인류학석사학위논문

불평등과 보복 가능성이 협력과 처벌에 미치는 영향

2018년 2월

서울대학교 대학원
인류학과 인류학전공
황 준

불평등과 보복 가능성이 협력과 처벌에 미치는 영향

지도교수 박 순 영

이 논문을 인류학석사 학위논문으로 제출함.

2017년 12월

서울대학교 대학원

인류학과 인류학 전공

황 준

황준의 인류학석사 학위논문을 인준함.

2018년 1월

위 원 장 강 윤 희 (인)

부위원장 박 순 영 (인)

위 원 최 정 규 (인)

국 문 초 록

무임승차자, 또는 배신자에 대한 사적 처벌은 집단 내에서 협력을 증진시킨다고 알려져 있다. 그러나 무임승차자가 아닌 협력적인 사람을 대상으로 하는 반사회적 처벌이 발생하거나, 과도한 처벌 사용으로 인해 협력으로 발생한 이익이 소멸되는 등, 협력의 진화에 있어 처벌이 갖는 한계 역시 여러 차례 지적된 바 있다. 본 연구에서는 특히 구성원 사이의 불평등과 처벌에 대한 보복 가능성이 공공재 게임에서 나타나는 협력과 처벌 행위에 미치는 영향에 주목하였다.

모든 참여자들에게 동등한 양의 자원을 지급하는 일반적인 공공재 게임에서와 달리 참여자들에게 자원을 차등적으로 지급한다면, 자신에게 불리한 불평등을 기피하고자 하는 성향으로 인해 더 적게 가진 사람이 더 많이 가진 사람을 반사회적으로, 즉 상대가 무임승차하지 않았음에도 불구하고 처벌하는 양상을 보다 선명하게 관찰할 수 있으리라 판단하였다. 이를 통해 반사회적 처벌이 불평등 기피 성향과 밀접한 연관을 가짐을 보이고자 하였다. 그러나 여기서 한 걸음 더 나아가, 인간이 살아가는 현실 세계에서는 단순히 누군가 더 많이 가지고 있다는 이유만으로 그 사람에게 해를 입히는 일이 쉽게 발생하지 않는다는 사실에도 주목하였다. 더 많이 가진 사람은 자신에게 피해를 입힌 사람에게 더 강하게 보복할 수 있기 때문이다. 즉 현실에서는 보복 가능성으로 인해 불평등 기피 성향에 기반을 둔 반사회적 처벌이 억제된다는 것이다. 그러나 참여자들에게 단 한 차례의 처벌 기회만을 부여하는 일반적인 공공재 게임은 이러한 보복 가능성을 원천적으로 배제하고 있다는 점에서 비현실적이라고 할 수 있다.

본 연구에서는 이러한 논의를 반영하여, 평등 조건과 불평등 조건, 보복 불가능 조건과 보복 가능 조건을 조합한 2*2 공공재 게임을 진행하였다. 참여자들은 먼저 매 라운드 모두에게 동등한 양의 토큰이 지급되는 평등 공공

재 게임을 수행한 후, 이 게임에서 최종적으로 획득한 수익에 따라 토큰이 차등적으로 지급되는 불평등 공공재 게임을 수행하였다. 또한 참여자들을 두 집단으로 나누어 집단 1에 속한 참여자들은 처벌이 1회만 가능한 보복 불가능 공공재 게임을, 집단 2에 속한 참여자들은 처벌이 발생할 경우 처벌 단계가 반복되어 복수의 처벌이 가능한 보복 가능 공공재 게임을 수행하였다.

실험 결과, 불평등 기피 성향에 기반을 둔 반사회적 처벌이 보복 가능성의 도입에 따라 억제되리라는 예측은 타당한 것으로 검증되었다. 그러나 실험을 기획하는 단계에서 연구자가 미처 고려하지 못했거나 예상과 어긋난 결과들을 관찰할 수 있었다. 먼저, 보복 가능성을 도입하자 반사회적 처벌과 더불어 무임승차자를 대상으로 하는 정당한 처벌 역시 크게 감소하였다. 즉, 보복이 가능한 보다 현실적인 상호작용 환경에서는 참여자들 사이의 사적 처벌이 거의 사용되지 않는 것으로 나타났다. 더 나아가, 평등 조건에서 참여자들의 협력 수준은 보복 가능할 때와 불가능할 때 차이가 없었지만 보복 불가능 조건에서 훨씬 더 많은 처벌이 사용되었다. 즉 보복 불가능-평등 조건에서는 지나치게 많은 자원이 처벌에 사용됨으로써 협력으로 인해 생산된 이익이 전부 파괴되었지만, 보복 가능-평등 조건에서는 처벌이 거의 발생하지 않았음에도 유사한 수준의 협력 수준이 유지되어 결과적으로 협력 이익이 보존되었다.

이와 같은 결과는 협력의 진화에 있어 개체들 간의 사적 처벌이 수행했을 역할을 새로운 각도에서 바라보게 해준다. 인간을 협력하게 만든 것은, 실제로 발생한 처벌보다는 서로에게 처벌당할 수도 있다는 두려움, 혹은 그 위험에 대한 인식이라는 것이다. 보복 가능성은 무임승차를 억제하여 높은 협력 수준의 유지에 기여하는 동시에, 과도한 처벌 사용을 억제하여 협력으로 발생한 이익의 보존에 기여하였다.

주요어 : 협력, 처벌, 불평등, 보복, 공공재 게임
학 번 : 2015 - 22542

목 차

* 이 논문은 2015년 정부재원(교육부 BK21플러스 사업비)으로 한국연구재단의 지원을 받아 연구되었음 (No. 21B20151813155)

표 목차

그림 목차

1. 서론

1.1. 연구 목적

무임승차자, 또는 배신자에 대한 사적 처벌(peer punishment)은 일반적으로 집단 내에서 협력을 증진시킬 수 있는 방법 중 하나로 여겨져 왔으며, 이는 다양한 연구들에서 여러 차례 밝혀진 바 있다(Clutton-Brock & Parker 1995, Dreber *et al.* 2008, Egas & Riedl 2008, Fehr & Gächter 2000; 2002 Henrich *et al.* 2006). 이때, 사적 처벌이란 중앙화된 기관이나 제도에 의해 수행되는 공적인 제재가 아닌, 개인이 직접 비용을 감수하여 특정한 상대에게 피해를 입히는 행위를 의미한다. 그러나 사적 처벌의 대상이 언제나 무임승차자로 한정되지는 않으며, 오히려 협력적인 사람을 대상으로 하는 반사회적 처벌의 빈도가 결코 낮지 않다는 연구 결과(Dreber & Rand 2012, Dreber *et al.* 2008, Herrmann *et al.* 2008, Janssen & Bushman 2008, Nikiforakis 2008, Rand & Nowak 2011)들은 처벌이 무임승차에 대한 반감뿐만 아니라 다른 동기들에 의해 행해질 가능성을 시사한다. 예컨대, 불평등 기피 성향(inequality aversion)은 인간과 영장류의 처벌 행위를 촉발하는 주요 요인으로 지목된 바 있다(Dawes *et al.* 2007, Johnson *et al.* 2009, Leimgruber *et al.* 2016, Raihani & McAuliffe 2012).

소비가 비경합적이고 배제 불가능한 공공재의 경우, 언제나 무임승차자가 협력자보다 더 많은 이익을 얻는 불평등이 발생하므로 무임승차자에 대한 처벌은 바로 이러한 불평등을 제거하기 위해 행해질 수 있다. 모든 구성원들에게 같은 양의 자원이 지급되는 기존의 공공재 게임 구조에서, 불평등의 발생은 무임승차 행위와 밀접한 관계를 가질 수밖에 없다. 따라서 무임승차자에 대한 처벌이 이루어질 때, 그것이 무임승차에 대한 반감에 의한 것인지 아니면 그로 인해 발생하는 불평등을 줄이기 위한 것인지 구분하기 힘들다는 한계가 존재한다. 이에 본 연구에서는 기존의 공공재 게임에서와는 달리 구성원들에게 자원을 차등적으로 지급함으로써 처벌을 야기하는 두 가지 동기(무

임승차에 대한 반감, 불평등 기피 성향)를 구분해보고자 했다. 자원을 차등적으로 지급할 경우 무임승차 여부와는 별개로 구성원들 사이에 불평등이 존재하게 된다. 이처럼 무임승차와 무관하게 불평등이 발생한 상황에서도 더 많이 가진 사람을 대상으로 하는 처벌이 발생한다면, 무임승차에 대한 반감보다는 불평등 기피 성향이 처벌의 동기로서 더 강하게 작용한다고 할 수 있을 것이다.

그러나 단순히 더 적게 가진 사람의 불평등 기피 성향을 자극하는 것을 넘어서, 불평등이 갖는 보다 근본적인 의미는 바로 더 많이 가진 사람이 더 적게 가진 사람에게 막대한 영향력을 미칠 수 있다는 것이다. 현실에서 어떠한 형태로든 더 많은 자원을 가진 사람은 쉽게 질투의 대상이 되지만 그렇다고 해서 그 질투로 인해 더 많이 가진 사람이 직접적으로 해를 입는 일은 쉽게 관찰되지 않는데, 그 이유는 더 많이 가진 사람은 자신에게 피해를 입힌 사람에게 훨씬 더 큰 보복을 가할 수 있기 때문이다. 즉 불평등 기피 성향으로 촉발될 수 있는 처벌이 보복 가능성으로 인해 억제된다는 것이다. 따라서 불평등이 인간의 행위에 미치는 영향을 보다 현실적으로 이해하기 위해서는 기존의 공공재 연구(Fehr & Gächter 2000; 2002)에서처럼 참여자들에게 단 한 차례의 처벌 기회만을 부여해서는 안 되며, 자신을 처벌한 사람에게 다시금 처벌을 가할 수 있게, 즉 보복이 가능하게끔 복수의 처벌 기회를 부여(Balafoutas *et al.* 2014, Denant-Boemont *et al.* 2007, Engelmann & Nikiforakis 2015, Fehl *et al.* 2012, Janssen & Bushman 2008, Nikiforakis 2008, Nikiforakis *et al.* 2012, Wolff 2012)해야 한다.

종합하자면, 연구자는 크게 두 가지 요인, 불평등과 보복 가능성이 인간의 협력과 처벌에 미치는 영향을 파악하기 위하여 상이한 조건에서 공공재 게임을 총 네 차례 (평등/불평등 * 보복 불가능/보복 가능) 수행하였다. 참여자들은 먼저 모두가 동등한 양의 토큰을 지급받는 평등 공공재 게임을 한 차례 수행한 후, 이 게임에서 얼마나 많은 수익을 얻었는지에 따라 차등적으로 토큰이 지급되는 불평등 공공재 게임을 수행하였다. 이를 통해 불평등의 도입이 인간의 협력과 처벌에 미치는 영향을 파악하고자 하였다. 특히, 불평등 기피 성향이 무임승차를 하지 않은 사람을 대상으로 하는 처벌 행위에 미치

는 영향에 주목하였다. 또한 참여자들은 두 집단으로 나뉘었는데, 이때 집단 1에 속한 참여자들은 처벌 기회가 단 한 차례만 주어져 처벌에 대한 보복이 불가능한 공공재 게임을, 집단 2에 속한 참여자들은 처벌이 발생할 경우 처벌 단계가 반복되어 처벌당한 사람이 처벌에 대한 보복 기회를 갖는 공공재 게임을 수행하였다. 이러한 두 집단의 비교를 통해 불평등과 더불어 보복 가능성이 협력과 처벌에 미치는 영향을 파악하고자 하였다. 이러한 실험 설계를 통해 본 연구에서는 만약 불평등 조건에서 불평등 기피 성향으로 인해 상대가 무임승차를 하지 않았음에도 처벌이 행해진다면, 보복 가능성의 도입에 의해 이러한 처벌이 억제되는지 검증해보고자 하였다.

