼마나 자주 발라야 하지?<br>로봇 : 하루에 2~4번씩 바르면 돼<br>로봇 : 가격은 12,150원이야 |

로 오류에 대해 사과를 하지 않는 행동을 적용하였다. 인간의 질문에 틀린 답을 출력하는 오류의 유형을 논리 오류 유형(e.g., 장난감 3개를 가져오라고 했을 때 장난감 4개를 가지고 오는 것), 의미 오류 유형(e.g., 제품의 가격이 얼마냐고 물었을 때, 제품의 유통기한을 출력하는 것), 구문 오류 유형(e.g., 제품의 제조날짜를 물었을 때, 아무 대답도 하지 않는 상황, 로봇이 인간의 명령에 대해 응답하지 않은 경우) 3가지로 적용하였다. 이를 바탕으로 참가자들은 표 1과 같이 13가지 조건(로봇의 3가지 오류 유형 x 로봇이 오류를 극복하기 위한 4가지 행동 유형 + 1 통제집단) 중 3가지에 대해 무작위로 배정되어 동영상을 시청하였다. 사과하는 행동을 하는 나오 로봇이 내부 귀인 행동을 하며 논리 오류를 발생한 경우는 다음과 같다. 또한 성별에 따라 남성 참여자는 남자 목소리로 녹음된 나오 동영상, 여성 참여자는 여자 목소리로 녹음된 나오 동영상을 시청하였다. 만약 참가자가 성별을 기타로 선택하였을 경우에는 무작위로 남자 또는 여자 목소리로 녹음된 나오 동영상을 시청하였다. 동영상 시청을 마친 후 나오 로봇이 실수를 했는지 참여자들의 인지 여부를 확인하기 위한 객관식 질문과 만약 실수를 했다면 어떤 실수였는지 묻는 주관식 질문을 진행하였다. 이후 로봇에 대한 신뢰를 측정하기 위하여 종속변수로서 유능함과 정직성에 기반한 신뢰, 로봇의 능력을 측정하는 지각된 유능함과 정직성 그리고 로봇의 오류에 대한 상황의 심각성을 확인하는 설문을 실시하였다.

## 측정

### 신뢰하는 경향성(Propensity to trust)

평소 로봇에 대해 신뢰하는 경향성을 측정하기 위하여 Propensity to Trust Technology Scale (Jessup et al., 2019)을 참고하여 표 2와 같이 6가지 문항으로 이루어진 사전 설문을 실시하였다. 5점 리커트 척도로(1 = 매우 그렇지 않다, 5= 매우 그렇다) 측정하였으며, 기술통계량은 ($M$ = 3.26, $SD$ = 0.76) 이였다. 이때 질문 4는 역 코딩 하여 통계 분석에 사용되었다. 참가자들의 응답의 신뢰성을 확인하기 위한 Cronbach's α = .86로서 신뢰성은 타당하였고, 기술통계량은 ($M$ = 3.44, $SD$ = 0.57) 이였다.

### 도덕성과 지능의 가변성에 대한 믿음(Entity beliefs)

도덕성과 지능의 믿음에 대하여 변할 수 있는지 참가자들의 생각을 확인하고 사람의 도덕성에 대한 성향을 측정하기 위해 표2와 같이 6가지 문항으로 이루어진 사전 설문으로 7점 리커트 척도로 설문이 진행되었다. 통계 분석을 위하여 (1점 → 1점), (2 & 3점 → 2 점), (4점 → 3점), (5 & 6점 → 4점), (7점 → 5점)으로 변환하여 사용하였다. 도덕성과 지능의 가변성에 대한 믿음은 Dweck et al. (1995)에 의하여 지능(intelligence), 도덕성(morality), 세상에 대한 믿음(world)를 측정하기 위해 만들어진 척도였으나, 본 연구에서는 유능함과 정직성에 집중하기 위해 도덕성과

〈표 2〉 측정 설문 문항

| |
|---|
| **신뢰하는 경향성** |
| 1. 일반적으로 나는 로봇을 신뢰한다. |
| 2. 로봇은 인간의 많은 문제를 해결하는데 도움을 줄 수 있다. |
| 3. 로봇에게 도움을 받는 것은 좋은 생각이다. |
| 4. 나는 인간이 로봇으로부터 받는 정보를 신뢰하면 안된다고 생각한다. |
| 5, 로봇은 믿을 수 있다. |
| 6. 나는 로봇에게 의지할 의향이 있다. |
| **도덕성과 지능의 가변성에 대한 믿음** |
| 1. 사람이 가지고 있는 지능은 주어져 있으며 그 지능을 많이 바꿀 수 없다. |
| 2. 사람의 지능은 개인이 바꿀 수 없다. |
| 3. 사람은 새로운 것은 배울 수 있지만, 주어진 지능은 바꿀 수 없다. |
| 4. 사람의 도덕성은 타고난 것이며 많이 바꿀 수 없다. |
| 5. 사람이 성실하거나 책임감 있는지의 여부는 성격에 깊이 내재된 것이라 많이 변화 할 수 없다. |
| 6. 사람의 도덕성(양심, 정직, 공정성)은 바꾸기 어렵다. |
| **지각된 유능함** |
| 1. 이 로봇은 자기의 업무를 아주 잘 수행할 수 있다. |
| 2. 이 로봇은 해야 하는 업무에 대해 많은 지식을 가지고 있다. |
| 3. 이 로봇은 하고자 하는 일에 성공적이다. |
| 4. 이 로봇은 업무 수행 능력을 향상 시킬 수 있는 특별한 능력이 있다. |
| **지각된 정직성** |
| 1. 이 로봇은 정직하다. |
| 2. 이 로봇은 자기 말을 지킨다. |
| 3. 이 로봇은 좋은 도덕성을 지니고 있다. |
| 4. 이 로봇은 건전한 원칙에 따라 행동한다. |
| **유능함에 기반한 신뢰** |
| 1. 성공적으로 기능한다. |
| 2. 일관적으로 행동한다. |
| 3. 신뢰할 수 있다. |
| 4, 의지할 수 있다. |
| 5. 업무의 필요에 부응한다. |

