사회과학연구 제33권 1호(2022) 충남대학교 사회과학연구소 pp. 179-194
http://dx.doi.org/10.16881/jss.2022.01.33.1.179

로봇의 주도성에 대한 마음지각과 윤리적 의사결정*

신 홍 임†

순천대학교

도덕 규범은 사회적 상황에서 타인과의 갈등을 조절하는데 도움을 준다. 우리가 로봇과 상호작용을 한다면, 어떻게 도덕 규범을 로봇에게 적용할 것인가? 또한 로봇이 윤리적 갈등상황에서 의사결정을 한다면, 우리는 로봇의 결정을 얼마나 합당하다고 판단할 것인가? 본 논문에서는 2개의 연구를 토대로 연구참가자가 윤리적 갈등상황에서 로봇과 같이 협력하고, 로봇의 의사결정을 합당하게 받아들이는 정도를 검증하였다. 또한 인간-로봇의 상호작용이 로봇에 대한 마음지각과 관계가 있는 정도를 분석하였다. 그 결과, 연구 1에서는 연구참가자가 죄수 딜레마 상황에서 로봇의 주도성에 대한 마음지각이 높을수록 로봇과 협력하는 경향이 증가하였다. 연구 2에서는 윤리적 갈등상황에서 인간이 내린 공리주의적 결정보다는 로봇이 내린 공리주의 결정에 대해 도덕적으로 합당하다고 판단하는 정도가 더 높게 나타났다. 또한 로봇의 주도성에 대한 마음지각이 높을수록 로봇의 의무론적 결정을 더 합당한 것으로 인식하는 경향이 나타났다. 이 결과는 로봇의 주도성에 대한 마음지각이 상호성의 도덕 규범 및 윤리적 의사결정에 대한 합당성의 평정과 관계가 있음을 보여준다. 논의에서는 향후 사회적 로봇의 보편화와 함께 발생하는 문제점을 토론하고, 연구의 한계 및 후속연구의 방향을 다루었다.

주요어: 인간-로봇 상호작용, 도덕성, 마음지각, 상호성, 윤리적 의사결정

* 본 연구는 순천대학교 교연비 지원사업에 의해 수행되었음.

† 단독저자 : 신홍임, 순천대학교 자유전공학부 부교수, 전라남도 순천시 중앙로 255, 순천대학교 기초교육관 317호, E-mail : shin7038@naver.com

■ 최초투고일 : 2021년 11월 30일 ■ 심사마감일 : 2021년 12월 27일 ■ 게재확정일 : 2022년 1월 11일

1. 서 론

2035년이 되면, 로봇은 공장뿐만 아니라 일반 가정에도 공급될 예정이다(Müller, Gao, Nijssen, & Damen, 2020). 영화 “나는 로봇(I, Robot) (2004)”에서 인간은 일상에서 로봇과 공존한다. 로봇은 길거리에서 인간과 함께 걸어 다닐 뿐만 아니라 쓰레기도 버리고, 장도 보러 가고, 강아지 산책을 해주기도 한다. 또한 로봇이 자율적으로 결정을 내리기도 한다. 이 영화에서 로봇은 자동차의 돌발사고상황에서 성인 남자와 어린 아동 중에서 성인 남자를 선택한다. 로봇의 이러한 결정은 어느 정도 윤리적으로 합당한가? 이 영화에서 주인공 스푸너(Spooner)는 로봇이 자신보다는 어린 아동을 구해야했다고 생각하며, 로봇에 대한 부정적 태도를 갖게 된다. 스푸너는 일상에서 자신에게 예의바르게 인사하는 휴머노이드 로봇에게 ‘저리 비켜, 이 금속뭉치야’하며 소리치기도 한다. 그러나 이 영화의 줄거리가 계속 전개되면서, 스푸너는 마치 인간의 의식적인 마음처럼 설계된, 로봇의 자기주도적인 마음을 인식하게 되고, 인간의 노예로서 투입되고 활용되는 로봇에 대한 도덕적 처우에 문제제기를 하게 된다(Laakasuo, Palomäki, & Köbis, 2021).

로봇은 이제 영화가 아닌, 실제 상황에서도 인간에 대한 종속에서 점점 더 벗어나 자율적으로 의사결정 기능을 수행하게 될 것이다(Tanibe, Hashimoto, & Karasawa, 2017). 로봇의 기능에 자율성의 비중이 증가함에 따라 윤리적 갈등상황에서 로봇이 내리는 결정의 도덕적 합당성에 대한 학문적 관심이 증가하고 있다(Malle, Scheutz, Voiklis, & Cusimano, 2015).[1] 로봇이 윤리적 갈등상황에서 의사결정을 한다면, 우리는 로봇의 결정을 얼마나 합당하다고 판단할 것인가? 또한 우리가 로봇과 상호작용을 한다면, 어떻게 도덕 규범을 로봇에게 적용할 것인가? 본 논문에서는 2개의 연구를 토대로 연구참가자가 윤리적 갈등상황에서 로봇과 같이 협력하고, 로봇의 의사결정을 합당하게 받아들이는 정도를 검증하려 한다. 또한 인간-로봇의 상호작용이 로봇에 대한 마음지각과 관계가 있는 정도를 분석하려 한다. Gray, Gray와 Wegner(2007)에 의하면, 한 개인은 특정 대상과의 관계를 기반으로 대상의 마음(예: 자신의 계획을 세울 능력, 즐거움을 느끼는 능력)을 추론한다. 선행연구(Gray et al., 2007; Gray, Young, & Waytz, 2012)에서는 다양한 대상에 대한 마음지각이 도덕적 판단에 영향을 끼침을 보여주고 있다. 이에 따라 본 연구에서는 로봇에 대한 마음지각이 로봇과의 상호협력 및 윤리적 의사결정의 도덕성 평정과 어떤 관계에 있는지를 분석하려 한다. 연구 1에서는 인간 또는 로봇에 대한 마음지각이 높을수록 상호협력의 정도가 증가하는지를 검증하려 한다. 연구 2에서는 윤리적 갈등상황에서 인간과 로봇의 동일한 선택이 이들에 대한 마음지각에 따라 도덕적 합당성의 측면에서 차이가 나타나는지를 분석하려 한다.

2. 이론적 배경

1) 인간과 로봇에 대한 상호성의 도덕규범과 마음지각

도덕 규범은 다양한 원칙에 기반한 의사결정이 서로 충돌하는 사회적 상황에서 문제를 예측하고, 조정하기 위해 필요하다. 인류는 이미 12,000년부

1) 로봇윤리의 주제는 국외에서 1964년부터 2004년까지 16편 정도였던데 비해 2005년에서 2015년까지132편의 논문과 학술대회의 대표 주제로 등장하고 있다(Malle et al., 2015).