1.2. 선행연구 검토

1.2.1. 협력의 진화에 있어 처벌의 기능과 그 한계

"인간 사회는 협력으로 가득하다(노왁 & 하이필드 2012)." 우리가 일상적으로 수행하는 아주 간단한 일들, 예를 들어 카페에서 커피를 사먹는 행위조차도 우리가 상상하기 힘든 대규모의 협력을 전제로 한다. 남미의 농장에서 커피콩을 재배한 농부, 그것을 국내로 유통한 수입업체, 그리고 그 콩을 갈아서 커피로 만든 카페 직원에 이르기까지(노왁 & 하이필드 2012), 협력의 범위는 손쉽게 국경을 초월한다.

그리고 인간의 대규모 협력은 반복되지 않는 일회적 상호작용 관계에서도, 혈연으로 이어지지 않은 사람들 사이에서도, 심지어 협력으로 인한 좋은 평판 축적 혹은 신호 전달 효과가 미미한 상황에서도 발생한다(Fehr & Gächter 2002)는 점에서 진화적 수수께끼라고 할 만하다. 무임승차자에 대한 처벌은 이러한 수수께끼에 대한 해답을 제시한다(Fehr & Gächter 2002). 처벌은 무임승차로 인해 발생하는 상대적 이익을 제거하여 무임승차자가 협력자보다 큰 이익을 얻는 것을 방지하고, 향후 상호작용에서 협력을 이끌어내는 역할을 수행한다. Fehr와 Gächter의 공공재 게임 실험 연구에서 참여자들은 처벌이 없는 환경에서는 완전한 무임승차를 지배적인 전략으

로 사용했지만 처벌이 있는 환경에서는 적극적으로 무임승차자를 처벌했으며 그로 인해 높은 수준의 협력이 유지될 수 있었다(Fehr & Gächter 2000).

그러나 처벌이 협력의 진화에 있어 주요한 원동력으로 작용했다는 주장을 비판하는 목소리도 결코 적지 않다. 먼저, 노왁과 하이필드(2012)는 처벌이 남용됨에 따라 그 비용으로 인해 협력으로 인한 이익이 소멸된다고 주장하였다. 실제로, 처벌이 가능한 환경에서 협력 수준은 증가했으나 개별 참여자가 처벌에 사용한 비용으로 인해 수익은 감소(Egas & Riedl 2008, Nikiforakis & Normann 2008)했으며, 오히려 개별 참여자의 처벌 사용 빈도는 수익과 반비례(Dreber *et al.* 2008)하는 것으로 나타났다. 간단하게 표현하자면, (진화적 경쟁에서) "승자는 처벌하지 않는다(노왁 & 하이필드 2012)"는 것이다. 한편, 이러한 비판을 수용하여 Boyd 등(2010)은 무조건적인 처벌 전략 대신 처벌을 수행하기 전에 다른 집단 구성원들의 처벌 의사를 미리 확인하고, 처벌자의 수가 많을 때에만 함께 처벌을 수행함으로써 처벌의 비용을 분산하고 협력의 이익을 보존하는 '조건적 처벌' 전략이 협력의 진화에 기여했을 가능성을 제기하였다.

처벌의 과도한 사용으로 인한 협력 이익의 소멸 외에도, 무임승차자가 아닌 협력자를 향한 반사회적 처벌(Dreber & Rand 2012, Hilbe & Traulsen 2012, Rand & Nowak 2011)이 결코 낮지 않은 빈도로 발생한다는 점 역시 처벌의 한계로 지적된 바 있다. Dreber와 Rand(2012)에 따르면, 처벌이 협력의 진화에 기여함을 보였던 기존 연구들은 협력자가 처벌의 대상이 될 수 없게끔 미리 제한하여 반사회적 처벌을 분석에 포함할 수 없었다는 한계를 갖는다. 만약 무임승차자가 협력자를 대상으로 가하는 반사회적 처벌이 빈번하게 발생한다면 협력자가 무임승차자에 비해 상대적으로 더 높은 적합도를 갖는다고 담보할 수 없게 된다. 특히 Rand와 Nowak(2011)의 시뮬레이션 실험에서 무임승차자와 독행자(loner)는 상대의 무임승차를 처벌로 응징하는 협력자가 집단 내로 침투하는 것을 방지하기 위하여 협력자에게 강도 높은 반사회적 처벌을 가하는 것으로 나타났다.

1.2.2. 불평등 기피 성향과 반사회적 처벌

공공재 게임이나 죄수의 딜레마 게임 등에서 처벌은 흔히 효율적이며, 값비싼 차감(costly reduction)의 형태를 갖는 것으로 상정된다. 처벌이 값비싼 차감이라는 것은 상대방에게 피해를 입히고 싶다면 나 또한 어느 정도 비용을 감수해야 한다는 것을, 효율적이라는 것은 내가 지불하는 비용보다 상대방이 입는 피해가 더 크다는 것을 의미한다. 일반적으로 게임 상에서 처벌은 스스로 n의 비용을 감수하여 지정한 상대의 자원을 3n 만큼 깎는 것으로 설정(Fehr & Gächter 2000)되어 있다.

몇몇 연구자들은 처벌의 이러한 특징에 주목하여 처벌의 동기가 상대의 협력 규범 위반뿐만 아니라 무임승차로 인한 불평등의 심화에도 있다는 '불평등 기피(inequality aversion) 가설'을 주장하였다(Dawes *et al.* 2007, Johnson *et al.* 2009, Leimgruber *et al.* 2016, Raihani & McAuliffe 2012). 처벌은 무임승차자의 향후 협력을 이끌어낼 수 있을 뿐만 아니라 그 효율성으로 인해 무임승차로 나보다 앞서 나간 상대를 다시 나와 같은 수준으로 끌어내리는, 다시 말해 적합도 격차를 조정하는 효과를 낳기 때문이다. 사람들은 자신이 배신을 당했다 하더라도 결과적으로 배신자에 비해 자신이 가진 돈이 더 많은 상황에서는 배신자를 잘 처벌하지 않았으며, 배신 행위로 인해 우열이 역전될 때 높은 빈도로 배신자를 처벌하는 것으로 나타났다(Raihani & McAuliffe 2012). 또한 정해진 양의 금액이 집단 내의 참여자들에게 무작위로 분배되는 '무작위 소득 게임'에서 사람들은 더 많이 가진 사람의 소득을 감소시키는 일에, 또는 적게 가진 사람의 소득을 증가시키는 일에 자신의 자원을 사용하는 것으로 나타났다(Dawes *et al.* 2007). 더 나아가 무작위 소득 게임에서 평등한 결과를 도출하기 위해 더 많은 자원을 투자한 사람이 공공재 게임에서도 무임승차를 처벌하는 데 더 많은 비용을 지불한다는 결과(Johnson *et al.* 2009) 역시 처벌의 동기에 불평등을 기피하는 성향이 있음을 잘 보여준다.

반면 처벌이 효율적이지 않은 상황(처벌의 비용과 피처벌자의 피해가 1:1)에서도 사람들이 결코 낮지 않은 빈도로 무임승차자를 처벌했다는 연구 결과(Bone & Raihani 2015)를 통해 협력 규범 위반(무임승차)에 대한 반감을

처벌의 동기에서 완전히 배제한 채 불평등 기피 성향만으로 처벌의 동기를 완전히 설명하는 것은 불가능하다는 비판이 제기되기도 했다. Marczyk(2017) 또한 양자 사이의 불평등이 어느 한 쪽에게 비용을 초래함으로써 발생한 것이 아니라면 불평등으로 인해 처벌이 촉발되지 않는다는 것을 보인 바 있다. 불평등 상황에서 처벌이 발생하는 것은 사람들이 불평등을 기피해서라기보다는 그 불평등으로 인해 손해를 입었기 때문이라는 것이다.

하지만 불평등이 어느 한 쪽에 손해를 입히지 않고 발생했다 하더라도, 진화적 관점에서 불평등의 지속은 적합도 손해로 이어질 수 있다. 다수로 이루어진 집단에 속한 두 사람 사이에 불평등이 발생하여 한 사람이 불리한 위치에 놓인다면, 비록 그로 인해 직접적인 손해를 입지 않았다 해도 이는 곧 집단 내에서 그의 지위가 한 단계 하락했다는 것을 의미하기 때문이다. 따라서 참여자들이 집단에 속하지 않은 채 오직 한 명의 상대와 단 한 차례 상호작용하고, 성과와 무관하게 고정된 금액의 보상을 받았던 Marczyk(2017)의 실험은 불평등이 진화적 맥락에서 사회적 동물들에게 미치는 영향을 온전히 반영하지 못했다는 한계를 갖는다.

또한 처벌의 동기에서 무임승차에 대한 반감을 완전히 배제할 수 없다는 Bone과 Raihani(2015)의 비판은 분명 타당하지만, 전체 처벌이 아닌 반사회적 처벌에 있어서는 불평등 기피 성향이 미치는 영향을 보다 확실하게 관찰할 수 있으리라 판단된다. 무임승차자를 대상으로 하는 정당한 처벌이라면 처벌의 동기에서 무임승차와 불평등 기피 성향을 분리하는 것이 어렵겠지만 협력자를 대상으로 하는 반사회적 처벌에서는 무임승차에 대한 반감을 처벌의 동기에서 배제할 수 있기 때문이다.

위의 분석을 종합하여 본 연구에서는 참여자들에게 자원을 평등하게 지급한 조건과 차등적으로 지급한 조건에서 처벌이 가능한 공공재 게임을 수행함으로써 불평등의 도입이 처벌에 미치는 영향을 파악하고자 하였다. 특히 불평등이 심화되면 불평등 기피 성향으로 인해 자원이 적은 사람의 반사회적 처벌이 매우 높은 빈도로 수행되리라 예측하였다.

1.2.3. 보복 가능성이 협력과 처벌에 미치는 영향

공공재 게임에서 처벌의 도입이 협력의 증진에 기여함을 보인 연구(Fehr & Gächter 2000; 2002)에서 처벌은 일반적으로 단 한 차례 수행되며, 누가 자신을 처벌했는지 알 수 없게 되어 있다. 이러한 처벌의 일회성과 익명성이 문제가 되는 이유는, 그로 인해 보복의 가능성이 원천적으로 봉쇄되기 때문이다. A가 B를 처벌하고자 하는 상황에서, B는 자신을 처벌한 것이 A라는 사실을 모르고, 그 처벌이 이루어진 후 B에게 더 이상 처벌할 기회가 주어지지 않는다면 A는 별다른 고민 없이 B를 처벌할 것이다. 즉 보복이 원천적으로 불가능한 게임 구조에서 사람들은 현실에서보다 쉽게 다른 사람을 처벌하게 된다.