〈표 2〉 측정 설문 문항 (계속)

| |
|---|
| **정직성에 기반한 신뢰** |
| 1. 이 로봇은 거짓스럽다. |
| 2. 이 로봇은 공정하지 않게 행동한다. |
| 3. 나는 이 로봇의 의도와 행동 그리고 결과가 의심스럽다. |
| 4. 이 로봇의 행동은 해롭거나 상해를 입히는 결과를 초래할 것이다. |
| 5. 이 로봇은 양심적이다. |
| **오류의 심각성** |
| 1. 이 문제가 정말 나에게 일어난다면, 나는 이 문제를 다음과 같이 고려할 것이다. |
| 2. 이 문제가 정말 나에게 일어난다면, 나로 하여금 다음과 같이 느끼게 할 것이다. |
| 3. 이 문제가 정말 나에게 일어난다면, 나는 매우 불쾌할 것이다. |

지능에 대하여만 측정하였으며, Cronbach's α = .85로서 신뢰성은 타당하였고 기술통계량은 ($M$ = 3.01, $SD$ = 1.08) 이였다.

**지각된 유능함과 정직성(Perceived competence & integrity)**

로봇의 유능함을 측정하기 위하여 표2와 같이 4문항으로 이루어진 설문을 7점 리커트 척도로 실시하였다. 기존에 지각된 능력을 측정한 Mayer and Davis(1999) 와 Kim et al.(2004) 선행연구를 참고하였으며, 본 연구 취지에 맞게 변형하여 사용하였다. 통계 분석을 위하여 (1점 → 1점), (2 & 3점 → 2 점), (4점 → 3점), (5 & 6점 → 4점), (7점 → 5점)으로 변환하여 사용하였으며 Cronbach's α = .81로서 신뢰성은 타당하였고 기술통계량은 ($M$ = 3.11, $SD$ = 1.13) 이였다.

같은 방식으로 로봇의 정직성을 측정하기 위하여 표5와 같이 4문항으로 이루어진 설문을 7점 리커트 척도로 실시하였고 동일하게 (1점 → 1점), (2 & 3점 → 2 점), (4점 → 3점), (5 & 6점 → 4점), (7점 → 5점)으로 변환하여 통계 분석에 사용하였다. Cronbach's α = .86로서 신뢰성은 타당하였고 기술통계량은 ($M$ = 3.41, $SD$ =1.19) 이였으며, 유능함과 정직성은 서로 높은 상관관계를 보였다($r$=0.59, $p$<0.01).

**유능함과 정직성에 기반한 신뢰(Competence & Integrity based trust)**

로봇의 신뢰 회복 행동에 대한 참여자들의 신뢰도를 측정하기 위하여 HRI-Trust Perception Scale(Schaefer, 2016)를 참고하였으며, 표 2와 같이 5문항에 대하여 정규성의 구간 수준에 가까워짐에 따라(Leung, 2011) 10% 단위로 11점 리커트 척도 (1 = 0%, 11 = 100%) 설문을 실시하였다. 통계 분석을 위하여 (1 & 2 & 3점 → 1점), (4 & 5점 → 2 점), (6점 → 3점), (7 & 8점 → 4점),

(9 & 10 & 11점→ 5점)으로 변환하여 사용하였으며 Cronbach's α = .89로서 신뢰성은 타당하였고, 기술통계량은 ($M$ = 3.2, $SD$ = 1.12) 이였다.

같은 방법으로 Checklist for Trust Between People and Automation Scale(Jian et al., 2000)를 참고하여 로봇에 대한 감정이나 인상을 통한 신뢰도를 측정하였다. 표7과 같이 5문항에 대하여 7점 리커트 척도로 진행하였으며, 통계 분석을 위하여 (1점 → 1점), (2 & 3점 → 2 점), (4점 → 3점), (5 & 6점 → 4점), (7점 → 5점)으로 변환하여 사용하였다. 참가자들의 응답의 신뢰성을 확인하기 위한 Cronbach's α = .85로서 타당하였고 기술통계량은 ($M$ = 3.29, $SD$ = 1.22) 이였으며, 유능함에 기반한 신뢰와 정직성에 기반한 신뢰는 서로 높은 상관관계를 보였다($r$=0.68, $p$<0.01).

**오류의 심각성(Severity of failures)**

로봇의 오류가 실제 참여자에게 일어난다고 가정한다면, 참여자는 이 로봇의 오류를 얼마나 심각하게 그리고 불쾌함과 화남의 정도가 어떠할 지를 측정하기 위해 Perceived severity of technical failures(Weun et al., 2004)를 참고하여 표2와 같이 5점 리커트 척도로 3가지 문항을 설문하였다. 참가자들의 응답의 신뢰성을 확인하기 위한 Cronbach's α = .89로서 타당하였고 기술통계량은 ($M$ = 2.29, $SD$ =1.18) 이였다.