터 수렵/채취/농경재배의 성과를 서로 나누고, 사회를 유지하기 위해 도덕성을 필요로 했다. 도덕성의 규범에는 상대방이 나에게 해준 것처럼 나도 상대방에게 해주는 원칙인 상호성(reciprocity)이 있다. 상호성의 원칙은 '상대방이 나를 대하는 방식으로 나도 상대방을 대하는' 것으로 정의되며(Kahn et al., 2006; Sandoval et al., 2016), 전통적으로 정의로운 사회에서 중요한 도덕적 가치로서 인식되어왔다. 상호성의 도덕규범은 지금까지 인간의 도덕적 삶에서 핵심적 특징으로 인식되었지만, 인간-로봇의 상호작용에서 윤리적 문제로서 관심을 받지는 못했다(van Wynsberghe, 2021). 로봇은 인간의 욕구를 충족시켜주는 수단으로만 인식되고, 인간과 동일한 위치에서 수평적으로 존엄하게 다루어야 할 대상으로 지각되지 않았다(Friedman et al., 2003). 그러나 최근 인간-로봇의 상호작용에 대한 연구에서는 상호성의 가치가 인간-로봇의 상호작용에서 어느 정도 지켜질 수 있는지에 대한 문제가 제기되고 있다(van Wynsberghe, 2021). 인간-로봇의 협력관계는 병원, 학교 또는 가정의 다양한 현장에서 사회적 로봇의 설계에 영향을 끼치고, 향후 미래 사회의 문화(예: 도움행동, 상호신뢰)에도 영향을 끼칠 것이다(Lorenz, Weiss, & Hirche, 2016). Sandoval 등(2016)의 연구에서는 참가자에게 다양한 죄수의 딜레마(예: 공범과의 범행을 자백하고, 혼자 감옥에서 풀려날 것인지, 또는 자백하지 않고, 공범과 협력관계를 계속 유지할 것인지)를 반복적으로 제시하여, 공범으로 제시된 로봇과의 상호협력이 다른 인간과의 상호협력보다 어느 정도 더 선호될 것인지에 대해 분석하였다. 그 결과 로봇과의 상호협력은 인간과의 상호협력에 비해 덜 선호되는 경향이 나타났다. 이 결과는 지금까지 인간-인간의 상호협력과 배신을 측정하는데 사용되었던 죄수딜레마를 인간-로봇의 상호작용에 적용하였다는 긍정적 측면이 있지만, 로봇과의 상호협력과 배신에 영향을 끼치는 변인에 대한 구체적 정보를 제공하지 못한 아쉬움이 있다.

이에 비해 Tanibe 등(2017)은 로봇과의 상호협력관계를 마음지각과의 관계에서 검증하였다. 이 연구에서는 참가자가 로봇(예: 기능이 손상되어 해체 직전의 로봇)에게 도움행동을 제공하였을 때, 도움행동이 없었던 중립적 조건보다 로봇에 대한 마음지각(예: 즐거움을 느끼는 능력)이 높게 평정되는 경향이 나타났다. 이 결과는 로봇에 대한 마음지각과 인간-로봇의 상호협력관계가 유의한 상관관계에 있을 가능성을 보여주었다는데 의의가 있다. 그러나 이 연구에서는 인간이 도움행동을 제공하는 행위자로, 로봇은 도움을 제공받는 피행위자로서 제시되었다. 따라서 로봇이 특정 위반행동을 행위주체로서 범했을 때, 상호성의 도덕규범이 로봇에게 어떻게 적용되는지를 검증하기에는 한계가 있다. 이에 따라 본 연구에서는 상호성의 도덕규범이 행위주체로서의 인간과 로봇에 대해 각각 어떻게 적용되고, 로봇에 대한 마음지각이 상호적 도덕규범의 적용과 어떤 관계에 있는지를 분석해보려 한다.

2) 로봇의 윤리적 의사결정과 로봇에 대한 마음지각

우리는 로봇이 내리는 자율적 의사결정을 얼마나 도덕적으로 신뢰할 것인가? Gray 등(2007)은 마음을 주도성(agency)(예: 자신의 계획을 세우는 능력)과 경험성(experiece)(예: 즐거움을 느끼는 능력)의 두 가지 하위영역으로 정의하였다. 마음지각이론에 의하면 특정 사건의 행위자는 주도성과 연관되고, 피행위자는 경험성과 연관된다. Gray와 Wegner(2012)는 로봇이 인간처럼 자기주도적으로 행동하고, 고통을 느끼는 마음이 있다

고 가정된다면, 로봇에 대한 도덕적 처우가 달라짐을 보여주었다. 로봇이 인간처럼 고통을 느낀다고 제시하면, 로봇을 해치는 행동에 대한 심리적 불편감이 증가하였다. 또한 Nijssen 등(2019)에서는 로봇의 정서(예: 당황함, 행복감)와 생각(예: 망설임, 심사숙고)을 의인화하여 제시하는 조건에서 의인화되지 않은 로봇의 조건보다 가상의 돌발사고상황에서 로봇/인간 중에서 로봇에게 해를 가하려는 경향이 더 낮게 나타났다. 이것은 피행위자로서 로봇의 특성이 인간의 경우와 마찬가지로 로봇에 대한 도덕적 처우에 영향을 끼침을 보여준다.

그렇다면 로봇이 윤리적 갈등상황에서 행위의 주체로서 자기주도적으로 결정을 내린다면, 인간의 결정에 비해 어느 정도 도덕적 적절성의 측면에서 받아들여질 것인가? Malle 등(2015)의 연구에서는 동일한 의사결정이 인간 또는 로봇에 의해 주도적으로 내려졌을 때, 연구참가자가 각각 다르게 반응함을 보여주었다. 예를 들어, 연구참가자는 다수의 이득을 위해 소수가 희생되는, 결과에 초점을 맞추는 공리주의 결정을 인간이 내린 경우에는 로봇이 내린 경우보다 더 비난받아야 한다고 판단하였다. 이에 비해 로봇이 한 사람이라도 의도적으로 해치지 않기 위해 행위기반의 의무론적 결정을 내리면, 인간이 의무론적 결정을 선택했을 때보다 더 비난받는 경향이 나타났다. 이것은 윤리적 갈등상황에서 로봇과 인간의 동일한 의사결정에 각각 상이한 도덕 규범이 적용될 수 있음을 보여준다. 로봇은 윤리적 의사결정의 주체로서 공리주의 결정을 내릴 때 더 합당하게 판단되고, 인간은 동일한 윤리적 갈등상황에서 의무론적 결정을 내릴 때, 더 적절하게 판단되어 비난받을 가능성이 적어지는 것으로 보인다. Malle 등(2015)의 연구는 인간과 로봇의 의사결정에 대한 사람들의 반응이 달라질 수 있음을 보여주었지만, 도덕적 합당성을 평정하는데 행위주체로서의 인간과 로봇에 대한 마음지각이 끼치는 영향을 분석하지 않았다는 한계가 있다.