그러나 실험실 밖의 현실에서 보복은 얼마든지 발생할 수 있다. 셰익스피어의 작품 "로미오와 줄리엣"에 등장하는 몬태규 가문과 캐퓰릿 가문, 그리고 춘추전국시대 오나라와 월나라 사이의 갈등은 보복이 꼬리에 꼬리를 물고, 심지어 한 세대를 넘어 다음 세대로 이어질 수 있음을 잘 보여준다. 이 사례들과 같은 연쇄적인 보복은 범문화적 현상으로 알려져 있으며(Ericksen & Horton 1992), 처벌을 전담하는 제도가 부재하거나 그 힘이 약할 때 더욱 빈번하게 발생하는 것으로 밝혀졌다(Elster 1990).

만약 공공재 게임 상황에서도 보복이 가능하다면, 처벌의 협력 증진 또는 무임승차 억제 효과는 상당히 감소할 것으로 예상할 수 있다. 처벌이 보복을 불러올 수 있다면, 보복에 대한 두려움으로 인해 처벌은 감소할 것이고, 처벌이 감소한다면 결과적으로 무임승차에 대한 억제력이 사라져 무임승차가 증가할 것이기 때문이다. 이러한 예측은 처벌에 대한 보복이 가능한 공공재 게임 실험(Nikiforakis 2008)과 3자 처벌 게임 실험(Balafoutas *et al.* 2014), 시뮬레이션 실험(Janssen & Bushman 2008, Wolff 2012)을 통해 확인된 바 있다. 특히 Wolff(2012)의 시뮬레이션 실험에서 처벌자(punisher)는 초기에는 협력하지 않지만 처벌을 받으면 협력하도록 전략을 바꾸는 소극적 협력자(reluctant cooperator)들을 처벌하여 협력 수준의 증진을 이끌어내지만, 그 이후 처벌에 대해 무조건 보복하는 보복자(retaliator)의 번성으로 인해 사라지는 것으로 나타났다. 즉 보복이 가능한

환경에서 처벌의 역할은 최초의 협력 발생을 유도하는 부싯돌에 지나지 않는다는 것이다.

그러나 위의 두 연구는 보복의 기회를 단 한 차례 부여했다는 점에서 처벌을 단 한 차례만 가능케 했던 기존의 처벌 연구들과 유사한 한계를 갖는다. 단 한 차례만 처벌이 가능한 상황에서 처벌이 협력을 증진하는 데 효과적이었던 것은 보복 당할 두려움 없이 처벌이 가능했기 때문이다. 마찬가지로, 보복이 단 한 차례만 가능하다면 다시 보복 당할 두려움 없이 자신을 처벌한 사람에게 보복할 수 있고, 그로 인해 처벌을 억제하는 효과를 낳을 수 있다. 그러나 처벌이 단 한 차례로 끝나는 것이 현실적이지 않듯, 보복이 단 한 차례로 끝나는 것 또한 현실적이지 않다. 따라서 현실적으로 처벌과 보복이 협력에 미치는 영향을 파악하기 위해서는, 춘추전국시대 오나라와 월나라 사이의 적대적 관계가 잘 보여주듯이 처벌이 보복을 낳고 보복이 또 다른 보복을 낳는 연쇄가 가능해야 한다. 이를 반영하여 처벌에 대한 보복뿐만 아니라 보복에 대한 보복이 가능하게끔 참여자들에게 복수의 처벌 기회를 부여할 경우 참여자들의 기여량은 증가(Fehl *et al.* 2012)하거나 처벌이 1회 가능할 때와 유사한 수준을 유지(Engelmann & Nikiforakis 2015)했지만, 처벌과 보복의 과도한 사용으로 인해 협력으로 발생한 이익이 소멸(Denant-Boemont *et al.* 2007, Nikiforakis *et al.* 2012)되는 것으로 나타났다.

위와 같은 이론적 배경을 바탕으로, 본 연구 또한 처벌이 1회 가능할 때와 여러 차례 가능할 때 공공재 게임에서 참여자들의 행위가 어떠한 차이를 보이는지 관찰하고자 했다. 다만 참여자가 매 라운드 같은 사람들과 상호작용하게끔 같은 조를 이루는 구성원들을 고정(동반자 조건)했던 연구(Engelmann & Nikiforakis 2015, Denant-Boemont *et al.* 2007, Nikiforakis *et al.* 2012)들과 달리 본 연구에서는 참여자들이 서로 다른 참여자들과 상호작용하게끔 매 라운드 조 구성원을 무작위로 재배치(이방인 조건)하였다. 기존 연구들과 달리 이방인 조건을 선택한 이유는 동반자 조건에서 보복이 가능한 공공재 게임을 진행할 경우 보복이 다음 라운드의 처벌이나 무임승차로 이연되어 처벌과 협력을 교란할 수 있다고 판단하였기 때문

이다.

보복이 가능할 경우, 동반자 조건에서 처벌을 받은 참여자는 이어지는 처벌 단계에서 곧바로 보복하는 것 외에도 여러 가지 방식으로 보복할 수 있다. 어차피 같은 사람들과 반복적으로 만난다는 것을 이용해 몇 라운드가 지난 후에 처벌하거나 그 다음 라운드의 기여 단계에서 무임승차를 선택함으로써 간접적으로 보복하는 것도 가능하다. 혹은 이러한 방법들을 전부 이용하여 보복할 수도 있다. 따라서 동반자 조건을 채택한 Denant-Boemont 외(2007), Engelmann과 Nikiforakis(2015) 그리고 Nikiforakis 외(2012)의 연구에서 나타난, 강도 높은 처벌과 보복으로 인해 협력 이익이 소멸되는 양상은 처벌과 보복 그 자체의 문제라기보다 보복이 처벌과 협력을 교란하여 발생한 문제일 가능성도 배제할 수 없는 것이다. 반면 이방인 조건에서는 이와 같은 교란을 쉽게 제거할 수 있다. 이번 라운드에서 받은 처벌에 대해 보복하는 방법은 단 하나, 이어지는 처벌 단계에서 자신을 처벌한 사람에게 보복하는 것뿐이다. 다음 라운드에서는 조가 무작위로 재배치되므로 이번 라운드에서 받은 처벌에 대한 보복을 그 다음 라운드로 이연하는 것이 불가능하기 때문이다.

1.3. 연구 가설

위와 같은 선행연구 검토를 통해 본 연구에서는 다음과 같은 질문들에 대한 답을 구하고자 한다. 먼저, '불평등 기피 성향은 무임승차와 관계없는 불평등 상황에서도 처벌을 이끌어내는가?' 만약 그렇다면, '기존의 공공재 게임에 어떠한 요인을 새로이 도입했을 때 불평등 기피 성향이 반사회적 처벌로 이어지는 것을 억제할 수 있을 것인가?'

이와 같은 연구 질문은 아주 단순한 문제의식으로부터 출발하였다. 그 문제의식은, 현실에서는 단순히 더 많이 가졌다는 이유만으로, 혹은 더 뛰어나다는 이유만으로 누군가에게 부당한 해를 입는 일을 쉽게 관찰할 수 없다는 것이었다. 물론, 현실에서도 더 많이 가졌거나 더 뛰어난 사람은 질투나 적의의 대상이 되기 쉽고, 이는 인간의 불평등 기피 성향을 반영한 것이라 할

수 있다. 그러나 이러한 질투와 적의가 직접적인 가해 행위로 이어지는 것은 전혀 다른 문제다. 오히려 그 사람에게 질투와 적의를 느끼면서도 막상 면전에서는 우호적인 자세를 취하거나 가까이 지내고자 노력하는 것이 더 현실에 가까운 모습이라고 할 수 있을 것이다.

그 이유는 보복 가능성에서 찾을 수 있다. 더 많이 가졌거나 더 뛰어난 사람에게 쉽게 해를 입힐 수 없는 이유는, 그 사람이 더 큰 보복을 가할 수 있기 때문이라는 것이다. 신체적 능력이나 물질적 부, 집단 내에서의 위세 등 어떠한 형태로든 더 많은 자원을 보유했다는 것은 그 사람이 더 큰 영향력을 행사할 수 있다는 것을 의미한다. 즉 인간이 살아가는 현실 세계에서 더 많이 가진 사람은 더 적게 가진 사람에게 질투의 대상인 동시에 적으로 돌리고 싶지 않은 사람이라는 것이다. 이러한 맥락에서, 불평등 기피 성향으로 인해 촉발되는 처벌을 다룬 기존의 실험 연구들이 대부분 불평등 상황에서 처벌 기회를 1회로 (보복이 불가능하도록) 제한한 것은 참여자들이 더 많이 가진 사람을 처벌하는 것에 대해 느낄 수 있는 심리적 부담감을 제거한 것이나 다름없다.

따라서 기존 공공재 게임의 구조에 보복 가능성을 추가하여 복수의 처벌 기회를 부여한다면, 불평등이 인간의 협력과 처벌 행위에 미치는 영향을 보다 현실에 가깝게 파악할 수 있으리라 예상된다. 또한 보복 가능성이 갖는 의미를 보다 분명하게 드러내기 위해서 보복이 불가능한 기존의 공공재-처벌 게임 구조에서 불평등을 도입했을 때와 보복이 가능한 공공재-처벌 게임 구조에서 불평등을 도입했을 때 협력과 처벌에서 어떠한 차이가 나타나는지 관찰하였다.

지금까지의 논의를 정리하여, 본 연구에서 검증하고자 하는 가설을 한 문장으로 정리하면 아래와 같다.

가설: 구성원간 불평등이 도입되었을 때 보복이 불가능하다면 불평등 기피 성향은 더 많이 가진 사람을 대상으로 하는 반사회적 처벌로 이어지지만, 보복이 가능하다면 불평등 기피 성향에 기반을 둔 반사회적 처벌은 억제될 것이다.

2. 연구 방법

2.1. 실험 개요

서울대학교 커뮤니티 사이트인 스누라이프를 통해 56명의 서울대학교 구성원(학부생, 대학원생 포함)을 모집하였다. 총 두 차례에 걸쳐 참여자를 모집했으며 먼저 모집된 32명을 집단 1, 후에 모집된 24명을 집단 2로 분류하였다. 집단 1은 남자 20명, 여자 12명으로 구성되었고, 평균연령은 24.91세였다. 집단 2는 남자 12명, 여자 12명으로 구성되었고, 평균연령은 25.46세였다.

