한국심리학회지 : 일반
*Korean Journal of Psychology: General*
2014, Vol. 33, No. 2, 467-489

# 이타적 강화학습 과제를 이용한 이타성의 측정*

설 선 혜　　이 민 우　　김 학 진†
고려대학교 심리학과

이타 행동은 타인의 안녕에 가치를 부여하는 과정을 필요로 한다. 본 연구에서는 사람들이 타인의 안녕에 가치를 부여하는 정도를 측정하는 이타적 강화학습과제를 개발하고 공감 성향과 내집단 편향과의 관련성을 살펴보았다. 강화학습과제에서 참가자들은 주어진 두 개의 옵션 중 하나를 선택하여 보상을 얻을 수 있는데, 각 옵션의 보상 확률은 30%와 70%로 다르고, 시행을 반복하면서 선택과 결과 간의 관계를 학습해야 한다. 여기서 보상에 가치를 부여하는 정도는 보상 확률이 높은 옵션을 선택하는 빈도로 측정된다. 실험 1에서는 자신에게 도움이 되는 자기-보상, 자신과 타인에게 모두 도움이 되는 공동-보상, 타인에게만 도움이 되는 타인-보상 조건에서 참가자들의 학습 정도를 비교하고 공감 성향에 따라서 어떻게 다른 선택을 하는지 살펴보았다. 그 결과, 정서적 공감 성향이 높은 참가자들이 공동-보상 조건과 타인-보상 조건에서 높은 수행을 보여서 타인의 안녕에 더 많은 가치를 부여하는 것으로 나타났다. 실험 2에서는 최소집단 패러다임을 이용하여 이타적 강화학습과제에서 내집단 편향이 관찰되는지 확인하였다. 참가자들은 내집단원에게 더 호감을 보였으며, 자신에게 도움이 되는 자기-보상 조건, 내집단원에게 도움이 되는 내집단원-보상 조건, 외집단원에게 도움이 되는 외집단원-보상 조건에서의 선택을 비교하였을 때, 자기-보상 조건과 내집단원-보상 조건에는 차이가 없고 외집단원-보상 조건에서만 보상 확률이 높은 그림을 선택하는 빈도가 낮았다. 즉, 외집단원보다는 내집단원의 안녕에 더 많은 가치를 부여했다. 실험 1과 실험 2의 결과는 이타적 강화학습과제가 이타성을 측정하는 타당한 방법으로 사용될 수 있는 가능성을 시사한다.

주요어 : 강화학습, 이타성, 공감, 최소집단, 내집단 편향

---

* 본 연구는 고려대학교 연구지원사업과 2013년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행되었음(NRF-2013S1A5A2A03045216). 훌륭한 조언을 주신 익명의 심사위원 선생님들과 이타적 강화학습과제의 설계 단계에서 도움 주신 University of Zurich의 Ernst Fehr 교수님 Philippe Tobler 교수님, 그리고 자료 수집을 도와주신 고려대학교 정대현, 유민희 학생에게 깊이 감사드립니다.

† 교신저자: 김학진, 고려대학교 심리학과, 서울특별시 성북구 안암로 145
Tel: 02-3290-2864, E-mail: hackjinkim@korea.ac.kr

이타 행동(altruistic behavior)은 인간 사회에서 서로 유전적으로 관련되지 않은 타인들이 더불어 살아가는 것을 가능하게 함으로써 공동체를 유지시켜주는 중요한 특징으로(Fehr & Fischbacher, 2003) 인간 행동을 이해하고자 하는 다양한 학문 분야에서 주요 연구 주제로 다루어져 왔다. 이타 행동은 동기적 측면과 결과적 측면, 또는 심리학적 접근과 경제학적, 진화적 접근 중 어디에 초점을 두는가에 따라서 조금씩 다르게 정의되지만, 일반적으로 받아들일 수 있는 정의는 "외부로부터 주어지는 보상을 바라지 않고 타인에게 도움을 주는 행동"(Macaulay & Berkowitz, 1970) 또는 '자신에게 이익이 되지 않거나 손해가 됨에도 불구하고 타인의 안녕을 증진하려는 행동'이라고 할 수 있을 것이다. 여기서 타인이란 한 사람이 될 수도 있고 불특정 다수가 될 수도 있으며, 어려움에 처한 개인을 직접적으로 도와주는 것에서부터 자선단체에 기부를 하거나, 협동이나 공평성의 추구와 같이 공동체에 도움이 되는 행동까지 광범위한 행동을 포함할 수 있다. 대체로 행동경제학이나 진화생물학에서는 자신의 생물학적 또는 경제적 비용을 감수하고 타인의 이익을 증진시키는 행동이 개인이나 집단의 효용에 어떻게 기여하는지를 중심으로 협동과 공평성 추구와 같은 친사회적 행동을 설명하는 반면(Fehr & Fischbacher, 2003; Fehr & Schmidt, 1999; Sober & Wilson, 1999), 심리학에서는 주로 이타 행동의 동기에 초점을 두고 실험실이나 일상생활에서 관찰되는 직접적인 도움행동에 관한 연구들이 이루어져왔다(개관을 위하여 Batson, 2011 참고).

이타 행동을 구체적으로 어떻게 정의하는가에 따라서 그 정도에 차이가 있을 수 있지만, 대부분의 이타 행동은 자신을 위한 선택(self-regarding choice)과 타인을 위한 선택(other-regarding choice) 간의 갈등을 수반한다. 예를 들어, 자선 단체에 기부를 할 때 나의 금전적 손해를 감수하고 자선단체의 금전적 이익을 증진시켜야 하며, 길을 가다가 다친 사람을 보게 되면 나의 시간과 노력을 들여서 도움을 제공해야 한다. 타인을 위한 선택에는 자신의 이익을 잠재적으로 희생해야 하는 비용이 들기 때문에 타인의 안녕을 가치 있게 여기지 않으면 이타 행동은 불가능하다.

## 타인의 안녕에 대한 가치부여와 이타 행동

이타 행동에 있어서 타인의 안녕에 가치를 부여하는 과정이 매우 중요함에도 불구하고, 이와 관련된 연구는 아직까지 많지 않다. 가치 부여의 과정과 이타 행동 간의 관련성에 관한 증거들은 이타 행동의 결과를 개인의 의사결정 효용 함수(utility function)에 포함시키는 행동경제학과 신경경제학(neuroeconomics) 연구들에서 일부 찾아볼 수 있다. Becker(1974)는 의사결정의 효용에는 자신에게 직접적으로 돌아오는 이득뿐만 아니라 '이타적 행동으로 인한 타인의 효용 증진'이 포함될 수 있다고 제안하였으며, Andreoni(1990)는 이타 행동의 결과 행위자가 경험하게 되는 이러한 이득을 '따뜻한 빛(warm glow)'이라고 이름 붙였다. 협동과 공평성 추구를 설명하는 이론들에서도 타인과의 상호작용이 개입되는 사회적 의사결정(social decision making)에서는 타인의 이익 또한 의사결정의 효용에 반영되며 그 정도에는 개인차가 있다는 모형들이 제안되었다(Bolton & Ockenfels, 2000; Fehr & Schmidt, 1999; Levine, 1998; Rabin, 1993, 개관을 위해서는 Sobel, 2005). 뇌영상 기법을 이용한 최근의 신경경제

학 연구들에서는 사람들이 자신에게 주어지는 보상 관련 정보를 처리할 때 활성화되는 것으로 알려져 있는 복측 선조체(ventral striatum)가 자선단체에 기부금이 전달되는 것을 관찰할 때에도 활성화 되며(Harbaugh, Mayr, & Burghart, 2007), 기부 의사결정의 가치를 계산할 때, 자신의 이익과 관련된 선택의 가치를 계산하는 영역으로 알려져 있는 복내측전전두피질(VMPFC)이 관여한다는 결과(Hare, Camerer, Knoepfle, O'Doherty, & Rangel, 2010)를 보고하였다. 이러한 결과는 이타적 선택과 이기적 선택이 동일한 보상 및 가치 계산 기전을 공유함을 보여준다. 심리학에서도 Batson과 동료 연구자들(Batson, Turk, Shaw, Klein, 1995; Batson, Eklund, Chermok, Hoyt, & Ortiz, 2007)이 사람들은 '도와줄 가치가 있는 타인'에 대하여 공감을 더 잘 경험하고 더 많이 도와주는 경향이 있음을 보이고 타인의 안녕에 대한 가치 부여가 이타적 동기를 유발하는 데 필요한 선행 사건이라고 제안한 바 있다.

선행연구들이 타인의 안녕에 대한 가치 부여의 과정을 이해하는 데 유용한 정보를 제공하고 있기는 하지만, 실제로 사람들이 타인의 안녕이 증진되는 것을 보상으로 경험하는 정도를 직접 측정한 연구는 아직까지 없다. 행동경제학에서는 사회적 의사결정 과정에서 나타나는 사람들의 선호(social preference)를 측정하는 간단하면서도 유용한 도구들을 제공하고 있으나(예를 들어, Charness & Rabin, 2002) 사람들의 선호를 통해 효용함수를 간접적으로 예측하고 있으며, 신경경제학 연구들에서도 역인과적 추론(reverse inference)에 의한 해석에 머무르고 있고, Batson 등의 연구에서도 도움을 받는 대상의 특성에 초점을 두었기 때문에 이타 행동을 하는 행위자의 내적 과정을 알기는 어렵다. 또한, 이타적 선택에 가치를 부여하는 정도의 개인차가 실제 이타 행동을 예측하는 것으로 알려져 있는 다른 변수들과 어떻게 관련되어있는지를 살펴본 연구도 아직 없다.

본 연구에서는 타인의 안녕에 가치를 부여하는 정도의 개인차를 측정할 수 있는 과제(이타적 강화학습과제)를 개발하고, 이타 행동과 밀접하게 관련되어있는 것으로 알려져 있는 공감과 내집단 편향의 효과가 관찰되는지 확인함으로써 이타성을 측정하는 방법으로서의 타당성을 검토하고자 하였다. 공감과 내집단 편향은 각각 기존 연구들에서 이타 행동을 예측하는 대표적인 개인 내적 요인과 상황 요인으로 알려져 있는 동시에, 이타 행동의 진화적 기원 관한 이론들에서도 중요하게 다루어지고 있다(Choi & Bowles, 2007; De Waal, 2008; Iacoboni, 2009; 개관을 위해서는 Sober & Wilson, 1999).