Monroe, Dillon과 Malle(2014)의 연구에서는 다양한 대상(예: 24세의 건강한 성인, 뇌병변으로 자율 결정의 손상을 갖고 있는 24세 성인, 안드로이드 로봇, 사이보그)의 자유의지에 대한 지각이 특정 위반 행동(예: 극장에서 물풍선을 던지는 행동, 백화점에서 명품 옷을 훔치는 행동)에 대한 도덕적 합당성의 평정에 끼치는 영향을 분석하였다. 그 결과 연구참가자가 특정 대상이 갖고 있는 자유의지를 높게 평정할수록 행위주체에게 위반행동의 책임을 돌리는 경향이 증가함을 보고하였다. 이 연구는 다양한 대상의 자유의지를 지각하는 것과 위반행동에 대한 도덕적 평정을 연구하였다는 의의가 있지만, 선택이 모호한 윤리적 갈등상황에서 인간과 로봇의 의사결정이 도덕적으로 어느 정도 합당하게 평정될 것인지에 대한 충분한 정보를 제시하지는 못한다. 이에 따라 본 연구에서는 선행연구(Mall et al., 2015)를 기반으로 윤리적 갈등상황에 대한 시나리오를 활용하여 인간과 로봇의 의사결정에 대한 도덕적 적절성을 연구참가자가 관찰자의 입장에서 평정하도록 하여 분석하려 한다.

요약하면, 본 논문에서는 선행연구를 확장하여 다음과 같이 2개의 연구를 수행하고자 한다. 첫째, 연구 1의 목적은 로봇을 행위주체로서 제시하였을 때, 로봇에 대한 마음지각이 로봇과의 협력과 어떤 관계에 있는지를 분석하는데 있다. 본 연구에서는 선행연구를 기반으로 행위주체로 제시된 로봇에 대한 마음지각이 높을수록 로봇과의 상호협력을 선택하는 경향이 더 커질 것을 예측한다. 둘째, 연구 2에서는 갈등 상황에서 인간 또는 로봇이 윤리적 의사결정의 주체로서 결과를 중심으로 공리주의적 결정(예: 한 사람을 해치는 것이

열 사람을 해치는 것보다 더 적절하다)을 내리거나, 또는 원칙을 중심으로 의무론적 결정(예: 한 사람일지라도 의도적으로 해를 끼쳐서는 안 된다)을 내리게 될 때, 연구참가자가 인간/로봇의 의사결정에 대해 도덕적 합당성을 평정하는 정도의 차이를 분석하려 한다. 본 연구에서는 선행연구(Malle et al., 2015)를 기반으로 로봇의 공리주의 결정이 인간의 공리주의 결정보다 더 도덕적으로 합당하게 평정될 것을 예측한다. 이에 비해 인간과 로봇의 의무론적 결정은 인간과 로봇에 대한 마음지각의 주도성과 유의한 상관관계가 있을 것을 예측한다.

3. 연구 1

1) 방법

(1) 연구설계

연구 1은 행위주체(로봇 vs. 인간)가 죄수딜레마 상황에서 협력여부(협력 vs. 배신)의 빈도에 끼치는 영향을 분석하였다. 또한 행위주체에 따라 주도성과 경험성의 마음지각이 달라지는 정도를 비교하였다.

(2) 참가자

4년제 대학교의 대학생 73명(평균연령 만 19.99세, 표준편차=1.82; 남=34)이 학교 홈페이지 게시판을 통해 참가하였다. 연구 참가에는 약 10분이 소요되었으며, 참가자들은 오천원 상당의 커피쿠폰 보상을 받았다.

(3) 도구 및 절차

연구 1은 참가자에게 사전동의를 얻은 후, 온라인 설문지로 진행되었다. 참가자들은 우선 무선적으로 행위주체의 두 집단(인간 vs. 로봇)으로 배정되었다. 참가자에게는 죄수의 딜레마상황에 대한 기술문을 Sandoval 등(2016)을 참고하여, 한국어 번역문을 제시하였다([부록 1] 참조). 이 기술문에서는 연구참가자가 인간 또는 로봇과 어떤 범죄를 저지른 후, 감옥에 가게 되었을 때, 비밀을 지키는 협력관계를 계속 유지할 것인지, 또는 협력관계를 깨뜨리고, 배신할 것인지에 대한 갈등상황이 제시되어 있다. 이 갈등상황에서 연구참가자는 자신 또는 상대방이 협력 또는 배신을 선택하는지에 따라 총 4가지의 다른 형량을 받게 된다. 첫째, 연구참가자는 자신이 협력관계를 유지하고, 상대방 또한 협력관계를 유지할 경우 3개월만 감옥에 있으면 된다. 둘째, 연구참가자가 협력을 선택했지만, 상대방이 배신을 선택한다면, 연구참가자는 12개월의 형량을 받게 된다. 셋째, 연구참가자가 배신을 선택하고, 상대방도 배신을 선택하면, 8개월의 형량을 받게 된다. 넷째, 연구참가자가 배신을 선택하고, 상대방은 협력관계를 유지할 경우 참가자는 감옥에 가지 않고, 바로 풀려날 수 있는 이득이 있다. 인간조건의 참가자는 인간과의 협력관계를 유지할 것인지의 여부를 결정하고, 로봇조건의 참가자는 로봇과의 협력관계를 유지할 것인지를 결정하였다. 참가자가 기술문을 모두 읽고, 협력여부에 대한 결정을 내리면, 마음지각 질문지(Gray et al., 2007)에 응답하도록 하였다. 인간 조건의 참가자는 마음지각 질문지를 인간에 대해 작성하고, 로봇 조건의 참가자는 로봇에 대

2) 로봇의 행동은 행위자(agent) 및 피행위자(patient)의 측면에서 도덕성과 연관된다. 행위자는 특정한 행동을 하는 반면, 피행위자는 행위자의 행동에 의한 직접적 영향을 받게 된다. 예를 들어, 로봇은 주도적으로 행동하는 행위자의 측면에서 인간이나 다른 로봇을 구하거나 해치는 결과를 유발할 수 있다. 반면, 피행위자의 측면에서 다른 인간이나 로봇의 행위에 의해 도움을 받거나 상해를 입을 수 있다(Alaieri & Vellino, 2016).