	집단 1 (9/26, 32명)	집단 2 (9/30, 24명)
전반부	보복 불가능 / 평등 공공재-처벌 게임 (20R)	보복 가능 / 평등 공공재-처벌 게임 (15R)
후반부	보복 불가능 / 불평등 공공재-처벌 게임 (20R)	보복 가능 / 불평등 공공재-처벌 게임 (20R)

〈표 2-1〉 실험 구조와 일정

집단 1 참여자는 2017년 9월 26일에 보복이 불가능한 공공재-처벌 게임을 모두에게 동등한 양의 토큰을 지급한 평등 조건에서 1회 수행한 후, 이때 획득한 수익에 따라 토큰을 차등적으로 지급한 불평등 조건에서 1회 수행하였다. 집단 2 참여자는 2017년 9월 30일에 처벌에 대한 보복이 가능한 공공재 게임을 집단 1과 같은 방식으로 총 두 차례 수행하였다. 집단 1에 속한 32명의 참여자들은 전반부의 평등 공공재-처벌 게임과 불평등 공공재-처벌 게임을 각각 20라운드 수행하였고, 집단 2에 속한 24명의 참여자들은 전반부의 평등 공공재-처벌 게임을 15라운드[1], 후반부의 불평등 공공재-처벌

1) 총 20R 수행된 다른 게임들과 달리 집단 2 전반부의 보복 가능-평등 공공재 게임을 15R 수행한 까닭은 총 실험 시간이 2시간으로 제한된 상황에서 보복 가능성의 도입으로 인해 이론적으로 실험 소요 시간이 무한하게 연장될 수 있기 때문이었다.

게임을 20라운드 수행하였다. 참여자들에게는 게임이 총 몇 라운드 진행되는지 알려주지 않았다. 이를 통해 일반적인 공공재-처벌 게임의 구조(집단 1 전반부)에서 보복 가능성이 추가되었을 때의 결과(집단 2 전반부)와 불평등이 추가되었을 때의 결과(집단 1 후반부), 보복 가능성과 비대칭성이 모두 추가되었을 때의 결과(집단 2 후반부)를 비교하였다(표 2-1).

실험에 소요된 시간은 두 집단 모두 두 시간을 넘지 않았다. 모든 실험이 종료된 후, 참여자들에게는 기본 참여비 5000원과 후반부 불평등 공공재-처벌 게임에서 획득한 금액(교환비 1:1)이 지급되었다.

2.2. 실험 설계

Z-tree 소프트웨어(Fischbacher 2007)를 사용하여 참여자 4명이 1개 조를 이루는 공공재-처벌 게임을 설계하였다. 참여자들은 서울대학교 사회과학대학의 전산실에서 컴퓨터를 이용해 공공재 게임을 수행하였으며 다른 참여자의 정체와 선택, 행동을 알 수 없도록 칸막이가 설치된 자리에 간격을 두고 배치되었다.

전반부의 평등 공공재-처벌 게임에서 참여자들은 모두 500 토큰을 동일하게 지급받았다. 반면 후반부의 불평등 공공재-처벌 게임에서 한 조를 이루는 4명의 참여자들은 각각 250/250/500/1000 토큰을 차등적으로 지급받았다(그림 2-1). 그리고 불평등 공공재-처벌 게임에서 누가 얼마를 받게 될지는 평등 공공재-처벌 게임에서 얼마나 많은 수익을 얻었는지에 따라 결정된다. 즉 전반부에서 가장 많은 수익을 올린 상위 25%에게 1000 토큰, 그 다음 상위 25%에게 500 토큰, 마지막 하위 50%에게 250 토큰을 지급했다.

〈그림 2-1〉 평등 공공재-처벌 게임(위)과 불평등 공공재-처벌 게임(아래)

이처럼 평등 조건에서 참여자들이 최종적으로 얻은 토큰은 뒤에 있을 불평등 조건에서 유리한 위치를 얻는 데 쓰이며 실제 보상금으로 전환되지 않는 반면, 불평등 조건에서 참여자들이 최종적으로 얻는 토큰은 1 단위당 1원으

로 게임이 끝난 후 참여자들에게 실제 돈으로 지급했다. 이러한 처리를 통해 참여자들에게 후반부 게임에서의 불평등에 명분이 있음을 납득시키고, 더 나아가 후반부의 게임에 사용되는 돈이 그저 주어진 것이 아니라 일정한 노력을 투여한 끝에 성취한 결과물이라는 소유의식을 바탕으로 의사결정에 임할 수 있게끔 만들고자 했다.

〈그림 2-2〉 기여 단계 의사결정 화면(전반부)

공공재-처벌 게임에서 참여자는 두 종류의 의사결정을 하게 된다. 하나는 공공재 생산을 위한 기여량을 결정하는 것이고 다른 하나는 공공재 생산을 통해 얻어진 수익을 분배한 후 다른 참여자들에 대한 처벌 여부와 그 강도를 결정하는 것이다. 참여자는 게임 상의 화면을 통해 같은 조에 속한 다른 세 명의 참여자들이 매 라운드 얼마를 지급받는지, 또 현재까지 축적한 개인 계정의 토큰은 얼마인지 알 수 있다. 참여자들은 매 라운드 자신에게 지급된

토큰의 범위 안에서 공공재 생산을 위해 기여할 토큰의 규모를 결정하게 된다(그림 2-2). 참여자가 사용하지 않은 토큰은 수익으로 개인 계정에 더해진다.

〈그림 2-3〉 차감 단계 의사결정 화면(전반부)

공공재 기여 단계에서, 같은 조에 속한 참여자 네 명이 모두 자신의 기여량을 결정하고 나면, 각 공공재 계정에 기여된 토큰의 총량은 두 배가 되어 참여자 네 명 모두에게 동등하게 분배되고 처벌 단계(실제 게임에서는 '처벌' 대신 '차감'으로 표현)가 이어진다(그림 2-3). 참여자는 기본적으로 지급된 토큰 중 공공재 계정에 기여하지 않은 토큰과 공공재 생산을 통해 획득한 토큰을 합한 범위 내(그림 2-3에서 '중간 결과')에서 같은 조에 속한 다른 참여자들을 처벌할 수 있다. 이때 참여자들은 같은 조에 속한 다른 구성원들이 공공재 계정에 얼마나 기여했고 얼마나 많은 이익을 얻었는지 알 수 있

으며 이러한 정보를 활용하여 처벌 여부를 결정할 수 있다. 처벌 단계에서 참여자는 아무도 처벌하지 않을 수도 있고 자신을 제외한 나머지 조원 셋 모두를 처벌할 수도 있다. 만약 참여자가 1토큰을 상대에게 지정하면, 상대의 총 재산은 3토큰 차감된다. 즉 처벌은 1:3의 효율을 따른다.

	총 재산	지급금	기여금	개인별 수익	손익	중간 결과	다른 조원 차감	차감 당함	최종 결과
당신	550	500	100	200	100	600	50	0	550
조원 1	600	500	100	200	100	600	0	0	600
조원 2	600	500	100	200	100	600	0	0	600
조원 3	450	500	100	200	100	600	0	150	450

	당신에게	조원 1에게	조원 2에게	조원 3에게
당신이	--	0	0	50
조원 1이	0	--	0	0
조원 2가	0	0	--	0
조원 3이	0	0	0	--

〈그림 2-4〉 차감 결과 확인 단계 (전반부)

처벌이 이행되면 참여자는 각 구성원이 처벌에 얼마나 많은 토큰을 썼고 또 얼마나 많은 처벌을 당했는지, 그리고 그 처벌의 주체와 대상이 누구인지 알 수 있다(그림 2-4). 보복이 불가능한 집단 1에서는 공공재, 각 참여자들

에게 처벌할 기회가 단 1회 주어진다. 따라서 집단 1에서는 처벌이 한 차례 수행되는 것으로 라운드가 종료되고, 다음 라운드의 게임이 시작된다.

〈그림 2-5〉 차감 반복 단계 의사결정 화면 (집단 2, 전반부)

그러나 보복이 가능한 집단 2에서 각 참여자들은 만약 자신이 속한 조에서 처벌 사건이 발생한다면 다른 참여자들을 처벌할 기회를 추가적으로 갖게 된다(그림 2-5). 처벌 단계가 반복되면, 같은 조에 속한 4명이 누구를 처벌하였고, 그 처벌에 얼마나 많은 토큰을 사용하였는지 공개된다. 이러한 정보를 확인 가능한 상태에서 참여자들은 또 다시 처벌(보복) 여부를 결정하게 된다. 따라서 앞선 단계에서 처벌을 당한 사람은 자신을 처벌한 사람에게 보복할 기회를 얻게 된다. 이렇게 하여 각 조에 속한 네 명 모두가 더 이상 다른 사람을 처벌하지 않을 때, 집단 2에서는 다음 라운드로 넘어가게 된다.

마지막으로, 앞에서 설명한 규칙에 따라 한 라운드가 종료되고 다음 라운드의 게임이 시작될 때 각 조의 구성원은 무작위로 재배치된다. 다만, 조를 무작위로 재배치한다고 해도 후반부의 불평등 공공재-처벌 게임에서 불평등 수준을 고려한 인구구조(250/250/500/1000)는 변하지 않는다. 또한 같은 조에 속한 네 명의 참여자는 매 라운드 당신(참여자 자신), 조원 1, 조원 2,

조원 3으로 표기되지만 한 라운드에서 다음 라운드로 넘어갈 때마다 무작위로 배치가 이루어지기 때문에 그 숫자만으로는 어떤 참여자가 자신과 상호작용을 한 이력이 있는 참여자인지 알 수 없다. 이처럼 매 라운드 조를 뒤섞고 개인 식별이 불가능하게 한 이유는 참여자들의 의사결정이 평판 또는 호혜성에 영향을 받는 것을 최대한 배제하기 위함이며, 앞서 서술했듯 보복이 협력과 처벌 행위를 교란하는 것을 막기 위함이다.

위와 같은 실험 설계를 간략하게 정리하여 실험의 진행 절차를 서술하면 다음과 같다.

1) 전반부의 평등 공공재-처벌 게임에서 모든 참여자에게는 매 라운드 시작과 동시에 500토큰이 지급된다.

2) 기여 단계에서 참여자는 자신에게 지급된 500토큰의 한도 내에서 공공재 계정에 얼마나 많은 토큰을 기여할 것인지 결정한다. 공공재 계정에 기여하지 않은 토큰은 개인 수익(총 재산)에 더해진다.

3) 같은 조에 속한 네 명이 기여한 토큰의 총합은 두 배가 되어 조원에게 균등하게 분배된다. 참여자는 자신을 포함하여 다른 조원들이 공공재 계정에 얼마나 기여했고, 그로 인해 얼마나 많은 수익을 얻었는지 알 수 있다.

4) 각 조원의 기여량과 수익이 공개된 상태에서, 참여자는 같은 조에 속한 다른 구성원을 처벌할 수 있다. 참여자는 이번 라운드의 기여 단계에서 획득한 토큰(지급된 500토큰 – 기여량 + 공공재 계정 수익)을 처벌에 사용할 수 있다. 처벌의 비용과 효과는 1:3의 비율을 따른다.

5) 처벌이 끝나면 같은 조의 구성원이 누구를 대상으로 얼마나 많은 토큰을 처벌에 사용했는지 알 수 있다. 보복이 불가능한 집단 1은 여기서 라운드가 종료된다.