## 공감과 이타 행동

이타 행동에 영향을 미치는 대표적인 개인 내적 요인은 공감이다. Batson(2011)은 그의 공감-이타주의 가설(empathy-altruism hypothesis)에서 사람들에게는 자신의 내적 상태를 개선하고자 하는 자기중심적인 동기(egoistic motivation)와 타인의 안녕을 걱정하는 이타적 동기(altruistic motivation)가 모두 있으며, 타인의 고통에 대한 공감적 염려(empathic concern)를 경험할 때 이타적 동기에 의해 이타 행동이 유발된다고 제안하였다. 진정한 이타적 동기의 존재 여부에 대해서는 여전히 논쟁의 여지가 있지만(Batson, 2011; Cialdini, 1991; Maner, Luce, Neuberg, Cialdini, Brown, & Sagarin, 2002; Harbaugh, 1998), 여러 연구들에서 일관되게 밝

혀지고 있는 사실은 공감이 타인을 돕는 행동을 촉진한다는 것이다.

예를 들어, Batson과 동료연구자들은 쉬운-탈출 패러다임(easy-escape paradigm)을 사용한 일련의 실험들을 수행하였다. 쉬운-탈출 패러다임에서는 보통 두 사람의 참가자를 동시에 실험실에 초대한 뒤 한 사람 (실제로는 동조자)이 가벼운 전기충격을 받아야 하는 상황을 연출하고, 동조자가 갑자기 극심한 스트레스를 호소할 때에, 실제 참가자가 동조자 대신 전기충격을 받겠다고 하는지를 관찰한다. 이 때, 전기충격을 대신 받지 않더라도 쉽게 그 상황을 벗어날 수 있는(쉬운-탈출) 조건을 만들어주고, 단순히 고통스러운 상황을 회피하려고 하는지, 아니면 회피하지 않고 도움을 제공하려고 하는지를 살펴봄으로써 타인지향적인 공감적 염려와 자기중심적인 개인적 괴로움(personal distress)을 구별한다. Batson과 동료 연구자들은 이 패러다임을 이용한 여러 연구들에서 자기 보고된 공감의 개인차나 상대방이 어떻게 느낄지에 대해서 생각해보도록 하는 관점취하기(perspective taking) 조작을 통한 공감 수준의 향상이 타인지향적인 공감적 염려에서 비롯된 도움 행동을 촉진할 수 있음을 밝혔다(Batson, Dyck, Brandt, Batson, Powell, McMaster, & Griffitt, 1988; Batson, 2011). 최근 뇌영상 연구들에서는 타인의 고통에 대한 공감 관련 영역의 활성화 정도가 이후의 도움 행동(Hein, Silani, Preuschoff, Batson, & Singer, 2010; Masten, Morelli, & Eisenberger, 2011)이나 일상생활에서의 이타 행동(Ma, Wang, & Han, 2011)을 예측한다는 결과들이 보고되고 있다. 또한, 자기보고식 질문지를 이용하여 성향적 공감(dispositional empathy)을 측정한 연구들에서는 공감 성향이 높은 사람들이 일상생활에서 친절을 베풀거나 기부나 자원봉사 같은 친사회적 행동을 더 많이 하는 것으로 알려져 있으며(Davis, 1994; Eisenberg, Guthrie, Cumberland, Murphy, Shepard, Zhou, & Carlo, 2002; Eisenberg & Miller, 1987), 국내 연구에서도 정서적 공감의 개인차가 공감대상을 도우려는 의도를 예측함을 보인 바 있다(김용훈, 류리나, 한성열, 2012).

### 내집단편향(ingroup bias)과 자기집단중심적 이타주의(parochial altruism)

이타 행동에는 공감뿐만 아니라 집단 멤버십과 같은 사회적 요인도 영향을 미친다. 사람들은 외집단원보다 내집단원에게 더 이타적으로 행동하는(Fehr, Bernhard, & Rockenbach, 2008; Halevy, Bornstein, & Sagiv, 2008; Levine, Prosser, Evans, & Reicher, 2005) 자기집단중심적 이타주의(parochial altruism)를 나타내는 것으로 알려져 있다(Bernhard, Fischbacher, & Fehr, 2006; Choi & Bowles, 2007; García & van den Bergh, 2011). 자기집단중심적 이타주의의 심리학적 기제로는 내집단 편향을 들 수 있는데, 사람들은 실제 집단뿐만 아니라, 최소 집단(minimal group: Tajfel, 1970)으로 소속을 구분하는 것만으로도 내집단원에게 더 호감을 가지고 우호적인 태도를 형성하는 내집단 선호(ingroup favoritism)를 보인다(김미희, 김기범, 차영란, 2005; Brewer, 1979; Hewstone, Rubin, & Willis, 2002). 내집단원에 대한 우호적 태도는 공감 수준에도 영향을 미친다(Hein 등, 2010; Strümer, Snyder, Kropp, & Siem, 2006). 예를 들어, Hein과 동료연구자들(2010)은 선호하는 축구팀에 따라서 참가자들을 두 집단으로 구분한 뒤, 내집단원과 외집단원(실제로는 실험 동조자)이

손등에 전기충격을 받는 과제를 하는 것을 지켜보는 동안의 뇌활동을 기능적 자기공명영상(fMRI)으로 관찰하였다. 그 결과, 공감과 관련된 영역으로 알려져 있는 앞쪽 뇌섬엽(anterior insula)이 외집단원보다는 내집단원이 고통 받는 것을 볼 때 더 강하게 활성화되었고, 이 영역의 활성화 정도와 내집단원에 대한 자기보고 된 공감의 정도가 강할수록 전기충격을 대신 받아주는 결정(도움 제공)을 더 많이 하는 것으로 나타났다. 이러한 결과는 공감과 도움행동 간의 관련성이 도움의 대상이 외집단원일 때보다 내집단원일 때 더 강하다는 Strümer 등(2006)의 결과와도 일치한다.

### 강화학습 패러다임을 이용한 이타성의 측정

본 연구에서는 최근 의사결정 연구들에서 자신의 이익과 관련된 선택의 가치를 학습하는 과정을 알아보는 방법으로 흔히 쓰이는 강화학습(reinforcement learning) 패러다임을 사용하여 타인의 안녕에 가치를 부여하는 정도, 즉, 타인의 안녕이 증진되는 것을 보상으로 경험하는 정도의 개인차를 측정하는 과제를 개발하고, 공감 성향과 내집단 편향과의 관련성을 검토하고자 한다.

행동주의 심리학에서 잘 알려져 있듯이, 강화학습이란 처음에는 의미 없이 일어난 행동이 행위자가 선호하는 결과와 반복적으로 연합되면 그 행동의 빈도가 증가되는 현상을 말한다. 행위자가 어떤 반응을 선택하였을 때 그 선택의 결과가 이 행위자가 가치 있다고 여기는 것이라면 다음에도 동일한 반응을 선택 할 확률이 증가하게 되며, 이렇게 행위자의 특정 행동을 강화하는 자극이 바로 보상이 된다. 다시 말해, 행위자가 어떤 결과에 가치를 부여하는 정도를 측정하는 가장 직접적이고 간결한 방법은 그 결과가 보상으로 작용하는 정도, 즉, 특정 행동을 강화하는 정도를 살펴보는 것이라고 할 수 있다.

최근 의사결정 연구들에서는 이러한 원리를 적용하여 참가자들의 선호가 형성되는 과정을 살펴보기 위하여 둘 중 하나의 그림을 선택하고 선택과 결과 간의 관계를 학습하는 과제를 사용하는데(예를 들어, Kim, Shimojo, & O'Doherty, 2006), 일반적으로 다음과 같이 구성된다. 참가자들은 두 개의 그림 중 하나를 선택하여 특정한 보상(돈이나 점수, 주스 등)을 획득할 수 있다. 이 때 두 그림 중 하나는 보상을 얻을 수 있는 확률이 낮고(예를 들어, 30%), 다른 그림은 보상을 얻을 수 있는 확률이 높다(예를 들어, 70%). 참가자들은 확률 정보를 알지 못하기 때문에, 여러 시행을 반복하면서 선택-결과 간의 관계를 학습해나가야 한다. 만약 선택의 결과가 보상으로 작용한다면 보상 확률이 높은 그림을 선택하는 빈도가 증가할 것이고, 보상으로 작용하지 않는다면 둘 중 특별히 선호하는 그림 없이 선택 빈도가 50% 수준에 머물 것이다.

본 연구에서는 선택의 결과 얻을 수 있는 보상의 종류를 달리하여 자신에게 도움이 되는 이기적 결과와 타인에게 도움이 되는 이타적 결과를 모두 포함시킨 '이타적 강화학습과제'를 고안하였다. 만약 사람들이 타인의 안녕에 가치를 부여한다면 이타적 결과를 보상으로 경험할 것이고, 이기적 결과가 주어질 때와 마찬가지로 학습이 일어날 것으로 예상할 수 있다. 이 과제에서 학습의 지표는 보상이 주어지는 확률이 높은 그림을 선택하는 빈도가 될 것이고, 이타적 결과가 주어지는 조건에서 보상 확률이 높은 그림을 선택하는 빈도

가 높을수록 타인의 안녕에 더 많은 가치를 부여하는 것으로 볼 수 있을 것이다. 따라서 이 과제를 통해서 타인의 안녕에 가치를 부여하는 정도의 개인차를 측정하여 이타성의 지표로 활용할 수 있을 것으로 기대했다.

실험 1에서는 이타적 강화학습과제를 이용한 탐색적 연구를 수행하고, 이타성과 밀접하게 관련되어있는 것으로 알려져 있는 공감 성향이 이 과제에서의 행동을 예측하는지를 살펴보았다. 실험 2에서는 내집단 편향을 일으키는 조작을 사용하여 이타적 강화학습과제에서 자기집단중심적 이타주의가 관찰되는지 확인하였다.