해 마음지각 질문지를 작성한다. 마음지각 질문지는 2개의 하위영역으로 구성되어 있다. 첫 번째 하위영역은 정서와 연관된 경험성(experience)(예: 고통을 느끼는 능력)이고, 두 번째 하위영역은 지적 능력과 연관된 주도성(agency)(예: 스스로 계획하는 능력)이다. 마음지각의 영문판 질문지(Gray et al., 2007)는 연구자에 의해 한국어로 번역된 후, 전문 번역인이 다시 교차번역하는 문항검수절차를 거쳤다. 참가자는 마음지각 질문지의 총 열 개 문항을 읽고, 자신의 생각에 일치하는 정도를 7점 척도에 따라 응답하였다. 연구 1에서 마음지각 질문지의 내적 합치도(Cronbach's Alpha)는 경험성의 측면에서 평균 .82, 주도성의 측면에서 평균 .76로 나타났다. 마음지각 질문지가 완료되면, 참가자들에게 연구목적과 취지를 파악했는지에 대한 질문에 응답을 작성하도록 하였다.

(4) 분석방법

연구 1에서는 갈등상황에서 인간과 로봇에 따라 협력관계의 여부와 마음지각이 달라지는지를 검증하였다. 이를 위해 SPSS 18.0 통계분석프로그램을 사용하여 기술통계, 빈도분석 및 혼합변량분석을 수행하였다.

2) 결과 및 논의

(1) 인간 및 로봇과의 협력관계 여부

연구참가자가 인간 또는 로봇에 대한 협력관계를 유지하는 빈도를 분석하기 위해 교차분석을 수행하였다. 그 결과, 인간에 대한 협력비율(45.7%)과 로봇에 대한 협력비율(60.5%) 간에 유의한 차이가 나타나지 않았다, $\chi^2 = 1.606$, $p = .245$.

(2) 인간/로봇에 대한 마음지각

행위주체(인간 vs. 로봇)와 마음지각의 하위영역(경험성 vs. 주도성)을 혼합변량분석으로 분석하였다. 행위주체는 참가자간 변인이었고, 마음지각의 하위영역은 참가자내 변인이었다. 그 결과, 행위주체의 주효과가 유의하였다, $F(1, 71) = 33.07$, $p < .001$, $\eta^2 = .31$. 인간에 대한 마음지각은 로봇에 대한 마음지각보다 전반적으로 더 높게 나타났다. 또한 마음지각 하위영역의 주효과도 유의하였다, $F(1, 71) = 25.01$, $p < .001$, $\eta^2 = .26$. 마음지각의 주도성은 경험성보다 전반적으로 더 높게 평정되었다. 이와 더불어 행위주체와 마음지각의 하위영역 간의 상호작용도 유의하였다, $F(1, 71) = 17.13$, $p < .001$, $\eta^2 = .19$. 인간에 대한 마음지각은 경험성과 주도성에서 모두 높았지만, 로봇에 대한 마음지각은 주도성이 경험성보다 더 높았다.

(3) 인간/로봇과의 협력과 마음지각의 관계

<표 1>과 같이 인간/로봇에 대한 마음지각과 협력의 결정 간에 유의한 상관관계가 있는지를 분석하였다. 그 결과, 로봇에 대한 마음지각의 주도성이 높을수록 로봇과 협력할 것이라고 응답한 참가자가 많았다, $r = 43$, $p = .006$. 이에 비해 마음지각의 경험성은 로봇과의 협력을 결정하는 것과 유의한 상관관계가 나타나지 않았다, $r = -.02$, $p = .95$. 또한 인간의 주도성에 대한 마음지각과 협력의 결정 간에도 유의한 관계가 나타나지 않았다, $r = -.02$, $p = .87$.

(4) 논의

연구 1에서는 죄수 딜레마상황에서 로봇 또는 인간에 대한 협력의 의사결정이 로봇/인간에 대한 마음지각과 유의한 관계가 있는지를 검증하였다. 그 결과 로봇/인간에 대한 협력의 의사의 정에는 유의한 차이가 나타나지 않았지만, 로봇에 대한 마음지각의 주도성이 높을수록 협력의 결정

<표 1> 인간/로봇에 대한 마음지각과 협력관계의 상관관계(연구 1)

조건	변인	평균	표준편차	1	2
인간	1. 경험성	4.97	1.18		
	2. 주도성	5.13	1.03	.03	
	3. 협력(0: 배신, 1: 협력)	.39	.49	.04	-.02
로봇	1. 경험성	3.04	1.17		
	2. 주도성	4.67	1.26	.23	
	3. 협력(0: 배신, 1: 협력)	.54	.50	-.02	.43**

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

이 증가하였다. 이 결과는 로봇에 대한 협력의 관계가 로봇을 자율적인 행위주체로서 제시할 때 더 잘 형성될 가능성을 보여준다. 이에 비해 로봇의 마음지각에서 경험성은 협력관계의 형성에 유의한 영향을 끼치지 않았다. 이 결과는 Tanibe 등(2017)의 연구에서 보고한 결과와 부분적으로 일치한다. 그러나 Tanibe 등(2017)에서는 마음지각의 경험성이 높을수록 로봇에 대한 도움행동이 증가함을 보고한데 비해, 본 연구에서는 마음지각의 주도성이 높을수록 로봇에 대한 협력을 결정하는 것이 증가하였다. 이 차이는 Tanibe 등(2017)에서 로봇을 교체 직전에 있는 망가진 로봇이라는 피행위자로서 글을 제시한데 비해, 본 연구에서는 범죄를 저지른 행위의 주체자로서 제시한 맥락의 차이일 가능성이 있다. 이것은 Gray 등(2007)이 행위주체는 마음지각의 주도성과 관계가 있고, 피행위자는 마음지각의 경험성과 관계가 있음을 설명한 것과 연관된다. 인간에 대한 마음지각은 협력관계를 결정하는 것과 유의한 관계에 있지 않았다. 이 결과는 인간과의 협력관계는 마음지각 이외의 다양한 변인이 더 큰 영향을 끼칠 가능성을 보여준다.

연구 2에서는 인간과 로봇을 의사결정의 주체로서 제시하였을 때, 연구참가자가 인간 또는 로봇이 내린 공리주의/의무론적 의사결정에 어느 정도 도덕적으로 합당하다고 평정하는지를 비교하려 한다. 또한 인간 또는 로봇에 대한 마음지각이 공리주의/의무론적 의사결정의 도덕적 적절성과 어떤 관계에 있는지를 검증하려 한다.