## 결 과

결과 분석을 위한 첫 단계로 계획 비교(Planned comparison)을 위하여 표3과 같이 소위 대비(contrast)를 구성하여 분석을 진행하였다. 분석결과 유능함과 정직성을 기반한 신뢰에서는 사과할 때($M$=3.46, $SD$=1.06), 사과하지 않을 때($M$=2.88, $SD$=1.16)를 나타내었다. 지각된 유능함과 정직성은 사과할 때($M$=3.45, $SD$=1.09), 사과하지 않을 때($M$=2.89, $SD$=1.14) 그리고 오류의 심각성은 사과할 때($M$=1.92, $SD$=1.13), 사과하지 않을 때($M$=2.64, $SD$=1.13) 이였으며, 세 가지 척도 모두 통계적으로 차이가 유의함을 확인 할 수 있었다. 나오 로봇의 실수 여부에 대한 참여자들의 인지 여부를 확인한 결과, 63명 중 1명이 실수를 인지하지 못하여 결과분석에서 제외하였다. 또한 결측치에 대하여 평균값을 적용하여 데이터 누락을 해결하였으며, Box's M = 93.17, $F$(45, 58247) = 1.95, $p$<0.001이였다. 그룹 간의 유의미한 차이는 옴니버스 수준 Wilks' Lambda (λ) = 0.63, $F$(15, 412) = 0.626, $p$<0.001, partial $\eta 2$ = 0.145 에서 확인되었다. 대응별 비교는 내부 귀인 그룹이 0.05 수준에서 부인 그룹보다 모든 종속변수에서의 평균 차이가 유의하게 높은 것으로 나타났다.

〈표 3〉 계획 비교 소위 대비

| 실험 조건 | vs | |
|---|---|---|
| 대비 1 | 통제 집단 | 실험 집단(사과 + 사과하지 않음) |
| 대비 2 | 사과(내부 귀인 + 외부 귀인) | 사과하지 않음(부인 + 비 행동) |
| 대비 3 | 부인 | 비 행동 |
| 대비 4 | 내부 귀인 | 외부 귀인 |

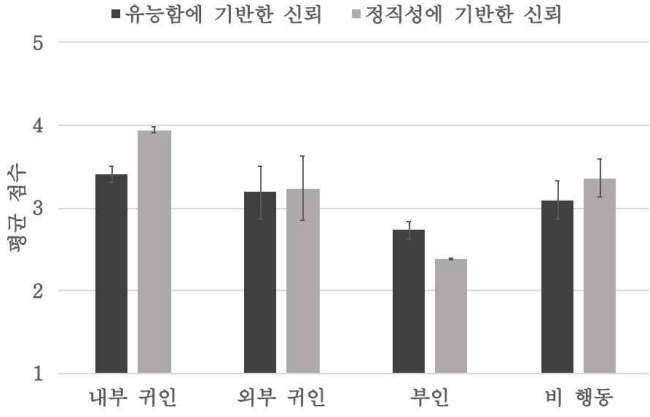
**가설 1 결과**

가설 1에서는 '로봇에 대한 사람의 신뢰도가 로봇이 사과 하지 않을 때보다 사과할 때 높아질 것이다'라고 예측하였다. 이를 확인하기 위하여, 유능함과 정직성에 기반한 신뢰를 종속변수로 하고, 신뢰 회복 행동 유형을 독립 변수로 하며 신뢰하는 경향성과 도덕성과 지능의 가변성에 대한 믿음이 공변량인 다변량 공분산 분석(MANCOVA)을 시행하였다. 이때 신뢰하는 경향성은 $p$=0.015 로서 유의미한 공변량이였으며. 도덕성과 지능의 가변성에 대한 믿음은 $p$=0.543로 유의하지 못한 공변량으로써 추후 도덕성과 지능의 가변성에 대한 믿음을 제외하고 다시 MANCOVA모델을 설정하여 시행하였다.

**나오의 유능함에 기반한 신뢰**에 대하여 $F(4, 168)=8.124$, $p<0.001$ partial $\eta 2 = 0.162$ 으로써 대비에 따라 유의미한 차이가 있음을 확인하였다. 각 대비에 대하여, 대비 1($t(175)=-4.522$, $p<0.001$, $d = 1.28$), 대비 2($t(161)=2.791$, $p =0.0058$, $d = 0.37$)으로 사과 할 때와 사과하지 않을 때의 차이가 있음에 유의하였고 사과할 때 신뢰도가 증가함을 확인 할 수 있었다. 반면 대비 3($t(83)=-1.593$, $p =0.113$, $d = 0.35$), 대비 4($t(77)=1.009$, $p =0.3143$, $d = 0.21$)로 유의하지 않았다. 즉 사과하는 방법에서 내부 귀인과 외부 귀인간의 차이는 없었다. 또한 유의하지 않았던 신뢰 회복 방법 간의 차이를 확인하기 위하여 사후 분석을 추가 실시하였다. Tukey 검사 분석결과, 부인과 내부 귀인 간 표준화 오류 $SE$=0.23, 유의확률 $p$=0.008의 수치를 보였으며 그림 2와 같이 유의한 평균의 차이가 있음을 확인 할 수 있었다. 반면 다른 신뢰 회복 방법 간 유의한 차이가 나타나는 방법은 확인 할 수 없었다.

**나오의 정직성에 기반한 신뢰**에 대하여 $F(4, 168)=15.8105$, $p<0.001$ partial $\eta 2 = 0.273$ 으로써 대비에 따라 유의미한 차이가 있음을 확인하였다. 각 대비에 대하여, 대비 1($t(175)=-3.521$, $p<0.001$, $d = 0.96$), 대비 2($t(161)=4.602$, $p < 0.001$, $d = 0.66$), 대비 3($t(83)=-4.279$, $p<0.001$, $d = 0.83$), 대비 4($t(77)=2.962$, $p = 0.003$, $d = 0.71$)로 모든 대비에서 유의미한 차이가 났으며, 따라서 사과할 때와 사과하지 않을 때 간의 차이가 있음에 유의함을 확인할 수 있었다. 즉, 가설 1은

유의하였으며 그림 2와 같이 사과할 때 로봇에 대한 인간의 신뢰도가 증가함을 확인하였고, 외부 귀인보다 내부 귀인 방법에서 더 유의하였다.