6) 보복이 가능한 집단 2의 참여자는 지난 단계의 처벌 정보가 공개된 상태에서 추가적인 처벌 기회를 얻는다. 처벌 규칙은 위와 동일하다. 이때 아무도 다른 사람을 처벌하지 않으면 라운드가 종료되지만, 처벌이 발생할 경우 계속해서 처벌 기회가 이어진다.

7) 한 라운드가 종료되면 조의 구성원을 무작위로 재배치하여 새로운 라운

드가 시작된다. 새로운 라운드의 진행 방식 역시 위의 1) ~ 6)과 동일하다. 평등 공공재-처벌 게임은 집단 1에서 총 20라운드, 집단 2에서 총 15라운드 진행된다. 참여자에게는 게임이 총 몇 라운드 진행되는지 알려주지 않는다.

8) 평등 공공재-처벌 게임이 종료되면 잠시 쉬는 시간을 갖고 불평등 공공재-처벌 게임이 진행된다. 불평등 공공재-처벌 게임에서 한 개 조는 매 라운드 1000토큰을 받는 한 명과 500토큰을 받는 한 명, 250토큰을 받는 두 명으로 구성된다. 누가 얼마나 많은 토큰을 받게 될 지는 전반부 게임의 결과에 따라 결정된다. 즉 평등 공공재-처벌 게임에서 가장 많은 수익을 얻은 상위 25%에게 1000토큰, 그 다음 25%에게 500토큰, 나머지 하위 50%에게 250토큰이 지급된다. 이를 제외한 나머지 진행방식은 전반부 공공재 게임과 동일하게 2) ~6) 단계를 따른다. 불평등 공공재-처벌 게임은 집단 1과 집단 2 모두에서 20라운드 진행된다. 이때도 참여자들은 게임이 총 몇 라운드를 진행되는지 알지 못한다.

2.3. 통계 분석

각 참여자가 주어진 환경에서 내린 의사결정을 변수로 하여 통계적 분석을 실시하였다. 먼저 집단 수준에서 각 조건(보복 불가능/보복 가능 * 평등/불평등)에 따라 협력과 처벌 행위에서 평균 차이가 유의하게 나타나는지 확인하기 위해 비모수적 통계방법의 하나인 Mann-Whitney의 U 검정을 실시하였다. 비모수적 통계방법은 표본 집단의 정규분포 가정이 성립하지 않아 t 검증의 사용이 불가능할 때 그 대안으로 사용되며, 본 연구에서 사용된 Mann-Whitney의 U 검정은 그 중에서도 대표적인 방법으로 알려져 있다. 본 연구에서는 각 참여자가 매 라운드 공공재 계정에 기여한 토큰 개수의 평균을 협력 수준의 지표로, 다른 참여자의 토큰을 차감하는 데 사용한 토큰 개수의 평균을 처벌 강도의 지표로 삼아 상이한 조건 하에서 나타나는 차이를 비교하였다.

여기에 추가적으로 차감에 사용된 토큰 개수와 관계없이 매 라운드 차감

행위가 발생한 횟수를 처벌 빈도의 지표로 삼아 처벌 강도와 더불어 처벌 행위를 비교하는 데 활용하였다. 처벌 행위를 이처럼 두 가지 측면에서 분석한 이유는 집단 2의 불평등 공공재-처벌 게임 결과 소수의 참여자(12명 중 2명)가 강도 높은 처벌을 집중적으로 사용하는 양상이 나타났기 때문이다. 이 경우 분명 다수(12명 중 9명)가 처벌에 참여하지 않았음에도 평균적인 처벌 강도만 놓고 비교한다면 통계적으로 유의한 차이가 없게 된다. 따라서 처벌 빈도를 변수에 추가함으로써 실제로 처벌의 사용이 유의하게 감소하였음을 보이고자 하였다.

더 나아가 개인의 처벌 강도에 영향을 미치는 요인과 그 영향력을 파악하기 위해 토빗 모형을 이용하여 회귀분석을 실시하였다. 이는 처벌 대상의 행위(무임승차)나 특성(더 많은 자원 보유)이 일정한 역치를 넘기지 않을 경우 처벌이 아예 발생하지 않을 것으로 예상되기 때문이다. 예를 들어 상대가 자신보다 적은 양을 기여했을 때 처벌을 한다고 가정해 보자. 그렇다면 상대가 1토큰 적게 기여했을 때보다 100토큰 적게 기여했을 때 더 강한 처벌을 가하리라고 예상할 수 있을 것이다. 그러나 반대로 만약 상대가 자신보다 많은 양을 기여했다면, 그 차이가 1토큰이건 100토큰이건 관계없이 처벌하지 않을 것이므로 처벌 강도는 0으로 기록된다. 따라서 종속변수인 처벌 강도는 관측치의 최소값이 0으로 제한된, 중도절단된(censored) 표본에 해당된다. 토빗 모형은 이처럼 종속변수가 중도절단된 표본의 형태를 가질 때 사용되는 회귀분석 방법이다. 토빗 모형에 기반을 둔 회귀분석을 통해 자신과 상대의 기여량 차이, 자신과 상대의 총 재산 차이, 전 라운드 자신이 받은 처벌 등의 변수가 종속변수인 처벌 강도를 설명하는 데 있어 과연 유의한지, 또 그 영향력은 얼마나 큰지 파악하고자 하였다.

마지막으로 본 연구에서는 SPSS v.23.0과 STATA(토빗 모형 회귀분석)를 사용하여 통계 분석을 실시하였음을 밝힌다.

3. 결과

3.1. 처벌 분석

3.1.1. 집단 수준에서의 처벌 비교분석

보복 불가능 조건에서는 참여자에게 처벌 기회가 한 차례만 주어졌으므로 그 단계에서 관찰된 처벌 행위를 분석 대상으로 삼았다. 그러나 보복 가능 조건에서는 같은 조 내에서 처벌이 발생할 경우, 처벌 단계가 반복되므로 최초의 처벌 기회와 반복된 처벌 기회 모두에서 관찰된 처벌 행위를 분석 대상으로 삼았다.

	보복 불가능	보복 가능
평등	102.68	18.59
불평등	41.85	20.88

〈표 3-1〉 각 조건별 처벌에 사용된 토큰

모든 조건을 통틀어 나타난 처벌 양상은 위의 표 3-1과 같다. 보복 불가능한 조건에서 참여자들이 매 라운드 처벌에 사용한 토큰은 평등할 때 102.68개, 불평등할 때 41.85개로 평등할 때에 비해 불평등할 때 통계적으로 유의하게 더 적은 토큰을 처벌에 활용(Mann-Whitney U test: $Z=-2.904$, $p<0.01$, $N=64$)하였다. 반대로 보복이 가능한 조건에서 참여자들이 매 라운드 처벌에 사용한 토큰은 평등할 때 18.59개, 불평등할 때 20.88개였으나 그 차이는 유의하지 않았다(Mann-Whitney U test: $Z=-1.274$, $p=0.203$, $N=48$). 또한 평등할 때와 불평등할 때 모두 보복 불가능 조건에서보다 보복 가능 조건에서 처벌이 더 약하게 이루어졌음을 알 수 있다(Mann-Whitney U test: $Z=-4.444$, $p<0.01$, $N=112$).

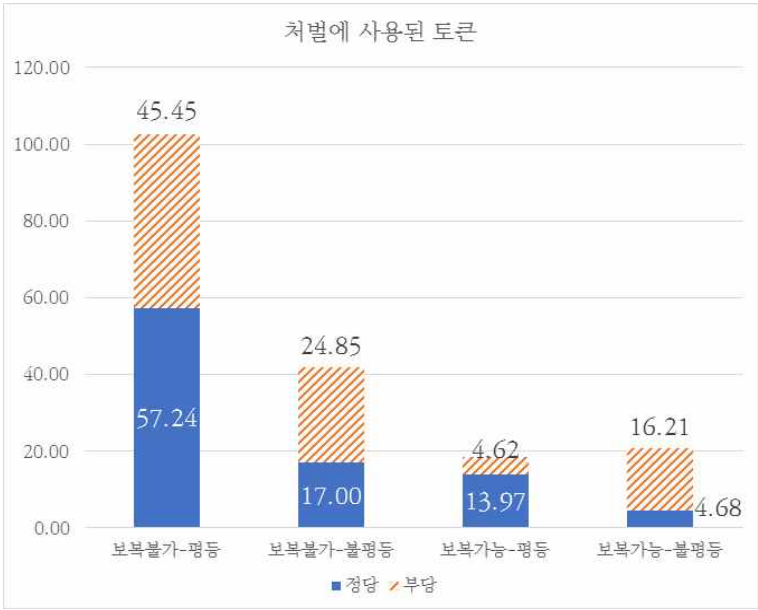
결과 1-1: 보복 불가능 조건에서, 평등할 때에 비해 불평등할 때 더 적은 양의 토큰이 처벌에 사용되었다.

결과 1-2: 보복 불가능 조건에 비해 보복 가능 조건에서 더 적은 양의 토큰이 처벌에 사용되었다.

처벌이 1회만 가능한 환경에서 불평등의 도입이 전체적인 처벌의 약화에 기여한다는 것은 이미 다른 연구들(김상인 2006, Prediger 2011)에서도 밝혀진 바 있다. 모든 참여자들이 같은 양의 토큰을 보유한 평등 조건에서는 상대적 우위를 차지하기 위한 경쟁이 치열하게 벌어지게 되고, 처벌은 이를 위한 전략적 수단으로서 사용(김상인 2006)될 수 있다. 반대로 참여자들에게 차등적으로 토큰을 지급하는 불평등 조건에서 처벌이 감소한 것은, 차등적인 토큰 지급으로 인해 위계가 처음부터 뚜렷하게 구분되어 참여자간의 지위 경쟁이 약화되었기 때문이라고 해석할 수 있다.

그러나 처벌에 대한 보복이 가능한 보다 현실적인 구조에서 처벌은 더 이상 지위 경쟁을 위한 전략적 수단으로 활용될 수 없다. 처벌을 통해 일시적으로 우위를 점했다 하더라도 보복이 이어진다면 그 우위가 지속된다고 보장할 수 없기 때문이다. 따라서 보복이 가능할 경우 보복이 불가능할 때에 비해 전체적으로 처벌이 크게 감소하였다는 결과와, 보복 가능 조건에서 평등할 때와 불평등할 때 처벌의 차이가 유의하지 않았다는 결과는 모두 보복이 가능할 때에는 지위 경쟁의 욕구가 실제 처벌 행위로 이어지지 않았다는 것을 의미한다.