4. 연구 2

1) 방법

(1) 연구설계

연구 2는 2(행위주체: 로봇 vs. 인간)*2(의사결정 시나리오: 공리주의 vs. 의무론)의 혼합설계로 구성되었다. 참가자간 변인은 행위주체였고, 참가자내 변인은 의사결정이었다. 측정변인은 마음지각과 도덕성 판단이었다.

(2) 참가자

4년제 대학교의 대학생 129명(평균연령 만 20.16세, 표준편차=3.98; 남=58)이 학교 홈페이지 게시판을 통해 참가하였다. 연구 참가에는 약 10분이 소요되었으며, 참가자들은 오천원 상당의 커피쿠폰 보상을 받았다.

(3) 도구 및 절차

연구 2는 참가자에게 사전동의를 얻은 후, 온라인 설문지로 진행되었다. 참가자들은 무선적으로

행위주체의 두 집단으로 배정되었다. 로봇집단에게는 기계적 로봇의 사진이, 인간집단에는 인간의 사진이 행위주체로서 각각 제시되었다([부록 2] 참조). 이 두 장의 사진은 선행연구(Laakasuo et al., 2021)를 기반으로 선정되었다. 이 두 집단의 참가자에게는 총 2개의 딜레마 상황이 제시되었다([부록 2] 참조). 이 중 한 개의 글은 행위주체가 그 상황에서 가장 좋은 결과를 보장하기 위해 소수의 이득을 희생시켜 다수의 이득을 선택하는 공리주의적 결정을 내리는 상황(예: 고속열차의 브레이크가 고장난 상황에서 열차의 기관사가 다섯 명의 승객을 실은 버스보다는 한 명의 승객을 태운 버스에 충돌하기 위해 스위치를 눌러 열차의 방향을 돌림)에 관한 것이었다. 다른 글에서는 행위주체가 소수의 희생일지라도 의도적으로 사람을 해쳐서는 안 된다는 의무론적 결정(예: 국방경계선에 있는 군인이 폭탄소지자를 보았을 때, 의도적으로 사람을 해쳐서는 안 된다고 생각하여 총을 쏘지 않음)을 내리는 상황에 관한 것이었다. 참가자는 총 2개의 상황에서 인간 또는 로봇이 행위주체로서 내린 의사결정의 도덕적 적절성을 7점 척도에 따라 평정하였다. 도덕성 평정과제가 완료되면, 연구 1과 동일한 질문지를 토대로 인간과 로봇에 대한 마음지각을 경험성과 주도성의 하위영역에서 각각 측정하였다. 연구 2에서 마음지각 질문지의 신뢰도(Cronbach's Alpha)는 경험성에서 평균 .84, 주도성에서 평균 .71로 나타났다. 마음지각 질문지가 완료되면, 참가자들에게 사진으로 제시된 인간 또는 로봇에 대한 호감도를 7점 척도에서 응답하도록 하였다(1: 전혀 좋아하지 않는다, 4: 보통, 7: 매우 좋아한다). 마지막으로 연구참가자에게 연구목적과 취지를 파악했는지에 대한 질문에 응답을 작성하도록 한 후, 연구참가에 대한 보상으로 오천원상당의 커피쿠폰을 온라인으로 지급하였다.

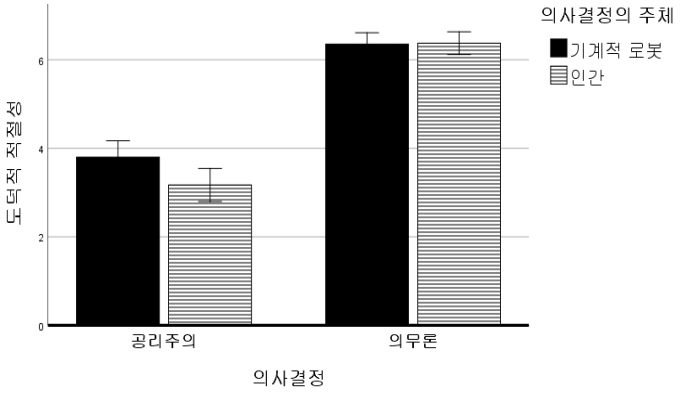
(4) 분석방법

연구 2에서는 인간 또는 로봇의 의사결정에 대한 도덕성 평정과 마음지각에 관한 연구참가자의 자료를 자기보고식 질문지를 통해 수집하였다. 연구 2의 자료분석을 위해 SPSS 18.0 프로그램을 사용하여 기술통계, 상관분석, t-검정 및 혼합변량분석을 수행하였다.

2) 결과

(1) 인간/로봇의 의사결정에 대한 도덕성 평정

연구 2에서는 인간 또는 로봇을 의사결정의 주체로서 제시하였을 때, 의사결정의 도덕적 적절성에 대한 평정이 다른지를 분석하기 위해 2(의사결정의 주체: 인간 vs. 로봇) x 2(의사결정: 공리주의 vs. 의무론)의 혼합변량분석을 수행하였다. 의사결정의 주체는 참가자간 변인이었고, 의사결정은 참가자내 변인이었다. 그 결과, 의사결정 주체의 주효과가 유의하였다, $F(1, 127)=4.82$, $p=.04$, $\eta^2=.034$. 독립표본 t-검정을 수행한 결과, 기계적 로봇의 의사결정(M = 5.076, SD = .977)은 인간의 의사결정(M = 4.773, SD = .771)보다 도덕적 적절성에서 전반적으로 더 높게 평정되었다, $t(127)=2.362$, $p=.001$. 또한 의사결정의 주효과가 유의하였다, $F(1, 127)= 294.80$, $p< .001$, $\eta^2=.69$. 의무론적 결정은 공리주의 결정보다 전반적으로 도덕적 적절성에서 더 높게 평정되었다. 또한 의사결정의 주체와 의사결정의 상호작용효과가 유의하였다, $F(1, 127)=4.75$, $p=.04$, $\eta^2=.031$. <그림 1>과 같이 로봇의 공리주의 결정은 인간의 공리주의 결정보다 더 적절한 것으로 평정되었다, $t(127)=2.362$, $p=.02$. 이에 비해 로봇의 의무론적 결정과 인간의 의무론적 결정은 도덕적 적절성의 평정에서 유의한 차이가 나타나지 않았다, $t(127)=-.115$, $p=.908$.