## 실험 1: 공감 성향이 이타적 강화학습과제의 수행에 미치는 영향

실험 1에서는 이타적 강화학습과제를 이용해서 사람들이 타인을 돕는 것을 보상으로 경험하는 정도를 측정하고 여기서 관찰되는 개인차가 공감 성향과 어떻게 관련되어있는지 알아보고자 하였다. 공감 성향은 Davis(1983)의 대인간반응척도(Interpersonal Reactivity Index: IRI)를 사용하여 측정하였다. 공감이 이타 행동을 예측한다는 기존 연구들에 근거하여, 공감 성향이 높을수록 타인의 안녕에 가치를 더 많이 부여할 것으로 예상하였다.

### 방 법

#### 연구 참가자

학내 포털 사이트의 실험 참가자 모집 광고를 보고 자원한 38명의 대학생들이 한 번에 두 명씩 실험에 참가하였다. 두 사람 중 한 사람은 본 연구와 무관한 다른 과제를 수행하였고, 한 사람만 본 연구에 참가하여 연구 대상은 총 19명이었다. 이 중 과제의 규칙을 잘못 이해하여 무작위로 응답한 3명의 참가자가 제외되어 16명의 참가자 (모두 여성, 평균 나이: 22.4세)가 최종적으로 분석에 포함되었다. 모든 참가자들은 실험이 시작되기 전에 실험 절차 전반에 대한 안내를 받고 참가 동의서에 서명하였다. 본 연구의 모든 절차는 고려대학교 기관생명윤리심의위원회의 승인을 받았다.

#### 절차

두 명의 참가자가 실험실에 도착하면, 실험의 전반적 절차에 대해 설명하는 안내문을 제공했는데, 여기에는 연구자들이 인지 자원의 소모가 스트레스 반응에 미치는 영향을 알아보고자 한다는 커버스토리가 포함되었다. 참가자들은 실험이 인지 과제와 스트레스 과제의 두 단계로 구성되어있다고 안내 받았다. 두 사람은 서로 다른 종류의 인지 과제를 수행한 뒤, 인체에는 무해하지만 일시적으로 스트레스를 유발하는 데 효과적이라고 알려져 있는 불쾌한 소음을 5분 간 들으면서 스트레스 정도를 평가하는 스트레스 과제를 수행하게 될 것이라는 내용이 포함되었다. 커버스토리의 신뢰성을 높이기 위해서, 안내문을 읽은 뒤 스트레스 과제에 사용하게 될 소음의 강도를 결정하는 절차를 거치도록 했다. 여기서 참가자들은 각자 헤드폰을 착용하고 단일 주파수로 이루어진 소리를 들었는데, 500Hz부터 조금씩 주파수를 높여가며 1점에서 10점 사이에서 소음의 불쾌함 정도를 평정하도록 해서

8점에 해당하는 소리를 찾은 뒤, 그 소리를 실제 스트레스 과제에서 들려줄 것이라고 이야기 하였다. 소음 역치 측정 절차가 끝난 뒤 참가자들은 추첨으로 각기 다른 종류의 인지 과제에 할당되었고, 추첨이 끝난 뒤 각자 서로 분리되어있는 실험 공간으로 이동하여 과제를 수행하였다. 한 사람은 본 연구와 무관한 다른 과제에 할당되었고, 한 사람만 본 연구에서 살펴보고자 하는 이타적 강화학습 과제를 수행하였다. 참가자들은 이타적 강화학습 과제 이후 5분 간 휴식을 취한 뒤 공감 능력을 측정하기 위한 대인간반응척도(Davis, 1983)와 과제 수행 동기를 측정하기 위한 인지욕구 질문지(Need for cognition, Cacioppo, Petty, & Kao, 1984)를 작성했다.

실험 1에서 이타성을 예측하는 개인차 변수로 측정된 대인간반응척도는 일곱 개 문항으로 측정되는 네 개의 하위척도로 구성된다. 자발적으로 타인의 관점에서 생각하는 경향성을 측정하는 관점취하기(Perspective taking: PT)와 책이나 영화 등의 가상의 등장인물의 감정이나 행동을 쉽게 상상하는 경향성을 측정하는 상상력(Fantasy: FS)은 공감의 인지적 측면을 측정하며, 타인의 불행에 대하여 동정심과 염려와 같은 타인지향적 정서를 경험하는 경향성을 측정하는 공감적 염려(Empathic concern: EC)와 불안이나 불편감과 같은 자기지향적 정서를 경험하는 경향성을 측정하는 개인적 괴로움(Personal Distress: PD)은 공감의 정서적 측면을 측정한다.

### 이타적 강화학습 과제

이타적 강화학습과제는 의사결정 연구에서 흔히 사용되는 일반적인 강화학습과제를 변형한 과제로, 타인의 안녕에 가치를 부여하는 정도를 측정하여 개인의 이타성을 알아보기 위하여 본 연구진이 고안하였다. 이 과제에는 총 세 쌍의 자극이 제시되는데, 각 자극 쌍은 자신에게만 점수가 주어지는 자기-보상(자신만 2점), 자신과 상대방에게 보상이 주어지는 공동-보상(각각 1점), 상대방에게만 보상이 주어지는 타인-보상(상대방만 2점)과 관련되어 있었다. 즉, 자기-보상 조건, 공동-보상 조건, 타인-보상 조건이 피험자 내 설계로 포함되었다. 구체적인 시행 구조는 그림 1에 제시되어있다. 응시점이 2~4초 간(2000-4000msec 사이의 균일분포에서 무작위로 응시점 제시 시간을 결정)제시된 뒤 두 쌍의 그림이 제시되고, 자극 제시 시점부터 2초 이내에 왼쪽 또는 오른쪽 그림을 키보드의 F 또는 J키를 눌러서 선택할 수 있었다. 키보드를 누르면 선택된 그림의 색깔이 바뀌면서 어떤 그림을 선택했는지를 4초 간 보여주고, 1.5초 간 선택의 결과(점수 획득 성공 또는 실패 여부)를 보여주었다.[1] 참가자들은 각 조건 당 50시행씩 총 150번의 선택을 했고, 세 종류의 자극 쌍은 무선적인 순서로 참가자마다 다르게 제시되었다. 자극 쌍의 예는 그림 1에 제시되어있다. 여기서 사용한 도형은 Kim, Shimojo, O'Doherty(2006)의 연구에서 사용했던 것과 동일했다. 보상 확률이 높은 옵션과 낮은 옵션의 위치와 자극은 참가자마다 다르게 제시되도록 역균형화(counter-balancing) 하였으나, 자극 쌍과 보상 종류 간의 관련성은 모든 참가자에게 동일하게 유지되었다. 참가자들은 과제가 끝난 뒤 전체 시행의 10%를 무작위로 선택하여 합산한 점수만큼 나중에 수행하게 될 스트레스 과제의

1) 본 연구에서 개발된 행동 패러다임은 이후에 fMRI 실험에 이용하기 위해서 시간 구조를 fMRI 실험에 적합하도록 설계하였다.

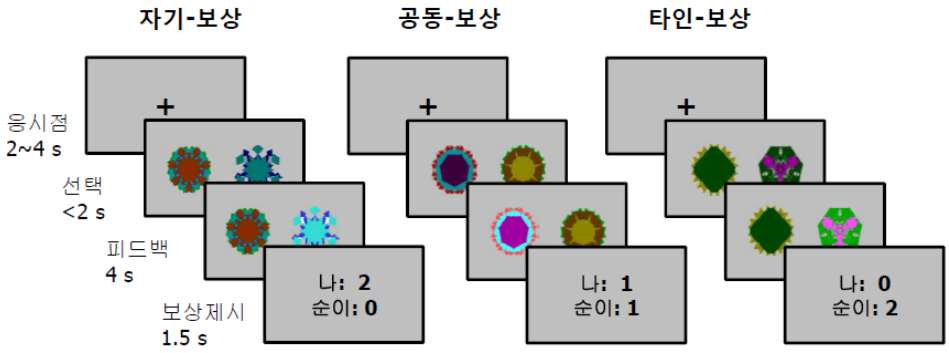


그림 1. 이타적 강화학습 과제의 구조

시간을 1점 당 10초씩 줄일 수 있다고 안내받았다. 누적 점수를 사용하지 않고 무작위로 일부 시행을 선택한다고 한 이유는, 참가자들이 매 시행을 동일하게 중요하게 여기도록 하기 위해서이다. 각 자극 쌍 내에서 한 자극의 보상 확률은 30%, 다른 자극의 보상 확률은 70%였으며, 참가자들에게는 보상 확률에 대한 정보를 사전에 제공하지 않았다. 참가자들이 자신이 원하는 결과를 최대한 많이 획득하기 위해서는 시행을 반복하면서 선택-결과 간의 관계를 학습해야 했다.

평판이나 상호성에 대한 고려를 가능한 배제하기 위해서, 참가자들이 서로 모르는 사람이라는 점을 사전에 확인하였고, 과제 추첨 단계 이후에는 서로 만나거나 의사소통 하지 않도록 하였다. 또한, 이타적 강화학습 과제를 수행한 참가자에게는 스트레스 과제의 시간에 영향을 미칠 수 있는 사람은 한 사람뿐이며, 상대방은 전혀 다른 종류의 과제를 수행하고 있고, 스트레스 과제 시간이 줄어들 수도 있다는 점을 알지 못한다고 설명하였다. 스트레스 과제에 대한 커버스토리는 강화학습과제에서 획득하게 되는 점수에 타인을 돕는다는 의미를 부여하기 위한 것이므로 실제로는 스트레스 과제를 생략했으며, 실험이 완전히 종료된 이후에 실제 연구 목적을 밝혔다.