다음으로, 자신보다 더 적은 토큰을 기여한 사람에 대한 차감 행위를 정당한 처벌로, 자신보다 더 많은 토큰을 기여한 사람에 대한 차감 행위를 부당한(반사회적) 처벌로 분류했을 때 정당한 처벌과 부당한 처벌이 각 조건에 따라 어떠한 차이를 보였는지 분석하였다(그림 3-1). 보복이 불가능할 때, 평등한 조건에서 참여자들은 평균적으로 정당한 처벌에 57.24개, 부당한 처벌에 45.45개의 토큰을 사용하였으나 이 차이는 통계적으로 유의하지 않았다(Wilcoxon signed rank test: $Z=-1.378$, $p=0.168$, $N=32$). 불평등한 조건에서는 평균적으로 정당한 처벌에 17.00개, 부당한 처벌에 24.85개의 토큰을 사용하였고, 이 차이 또한 유의하지 않았다(Wilcoxon signed rank test: $Z=-1.520$, $p=0.128$, $N=32$). 즉 보복이 불가능한 조건에서는 참여자들이 사용한 정당한 처벌과 부당한 처벌에서 통계적으로 유의한 차이를 발견할 수 없었다.



〈그림 3-1〉 처벌에 사용된 토큰

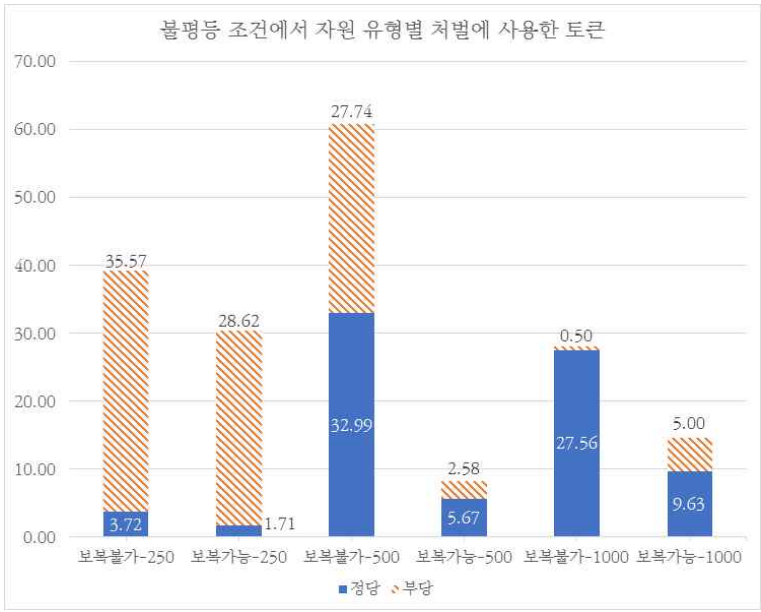
한편 보복이 가능할 경우, 평등 조건에서 참가자들은 평균적으로 정당한 처벌에 13.97개, 부당한 처벌에 4.62개의 토큰을 사용하여 정당한 처벌을 더 강하게 사용하는 것으로 나타났으며 이 차이는 통계적으로 유의하였다(Wilcoxon signed rank test: $Z=-2.277$, $p<0.05$, $N=24$). 반대로 불평등한 조건에서는 평균적으로 정당한 처벌에 4.68개, 부당한 처벌에 16.21개의 토큰을 사용하였지만 이 차이는 통계적으로 유의하지 않았다(Wilcoxon signed rank test: $Z=-0.261$, $p=0.794$, $N=24$). 즉 오직 보복 가능-평등 조건에서만 참여자들은 부당한 처벌보다 정당한 처벌에 더 많은 토큰을 사용하는 것으로 나타났으며, 다른 세 조건에서는 참여자들이 정당한 처벌과 부당한 처벌에 사용한 토큰이 통계적으로 유의한 수준에서 차이를 보이지 않았다.

결과 1-3: 보복 가능-평등 조건에서 참여자들은 부당한 처벌보다 정당한 처벌에 더 많은 토큰을 사용하였다.

3.1.2. 불평등 조건에서 자원 유형별 처벌 비교분석

이어서 불평등 조건에서 발생한 처벌 행위를 각 자원 유형별로 분석해보았다. 반사회적 처벌의 기저에 불평등 기피 성향이 작용할 것이라는 본 연구의 가정에 따르면, 자원의 분포가 불평등한 환경에서는 자원이 적은 사람이 부당한 처벌을 더 강하게 사용할 것으로 예측된다.

불평등 조건에서 각 자원 유형에 속한 참여자들이 정당한 처벌과 부당한 처벌에 사용한 토큰을 살펴보면 이러한 예측의 타당성을 어느 정도 확인할 수 있다(그림 3-2). 먼저 매 라운드 가장 적은 토큰을 지급받은 250유형은 보복이 불가능할 때 부당한 처벌에 35.57개, 정당한 처벌에 3.72개의 토큰을 평균적으로 사용(Wilcoxon signed rank test: $Z=-6.655$, $p<0.01$, $N=320$)했고, 보복이 가능할 때 부당한 처벌에 28.62개, 정당한 처벌에 1.71개의 토큰을 평균적으로 사용(Wilcoxon signed rank test: $Z=-3.909$, $p<0.01$, $N=240$)하여 보복 가능 여부와 관계없이 정당한 처벌보다 부당한 처벌을 강하게 사용했음을 알 수 있다.



〈그림 3-2〉 불평등 조건에서 자원 유형별 처벌

결과 2-1: 250유형은 보복 가능 여부와 관계없이 정당한 처벌보다 부당한 처벌에 통계적으로 유의한 수준에서 더 많은 토큰을 사용하였다.

다음으로 불평등 조건에서 중간 지위를 차지하는 500유형은 보복이 불가능할 때 부당한 처벌에 27.74개, 정당한 처벌에 32.99개의 토큰을 평균적으로 사용하였으며 그 차이는 유의하지 않았다(Wilcoxon signed rank test: $Z=-0.465$, $p=0.642$, $N=160$). 보복이 가능할 때에는 부당한 처벌에 2.58개, 정당한 처벌에 5.67개의 토큰을 사용하였는데, 이 차이 또한 유의하지 않았다(Wilcoxon signed rank test: $Z=-0.385$, $p=0.700$, $N=120$).

결과 2-2: 500유형이 정당한 처벌과 부당한 처벌에 사용한 토큰의 수는 보복 불가능 조건과 보복 가능 조건 모두에서 통계적으로 유의한 차이를 보이지 않았다.

마지막으로 매 라운드 가장 많은 토큰을 지급받은 1000유형은 보복이 불가능할 때 부당한 처벌에 0.50개, 정당한 처벌에 27.56개의 토큰을 평균적으로 사용하였고 이 차이는 유의하였다(Wilcoxon signed rank test: $Z=-4.95$, $p<0.01$, $N=160$). 보복이 가능할 때에는 부당한 처벌에 5.00개, 정당한 처벌에 9.63개의 토큰을 평균적으로 사용하였으나 이 차이는 유의하지 않았다(Wilcoxon signed rank test: $Z=-1.138$, $p=0.255$, $N=120$).

결과 2-3: 1000유형은 보복 불가능 조건에서만 부당한 처벌보다 정당한 처벌에 통계적으로 유의한 수준에서 더 많은 토큰을 사용하였다.

위의 세 결과(결과 2-1, 2-2, 2-3)를 종합하면, 자원의 분포가 불평등한 환경에서는 가진 자원이 적을수록 반사회적 처벌을 더 강하게 사용하는 것을 알 수 있다. 위에서 예측한 것처럼, 불평등 기피 성향이 반사회적 처벌의 사용에 강한 영향을 미친 것이다.

더 나아가 표 3-2를 보면 불평등 기피 성향과 반사회적 처벌 사이의 관계를 보다 분명하게 파악할 수 있다. 표 3-2는 부당한 처벌의 경우, 누가 누구를 대상으로 부당한 처벌에 얼마나 많은 토큰을 사용했는지 보여준다. 먼저 보복이 불가능할 때, 250유형은 부당한 처벌에 사용한 35.57개의 토큰 중 27.50개를, 500유형은 부당한 처벌에 사용한 27.74개의 토큰 전부를 자신보다 더 많은 자원을 보유한 1000유형을 대상으로 사용한 것을 알 수 있다. 이러한 경향은 보복 가능 조건에서도 마찬가지로 관찰되었다. 보복 가능 조건에서 250유형이 부당한 처벌에 사용한 28.62개의 토큰 중 26.65개가, 500유형이 부당한 처벌에 사용한 2.58개의 토큰 전부가 1000유형을 대상으로 하였다.

부당한 처벌 (강도)			처벌 대상			합계
			250	500	1000	
처벌 주체	250	보복 불가능	3.03	5.04	**27.50**	35.57
		보복 가능	1.33	0.64	**26.65**	28.62
	500	보복 불가능	0.00	-	**27.74**	27.74
		보복 가능	0.00	-	**2.58**	2.58
	1000	보복 불가능	0.00	0.50	-	0.50
		보복 가능	2.08	2.92	-	5.00

〈표 3-2〉 각 자원 유형별 부당한 처벌에 사용한 토큰

결과 2-4: 불평등 조건에서 발생하는 반사회적 처벌은 주로 처벌자 자신보다 더 많은 자원을 보유한 사람에게 사용된다.

추가적으로, 250유형의 반사회적 처벌이 500유형을 대상으로는 거의 사용되지 않았다(5.04/35.57, 0.64/28.62)는 사실에 주목할 필요가 있다. 불평등한 구조에서 500유형은 250유형에 비해 한 단계 높은 지위를 차지하고 있음에도 불구하고 250유형의 불평등 기피 성향을 자극하지 않았다. 이는 불평등한 구조에서 나타나는 반사회적 처벌을 지위경쟁 욕구로 설명할 수 없으며, 불평등 기피 성향으로 설명될 수밖에 없는 이유를 잘 보여준다고 할 수 있다.

만약 불평등 구조에서 250유형이 지위경쟁의 욕구에 의해 반사회적 처벌을 사용한다면, 1000유형보다는 500유형을 대상으로 하는 처벌에 집중하는 편이 더 효과적이다. 스스로 더 높은 지위에 올라서고자 한다면, 자신과 멀리 떨어져 있는 사람보다는 가까운 사람을 끌어내리는 것이 더 쉽기 때문이다. 그러나 정해진 인구구조(250/250/500/1000)에서 250유형이 500유형을 처벌한다면, 250유형과 500유형 사이의 격차는 좁혀지지만 1000유형과 250유형, 500유형 사이의 격차는 벌어진다는 점에서 결과적으로 불평등은 오히려 심화된다. 반대로 250유형이 1000유형을 처벌하는 일은 250유형의 지위 상승에는 보탬이 되지 못하지만, 1000유형과 500유형, 250유형 사이의 격차를 좁힌다는 면에서 불평등을 완화하는 결과를 낳는다고 할 수 있다. 따라서 500유형 대신 1000유형에게 250유형의 반사회적 처벌이 집중되었다는 것은 불평등 구조에서 나타나는 반사회적 처벌의 목적이 처벌 사용자의 개인적인 지위 상승보다는 불평등의 완화에 있음을 잘 보여준다고 할 수 있다.