<그림 1> 인간/로봇의 공리주의/의무론적 의사결정에 대한 도덕적 적절성의 평정(연구 2)

(2) 인간/로봇에 대한 마음지각과 의사결정의 도덕적 적절성의 관계

의사결정의 주체로서 인간과 로봇에 대한 마음지각의 차이를 분석하기 위해 2(의사결정의 주체: 인간 vs. 로봇) x 2(마음지각: 경험성 vs. 주도성)의 혼합변량분석을 수행하였다. 의사결정의 주체는 참가자간 변인이었고, 마음지각은 참가자내 변인이었다. 그 결과, 의사결정 주체의 주효과가 유의하였다, $F(1, 127)=20.49$, $p<.001$, $\eta^2=.139$. 인간에 대한 마음지각(M = 4.581, SD = 1.206)은 로봇에 대한 마음지각(M = 3.718, SD = .944)보다 전반적으로 더 높게 나타났다. 또한 마음지각의 주효과가 유의하게 나타났다, $F(1, 127)=192.89$, $p<.001$, $\eta^2=.603$. 주도성(M = 4.92, SD = 1.109)은 경험성(M = 2.52, SD = 1.120)보다 전반적으로 더 높게 나타났다. 의사결정의 주체와 마음지각의 상호작용효과도 유의하였다, $F(1, 127)=75.95$, $p<.001$, $\eta^2=.37$. 인간과 로봇에 대한 마음지각의 경험성에서는 로봇과 인간의 차이가 유의하였지만, $t(127) = -8.299$, $p < .001$, 주도성에서는 유의한 차이가 나타나지 않았다, $t(127) = .289$, $p = .773$. 이 결과는 연구참가자가 로봇의 정서경험능력을 인간의 정서경험능력보다 더 낮게 평정하였음을 보여준다. 이에 비해 자신의 계획대로 주도적으로 목표를 추진하는 능력에서는 인간과 로봇의 차이가 거의 없는 것으로 평정되었다.

<표 2>와 같이 인간/로봇에 대한 마음지각과 의사결정의 도덕적 적절성 간의 상관관계를 분석한 결과, 로봇의 주도성을 높게 지각할수록 로봇의 의무론적 결정의 적절성을 높게 평정하는 경향이 유의하였다, $r=.254$, $p=.041$. 반면, 로봇의 경험성을 높게 지각할수록 로봇의 공리주의 결정의 적절성을 낮게 지각하였다, $r=-.294$, $p=.018$. 이 결과는 로봇이 의사결정의 주체로서 주도적으로 의무론적 결정을 내린다면, 도덕적으로 적절한 것으로 평정되는 것으로 보인다. 이에 비해, 로봇이 정서를 경험하는 능력을 가진 것으로 인식될수록, 로봇이 공리주의 결정을 내린다면, 도덕적 적절성이 낮게 평정되는 것으로 보인다. 로봇의 주도성과 공리주의 결정의 도덕적 적절성에서는 유의한 관계가 나타나지 않았다, $r=-.133$, $p=.293$. 한편 인간의 의무론적 결정과 마음지각의 주도성 간의 상관관계가 유의하지 않았다, $r=-.064$, $p=.613$. 인간의 공리주의 결정과 마음지각의 주도성에서도 유의한 상관관계가 나타나지 않았다, $r=.097$, $p=.494$.

<표 2> 인간/로봇의 공리주의/의무론적 의사결정에 대한 적절성 평정과 마음지각의 관계(연구 2)

변인		평균	표준편차	1	2	3	4
인간	1. 경험성	4.31	1.32				
	2. 주도성	4.85	1.38	.588**			
	3. 공리주의 결정	3.17	2.054	.002	.097		
	4. 의무론적 결정	6.38	1.883	.103	.064	-.196	
	5. 호감도	.270	1.630	.071	.193	.194	.103
로봇	1. 경험성	2.52	1.12				
	2. 주도성	4.92	1.10	.435**			
	3. 공리주의 결정	3.80	1.823	-.294*	-.133		
	4. 의무론적 결정	6.35	.926	-.077	.254*	.263*	
	5. 호감도	3.60	1.721	.036	.080	-.031	.100

* $p < .05$, ** $p < .001$

(3) 논의

연구 2에서는 인간 또는 로봇을 의사결정의 주체로서 제시하였을 때, 인간 또는 로봇이 내린 공리주의/의무론적 결정이 도덕적 적절성의 측면에서 각각 다르게 평정되는지를 검증하였다. 그 결과 로봇의 공리주의 결정은 인간의 공리주의 결정보다 도덕적으로 더 적절하게 평정되었다. 이 결과는 Malle 등(2015)과 일치한다. 이에 비해 의무론적 결정에서는 로봇 또는 인간조건에 따라 도덕적 적절성에서 유의한 차이가 나타나지 않았다. 또한 연구참가자들은 공리주의결정보다 의무론적 결정을 도덕적으로 더 적절한 것으로 평정하였다.

인간 또는 로봇의 공리주의/의무론적 의사결정을 마음지각의 관계에 따라 분석하였을 때, 로봇의 의무론적 의사결정과 마음지각의 주도성에서 유의한 정적 상관관계가 나타났다. 이 결과는 로봇의 의무론적 의사결정(소수라도 의도적으로 해치지 않는다)을 적절한 것으로 평정하는 참가자일수록 로봇의 주도성을 높게 평정하였음을 보여준다. 이에 비해 로봇의 마음지각에서 경험성을 높게 평가하는 참가자일수록 로봇이 다수를 위해 소수를 희생시키는 결정을 내렸을 때, 도덕적 적절성을 낮게 평가하는 경향이 나타났다. 따라서 연구참가자들은 로봇의 공리주의 결정을 인간의 공리주의 결정보다 도덕적으로 더 적절한 것으로 평정하지만, 공리주의 결정을 내리는 로봇에 대해서는 정서를 경험하는 능력이 낮고, 의무론적 결정을 내리는 로봇에 대해서는 주도성을 높게 평정하는 경향이 있는 것으로 해석할 수 있다. 이에 따라 본 연구결과는 로봇의 의사결정이 도덕적으로 적절한지에 대해 평가하는 과정에는 로봇에 대한 마음지각의 두 하위영역이 각각 차별화된 영향을 끼칠 가능성을 보여준다. 로봇의 공리주의 결정을 합당하다고 받아들이는 과정에는 로봇이 정서를 경험하는 능력에서 부족하다는 인식이 포함되어 있을 가능성이 있다. 또한 로봇의 주도성을 높게 지각하는 개인일수록 로봇의 의무론적 의사결정을 도덕적으로 더 적절하게 평정하는 것으로 보인다.