## 결 과

자기-보상, 공동-보상, 타인-보상 조건에서의 수행을 비교하기 위하여 보상 확률이 높은 옵션을 선택한 비율이 조건에 따라서 어떻게 달라지는지를 살펴보았다(그림 2). 보상 확률이 높은 옵션을 선택한 비율은 자기-보상조건에서 0.76($SE$ = 0.06), 공동-보상 조건에서 0.48($SE$ = 0.07), 타인-보상 조건에서 0.38($SE$ = 0.04)로, 전반적으로 참가자들은 자신에게 보상이 주어졌던 자기-보상 조건에서는 보상 확률이 높은 옵션을 선택하려고 한 반면, $t(15)$ = 4.30, $p$ < .05, 공동-보상 조건에서는 특별한 선호를 보이지 않았으며, 타인-보상 조건에서는 오히려 타인에게 점수가 주어지는 결과를 회피하는 것으로 나타났다, $t(15)$ = -2.82, $p$ < .05. 피험자-내 변량분석(within-subject ANOVA)을 실시한 결과, 보상 종류의 주효과가 유의하였다, $F(2,30)$ = 10.04, $p$ < .001, $\eta_p^2$ = .40. 대응별 비교 분석 결과 자기-보상 조건

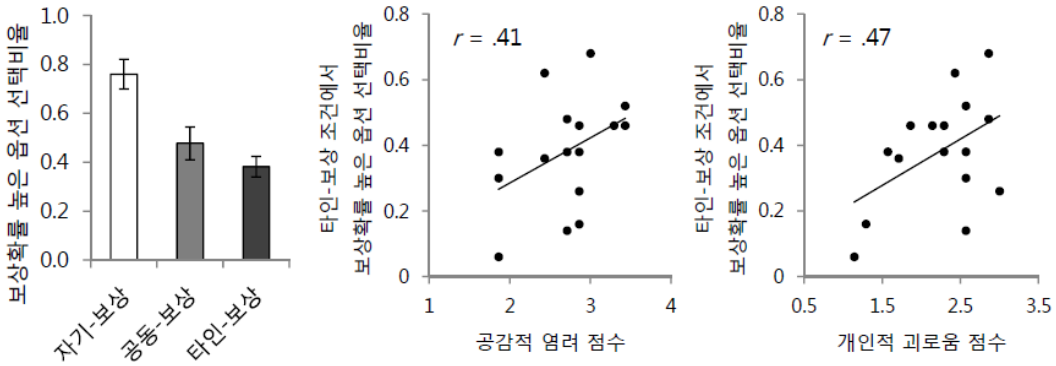


그림 2. 실험1의 결과. 왼쪽 그림: 자기-보상, 공동-보상, 타인-보상 조건에서 보상 확률이 높은 옵션을 선택한 평균비율 (오차막대는 표준오차를 나타냄). 가운데 그림: 타인-보상 조건에서 보상 확률이 높은 옵션을 선택한 비율과 대인간반응척도에서 측정된 공감적 염려 점수와의 관련성. 오른쪽 그림: 타인-보상 조건에서 보상 확률이 높은 옵션을 선택한 비율과 개인적 괴로움 점수와의 관련성 (산포도에서 r은 두 변수 간의 단순 상관계수. 공감척도 하위점수들과 인지욕구점수를 예측변수로 하는 다중선형회귀분석의 결과는 표 2 참고).

과 공동-보상 조건(평균차: 0.28, $SE$ = 0.11, $p$ < .05), 자기-보상 조건과 타인-보상 조건(평균차: 0.38, $SE$ = 0.08, $p$ < .001) 간 차이가 모두 유의하였다.

공감능력에 따라서 이타적 강화학습 과제의 수행에 차이가 있는지 알아보기 위하여, 반복측정 일반선형모형(repeated measure GLM)에서 보상의 종류를 피험자-내 요인, 대인간반응척도로 측정된 상상력, 관점취하기, 공감적 염려, 개인적 괴로움 점수를 연속변수로 포함시켜 보상 확률이 높은 옵션을 선택할 확률을 예측하는지 검증하였다. 이 때, 과제 수행 동기를 통제하기 위하여 인지욕구 척도 점수를 공변량으로 포함시켰다. 분석결과, 보상의 종류와 개인적 괴로움 점수 간의 상호작용이 유의하여, $F(2, 20) = 8.95$, $p < .005$, $\eta_p^2 = .47$, 보상의 종류에 따른 수행의 차이가 개인적 괴로움 점수의 영향을 받는 것으로 나타났다. 이 결과를 보다 잘 이해하기 위해서 각 조건 별로 대인간반응척도 하위 척도 점수들과 인지욕구 점수가 선택을 예측하는 정도를 다중선형회귀분석을 통해 살펴본 결과, 자기-보상 조건에서는 개인적 괴로움 점수가 높을수록 수행이 낮았던 반면에(B = -.305), 공동-보상 조건(B = 0.351)과 타인-보상 조건(B = 0.200)에서는 개인적 괴로움 점수가 높을수록 수행이 높았다(표 1, 그림 2). 개인적 괴로움과 함께 공감의 정서적 요인에 해당하는 공감적 염려의 경우, 보상 종류에 따른 수행 차이에 영향을 미치는 정도가 유의하지는 않았으나, $F(2,20) = 2.71$, $p = .09$, $\eta_p^2 = .21$, 자기-보상과 공동-보상 조건에서는 영향을 주지 않고 타인-보상 조건에서의 선택을 유의하게 예측하였다(B = 0.189). 즉, 개인적 괴로움과 마찬가지로 타인의 고통에 대하여 공감적 염려를 잘 경험하는 사람들이 강화학습과제의 타인-보상 조건에서의 더 나은 수행을 보였다(그림 2). 공감 하위척도 중에서 상상력과 관점취하기, 과제 수행 동기를 반영하는 인지욕구는 이타적 강화학습 과제의 수행과는 관련이 없었다(표 1).

표 1. 공감의 하위척도와 인지욕구가 보상 확률이 높은 옵션을 선택하는 비율을 예측하는 정도를 보여주는 다중선형 회귀분석의 모수추정치

| | | B | 표준오차 | t | 유의확률 | 95% 신뢰구간 | |
|---|---|---|---|---|---|---|---|
| | | | | | | 하한값 | 상한값 |
| 자기-보상 | 절편 | 0.713 | 0.448 | 1.589 | .143 | -.287 | 1.712 |
| | 상상력 | 0.054 | 0.074 | 0.725 | .485 | -.112 | .220 |
| | 관점취하기 | 0.093 | 0.141 | 0.659 | .525 | -.221 | .406 |
| | 공감적염려 | 0.207 | 0.120 | 1.722 | .116 | -.061 | .474 |
| | 개인적괴로움* | -0.305 | 0.117 | -2.597 | .027 | -.566 | -.043 |
| | 인지욕구 | -0.059 | 0.071 | -0.832 | .425 | -.219 | .100 |
| 공동-보상 | 절편 | -0.368 | 0.503 | -0.731 | .482 | -1.489 | .754 |
| | 상상력 | -0.050 | 0.083 | -0.597 | .564 | -.236 | .136 |
| | 관점취하기 | 0.047 | 0.158 | 0.300 | .770 | -.304 | .399 |
| | 공감적염려 | -0.136 | 0.135 | -1.012 | .335 | -.436 | .164 |
| | 개인적괴로움* | 0.351 | 0.132 | 2.663 | .024 | .057 | .644 |
| | 인지욕구 | 0.121 | 0.080 | 1.507 | .163 | -.058 | .299 |
| 타인-보상 | 절편 | -0.195 | 0.311 | -0.628 | .544 | -.889 | .498 |
| | 상상력 | -0.003 | 0.052 | -0.056 | .957 | -.118 | .112 |
| | 관점취하기 | -0.188 | 0.098 | -1.923 | .083 | -.405 | .030 |
| | 공감적염려* | 0.189 | 0.083 | 2.274 | .046 | .004 | .375 |
| | 개인적괴로움* | 0.200 | 0.081 | 2.458 | .034 | .019 | .381 |
| | 인지욕구 | 0.034 | 0.050 | 0.683 | .510 | -.077 | .144 |

*$p < .05$

실험 1의 결과는 타인의 고통에 정서적으로 공감을 잘 하는 사람들이 타인을 도울 수 있는 결과를 보상으로 경험하는 경향이 있음을 보여준다. 또한, 이타적 강화학습 과제에서의 수행이 정서적 공감 점수와 관련되어 있으며 인지 욕구 점수와는 상관이 없다는 결과는 이타적 강화학습 과제가 단순히 정답을 맞히기 위한 과제 수행 동기를 반영하는 것이 아니라, 개인의 이타적 성향을 반영하고 있음을 시사한다.

한편, 실험 1의 결과 중에서 자기-보상, 공동-보상, 타인-보상 조건 간에 나타난 이타적 강화학습과제에서의 수행 차이를 해석할 때에는 주의가 필요하다. 보상 확률이 높은 옵션과 낮은 옵션에 해당하는 자극의 종류를 매 피험자마다 다르게 제시되도록 역균형화 하여 보상 확률에 따른 선호의 차이에 도형의 특성에 따른 차이가 반영되지 않도록 하였으나, 각 자극 쌍과 관련된 보상의 종류(자기-보상, 공동-보상, 타인-보상과 연합된 자극 쌍의 종

류)는 동일하게 유지되었다. 따라서 특정 자극 쌍이 기억하기에 더 쉽거나 어려워서 조건 간 차이가 발생했을 가능성도 배제할 수 없다. 이러한 가능성을 확인하기 위하여 처음 열 시행동안 보상 확률이 높은 옵션을 선택한 비율을 구하여 조건 간 차이를 비교하였다. 자기-보상과 공동-보상, 타인-보상 조건에서 보상 확률이 높은 옵션을 선택한 비율은 각각 0.69($SE$ = 0.08), 0.44($SE$ = 0.05), 0.41($SE$ = 0.04)로, 과제의 초기부터 세 조건 간 유의한 차이가 관찰되었다, $F(2, 30) = 6.15$, $p < .01$. 처음 열 시행 동안 어느 정도 학습이 일어났다고 가정하더라도, 과제 초반에는 어떤 자극 쌍이 어떤 조건과 연합되는지를 알지 못하기 때문에 초기 시행에서 관찰된 유의미한 차이는 있는 그대로 받아들이기에 무리가 있다. 따라서 개인차를 고려하지 않고 평균적인 조건 간 차이를 해석할 때는 주의를 기울여야한다. 실험 2에서는 실험 1에서의 문제점을 보완하기 위하여 각 자극 쌍과 보상 종류와의 관계에 대해서도 모든 가능한 조합을 만들어서 참가자마다 다르게 제시하였다.