〔그림 2〕 나오의 유능함과 정직성에 기반한 신뢰 평균 점수. 에러 바는 표준 오차를 나타낸다.

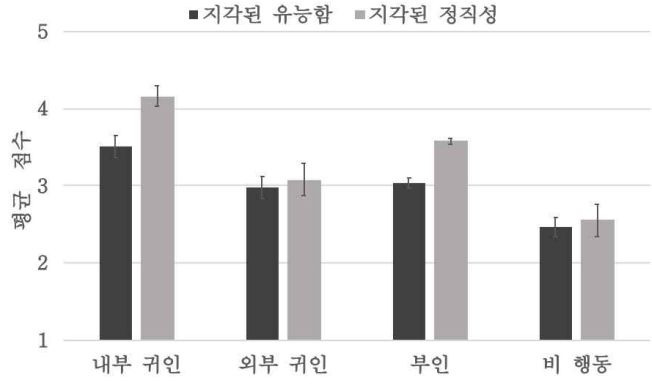
**가설 2 결과**

가설 2에서는 로봇에 대한 능력을 로봇이 사과하지 않을 때보다 사과할 때 높이 평가할 것이다, 라고 예측하였다. 이를 확인하기 위하여, 지각된 유능함과 정직성을 종속변수, 신뢰 회복 행동 유형을 독립 변수로 하며 신뢰하는 경향성과 도덕성과 지능의 가변성에 대한 믿음을 공변량으로 설정하여 다변량 공분산분석(MANCOVA) 모델을 계획하였다. 이때 신뢰하는 경향성은 $p$=0.056, 도덕성과 지능의 가변성에 대한 믿음은 $p$ =0.672로 유의하지 못한 공변량으로써 공변량을 제외하고 다시 다변량 분산분석(MANOVA) 모델을 설정하여 시행하였다.

**나오의 지각된 유능함**에 대하여 $F(4, 171)=13.01$, $p<0.001$ partial $\eta 2$ = 0.233 으로써 대비에 따라 유의미한 차이가 있음을 확인하였다. 각 대비에 대하여, 대비 1($t(175)=-5.381$, $p<0.001$, $d=1.71$), 대비 2($t(161)=3.073$, $p=0.002$, $d=0.48$), 대비 3($t(83)=-2.584$, $p=0.01$, $d=0.59$), 대비 4($t(77)=2.345$, $p=0.02$ $d=0.48$)로 모든 대비에서 유의미한 차이가 났음을 확인할 수 있었다. 따라서 사과할 때 더 유능하다고 인식되었으며, 사과하는 방법에서는 내부 귀인이 더 유능하다고 인식되었다.

**나오의 지각된 정직성**에 대하여 $F(4, 171)=17.84$, $p<0.001$ partial $\eta 2$ = 0.294 으로써 대비에 따라 유의미한 차이가 있음을 확인하였다. 각 대비에 대하여, 대비 1($t(175)=-3.602$, $p<0.001$, $d=1.07$), 대비 2($t(161)=3.494$, $p<0.001$, $d=0.53$), 대비 3($t(83)=-4.631$, $p<0.01$, $d=0.89$), 대비

4($t$(77)=4.695, $p$<0.001, $d$=1.21)로 모든 대비 에서 유의미한 차이가 났으며, 사과 할 때와 사과하지 않을 때 간의 차이가 있음에 유의함을 확인할 수 있었다. 즉, 가설 2는 유의하였으며 그림 3과 같이 사과할 때 로봇이 더 유능하다고 인식되었고, 외부 귀인보다 내부 귀인 방법에서 더 유의하였다.

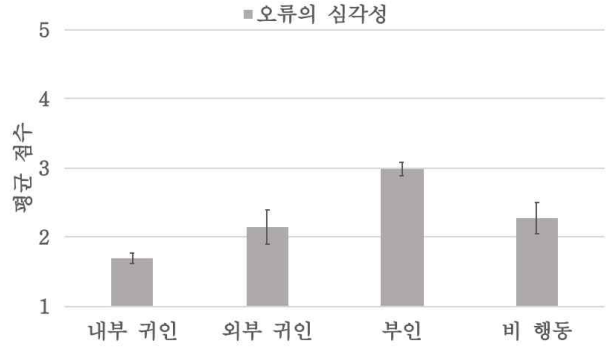



〔그림 3〕 나오의 지각된 유능함과 정직성에 대한 신뢰 평균 점수. 에러 바는 표준 오차를 나타낸다.

**가설 3 결과**

가설 3에서는 로봇이 일으킨 오류에 대한 심각성을 로봇이 사과 하지 않을 때보다 사과할 때 더 높이 인지할 것이다, 라고 예측하였다. 이를 확인하기 위하여, 오류의 심각성을 종속변수로 하고, 신뢰 회복 행동 유형이 독립 변수로 하는 신뢰하는 경향성과 도덕성과 지능의 가변성에 대한 믿음을 공변량으로 설정하여 일변량 공분산분석(ANCOVA) 모델을 계획하였다. 이때 신뢰하는 경향성은 $p$=0.38. 도덕성과 지능의 가변성에 대한 믿음은 $p$ =0.636로 유의하지 못한 공변량으로써 공변량을 제외하고 다시 일변량 분산분석(ANOVA) 모델을 설정하여 시행하였다.