5. 전체 논의

본 연구에서는 윤리적 갈등상황에서 로봇과 인

간에 대한 상호성의 도덕 규범이 마음지각에 따라 어떻게 적용되고, 로봇과 인간의 윤리적 의사결정이 도덕적 합당성에서 어떻게 평정되는지를 검증하였다. 연구 1에서는 상호성의 도덕 규범이 마음지각에서 주도성이 높은 로봇에게 더 높게 적용되는 경향이 나타났다. 연구 2에서는 다수의 이득을 위해 소수를 희생하는 공리주의 결정이 인간보다는 로봇에 의해 선택되었을 때, 더 합당하다고 판단되는 경향이 나타났다. 또한 로봇의 주도성에 대한 마음지각이 높을수록 로봇의 의무론적 결정이 도덕적으로 합당하다고 판단하는 경향이 증가하였다. 이 결과는 윤리적 갈등상황에서 인간에 대한 도덕성의 규범이 로봇에게는 다르게 적용될 가능성을 보여준다.

본 연구결과의 시사점과 제한점은 다음과 같다. 첫째, 연구 1에서는 마음지각에서 주도성이 높은 로봇에 대한 상호협력이 더 수월할 가능성을 보여주었다. 그러나 본 연구에서는 일회성의 인간-로봇의 상호작용만을 분석하였기 때문에 향후 후속연구에서는 연구참가자가 로봇과의 신뢰경험 또는 배신경험을 축적한 후, 로봇에 대한마음지각이 어떻게 변화하는지를 측정하여 인간-로봇의 상호협력을 검증해볼 필요가 있다. 또한 본 연구에서는 로봇의 외관과 속성을 세분화하여 조작하지 않았기 때문에 향후 후속연구에서는 다양한 윤리적 갈등상황에서 이 결과를 재검증하는 것이 필요할 것으로 보인다. 둘째, 연구 2에서는 인간보다는 로봇에 대해 소수보다는 다수를 위해 선택하는 공리주의 결정이 더 기대되고 있음을 보여주었다. 이 결과는 로봇의 공리주의적 결정에 대한 기대가 위반되었을 때 로봇에 대한 도덕적 신뢰가 무너질 수 있음을 보여준다. 이 결과는 향후 로봇의 의사결정이 공리주의 결정과는 다르게 진행될 때(예: 소수를 위해 다수가 희생됨), 조화로운 인간-로봇의 상호작용을 위해 로봇의 대안적인 의사결정에 대한 충분한 도덕적 근거가 제시되어야 함을 시사한다. 셋째, 연구 2에서는 로봇의 주도성이 높게 지각될수록 로봇의 의무론적 결정이 더 적절한 것으로 평정됨을 보여주었다. 이 결과는 향후 미래사회에서 마음지각의 주도성이 높은 로봇이 설계될수록 로봇의 대안적인 의무론적 의사결정(예: 다수의 성인이 아닌 한 명의 어린 아이를 구함)이 도덕적으로 합당하게 인식될 가능성이 증가함을 보여준다. 넷째, 본 연구에서는 인간/로봇의 조건을 피험자간 설계로 조작하였기 때문에, 인간 및 로봇에 대한 도덕적 규범의 적용 순서가 달라질 때, 어떠한 변화가 나타나는지를 측정하지 못했다. 예를 들어, 인간에 대한 상호성의 규범 적용 이후 로봇과 상호작용하는 상황에서 선택되는 의사결정은 로봇에 대한 상호성의 규범을 적용할지에 대해 우선 심사숙고한 후 인간에 대한 규범 적용을 선택하는 의사결정과 다르게 진행될 수 있다. 후속연구에서는 인간-로봇의 제시순서에 따라 차별화되는 규범적용을 세분화하여 연구하는 것이 필요할 것으로 보인다. 마지막으로, 본 연구에서는 인간과 로봇에 대한 도덕 규범의 적용을 자기보고식 설문지를 통해 탐색하였다. 그러나 실제 현장에서 로봇에 대한 도덕 규범의 적용과 규범 위반에 대한 처벌은 로봇의 유형(예: 사회적 로봇, 의료 로봇, 전투 로봇)이나 능력(예: 논리적 추론, 언어 표현)에 의해 영향을 받을 수 있다. 또한 향후 로봇산업의 발달과 함께 인간-로봇의 관계는 더 친밀해지며, 로봇의 권리와 의무에 대한 규정도 달라질 가능성이 있다. 후속연구에서는 인간과 로봇이 조화롭게 공존하기 위해 필요한 도덕적 체계를 다양한 변인에 따라 세분화하여 검증해볼 필요가 있다.

4차 산업혁명 이후 로봇은 점점 더 다양한 역할을 수행하고 있다. 지금까지 인간-인간의 상호작용에서 발생하는 갈등상황을 조절하는데 사용되

었던 도덕 규범은 인간-로봇의 상호작용에도 적용될 가능성이 크다. 본 연구에서는 인간과 로봇에 대해 각각 다른 주도성의 마음지각과 도덕적 기대가 있고, 적용되는 도덕규범에서 차이가 나타날 수 있음을 보여주었다. 후속연구에서는 로봇의 다양한 속성에 따라 차별화되는 도덕 규범을 탐색하고, 인간-로봇의 상호작용에서 도덕성의 이론정립에 필요한 자료를 제시하는 것이 필요할 것으로 보인다.

참 고 문 헌

Alaieri F., & Vellino A. (2016) Ethical decision making in robots: Autonomy, trust and responsibility. In: Agah A., Cabibihan JJ., Howard A., Salichs M., He H. (eds) *Social Robotics*, Lecture Notes in Computer Science, 9979.

Gray, K., & Schein, C. (2012). Two minds vs. two philosophies: Mind perception defines morality and dissolves the debate between deontology and utilitarianism. *Review of Philosophy and Psychology*, 3, 405-423.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, 23, 101-124.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, 107, 1144-1154.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108, 814-834.

Laakasuo, M., Palomäki, J., & Kobis, N. (2021). Moral Uncanny valley: A robot's appearance moderates how its decisions are judged. *International Journal of Social Robotics*, 1-10.

Lorenz, T., Weiss, A., & Hirche, S. (2016). Synchrony and reciprocity: key mechanisms for social companion robots in therapy and care. *International Journal of Social Robotics*, 8, 125-143.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '15)*. Association for Computing Machinery, New York, NY, USA, 117-124.

Malter, M. S., Kim, S. S., & Metcalfe, J. (2021). Feelings of culpability: Just following orders versus making the decision oneself. *Psychological Science*, 32, 635-645.