3.1.3. 보복 가능-불평등 조건에서 반사회적 처벌의 억제

한편, 본 연구에서는 보복 가능성의 도입이 불평등 조건에서 불평등 기피 성향에 기반을 둔 반사회적 처벌 행위를 강하게 억제할 것이라 예측한 바 있다. 500유형이 1000유형을 대상으로 가한 부당한 처벌(이하 500-1000 부당처벌)의 강도는 보복 불가능할 때 27.74개에서 보복 가능할 때 2.58개로 유의하게 감소하였다(Mann-Whitney U test: $Z=-2.377$, $p=0.01$, $N=14$). 그러나 250유형이 1000유형을 대상으로 가한 부당한 처벌(이하 250-1000 부당처벌)의 강도는 보복 불가능할 때 27.50개, 보복 가능할 때 26.65개로 이 차이는 통계적으로 유의하지 않았다(Mann-Whitney U test: $Z=-1.514$, $p=0.09$, $N=28$).

결과 3-1: 불평등 조건에서 500-1000 부당처벌에 사용된 토큰의 수는 보복이 불가능할 때에 비해 보복이 가능할 때 유의하게 감소하였다.

결과 3-2: 불평등 조건에서 250-1000 부당처벌에 사용된 토큰의 수는 보복이 불가능할 때와 가능할 때 유의한 차이를 보이지 않았다.

보복 가능성을 도입했음에도 불구하고 250-1000 부당처벌의 강도가 감소하지 않았다는 결과는, 본 연구에서 제기한 가설을 정면으로 반박한다는 점에서 문제가 된다. 다만 처벌에 사용된 토큰의 수(처벌 강도)가 아닌 처벌 사건이 발생한 횟수(처벌 빈도)를 기준으로 비교할 경우(표 3-3), 매 라운드 평균적으로 발생한 250-1000 부당처벌의 빈도는 보복 불가능할 때 0.23회, 보복 가능할 때 0.08회로 유의하게 감소(Mann-Whitney U test: Z=-1.918, p=0.04, N=28)하였음을 알 수 있다.

결과 3-3: 불평등 조건에서 250-1000 부당처벌의 빈도는 보복이 불가능할 때에 비해 보복이 가능할 때 유의하게 감소하였다.

부당한 처벌 (빈도)			처벌 대상			합계
			250	500	1000	
처벌 주체	250	보복 불가능	0.02	0.07	**0.23**	0.30
		보복 가능	0.03	0.03	**0.08**	0.14
	500	보복 불가능	0.00	-	0.26	0.26
		보복 가능	0.00	-	0.05	0.05
	1000	보복 불가능	0.00	0.02	-	0.01
		보복 가능	0.03	0.02	-	0.04

〈표 3-3〉 각 자원 유형별 부당한 처벌의 발생 빈도

처벌 강도는 감소하지 않았지만 처벌 빈도는 감소했다는 결과는 보복 가능-불평등 조건에서 250-1000 부당처벌이 거의 발생하지 않았지만, 발생한 경우 그 강도가 상당히 높았다는 것을 의미한다. 아래의 표 3-4를 보면, 보복 가능-불평등 조건에서 12명의 250유형 참여자들 중 단 3명(3, 8, 10)만이 1000유형에게 반사회적 처벌을 사용했으며, 그 중에서도 2명(8, 10)이 특히 많은 토큰을 사용했음을 알 수 있다. 나머지 9명은 250-1000 부당처벌에 단 하나의 토큰도 사용하지 않았다. 즉 소수의 참여자들이 사용한 높은

강도의 처벌이 전체적인 평균 상승을 견인하였고, 그로 인해 처벌 강도 측면에서는 보복 불가능 조건과 보복 가능 조건에서의 차이가 발견되지 않았던 것이다.

참여자	1	2	3	4	5	6	7	8	9	10	11	12
250-1000 부당처벌	0	0	17.5	0	0	0	0	99.25	0	203	0	0

〈표 3-4〉 개인별 250-1000 부당처벌에 사용한 토큰의 수

보복 가능-불평등 조건에서 2명의 250유형 참여자가 이처럼 많은 양의 토큰을 사용하여 1000유형을 처벌한 이유에 대해서는 아래의 표 3-5를 참고하여 다음과 같은 추론해볼 수 있다. 표를 보면 보복 가능-불평등 조건에서 발생한 250-1000 부당처벌의 강도는 대부분 1000유형의 토큰 보유량을 0, 혹은 그와 비슷한 수준(0±100)으로 만드는 데 맞춰져 있다[2]. 보복 가능 조건에서 처벌로 인해 토큰 보유량이 0이 될 경우 처벌당한 사람은 이어지는 처벌(보복) 단계에서 처벌에 토큰을 사용할 수 없게 된다. 따라서 보복 가능 조건에서 소수의 250유형은 이어지는 보복의 위험을 완전히 제거하기 위해서 1000유형을 대상으로 강도 높은 부당처벌을 가했을 것이라는 해석이 가능하다. 이와는 대조적으로 보복 불가능-불평등 조건에서는 250-1000 부당처벌이 활발하게 발생했음에도 불구하고 보복 가능-불평등 조건에서처럼 1000유형의 토큰 보유량을 0으로 만드는 처벌 유형은 거의 발견되지 않았다는 사실 역시 이러한 해석을 뒷받침해주고 있다.

매 라운드 지급되는 토큰과 기여를 통해 획득한 토큰을 합한 규모 내에서만 처벌의 강도를 결정할 수 있는 게임의 구조를 이용하여, 보복 가능 조건에서 2명의 250유형 참여자는 1000유형을 대상으로 토큰 보유량을 0으로 만드는 강도 높은 부당처벌을 가하였다. 이 경우 1000유형의 보복이 불가능해진다는 점에서 이러한 처벌은 비록 보복 가능 조건에서 발생했지만 오히려 보복 불가능 조건에서의 처벌과 더 유사하다고 할 수 있다. 보복 가능 조건에서 발생한 250-1000 부당처벌은 오로지 1000유형의 보복이 불가능할 때

2) 1000유형 토큰 보유량-처벌 강도*3

에만 발생했기 때문이다. 다른 관점에서 본다면, 보복 가능 조건에서 보복이 불가능한 상태를 만들기 위해 부당처벌의 강도가 비약적으로 높아질 수밖에 없었다는 해석도 가능하다. 따라서 게임의 구조를 이용하여 보복 불가능 상태를 만들어낸 이와 같은 처벌을 제외하고 나면 250-1000 부당처벌이 거의 발생하지 않았다는 결과, 또 빈도의 측면에서도 보복 가능성이 도입되자 250-1000 부당처벌이 감소하였다는 본 연구의 결과는 여전히 최초의 예측, 즉 불평등 기피 성향으로 인한 반사회적 처벌이 보복 가능성의 도입으로 인해 억제될 것이라는 예측을 지지한다고 할 수 있다.

〈8번 참여자〉

라운드	12	13	14	17	19
처벌 강도	525	350	350	360	400
1000유형 토큰 보유량 - 처벌 전	1175	1055	1100	1090	1165
1000유형 토큰 보유량 - 처벌 후	-400	5	50	10	-35

〈10번 참여자〉

라운드	6	8	9	10	12	13	14	17	18	19
처벌 강도	500	400	360	400	400	400	400	400	400	400
1000유형 토큰 보유량 - 처벌 전	1125	1200	1100	1150	1175	1200	1150	1125	1175	1100
1000유형 토큰 보유량 - 처벌 후	-375	0	20	-50	-25	0	-50	-75	-25	-100

〈표 3-5〉 8번과 10번 참여자의 라운드 별 처벌 강도와
처벌 대상(1000 유형)의 해당 라운드 토큰 보유량 변화

3.1.4. 보복 가능-불평등 조건에서 정당한 처벌의 억제

위의 분석을 통해 본 연구가 예측하였던 것처럼 보복 가능성의 도입이 불평등 조건에서 발생하는, 더 많이 가진 사람을 대상으로 하는 반사회적 처벌의 발생을 억제함을 확인할 수 있었다. 이를 확인하기 위해 지금까지는 무임

승차자가 아닌, 협력적인 사람을 대상으로 하는 반사회적 처벌에 초점을 맞춰 분석을 진행해 왔다. 여기서는 무임승차자를 대상으로 하는 정당한 처벌은 불평등 조건과 보복 가능 조건의 도입에 의해 어떠한 차이를 보이는지 확인해보고자 한다.

정당한 처벌(강도)		보복 불가능	보복 가능
평등 조건		57.24	13.97
불평등 조건	전체	17.00	4.68
	250유형	3.72	1.71
	500유형	32.99	5.67
	1000유형	27.56	9.63

〈표 3-6〉 각 조건에서 정당한 처벌에 사용한 토큰의 수

표 3-6에 따르면 평등 조건과 불평등 조건 모두에서 보복 가능성이 도입되자 참여자가 평균적으로 정당한 처벌에 사용한 토큰의 수가 감소한 것을 확인할 수 있다. 이러한 차이, 즉 보복 가능성의 도입으로 인한 정당한 처벌의 감소는 평등 조건에서는 통계적으로 유의(Mann-Whitney U test: $Z=-3.596$, $p<0.01$, $N=56$)했지만, 불평등 조건에서는 유의하지 않았다(Mann-Whitney U test: $Z=-0.858$, $p=0.391$, $N=56$). 비록 불평등 조건에서는 그 차이가 유의하지 않았지만, 이러한 결과는 보복 가능성의 도입이 부당한 처벌만을 선별적으로 억제하지 않았다는 것을 보여준다. 즉 보복이 가능해지자, 처벌의 정당성 여부와 관계없이 참여자들간의 처벌 사용이 전체적으로 감소한 것이다.

결과 4-1: 보복 가능성이 도입되자 평등 조건과 불평등 조건 모두에서 정당한 처벌에 사용된 토큰의 수는 감소하였으나, 그 차이는 평등 조건에서만 유의하였다.

다음으로, 불평등 조건에서 어떠한 자원 유형의 참여자가 정당한 처벌에 더 많은 토큰을 사용했는지 살펴보았다[3]. 먼저 보복 불가능-불평등 구조에서

250유형보다는 500유형과 1000유형이 정당한 처벌에 더 많은 토큰을 사용하는 것으로 나타났으며 그 차이는 통계적으로 유의하였다(Mann-Whitney U test: Z=-1.756, p=0.05, N=32). 마찬가지로, 보복가능-불평등 조건에서도 500유형과 1000유형이 평균적으로 정당한 처벌에 사용한 토큰의 수는 250유형에 비해 더 많았고 그 차이 또한 통계적으로 유의하였다(Mann-Whitney U test: Z=-1.949, p=0.03, N=24). 어떠한 유형으로든 참여자들 사이에 차이(heterogeneity)가 존재할 경우, 처벌 비용의 부담이 상대적으로 적은 쪽이 더 적극적으로 처벌에 참여한다는 것은 기존 연구들(Tan 2008, Bone *et al.* 2015, Przepiorka & Diekmann 2013)에서 밝혀진 바 있기에 이러한 결과는 이를 재현한 것이라 할 수 있다.

결과 4-2: 보복 불가능-불평등 조건과 보복 가능-불평등 조건에서 상대적으로 더 많은 자원을 보유한 참여자(500유형/1000유형)가 정당한 처벌에 더 많은 토큰을 사용하였다.