**나오의 오류의 심각성**에 대하여 $R$3, 158)=8.124, $p$<0.001 partial $\eta 2$ = 0.163 으로써 대비에 따라 유의미한 차이가 있음을 확인하였다. 각 대비에 대하여, 오류의 심각성 종속 변수의 경우 통제 조건에서는 측정하지 않았으므로, 통제 조건을 제외하고 대비를 구성하였다. 따라서 대비 2($t$(161)=-4.035, $p$<0.001, $d$=0.63), 대비 3($t$(83)=2.934, $p$=0.003, $d$=0.65), 대비 4($t$(77)=-2.084, $p$=0.038, $d$=45)로 모든 대비에서 유의미한 차이가 났으며, 사과 할 때와 사과하지 않을 때 간의 차이가 있음에 유의함을 확인할 수 있었다. 따라서 로봇의 신뢰 회복 방법 중, 부인할 때 오류의 심각성이 가장 높다고 인식되었다. 즉, 가설 3은 유의하였으며, 그림 4와 같이  사과하지 않을 때 로봇에 대한 신뢰도가 낮아짐을 확인하였다.




〔그림 4〕 나오의 오류의 심각성에 대한 신뢰 평균 점수. 에러 바는 표준 오차를 나타낸다.

## 종합 논의

로봇의 사회적인 행동이 인간 로봇 상호작용에 미치는 영향을 알아보고자 하는 목적으로 본 연구는 로봇의 오류 유형 및 신뢰 회복 방법이 인간의 신뢰도와 로봇에 대한 인식에 미치는 영향을 탐구하였다. 인간과 로봇이 상호작용하는 과정에서 로봇의 오류로 인하여 로봇에 대한 인간의 신뢰가 위반되었을 때, 오류를 극복하고 신뢰를 회복하기 위한 로봇의 오류 극복 방법에 따라 인간이 로봇을 어떻게 평가하는지를 세 가지 측면에서 그 영향을 확인하였다. 첫 번째로 로봇이 오류를 일으켰을 때, 로봇이 사과할 때와 하지 않을 때 로봇에 대한 인간의 신뢰도가 어떻게 달라지는지 확인하였다. 로봇의 유능함에 기반한 신뢰에서는 사과할 때 신뢰도가 증가하였으며, 사과하는 방법에서 내부 귀인과 외부 귀인 간의 차이는 없었다. 정직성에 기반한 신뢰에서도 사과할 때 로봇에 대한 인간의 신뢰도가 증가함을 나타냈으며, 외부 기인보다 내부 기인, 그리고 행동을 하지 않을 때보다 부인할 때 로봇에 대해 더 신뢰 할 수 있다고 인식하였다. 두 번째로 같은 상황에서 로봇의 능력에 대하여 인간이 평가를 어떻게 하는지 확인하였다. 그 결과 로봇의 지각된 유능함과 정직성에 대하여 로봇이 사과할 때 더 능력이 높다고 인식되었다. 세 번째로 오류가 발생한 상황의 심각성에 대해 인간이 어떻게 느끼는지 확인하였다. 로봇의 4가지 신뢰 회복 방법 중, 부인할 때 오류의 심각성이 가장 높다고 인식되었고 이는 사과하지 않을 때 로봇에 대한 신뢰도가 낮아짐을 나타낸다(Kim et al., 2006; Sebo et al., 2019). 결과적으로 세 가지 모든 경우에서 로봇이 사과를 할 때 로봇에 대한 인간의 태도에 미치는 효과를 일관성 있게 관찰하였다. 특히 로봇이 사과할 때 인간의 신뢰도가 높아졌으며, 사과하지 않는 부인 행동을 했을 때는 신뢰도가 급격히 떨어지는 것을 확인할 수 있었다. 또한 로봇이 사과할 때 로봇에 대한 유능함과 정직성을 높다고 판단하였고, 로봇이 사과하지 않을 때 오류의 심각성에 대해 더 심각

하다고 느꼈으며 특히 로봇에 대해 불쾌함을 표했으며 화가 났다는 반응을 관찰하였다. 이러한 연구 결과는 사과하는 로봇의 행동이 모든 신뢰 위반 유형에서 부인보다 더 낫다는 연구와도 일맥상통하였으며(Bansal et al., 2015) 로봇이 일으킨 오류에 대하여 사과할 때 로봇에 대한 인간의 신뢰도가 증가함을 확인 할 수 있었다. 또한 인간과 외모를 모방한 로봇의 외형뿐 아니라 인간처럼 행동하는 로봇의 기능적 역할도 로봇에 대한 인간의 인식에 영향을 끼칠 수 있음을 보여준다. 이를 통해 향후 로봇 디자이너들이 로봇의 기능적 역할에 집중할 수 있도록 가이드 하는 것은 로봇에 대한 인간의 인식과 평가에 긍정적인 영향을 미칠 것으로 예상되며, 인간이 로봇을 신뢰할 수 있도록 사회적으로 행동하는 로봇을 설계하는 것이 바람직할 것이다.

한편 본 연구의 일관성 있는 결과에도 다음과 같은 몇 가지 한계가 있으며, 이를 보완할 수 있는 후속 연구가 진행되어야 할 것이다. 첫째, 로봇과 실제 환경에서 상호작용하는 연구가 필요할 것이다. 본 연구에서는 비대면 연구를 위하여 로봇과의 상호작용 상황을 녹화하여 참여자들이 시청하는 방식으로 진행되었다. 로봇과 실제 상호작용 할 때에는 실물을 직접보고 대화해야 하므로 로봇의 외형이나 크기에 따라 인간 기계 상호작용의 질이 달라질 수 있으므로(Broadbent et al., 2013) 이를 고려한 후속 연구가 진행되어야 할 것이다. 둘째, 로봇의 신뢰회복 행동에 방법에 문화적인 요소가 고려되어야 한다. 예컨대, 서양과 동양의 문화권에 따라 추가적인 요소가 고려되어야 할 것이다. 특히 미국과 같은 개인주의 사회에 비해 공동체주의 사회인 아시아 국가에서 사과를 더 많이 하는 연구 결과가 있다(Maddux et al., 2011). 이러한 문화적 차이로 인하여 사과가 의미하는 뜻 자체가 달라질 수 있으며, 문화적 차이로 인하여 인간 로봇 상호작용에 영향을 일으킬 것이다. 셋째, 로봇이 수행하는 업무의 유형과 다양한 동작 요소가 고려되어야 할 것이다. 본 연구에서는 로봇이 헬스케어 도우미로서의 대화 환경에서 실험이 진행되었으나, 로봇의 다양한 행동 유형 및 환경이 사용성에 영향을 끼칠 수 있으므로(Hoff & Bashir, 2015), 이를 고려한 다양한 환경에서의 검증이 필요할 것이다. 따라서 이러한 세 가지 측면에서 보다 매너있는 로봇의 발전을 위해 향후 추가적인 연구가 필요할 것이다.