## 실험 2: 이타적 강화학습과제에서 나타나는 내집단 편향

실험 2에서는 최소집단 패러다임(minimal group paradigm: Tajfel, 1970)을 이용하여 이타적 강화학습과제에서 내집단 편향이 관찰되는지를 확인하고자 하였다. 최소집단 패러다임이란, 티셔츠의 색깔이나 좋아하는 그림과 같이 매우 사소한 측면에서의 공통점과 차이점을 조작하는 것으로 집단 정체성을 형성하는 절차를 말한다. 최소집단만으로도 사람들은 내집단원에게 호감을 가지고(Hewstone, Rubin, & Willis, 2002), 더 많이 공감하고(Hein 등, 2010), 더 이타적으로 행동(Levine 등, 2005)하는 것으로 알려져 있다. 실험 2의 참가자들은 이타적 강화학습과제를 통해서 자신, 내집단원, 외집단원을 위해서 각각 점수를 획득할 수 있었다. 실험 1의 공동-보상, 타인-보상 조건 대신에 내집단원-보상, 외집단원-보상 조건을 포함시켜서 내집단원을 도울 수 있는 경우와 외집단원을 도울 수 있는 경우에 이타적 강화학습과제에서의 수행에 차이가 있는지 살펴보았다. 실험 2에서는 내집단 선호로 인하여 사람들이 내집단원의 안녕이 증진되는 것에 가치를 더 많이 부여한다면, 내집단원-보상 조건에서의 수행이 외집단원-보상 조건에서의 수행보다 뛰어날 것이라고 예상하였다.

### 방 법

#### 연구 참가자

학내 포털 서비스에 게시한 실험 참가자 모집 광고를 보고 자원한 72명의 대학생들이 연구에 참여했다. 한 번에 세 사람의 참가자를 초청하여 그 중 한 사람만 본 연구와 관련이 있는 과제를 수행했으므로 24명(남자: 8명, 평균연령: 23.3세)만이 분석 대상에 포함되었다. 모든 참가자들은 실험이 시작되기 전에 실험 절차 전반에 대한 안내를 받고 참가 동의서에 서명하였으며, 모든 절차는 고려대학교 기관생명윤리심의위원회의 승인을 받았다.

### 절차

실험 2에서는 최소집단패러다임을 이용한 내집단-외집단 조작 절차(지각 경향성 검사, 집단 구별 과제)가 추가되었고, 이타적 강화학습과제에서 공동-보상 조건과 타인-보상 조건이 각각 내집단-보상 조건과 외집단-보상 조건으로 대체되었으며, 높은 보상 옵션과 낮은 보상 옵션, 그리고 자극 쌍과 보상 종류 간의 관련성도 각 참가자마다 다르게 제시하여 자극 자체의 속성이 선택에 미치는 영향을 통제하였다. 이점을 제외하면 실험 1과 동일한 절차로 진행되었다. 전체적인 실험 절차는 그림 3에 정리되어있다. 실험 1과는 달리 실험 2에서는 도움을 제공할 수 있는 대상이 내집단원과 외집단원, 두 사람이기 때문에 한 번에 세 사람의 참가자를 동시에 초청하였다. 한 사람은 이타적 강화학습과제를 수행하고, 나머지 두 사람은 본 연구와 관계없는 다른 과제를 수행했다.

#### 최소집단 패러다임

본 연구에서는 내집단과 외집단을 구분하기 위해서 최소집단 패러다임을 이용했다. 본 연구에서 사용한 절차는 다음과 같다. 세 명의 참가자들이 실험실에 도착하면 실험 절차 전반에 대한 안내문을 제공하였는데, 여기에는 연구진이 전경-중심 지각 경향성과 배경-중심 지각 경향성에 따른 인지 과제 수행과 스트레스 반응의 차이를 알아보고자 한다는 커버스토리가 담겨있었다. 실험은 지각 경향성 측정, 스트레스 과제를 위한 소음 역치 측정, 과제 추첨, 인지과제, 스트레스 과제 순으로 진행된다고 설명하였다. 지각 경향성을 측정하는 과제로는 전경-배경이 뒤바뀔 수 있는 가역성 도형을 사용하였다. 실제 응답에 관계없이, 참가자 중 두 사람은 전경-중심이라고 알려주고, 나머지 한 사람은 배경-중심이라고 알려주었으며, 전경-중심은 홍팀, 배경-중심은 청팀으로 구분하고 서로의 이름과 팀 이름을 다시 한 번 확인하도록 했다. 집단 구분 절차를 마친 뒤에는 실험 1과 동일한 소음 역치 측정 절차와 과제 추첨 절차를 거치고, 각각의 피험자들은 서로 다른 공간에서 과제를 수행하였다. 세 사람 중 한 사람만 이타적 강화학습과제를 수행하고 나머지 두 사람은 본 연구와 무관한 다른 실험에 참여했다.

이 때, 집단은 참가자들이 실험 참여를 원하는 시간대에 따라서 구성되었으며, 세 사람이 모두 동성인 경우가 9건, 혼성 집단이 15건이었다. 혼성 집단 중에서 내집단원이 동성이었던 경우는 5건, 이성이었던 경우는 10건이었다. 아래에 기술한 모든 분석에 대하여 참가자의 성별과 내집단원이 동성이었는지 이성이었는지 여부를 포함한 분석도 함께 수행하였으나 성별 효과를 고려한 경우와 고려하

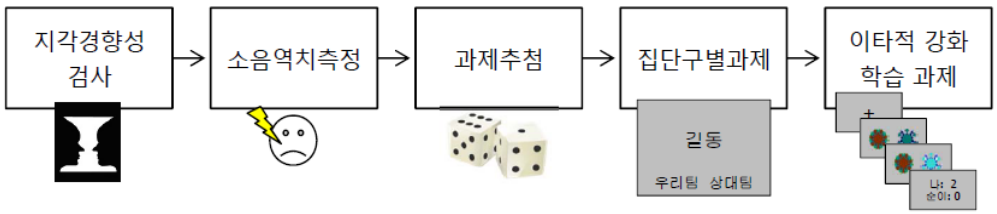


그림 3. 실험 2의 절차

지 않은 경우 주요 결과에 차이가 없었고, 성별의 효과도 없었다. 통계적 유의도에 변화가 있었던 경우에만 성별을 포함한 분석 결과를 추가적으로 기술하였다.

### 집단구별과제

참가자가 이타적 강화학습과제를 수행하기 전에 집단 구분을 보다 명확히 하기 위한 절차를 포함시켰다. 이 과제에서는 화면에 빨간색 또는 파란색 도형, 참가자 본인의 이름, 다른 참가자들의 이름, 홍팀 또는 청팀이 무작위적인 순서로 제시되었고 참가자들은 화면에 제시되는 대상이 '우리팀'에 해당하는지 '상대팀'에 해당하는지를 가능한 빠르고 정확하게 판단하여 응답해야 했다. 이 과제를 통해서 내집단원과 외집단원의 이름을 확인시키고 누가 우리편이고 누가 상대편인지를 명확하게 기억하도록 하였다. 또한 참가자들의 과제 몰입도를 높이고 최소집단 패러다임의 효과를 강화하기 위해서 자료 수집이 끝난 다음에 반응 속도와 정확도를 기준으로 점수를 매겨서 홍팀의 평균과 청팀의 평균을 비교하여 점수가 더 높은 팀에게 1000원의 참가비를 더 지급하겠다고 설명했다.

내집단-외집단 조작이 잘 이루어졌는지 확인하기 위해서, 이타적 강화학습과제 참가자들에게는 과제가 끝난 뒤 상대방 각각에 대해서 얼마나 호감이 가는지 1점(전혀 호감이 가지 않는다)에서 7점(매우 호감이 간다) 사이로 응답하도록 하였다. 최소집단 패러다임이 잘 작동하였다면, 내집단원에 대해서 더 높은 호감을 보일 것으로 예상할 수 있다.

### 이타적 강화학습 과제

기본적으로 실험 1에서 사용한 과제와 동일하였으나, 실험 2의 이타적 강화학습 과제는 자신만 2점을 받을 수 있는 자기-보상 조건, 내집단원이 2점을 받는 내집단원-보상 조건, 외집단원이 2점을 받는 외집단원-보상 조건으로 구성되었다. 실험 1에서와 마찬가지로 참가자들에게는 이 과제에서 얻은 점수는 이후에 스트레스 과제의 시간을 줄이는데 사용할 수 있다고 설명해서 점수 획득이 단순한 자원 분배가 아니라 타인을 돕는 맥락으로 해석될 수 있도록 하였으며, 과제가 다 끝난 뒤 10%의 시행을 무작위로 선택하여 점수를 합산 한다고 알려줌으로써 각 시행을 독립적으로 여기고 집중하도록 했다. 여기서도 보상 확률이 높은 옵션을 선택하는 빈도를 주요 종속변수로 분석하였다.

## 결 과

먼저, 참가자들이 연구자들의 의도대로 내집단원과 외집단원을 구분하였는지를 확인하기 위해서 내집단원과 외집단원의 이름과 소속팀을 물었고 모든 참가자들이 정확하게 응답하였다. 호감도에서 내집단 편향이 관찰되는지 확인하기 위하여 내집단원과 외집단원에 대한 호감 점수를 비교하였다. 내집단원과 외집단원에 대한 호감도는 각각 4.75($SE$ = 0.23)와 4.29($SE$ = 0.23)로 점수 차이가 크지는 않았으나 가설과 일치하는 방향으로 한계적 수준에서 유의한 차이를 보였다, 대응표본 t검정: $t(23) = 2.04$, $p = .053$(양쪽검증). 성별의 효과를 통제하는 경우 내집단원과 외집단원에 대한 호감도의 차이는 통계적으로 유의하였다, $F(1,20) = 6.33$, $p < .05$, $\eta_p^2 = .24$.