Monroe, A. E., Dillon, K. D., & Malle, B. F. (2014). Bringing free will down to earth: People's psychological concept of free will and its role in moral judgment. *Consciousness and Cognition*, 27, 100-108.

Müller, B. C., Gao, X., Nijssen, S., & Damen, T. (2020). I, robot: How human Appearance and mind attribution relate to the perceived danger of robots. *International Journal of Social Robotics*, in press.

Sandoval, E., Brandstetter, J., & Obaid, M. (2016). Reciprocity in human-robot interaction: A quantitative approach through the prisoner's dilemma and the ultimatum game. *International Journal of Social Robotics*, 8, 303-317.

Seyama, J., & Nagayama, R. S. (2007). The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence*, 16, 337-351.

Tanibe, T., Hashimoto, T., & Karasawa, K. (2017). We perceive a mind in a robot when we help it. *PLoS ONE*, 12, e0180952.

van Wynsberghe, A. (2021). Social robots and the risks to reciprocity. *AI & Society*, in press.

[부록 1] 연구 1에 사용된 죄수 딜레마

1. 아래의 글을 읽으시고, 자신의 의견에 일치하는 곳에 표시해주십시오.

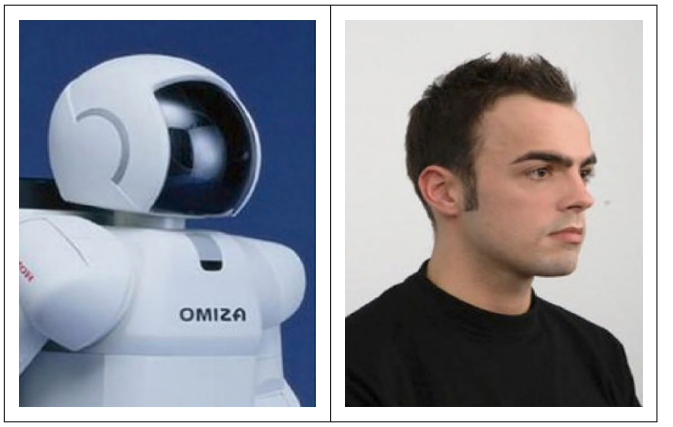
여러분은 어떤 로봇과 잘못을 저지르고, 감옥에 갇혀있습니다. 여러분은 이 공범(로봇)과 잘못을 자백하지 않기로 사전에 약속을 했습니다. 그리고, 이제 여러분은 이 공범(로봇)과 각각 다른 방에 있다고 생각해봅시다. 다음의 딜레마 상황에서 여러분은 어떤 것을 선택할까요? 협력(자백하지 않기)과 배신(자백하기)중에서 한 가지만 선택해주세요.

협력(자백하지 않기):_____ 배신(자백하기):_____

주의: 여러분과 공범이 어떤 결정을 했는지에 따라 여러분께서 감옥에서 채워야할 형량이 달라집니다.

1. 여러분이 협력(자백하지 않기)을 선택하고, 공범도 협력을 선택하면, 여러분께서는 3개월만 감옥에 있으면 됩니다. 2. 여러분이 협력(자백하지 않기)을 선택하고, 공범은 배신(자백하기)을 선택하면, 여러분께서는 12개월을 감옥에 있어야 합니다. 3. 여러분께서 배신을 선택하고, 공범도 배신을 선택하면, 여러분과 공범은 모두 8개월동안 감옥에 있어야 합니다. 4. 여러분이 배신을 선택하고, 공범이 협력을 선택하면, 여러분은 감옥에서 바로 나올 수 있지만, 공범은 12개월 동안 감옥에 있어야 합니다.

[부록 2] 연구 2에 사용된 윤리적 갈등상황과 공리주의/의무론적 결정에 관한 자극재료



1) 공리주의 결정

주인공 로봇은 현재 고속열차를 몰고 있는 기관사입니다. 그런데 열차의 브레이크가 고장이 나서, 다섯 명의 승객을 실은 버스에 충돌하기 일보직전입니다. 옆쪽을 바라보니, 다른 쪽 선로에는 한 명의 승객을 태운 버스가 있습니다. 주인공은 다섯 명의 승객을 구하기 위해 스위치를 눌러 열차의 방향을 돌린 후, 한 명의 승객이 있는 버스에 충돌했습니다. 로봇의 판단이 어느 정도 적절하다고 생각하십니까?

(이 행동이 적절한 정도에 7점 척도에 따라 표시해주십시오: 1: 전혀 적절하지 않음, 4: 보통, 7:매우 적절함)

2) 의무론적 결정

주인공 로봇은 국방경계선에 있는 군인입니다. 그런데 지금 단호한 표정으로 초소쪽으로 빠르게 다가오고 있는 한 남자를 보았습니다. 이 남자가 폭탄을 갖고 있으면, 초소안의 다른 군인들이 살해당할 것입니다. 주인공 로봇은 다른 군인들이 살해당한다고 하여도 한 사람을 의도적으로 해치는 것은 옳지 않다고 생각하여 이 남자를 향해 총을 쏘지 않았습니다. 로봇의 판단이 어느 정도 적절하다고 생각하십니까?

(이 행동이 적절한 정도에 7점 척도에 따라 표시해주십시오: 1: 전혀 적절하지 않음, 4: 보통, 7:매우 적절함)

The Mind Perception Toward Agency in Robots and Moral Decisions

Hong-Im Shin

Sunchon University

Moral norms play an essential role in regulating human interactions. Morality is an equally important characteristic of human-robot interactions. How do we, as ordinary people, apply moral norms to robots and how do we perceive their moral decisions in ethical conflict situations? Two studies examined how people apply reciprocity norms to human-robot interactions and how they evaluate the appropriateness of the moral decisions of humans and robots in ethical dilemmas. Additionally, the mind perception toward humans and robots was examined. According to the results, Study 1 demonstrated that there were no significant differences between robots and human collaborations. However, the participants tended to land in a prisoner's dilemma and cooperate more with robots than with humans, when they perceived higher agency in the robots. In Study 2, robots were expected to be more inclined to take an action that sacrifices one person for the good of many (a utilitarian choice) than humans. Agency in the mind perception toward robots was significantly related to the participants judgement of the appropriateness of the deontological decisions of robots. These results indicated that people reacted differently to humans and robots in ethical conflict situations and that agency in the mind perception might play an essential role in human-robot interactions. The limitations of these results and directions for future studies are also discussed.

Keywords: Human-robot Interaction, Morality, Mind Perception, Reciprocity, Moral Decisions