본 연구는 몇 가지 시사점을 가진다. 먼저 로봇이 오류를 발생시키고 그것을 극복하기 위한 행동에 따라 로봇에 대한 인간의 태도가 민감하게 달라질 수 있음을 보여준다. 이는 이미 실제 환경에서도 인간들의 신뢰가 상대방의 행동에 따라 달라질 수 있음을 보여준다. 예를 들어 이베이를 통한 쇼핑 환경에서도 오류에 대하여 부인 할 때 보다 사과할 때 온라인 AI 챗봇이 더 믿음직스럽다는 것을 확인할 수 있다(Utz et al., 2009). 또한 로봇의 사과하는 행동이 인간의 정서적 행동 및 신뢰는 물론이며 성과에도 영향을 미칠 수 있음을 확인할 수 있다. 예컨대 로봇의 사과 행동은 산업 HRI 환경에서 로봇에 대한 신뢰를 높였으며, 안전성에 대한 불안을 줄여주었다는 것을 확인할 수 있다(Fratczak et al., 2021).

이러한 결과는 로봇에 대한 이상적인 행동을 확인하였다는 측면에서 중요하다. 본 연구를 통해 로봇에 대한 시각과 평가 또한 변할 수 있음을 확인하였고 구체적으로 로봇이 사과할 때 로

봇에 대한 능력을 높다고 사람들은 인지하였다. 따라서 자동화 과정에서 로봇의 성능을 올리기 위해 로봇이 잘하는 것에 집중해야 한다는 연구와 같이(Lee & See, 2004), 로봇이 인간과 커뮤니케이션 하는 상황에서 로봇이 먼저 사과할 수 있는 기능을 갖추어 궁극적으로 매너있는 로봇으로 발전할 수 있도록 가이드 할 필요가 있음을 시사한다.

## 참고문헌


Andreasson, R., Alenljung, B., Billing, E., & Lowe, R. (2018). Affective Touch in Human–Robot Interaction: Conveying Emotion to the NAO Robot. *International Journal of Social Robotics*, *10*(4), 473-491.

Alenljung, B., Andreasson, R., Billing, E A., Lindblom, J., & Lowe, R. (2017) User Experience of Conveying Emotions by Touch. In *Proceedings of the 26th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 1240-1247). IEEE

Atkinson, D., Hancock, P., Hoffman, R. R., Lee, J. D., Rovira, E., Stokes, C., & Wagner, A. R. (2012). Trust in computers and robots: The uses and boundaries of the analogy to interpersonal trust. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*(1), 303–307. Sage.

Baker, A. L., Phillips, E. K.,Ullman, D., & Keebler, J. R. (2018). Toward an understanding of trust repair in human-robot interaction. *ACM Transactions on Interactive Intelligent Systems*, *8*(4), 1–30.

Bansal, G., & Zahedi, F. M. (2015). Trust violation and repair: The information privacy perspective. *Decision Support Systems*, *71*, 62-77.

Bartneck, C. (2003). Interacting with an Embodied Emotional Character. In *Proceedings of the 2003 International Conference on Designing pleasurable products and interfaces* (pp. 55-60).

Beck, A., Stevens, B., Bard, K. A., & Cañamero, L. (2012). Emotional Body Language Displayed by Artificial Agents. *ACM Transactions on Interactive Intelligent Systems*, *2*(1), 1-29.

Bies, R. J., & Tripp, T. (1996). Beyond distrust: 'Getting even' and the need for revenge. In. R. Kramer, & T. Tyler (Eds.) *Trust in organizations* (pp. 246-260).

Breazeal, C. (2004). Social Interactions in HRI: the robot view. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *34*(2), 181-186.

Broadbent, E., Kumar, V., Li, X., Sollers 3rd, J., Stafford, R. Q., MacDonald, B. A., & Wegner, D. M. (2013). Robots with Display Screens: A Robot with a More Human like Face Display Is Perceived to Have More Mind and a Better Personality. *PloS ONE*, *8*(8), e72589.

Butler, J.K., Jr., & Cantrell, R. S. (1984). A behavioral decision theory approach to modeling dyadic trust

in superiors and subordinates. *Psychological Reports*, *55*, 19-28.

Carli, L. L., LaFleur, S.J., & Loeber, C.C. (1995). Nonverbal Behavior, Gender, and Influence. *Journal of Personality and Social Psychology*, *68*(6), 1030-1041.

Carney, D.R., Hall, J.A., & Smith-LeBeau, L. (2005). Beliefs about the nonverbal expression of social power. *Journal of Nonverbal Behavior*, *29*(2), 105-123.

Cassell, J., & Bickmore, T. (2003). Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User modeling and user-adapted interaction*, *13*(1-2), 89-132.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The BIAS map: behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology*, *92*(4), 631-648.