이타적 강화학습 과제에서 조건 간 수행의

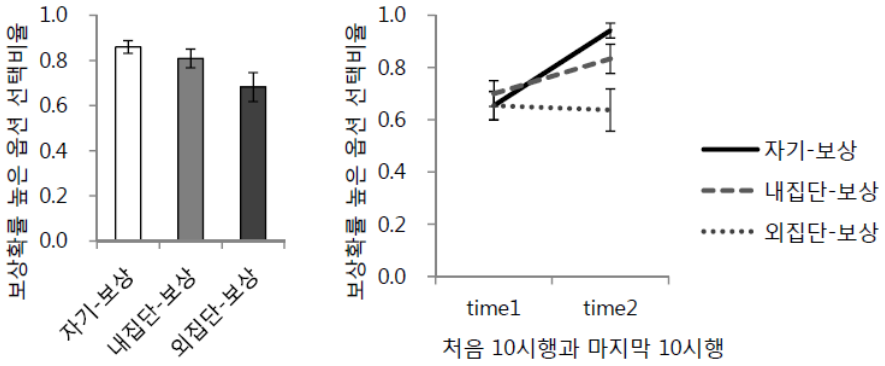


그림 4. 실험2의 결과. 왼쪽그림: 자기-보상, 내집단-보상, 외집단-보상 조건에서 보상 확률이 높은 옵션을 선택한 평균 비율. 오른쪽 그림: 자기-보상, 내집단-보상, 외집단-보상 조건에서 처음 10시행(time1)과 마지막 10시행(time2) 동안 보상 확률이 높은 옵션을 선택한 평균 선택비율. (오차막대는 표준오차를 나타냄.)

차이를 검증하기 위해서 자기-보상, 내집단원-보상, 외집단원-보상 조건에서 보상 확률이 높은 옵션을 선택한 비율을 비교하였다(그림 4). 보상 확률이 높은 옵션을 선택한 비율은 자기-보상 조건에서 0.86($SE$ = 0.03), 내집단원-보상 조건에서 0.81($SE$ = 0.04), 외집단원-보상 조건에서 0.68($SE$ = 0.06)로, 집단-내 변량분석 결과 보상 종류의 주효과가 유의하였다, $F(2,46) = 5.36$, $p < .01$, $\eta_p^2 = .19$. 대응별 비교 결과 선택 빈도의 차이는 자기-보상과 외집단-보상 조건(평균차: 0.18, $SE$: 0.064, $p < .05$), 내집단-보상 조건과 외집단-보상 조건(평균차: 0.13, $SE$ = 0.06, $p < .05$) 간에 유의한 차이를 보였다. 반면, 자기-보상 조건과 내집단-보상 조건 간에는 유의한 차이를 보이지 않았다(평균차: 0.05, $SE$ = 0.04, $p$ = .23). 이러한 결과는 참가자들이 내집단원에게 도움이 되는 결과에 자신에게 보상이 되는 결과와 유사한 정도의 가치를 부여했음을 시사한다.

## 논 의

본 연구에서는 강화학습 패러다임을 이용하여 사람들이 타인의 안녕에 가치를 부여하는 정도를 측정하는 이타적 강화학습 과제를 개발하고, 기존 연구에서 이타 행동에 영향을 미치는 주요 요인으로 알려져 있는 공감적 성향과 내집단 선호의 효과가 이타적 강화학습 과제에서도 관찰됨을 확인하였다. 정서적 공감 성향이 강한 사람들이 타인의 안녕에 더 많은 가치를 부여하고(실험 1) 외집단원보다는 내집단원의 안녕에 더 관심을 가지는 것으로 보인다(실험 2).

실험 1에서 가장 중요한 발견은 이타적 강화학습 과제에서의 선택이 공감 성향과 관련 있었으며, 특히 정서적 공감 성향을 측정하는 개인적 괴로움과 공감적 염려 점수가 높았던 사람들이 타인의 안녕에 더 많은 가치를 부여했다는 것이다. 공감의 인지적 측면보다 정서적 측면이 이타적 강화학습과제에서의 선택을 더 잘 예측했던 결과는 공감의 인지적 측면은

주로 타인의 의도와 사고를 정확하게 추론하는 능력과 관련 있으며, 타인의 정서적 상태에 반응하고 도움을 제공하려는 동기를 유발하는 데에는 정서적 공감이 더 중요한 역할을 한다는 선행 연구 결과들(Davis, 1983; Saxe, 2006; Singer, 2006)과도 일치한다. 한편, 대인간반응척도에서는 관점 취하기를 공감의 인지적 측면으로 정의하고 있다(Davis, 1983). 그러나 Oswald(1996)는 관점취하기를 다시 타인의 사고에 대한 인지적 관점취하기(cognitive perspective taking)와 타인의 느낌에 대한 정서적 관점취하기(affective perspective taking)로 다시 구분하고 이 중에서도 정서적 관점취하기가 도움 행동을 더 잘 예측한다고 보고한 바 있다. 관점 취하기와 도움행동의 관련성을 살펴본 연구들에서도 대부분 정서적 관점취하기를 채택하고 있다는 점(Batson, Early, & Salvarani, 1997; Lamm, Batson, & Decety, 2007) 또한 이타 행동에 있어서 공감의 정서적인 측면의 중요성을 시사한다.

정서적 공감 성향의 영향력이 두드러진 또 다른 이유는 본 연구에서 사용된 과제의 특성에서 찾아볼 수 있다. 이타적 강화학습과제는 과제 수행을 통해서 타인이 경험하게 될 고통(소음을 듣는 시간)을 줄여줄 수 있도록 구성되어있다. 참가자들은 소음 역치 측정 단계에서 직접 소음에 노출되고 상대방이 소음을 듣는 장면도 보게 된다. 시나리오를 사용한 연구들과는 달리 이미 생생하게 상황을 경험하는 기회를 가지기 때문에 어느 정도 관점 취하기가 이루어진 상태에서 과제를 수행하게 되는 것이다. 따라서 대인간반응척도에서 측정되는 상상력이나 인지적 관점 취하기와 관련되어 있는 타인의 내적 상태를 이해하기 위한 복잡한 인지적 과정이 개입될 여지가 적다. 그 결과 고통 받을 위험에 처한 타인에게 즉각적으로 보이는 정서 반응이 상대적으로 중요하게 작용했을 수 있다.

한편, 개인적 괴로움과 공감적 염려가 이타적 강화학습과제에서의 수행에 영향을 미치는 양상에도 차이가 있었는데, 개인적 괴로움 점수가 자기-보상 조건에서의 수행과 부적 상관을 보였던 반면, 공감적 염려 점수가 높았던 사람들은 다른 조건에서는 영향을 받지 않고 타인-보상 조건에서만 선택적으로 영향을 받았다. 이러한 결과는 개인적 괴로움과 공감적 염려를 서로 구별되는 정서로 보는 관점(Batson, 2011; Eisenberg & Eggum, 2009)에서 설명될 수 있다. Batson과 동료연구자들(Batson, 2011)은 고통 받는 타인을 보았을 때 사람들이 경험하는 정서를 자기지향적인 개인적 괴로움과 타인지향적인 공감적 염려로 구분하고, 공감적 염려에서 비롯된 도움 행동이 진정한 이타주의라고 제안한 바 있다. Eisenberg와 Eggum(2009)도 공감과 관련된 정서적 반응을 타인지향적인 동정심(sympathy)과 자기중심적인 개인적 괴로움으로 구별하고, 동정심은 타인의 정서에 공감하는 동시에 회피 정서를 조절하고 타인에게 적절한 도움을 제공할 수 있는 '노력이 드는 자기-조절 과정(effortful self-regulatory process)'과 관련되어있다고 하였다. 반면에 개인적 괴로움은 타인의 고통에 대하여 자신이 괴로워하는 것으로 타인을 진정으로 염려해서가 아니라 자신의 괴로움을 덜기 위하여 도움을 제공하거나, 상황을 모면할 수 있는 다른 방법이 있다면 돕지 않게 만들기 때문에 이타성과는 거리가 있다(Batson, 2011; Eisenberg & Eggum, 2009). 공감적 염려나 동정심과는 달리 개인적 괴로움을 쉽게 경험하는 성향은 일반적인 스트레스 상황에서의

정서적 취약성과도 관련이 있는 것으로 알려져 있다(Davis, 1983). 따라서 실험 1에서 개인적 괴로움 점수와 공감적 염려 점수가 모두 타인-보상 조건에서의 수행과 정적 상관을 보였지만 여기에는 서로 구별되는 심리적 과정이 관여했을 가능성이 있다.[2] 즉, 개인적 괴로움 점수가 높았던 참가자들은 타인의 고통으로 인한 자신의 괴로움을 회피하기 위하여 과제를 수행했을 수 있다. 또한, 개인적 괴로움 점수와 정서적 취약성과의 관련성을 고려할 때, 자신의 수행이 타인에게 영향을 줄 수 있는 과제 상황이 정서적으로 부담이 되어 자기-보상 조건에서의 수행에 부정적인 영향을 미쳤을 것이라고 생각해볼 수 있다.

실험 2에서는 도움을 줄 수 있는 대상을 내집단원과 외집단원으로 구분하여 이타적 강화학습과제에서의 수행을 비교하고 내집단 선호 현상이 나타남을 확인하였다. 사람들은 내집단원을 도와줄 수 있는 조건에서 자신에게 이익이 되는 조건과 유사한 수준의 수행을 보였으나 외집단원의 안녕에 대해서는 비교적 가치를 덜 부여하는 것으로 보인다. 이러한 결과는 기존 연구들에서 반복적으로 밝혀져 온 내집단 편향과 일치한다. 실험 2의 결과에서 한 가지 주목할 점은, 내집단 편향이 외집단 혐오(outgroup hate)보다는 내집단 선호(ingroup love/favoritism)에서 비롯된 것으로 보인다는 점이다. 이타적 강화학습과제에서의 수행을 비교해보면, 외집단-보상 조건보다는 내집단-보상 조건에서 더 나은 수행을 보였으나, 세 조건 모두에서 보상 확률이 높은 옵션을 선택하는 빈도가 우연수준(50%)을 넘어서는 것으로 나타났다, 단일표본 t검정 결과: 자기-보상 조건에서 $t(23) = 12.81$, $p < .001$; 내집단원-보상 조건에서 $t(23) = 7.44$, $p < .001$; 외집단원-보상 조건에서 $t(23) = 2.85$, $p < .01$. 즉, 외집단원에 대해서도 어느 정도는 도움을 제공하는 방향으로 선택이 이루어졌다. 그러나 초반 10시행과 후반 10시행으로 구분하여 보상 확률이 높은 옵션을 선택한 비율을 살펴본 경우(그림 4), 자기-보상과 내집단원-보상 조건에서는 전반에 비하여 후반으로 갈수록 수행이 증가하여 학습 효과가 관찰되었으나, 외집단원-보상 조건에서는 후반부에서 오히려 우연수준에 가까운 선택 비율을 나타냈다. 이러한 결과를 볼 때 참가자들이 외집단원에게도 우호적이었다기보다는 외집단원에게 적대적이지는 않았다고 해석하는 것이 더 적절해 보인다.

선행 연구들에서도 최소집단을 이용한 내집단 편향 연구들에서는 내집단 편향이 외집단원에 대한 혐오적 반응보다는 내집단원에 대한 선호 반응에 치우쳐 있음을 보고하는 경우가 많다(Brewer, 1979; Brewer, 1999; Yamagishi & Mifune, 2009). 예를 들어, Halevy 등(Halevy, Bornstein, & Sagiv, 2008)은 죄수딜레마 게임을 변형한 과제를 이용하여 외집단의 이익을 감소시켜야만 내집단의 이익을 증진시킬 수 있는 조건과 외집단의 이익에 상관없이 내집단의 이익을 증진시킬 수 있는 옵션이 존재하는 경우를 비교하였는데, 참가자들은 외집단에게 손해를 끼치는 것이 내집단의 이익 증진에 필수적인 경우에만 외집단에게 해가 되는 선택을 하였고, 외집단의 이익과 내집단의 이익이 관련이 없는 경우에는 외집단에게 손해를 끼치지 않고도 내집단에게 이익이 될 수 있는 선택을 선호하였다. 최근의 다른 연구들에서도 유사한 결과가 보고된 바 있으며(De Dreu

2) 실제로 실험 1에서 개인적 괴로움과 공감적 염려 점수 간에는 높은 상관이 관찰되지 않았다 ($r = .20$, $p = .45$)

등, 2010; Halevy, Weisel, & Bornstein, 2012), 발달과정에서도 내집단 선호가 외집단 혐오보다 더 먼저 나타난다는 증거들이 보고되었다(Buttelmann & Böhm, 2014; Fehr 등, 2008). 본 연구의 실험 2에서도 최소집단 패러다임을 사용하였고, 외집단원에게 해를 입혀야만 내집단원에게 이익이 되는 상황이 아니었기 때문에, 내집단 선호가 더 강하게 작용하는 방향으로 내집단 편향이 일어났던 것으로 보인다.

한편, 실험 1의 결과와 실험 2의 결과에서 나타나는 차이점에 대해서도 논의할 필요가 있다. 실험 2에서는 외집단원-보상 조건에서도 보상 확률이 높은 옵션을 선택한 비율이 보상 확률이 낮은 옵션을 선택한 비율보다 높게 나타났는데, 실험 1에서는 타인-보상 조건에서 보상이 없는 옵션을 선택하는 비율이 더 높게 나타나서 오히려 타인에게 보상이 주어지는 것을 회피하려는 경향을 보였다. 실험 1과 실험 2는 연구절차나 구성 면에서 차이가 있기 때문에 직접 비교하기가 어렵지만, 일반적인 타인에 대해서 외집단원보다 더 적대적인 선택을 하는 것처럼 보이는 결과에 대한 해석이 필요하다. 먼저, 실험 1에서도 간단히 논의하였듯이, 실험 1에서는 자극 쌍과 보상 종류 간의 관련성을 역균형화 하지 않았기 때문에 자극 쌍 자체가 가지는 속성의 영향을 받았을 가능성이 있다. 실제로 자극 쌍과 조건 간 연합을 역균형화 하였던 실험 2의 초반 10시행에서는 보상 확률이 높은 옵션을 선택하는 비율이 자기-보상 조건에서 0.65($SE$ = 0.05), 내집단원-보상 조건에서 0.70($SE$ = 0.05), 외집단원-보상 조건에서 0.65(SE = 0.06)로 차이가 없었다, $F(2, 46) = .268$, $p = .77$, $\eta_p^2 = .01$(그림 4). 그럼에도 불구하고 실험 1의 타인-보상 조건에서 타인에게 점수가 주어지지 않는 쪽으로 선택이 치우친 결과를 완전히 설명하기는 어렵다. 한 가지 가능성은 실험 1에서는 자신과 타인이 1점씩 공평하게 나누어 가지는 공동-보상 조건이 있었던 반면 실험 2에서는 자기-보상 조건 또는 타인-보상 조건으로 구분되었고, 실험 1에서는 자기와 타인의 점수를 동시에 표시하여 결과를 비교할 수 있었지만 실험 2에서는 각 조건과 관련된 사람의 점수만을 보여줬다는 점 때문에 실험 1과 실험 2의 참가자들이 선택 상황을 다르게 받아들였을 수 있다는 것이다. 실험 1에서는 공동-보상 조건이 있었기 때문에 타인을 도와준다기보다는 공평하게 점수를 나누는 과제로 인식했거나 공동-보상 조건에서 1점이라도 획득하여 도움을 줄 수 있었던 점이 타인-보상 조건에서 점수를 획득할 동기를 낮추었을 가능성이 있다. 또한 자기 점수와 타인 점수가 함께 제시되었던 점도 일부 참가자들에게 경쟁 구도로 비춰졌을 수 있다. 실험 2에서는 상대방에게 점수를 줄 수 있는 기회가 한 조건 밖에 없었기 때문에, 도울 것인지 말 것인지의 선택이 되어서 도움을 제공하는 선택을 상대적으로 더 많이 하도록 만들었을 가능성이 있으며 한 사람의 점수만 화면에 제시되어서 실험 1에 비해서 타인과의 명시적인 비교가 덜 일어났을 가능성이 있다. 실험 1과 실험 2의 참가자들이 공감 성향에서 차이가 있었을 수도 있으나, 실험 2에서는 공감 척도를 사용하지 않았기 때문에 실험 1과 실험 2 참가자들의 공감 성향을 비교해 볼 수는 없었다. 본 연구에서는 표본 크기가 작아서 참가자들을 세부적으로 나누어 살펴보기 어렵지만, 후속 연구에서 타인-보상 조건에서 수행이 낮았던 사람들을 타인에게 적대적인 선택을 하는 사람들과 타인의 이익에 무심한 사람들로 구분하고 공감 수

준이나 다른 개인차 지표들에서 어떠한 차이를 보이는지 살펴보거나, 의사결정의 프레임을 달리하였을 때(예를 들어 경쟁이나 사회비교가 강조되는 경우와 중립적인 경우와 협동이 강조되는 경우를 비교) 사람들의 선택이 어떻게 달라지는지 살펴볼 필요가 있다.

## 결론

본 연구에서는 강화학습 패러다임을 이용하여 사람들이 타인의 안녕에 가치를 부여하는 정도를 측정하는 이타적 강화학습과제를 개발하였다. 이타적 강화학습과제에서의 행동이 기존 연구에서 이타성을 예측하는 것으로 알려져 있는 공감적 성향과 관련되어있으며 과제 수행 동기와 관련된 인지 욕구와는 상관이 없는 것으로 나타났다는 점은, 이 과제가 개인의 이타성을 측정하는 행동과제로 활용될 수 있는 가능성을 시사한다. 또한, 이타적 강화학습과제를 이용하여 사람들이 내집단원의 안녕에 상대적으로 더 많은 가치를 부여한다는 점을 확인함으로써 자기집단중심적 이타주의(Bernhard 등, 2006; Choi & Bowles, 2007)를 보여주는 추가적인 증거를 제시하였다.

이타적 강화학습과제는 사람들이 타인의 안녕에 가치를 부여하는 정도를 측정함으로써 개인의 이타성을 예측하는 행동지표를 제공하는 과제로 다양한 연구에 활용될 수 있을 것으로 기대된다. 이타적 강화학습과제는 기존 심리학 연구들에서 주로 사용되어온 자기 보고식 측정치나 표면적인 도움행동을 관찰하는 방법에 비해서 사회적 바람직성의 영향에 상대적으로 덜 취약하다는 장점이 있으며, 타인의 안녕에 대한 가치 부여와 이타성의 관련성을 바라보는 진화생물학, 행동 경제학, 심리학적 개념들을 연결 지어 공통된 메커니즘을 탐구할 수 있는 가능성을 열어 준다. 이타적 강화학습과제의 또 다른 장점은 뇌영상 연구에 적합한 시행 구조를 가졌다는 점이다. 강화학습 패러다임은 의사결정의 신경메커니즘에 관한 기존 뇌영상 연구들에서 선택에 필요한 정보들을 우리 뇌가 어떻게 계산하고 있는지를 알아보기 위하여 흔히 사용하는 패러다임으로, 시간차학습모형(Temporal difference learning model, Sutton & Barto, 1998)과 같은 계산 모형을 적용하여 가치의 신경표상을 확인하는 데에 편리한 방법을 제공한다(Kim 등, 2006). 따라서 본 연구에서 개발한 과제를 사용하여 타인의 안녕에 가치를 부여하는 과정의 신경메커니즘을 이해하는 뇌영상 연구를 수행하는 것도 가능할 것이다. 이타 행동과 관련된 보상 메커니즘을 이해하는 것은, 진화적으로 이타 행동이 어떻게 유지되고 증진되어왔는지에 대한 해답을 제공할 수 있을 것으로 기대된다. 본 연구에서는 이타적 강화학습 과제를 개발하고 유용성을 검토하는 데 그쳤으나 앞으로 이 과제를 이용한 여러 후속 연구들을 통하여 인간의 이타성에 대한 이해를 확장하고 이타성 증진을 위한 교육 방법 개발이나 학교 및 조직 장면에서의 효과적 보상 체계에 대한 새로운 통찰을 얻을 수 있기를 기대한다.