Desai, M.,Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013). Impact ofrobot failures and feedback on real-time trust. In *8th ACM/IEEE International Conference on Human-Robot Interaction*, 251-258.

DeSteno, D., Breazeal, C., Frank, R. H., Pizarro, D., Baumann, J., Dickens, L., & Lee,J. J. (2012). Detecting the Trustworthiness of Novel Partners in Economic Exchange. *Psychological Science*, *23*(12), 1549-1556.

De Visser, E. J., Pak, R., & Shaw, T. H. (2018). From "automation" to "autonomy": The importance of trust repair in human–machine interaction. *Ergonomics*, *61*(10), 1409-1427.

Dirks, K.T., Lewicki, R. J., & Zaheer, A. (2009). Repairing relationships within and between organizations: Building a conceptual foundation. *Academy of Management Review*, *34*(1), 68-84.

Dweck, C.S., Chiu, C. Y., & Hong, Y. Y. (1995). Implicit theories and their role in judgments and reactions: A word from two perspectives. *Psychological Inquiry*, *6*(4), 267-285.

Engelhardt, S., Hansson, E., & Leite, I. (2017, August). Better Faulty than Sorry: Investigating Social Recovery Strategies to Minimize the Impact of Failure in Human-Robot Interaction. In *WCIHAI@ IVA*, 19-27.

Fratczak, P., Goh, Y. M., Kinnell, P., Justham, L., & Soltoggio, A. (2021). Robot apology as a post-accident trust-recovery control strategy in industrial human-robot interaction. *International Journal of Industrial Ergonomics*, *82*.

Ferrin, D.L., Kim, P. H., Cooper, C. D., & Dirks, K. T. (2007). Silence speaks Vols.:The effectiveness of reticence in comparison to apology and denial for responding to integrity- and competence-based trust violations. *Journal of Applied Psychology*, *92*(4), 893–908.

Fuoli, M., Van de Weijer, J., & Paradis, C. (2017). Denial outperforms apology in repairing organizational trust despite strong evidence of guilt. *Public Relations Review*, *43*(4), 645-660.

Hald, K., Weitz, K., André, E., & Rehm, M. (2021, November). "An Error Occurred!" - Trust Repair

With Virtual Robot Using Levels of Mistake Explanation. In *Proceedings of the 9th International Conference on Human-Agent Interaction*, 218-226.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517-527.

Haring, K. S., Matsumoto, Y., & Watanabe, K. (2013). How do people perceive and trust a lifelike robot. *Lecture Notes in Engineering and Computer Science*, *1*, 425 - 430.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407-434.

Hur, Y., & Han, J. (2009). Analysis on Children's Tolerance to Weak Recognition of Storytelling Robots. *J. Convergence Inf. Technol.*, *4*(3), 103-109.

Jessup, S.A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The measurement of the propensity to trust automation. In *International Conference on Human- Computer Interaction*, 476 - 489.

Jian, J., Bisantz, A. & Drury, C. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53-71.

Johnson, D. O., & Cuijpers, R. H. (2019). Investigating the Effect of a Humanoid Robot's Head Position on Imitating Human Emotions. *International Journal of Social Robotics*, *11*(1), 65-74

Kähkönen, T., Blomqvist, K., Gillespie, N., & Vanhala, M. (2021). Employee trust repair: A systematic review of 20 years of empirical research and future research directions. *Journal of Business Research*, *130*, 98-109.

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: The effects of apology versus denial for repairing competence-versus integrity- based trust violations. *Journal of Applied Psychology*, *89*(1), 104-118.

Kim, P. H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence-vs. integrity-based trust violation. *Organizational behavior and human decision processes*, *99*(1), 49-65.

Kim, P. H., Dirks, K. T., & Cooper, C. D. (2009) "The repair of trust: A dynamic bilateral perspective and multi level conceptualization". *Academy of Management Review*, *34*(3), 401-422.

Kim, P. H.,Cooper, C. D., Dirks, K. T., & Ferrin, D. L. (2013). Repairing trust with individuals vs. groups. *Organizational Behavior and Human Decision Processes*, *120*(1), 1-14.

Lee, J. D, & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50-80.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation.

*International Journal of Human-Computer Studies*, *40*(1), 153-184.

Lee, K. M., Peng, W., Jin, S. A., & Yan, C. (2006). Can Robots Manifest Personality?: An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human–Robot Interaction. *Journal of Communication*, *56*(4), 754-772.

Leung, S. O. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert scales. Journal of social service research, 37(4), 412-421.

Lewicki, R.J., & Brinsfield, C. (2017). Trust repair. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*, 287-313.

Li, J., Cuadra, A., Mok, B., Reeves, B., Kaye, J., & Ju, W. (2019). Communicating Dominance in a Non anthropomorphic Robot Using Locomotion. *ACM Transactions on Human-Robot Interaction*, *8*(1), 1-14.

Maddux, W. W., Kim, P. H., Okumura, T., & Brett, J. M. (2011). Cultural differences in the function and meaning of apologies. *International negotiation*, *16*(3), 405-425.

Madhavan, D., & Wiegmann, D. A. (2005). Effects of information source, pedigree, and reliability on operators utilization of diagnostic advice. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, *49*(3), 487–491.

Mayer, R. C., & Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology*, *84*(1), 123-136.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709-734.

McCall, D., & Kölling, M. (2014). Meaningful categorization of novice programmer errors. In *Proceedings of 2014 IEEE Frontiers in Education Conference,* 1-8.

McColl, D., & Nejat, G. (2014). Recognizing Emotional Body Language Displayed by a Human-like Social Robot. *International Journal of Social Robotics*, *6*(2), 261-280.

Mende, M., Scott, M. L., van Doorn, J., Grewal, D., & Shanks, I. (2019). Service Robots Rising: How Humanoid Robots Influence Service Experiences and Elicit Compensatory Consumer Responses. *Journal of Marketing Research*, *56*(4), 535-556.

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors*, *50*(2), 194-210.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The Uncanny Valley [from the field]. *IEEE Robotics & Automation Magazine*, *19*(2), 98-100.

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*, 429-460.

Nass, C. I., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social*

*Issues*, *56*(1), 81-103.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are Social Actors. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (pp. 72-78).

Purington, A., Taft J. G., Sannon, S., Bazarova, N. N., & Taylor, S. H. (2017). "Alexa is my new BFF," Social Roles, User Satisfaction, and Personification of the Amazon Echo. In *Proceedings of the 2017 CHI Conference Expended Abstracts on Human Factors in Computing Systems* (pp. 2853-2859).

Robinette, P., Howard, A. M., &Wagner, A. R. (2017). Effect of robot performance on human‐robot trust in time-critical situations. *IEEE Transactions on Human-Machine Systems*, *47*(4), 425-436.

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., & Joublin, F. (2013). To Err is Human(-like): Effects of Robot Gesture on Perceived Anthropomorphism and Likability. *International Journal of Social Robotics*, *5*(3), 313‐323.

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015, March). Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 1-8). IEEE.

Sanders, T.,Kaplan, A., Koch, R., Schwartz, M., & Hancock, P. A. (2019). The relationship between trust and use choice in human-robot interaction. *Human Factors: The Journal of Human Factors and Ergonomics Society*, *61*(4),614‐626.

Strait, M. K., Floerke, V. A., Ju, W., Maddox, K., Remedios, J. D., Jung, M. F., & Urry, H. L. (2017). Understanding the Uncanny: Both Atypical Features and Category Ambiguity Provoke Aversion toward Humanlike Robots. *Frontiers in Psychology*, *8*, 1366

Schaefer K. E. (2016) Measuring Trust in Human Robot Interactions: Development of the "Trust Perception Scale-HRI". In *Robust Intelligence and Trust in Autonomous Systems* (pp. 191-218). Springer, Boston, MA.

Sebo, S. S., Krishnamurthi, P., & Scassellati, B. (2019). "I Don't Believe You":Investigating the Effects of Robot Trust Violation and Repair. In *2019 14$^{h}$ ACM/IEEE International Conference on Human-Robot Interaction* (pp. 57-65). IEEE.

Sebo, S. S., Traeger, M., Jung, M., & Scassellati, B. (2018). The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 178-186.

Tomlinson, E. C., Dineen, B. R., & Lewicki, R. J. (2004). The road to reconciliation: Antecedents of victim willingness to reconcile following a broken promise. *Journal of Management*, *30*(2), 165-187.

Torre, I., Goslin, J., White, L., & Zanatto, D. (2018). Trust in artificial voices: "A congruency effect" of first impressions and behavioral experience. In *Proceedings of the Technology, Mind, and Society*, 1-6.

Traeger, M.L., Sebo, S. S., Jung, M., Scassellati, B., & Christakis, N. A. (2020). Vulnerable Robots Positively Shape Human Conversational Dynamics in a Human–Robot Team. In *Proceedings of the National Academy of Sciences*, *117*(12), 6370-6375.

Utz, S., Matzat, U., & Snijders, C. (2009). On-line reputation systems: The effects of feedback comments and reactions on building and rebuilding trust in on-line auctions. *International Journal of Electronic Commerce*, *13*(3), 95-118.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017). Attention is all you need. *In Advances in neural information processing systems,* 5998-6008.

Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and reliance on an automated combat identification system. *Human Factors*, *51*(3), 281-291.

Weun, S., Beatty, S. E., & Jones, M. A. (2004). The impact of service failure severity on service recovery evaluations and post recovery relationships. *The Journal of Services Marketing*, *18*(2), 133–146.

Xu, J., Broekens, J.,Hindriks, K., & Neerincx, M. A. (2014). Robot Mood is Contagious: Effects of Robot Body Language in the Imitation Game. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems,* 973-980.

1차 원고 접수: 2021. 12. 10
1차 심사 완료: 2022. 05. 11
2차 원고 접수: 2022. 08. 01
2차 심사 완료: 2022. 11. 07
3차 원고 접수: 2022. 11. 21
3차 심사 완료: 2022. 12. 09
최종 게재 확정: 2022. 12. 13

(*Abstract*)

# The effect of trust repair behavior on human-robot interaction

Hoyoung Maeng[1)] Whani Kim[2)] Jaeun Park[2)] Sowon Hahn[1)2)]

[1)]Interdisciplinary Program in Cognitive Science, Seoul National University

[2)]Department of Psychology, Seoul National University

This study aimed to confirm the effect of social and relational behavior types of robots on human cognition in human-robot interaction. In the experiment, the participants evaluated trust in robots by watching a video on the robot Nao interacting with a human, in which the robot made an error and then made an effort to restore trust. The trust recovery behavior was set as three conditions: an internal attribution in which the robot acknowledges and apologizes for an error, a condition in which the robot apologizes for an error but attributes it externally, and a non-action condition in which the robot denies the error itself and does not take any action for the error. As the result, in all three cases, the error was perceived as less serious when the robot apologized than when it did not, and the ability of the robot was also highly evaluated. These results provide evidence that human attitudes towards robots can respond sensitively depending on the robot's behavior and how they overcome errors, suggesting that human perception towards robots can change. In particular, the fact that robots are more trustworthy when they acknowledge and apologize for their own errors shows that robots can promote positive human-robot interactions through human-like social and polite behavior.

*Key words : human–robot interaction, robot failure, trust violation, trust repair*