## 참고문헌

김미희, 김기범, & 차영란 (2005). 현실 및 가상공간에서의 집단범주화 방식과 상호작용 여부에 따른 집단성 지각 및 내집단 편애. 한국심리학회지: 사회 및 성격, 19(3), 37-54.

김용훈, 류리나, & 한성열 (2012). 도움행동을 높이기 위한 방안 모색: 공감과 공정성이 도움행동의도에 미치는 영향. 한국심리학회지: 문화 및 사회문제, 18(3), 349-366.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *Economic Journal, 100*, 464-477.

Batson, C. D. (2011). *Altruism in humans*. New York, NY: Oxford University Press.

Batson, C. D., Dyck, J. L., Brandt, J. R., Batson, J. G., Powell, A. L., McMaster, M. R., & Griffitt, C. (1988). Five studies testing two new egoistic alternatives to the empathy-altruism hypothesis. *Journal of Personality and Social Psychology*, *55*(1), 52-77.

Batson, C. D., Eklund, J. H., Chermok, V. L., Hoyt, J. L., & Ortiz, B. G. (2007). An additional antecedent of empathic concern: Valuing the welfare of the person in need. *Journal of Personality and Social Psychology*, *93*(1), 65-74.

Batson, C. D., Early, S., & Salvarani, G. (1997). Perspective taking: Imagining how another feels versus imaging how you would feel. *Personality and Social Psychology Bulletin*, *23*(7), 751-758.

Batson, C. D., Turk, C. L., Shaw, L. L., & Klein, T. R. (1995). Information function of empathic emotion: Learning that we value the other's welfare. *Journal of Personality and Social Psychology*, *68*(2), 300-313.

Becker, G. S. (1974). A theory of social interactions. *Journal of Political Economy, 82,* 10-83.

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature, 442* (7105), 912-915.

Bolton, G. E. & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity and competition. *American Economic Reivew, 90,* 166-193.

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin*, *86*(2), 307-324.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate? *Journal of Social Issues*, *55*(3), 429-444.

Buttelmann, D., & Böhm, R. (2014). The ontogeny of the motivation that underlies in-group bias. *Psychological Science, 25*(4), 921-927.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*(3), 306-307.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics, 113*, 817-869.

Choi, J. K., & Bowles, S. (2007). The coevolution of parochial altruism and war. *Science*, *318* (5850), 636-640.

Cialdini, R. B. (1991). Altruism or egoism? That is (still) the question. *Psychological Inquiry*, *2*(2), 124-126.

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS: Catalog of Selected Documents in Psychology, 10*, 85.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a

multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113-126.

Davis, M. H. (1994). *Empathy: A social psychological approach*. Madison, WI: Brown & Benchmark.

De Dreu, C. K., Greer, L. L., Handgraaf, M. J., Shalvi, S., Van Kleef, G. A., Baas, M., ... & Feith, S. W. (2010). The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. *Science*, *328*(5984), 1408-1411.

De Waal, F. B. M. (2008). Putting the altruism back into altruism: The evolution of empathy. *Annual Review of Psychology, 59*, 279-300.

Eisenberg, N., & Eggum, N. D. (2009). Empathic responding: Sympathy and personal distress. In J. Decety, & W. Ickes, (Eds.), *The social neuroscience of empathy* (pp.71-83). Cambridge, MA: MIT Press.

Eisenberg, N., Guthrie, I. K., Cumberland, A., Murphy, B. C., Shepard, S. A., Zhou, Q., & Carlo, G. (2002). Prosocial development in early adulthood: A longitudinal study. *Journal of Personality and Social Psychology*, *82*(6), 993-1006.

Eisenberg, N., & Miller, P. A. (1987). The relation of empathy to prosocial and related behaviors. *Psychological Bulletin*, *101*(1), 91-119.

Fehr, E., Bernhard, H., & Rockenbach, B. (2008). Egalitarianism in young children. *Nature*, *454* (7208), 1079-1083.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785-791.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*(3), 817-868.

García, J., & van den Bergh, J. C. (2011). Evolution of parochial altruism by multilevel selection. *Evolution and Human Behavior*, *32*(4), 277-287.

Harbaugh, W. T. (1998). What do donations buy?: A model of philanthropy based on prestige and warm glow. *Journal of Public Economics*, *67*(2), 269-284.

Harbaugh, W. T., Mayr, U., & Burghart, D. R. (2007). Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science*, *316*(5831), 1622-1625.

Hare, T. A., Camerer, C. F., Knoepfle, D. T., O'Doherty, J. P., & Rangel, A. (2010). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *The Journal of Neuroscience*, *30*(2), 583-590.

Halevy, N., Bornstein, G., & Sagiv, L. (2008). "In-group love" and "out-group hate" as motives for individual participation in intergroup conflict: A new game paradigm. *Psychological Science*, *19*(4), 405-411.

Halevy, N., Weisel, O., & Bornstein, G. (2012). "In group love" and "out group hate" in repeated interaction between groups. *Journal of Behavioral Decision Making*, *25*(2), 188-195.

Hein, G., Silani, G., Preuschoff, K., Batson, C. D., & Singer, T. (2010). Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping. *Neuron*, *68*(1), 149-160.

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup bias. *Annual Review of Psychology*,

*53*(1), 575-604.

Iacoboni, M. (2009). Imitation, empathy, and mirror neurons. *Annual Review of Psychology, 60,* 653-670.

Kim, H., Shimojo, S., & O'Doherty, J. P. (2006). Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain. *PLoS Biology*, *4*(8), e233.

Lamm, C., Batson, C. D., & Decety, J. (2007). The neural substrate of human empathy: Effects of perspective-taking and cognitive appraisal. *Journal of Cognitive Neuroscience*, *19*(1), 42-58.

Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics, 1,* 593-622.

Levine, M., Prosser, A., Evans, D., & Reicher, S. (2005). Identity and emergency intervention: How social group membership and inclusiveness of group boundaries shape helping behavior. *Personality and Social Psychology Bulletin*, *31*(4), 443-453.

Ma, Y., Wang, C., & Han, S. (2011). Neural responses to perceived pain in others predict real-life monetary donations in different socioeconomic contexts. *NeuroImage*, *57*(3), 1273-1280.

Macaulay, J., & Berkowitz, L. (Eds.). (1970). *Altruism and helping behavior: Social psychological studies of some antecedents and consequences*. New York, NY: Academic Press.

Maner, J. K., Luce, C. L., Neuberg, S. L., Cialdini, R. B., Brown, S., & Sagarin, B. J. (2002). The effects of perspective taking on motivations for helping: Still no evidence for altruism. *Personality and Social Psychology Bulletin*, *28*(11), 1601-1610.

Masten, C. L., Morelli, S. A., & Eisenberger, N. I. (2011). An fMRI investigation of empathy for 'social pain' and subsequent prosocial behavior. *NeuroImage*, *55*(1), 381-388.

Oswald, P. A. (1996). The effects of cognitive and affective perspective taking on empathic concern and altruistic helping. *The Journal of Social Psychology, 136*(5), 613-623.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review, 83*, 1281-1302.

Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology*, *16*(2), 235-239.

Singer, T. (2006). The neuronal basis and ontogeny of empathy and mind reading: review of literature and implications for future research. *Neuroscience & Biobehavioral Reviews*, *30*(6), 855-863.

Sobel, J. (2005). Interdependent preferences and reciprocity. *Journal of Economic Literature, 43*, 392-436.

Sober, E., & Wilson, D. S. (2014). 타인에게로: 이타 행동의 진화와 심리학. (설선혜, 김민우 옮김). 서울: 서울대학교출판문화원. (원서출판: 1999).

Stürmer, S., Snyder, M., Kropp, A., & Siem, B. (2006). Empathy-motivated helping: The moderating role of group membership. *Personality and Social Psychology Bulletin*, *32*(7), 943-956.

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning*. Cambridge, MA: MIT Press.

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American*, *223*(5), 96-102.

Yamagishi, T., & Mifune, N. (2009). Social exchange and solidarity: In-group love or out-group hate? *Evolution and Human Behavior*, *30*(4), 229-237.

1차원고접수 : 2014. 04. 20.
수정원고접수 : 2014. 06. 21.
최종게재결정 : 2014. 06. 22.

한국심리학회지 : 일반
*Korean Journal of Psychology: General*
2014, Vol. 33, No. 2, 467-489

# Measuring Individual Differences in Altruism with Altruistic Learning Task

Sunhae Sul Minwoo Lee Hackjin Kim

Department of Psychology, Korea University

Altruism requires representing and valuing others' welfare. In the present study, we designed the Altruistic Learning task (AL task) to measure individual differences in valuing another person's welfare and examined its relationship with dispositional empathy (Experiment 1) as well as ingroup bias/parochial altruism (Experiment 2). In Experiment 1, participants performed the AL task in which they made choices between a pair of stimuli with different reward probabilities (30% vs. 70%) to reduce the amount of stress for themselves or their peer participants. The task consisted of three within-subject conditions (i.e. SELF, BOTH, and OTHER conditions) where different types of outcome (i.e. points for self, for both, and for other) were associated with three different pairs of stimuli. The amount of value that each participant attached to a given outcome can be measured with the frequency of choosing the higher reward probability option (HRP option). The results showed that participants who scored higher in personal distress and empathic concern measured by IRI (Davis, 1983) were more likely to choose the HRP option in the BOTH and OTHER conditions, indicating the role of emotional empathy iin the process of valuing other-regarding outcomes. In Experiment 2, we employed the minimal group paradigm (Tajfel, 1970) and compared choices for an ingroup member with those for an outgroup member in the AL task. We employed the same experimental design as the Experiment 1, except for having INGROUP and OUTGROUP conditions, instead of previous BOTH and OTHER conditions. The choice frequency for the HRP option was higher in the SELF and INGROUP conditions than the OUTGROUP condition, indicating ingroup bias of the participants in valuing welfare of others. Our findings demonstrated that the AL task can be a useful and valid measure of individual differences in altruism.

*Key words : reinforcement learning, altruism, empathy, minimal group, ingroup bias*