마지막으로 각 자원 유형별 참여자가 정당한 처벌에 사용한 토큰의 수가 보복 불가능 조건과 보복 가능 조건에서 어떠한 차이를 보였는지 살펴보면, 먼저 250유형이 정당한 처벌에 사용한 토큰의 수는 보복 불가능 조건에 비

3) 본격적인 분석에 앞서, 불평등 조건에서 자원 유형별로 정당한 처벌의 사용을 비교분석함에 있어 본 연구의 실험 구조가 갖는 한계를 지적하고 넘어가고자 한다. 본 연구에서는 처벌 주체와 처벌 대상이 공공재 계정에 기여한 토큰의 절대적 차이를 기준으로 처벌의 정당성 여부를 판단한다고 밝힌 바 있다. 즉 상대가 나보다 많은 토큰을 기여했을 때 처벌하는 것은 반사회적 처벌로, 상대가 나보다 적은 토큰을 기여했을 때 처벌하는 것은 정당한 처벌로 분류하였다는 것이다. 이러한 기준은 모든 참여자가 공공재 계정에 기여할 수 있는 토큰의 최댓값이 동일(500개)한 평등 조건에서는 큰 문제가 되지 않는다. 그러나 불평등 조건에서는 적은 양의 토큰을 지급받는 참여자에게 정당한 처벌을 할 기회가 구조적으로 제한된다는 점에서 이 기준은 문제가 된다. 250유형은 자신이 얼마나 많은 토큰을 공공재 계정에 기여하든 500유형과 1000유형의 기여량이 250토큰 이상일 경우 이들에게 정당한 처벌을 가할 수 없는 반면, 1000유형은 자신이 충분히 많은 토큰을 기여한다면 250유형과 500유형이 얼마나 많은 토큰을 기여하든 관계없이 이들에게 정당한 처벌을 가할 수 있기 때문이다. 이때, 절대적인 기여량 차이 대신 기여 가능한 토큰의 최댓값에서 실제로 기여한 토큰이 차지하는 비중의 상대적 차이를 기준으로 처벌의 정당성 여부를 살피는 것도 대안으로 생각해볼 수 있다. 그러나 모두가 같은 비율로 기여했을 때 250유형은 공공재 생산으로 인해 이익을 얻지만 1000유형은 이익을 얻지 못한다는 점에서 이러한 기준은 문제가 된다.

해 보복 가능 조건에서 평균적으로 감소하긴 했으나 그 차이는 유의하지 않았다(Mann-Whitney U test: Z=-1.056, p=0.291, N=560). 이는 애초에 250유형이 정당한 처벌에 사용한 토큰의 수가 보복 불가능 조건에서도 상당히 적었기 때문이라고 볼 수 있다. 반면 500유형과 1000유형은 보복 불가능 조건보다 보복 가능 조건에서 평균적으로 더 적은 수의 토큰을 정당한 처벌에 사용하였으며 그 차이는 모두 유의하였다(500유형 Mann-Whitney U test: Z=-3.455, p<0.01, N=280 / 1000유형 Mann-Whitney U test: Z=-2.421, p=0.02, N=280).

결과 4-3: 불평등 조건에서 500유형과 1000유형이 정당한 처벌에 사용한 토큰의 수는 보복이 불가능할 때보다 보복이 가능할 때 감소하였다.

비록 500유형과 1000유형이 250유형에 비해 정당한 처벌에 더 많은 토큰을 사용하긴 했지만 보복 가능성이 도입되자 전체적으로 그 수준이 크게 감소하였다는 것은 협력의 진화에 있어 처벌이 기여한 바가 그다지 크지 않았을 가능성을 시사한다. 보복 가능성이 도입되자 불평등 조건과 평등 조건 모두에서 반사회적 처벌과 더불어 무임승차자에 대한 처벌 또한 크게 감소하였기 때문이다. 처벌 기회가 단 한 차례 주어지는 게임 구조의 비현실성은, 이미 본 연구의 선행연구 검토에서도 지적한 바 있다. 그렇다면 자신의 처벌이 상대의 보복으로 이어질 수 있는 현실적인 상황에서, 사람들은 더 많이 가진 사람을 끌어내리기 위한 반사회적 처벌도, 무임승차자가 더 큰 이익을 얻는 것을 방지하기 위한 정당한 처벌도 잘 사용하지 않는다는 것이다. 이러한 결과와 더불어 아래의 3장 2절에서 다루어질 참여자들의 기여량 분석을 통해, 처벌 대신 처벌을 당할 수 있다는 두려움이 협력의 진화에 기여하였을 가능성을 확인할 수 있었다. 이에 대해서는 3장 2절에서 더욱 자세하게 논의해보고자 한다.

3.1.5. 개인 수준에서의 처벌 회귀분석

독립변수	보복불가-평등	보복불가-불평등	보복가능-평등	보복가능-불평등
(상수)	-263.548 ***	-286.571 ***	-192.211 ***	-674.281 ***
	(22.552)	(37.566)	(29.484)	(89.823)
(1) 처벌자 기여 +	.804 ***	.301 ***	.500 ***	.542 **
	(.072)	(.063)	(.083)	(.161)
(2) 처벌자 기여 -	.172 *	-.027	-.306 *	-.867 **
	(.075)	(.046)	(.131)	(.337)
(3) 처벌자 재산 +	-.032 ***	-.071 ***	-.057 **	-.041 **
	(.006)	(.016)	(.017)	(.012)
(4) 처벌자 재산 -	.032 ***	.039 ***	-.002	.063 ***
	(.007)	(.032)	(.018)	(.016)
(5) 지난 R 처벌 받음	.017	-.037	.044	.052
	(.012)	(.032)	(.038)	(.086)
F	27.23 ***	6.37 ***	8.73 ***	5.82 ***
N	1920	1920	1080	1440

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

〈표 3-7〉 각 조건별 최초 단계 처벌에 사용된 토큰의 수 토빗 회귀분석

집단 수준에서 실시한 평균 비교분석에 이어, 개인 수준에서 처벌의 강도(처벌에 사용한 토큰의 수)를 결정하는 데 영향을 미친 요인을 각 조건 별로 확인하기 위해 토빗 회귀분석을 실시하였다(표 3-7). 다양한 변수들 가운데 본 연구에서는 참여자 자신과 처벌 대상의 기여량 차이(처벌자가 더 많이 기여했을 때, 처벌 대상이 더 많이 기여했을 때), 지금까지 획득한 총 토큰 수의 차이(처벌자가 더 많이 획득했을 때, 처벌 대상이 더 많이 기여했을 때), 지난 라운드 자신이 받은 처벌 강도가 각 라운드에서 참3여자의 처벌 의사결정에 영향을 미칠 수 있을 것으로 예상하였고 각 변수들의 영향력을 분석하였다. 그러나 앞에서와는 달리, 이번 토빗 회귀분석에 있어서는 보복 가능 조건에서 첫 번째 처벌 단계에서 발생한 처벌만을 변수로 지정하여 분석을 진행하였다. 반복된 처벌 단계에는 기여 단계가 반복되는 일 없이 바로 처벌 의사결정이 이루어지므로 자신과 상대방 사이의 기여량 차이라는 독립변수가 존재하지 않기 때문이다. 처벌이 기여량과 어떠한 상관관계를 갖는지 확인하

기 위해서, 본 연구에서는 기여 단계 직후에 이루어지는 최초의 처벌 단계에서 나타난 행위만을 변수로 설정하여 토빗 회귀분석을 실시하였다.

먼저, 지난 라운드 참여자가 받은 처벌의 강도는 그 어떤 조건에서도 현재 라운드의 처벌 의사결정에 유의한 영향을 미치지 못하는 것으로 나타났다. 이러한 결과는, 참여자가 매 라운드 서로 다른 사람들과 상호작용하는 이방인 조건의 도입을 통해 지난 라운드 자신이 받은 처벌에 대한 보복이 현재 라운드의 처벌로 이어지는 교란 효과의 발생을 성공적으로 차단할 수 있었다는 것을 의미한다.

결과 5-1: 모든 조건에서, 지난 라운드에 받은 처벌은 개인의 처벌 의사결정에 유의한 영향을 미치지 않았다.

다음으로 처벌자 자신이 처벌 대상보다 더 많은 토큰을 기여했을 때 기여량의 차이(처벌자 기여 +)는 모든 조건에서 처벌 강도와 유의한 양의 상관관계를 갖는 것으로 확인되었다. 즉 모든 조건에서 상대의 무임승차는 개인의 처벌 의사결정에 큰 영향을 미쳤다. 이는 무임승차자에 대한 처벌을 통해 협력이 유지될 수 있었다는, 처벌에 대한 기존의 관점을 지지하는 결과라고 할 수 있다.

그러나 처벌자 자신보다 처벌 대상이 더 많은 토큰을 기여했을 때 기여량의 차이(처벌자 기여 -)가 보복 불가능-평등 조건에서 처벌 강도와 유의한 상관관계를 가졌다는 결과는 문제가 된다. 상대가 자신보다 더 많이 기여할수록 더 많이 처벌하는, 반사회적 동기에 의한 처벌 발생을 증명하기 때문이다. 따라서 보복 불가능-평등 조건에서 처벌자의 기여량이 더 많았을 때와 더 적었을 때 그 차이가 모두 처벌 강도와 유의한 양의 상관관계를 가졌다는 것은 결국 처벌 대상이 무임승차자로 한정되지 않았다는 것을 의미한다. 그러나 처벌 대상이 더 많이 기여했을 때 기여량의 차이는 보복 불가능-불평등 조건에서 처벌 강도와 유의한 상관관계를 보이지 않았으며, 보복이 가능한 두 조건(평등/불평등)에서는 그것이 처벌 강도와 유의한 음의 상관관계, 즉 상대가 더 많이 기여할수록 더 적게 처벌하는 양상을 보였다.

결과 5-2: 모든 조건에서 무임승차자는 자신과 상대방 사이의 기여량 차이가 클수록 더 강하게 처벌받았다.

더 나아가, 본 연구의 공공재 게임에서 참여자는 다른 조원들이 지금까지 얼마나 많은 토큰을 획득하였는지 확인할 수 있다. 그렇다면 양자 사이의 재산 (현재까지 획득한 토큰의 총 개수) 차이 또한 처벌 의사결정에 유의한 영향을 끼쳤을 것으로 예상할 수 있다. 분석 결과, 모든 조건에서 참여자는 자신이 상대보다 더 많은 토큰을 보유하고 있을 때 그 재산 차이(처벌자 재산 +)가 크면 클수록 상대를 더 적게 처벌하는 것으로 나타났다. 이는 다른 조건이 동등하다면, 자신보다 확연하게 적은 재산을 보유하고 있는 사람보다는 자신과 비슷한 재산을 보유한 사람을 처벌하는 경향이 있다는 것을 의미한다.

반대로 자신보다 상대가 더 많은 토큰을 보유하고 있을 때 그 재산 차이(처벌자 재산 -)가 크면 클수록 상대를 더 많이 처벌하는 경향 또한 보복 가능-평등 조건을 제외한 나머지 세 조건에서 발견되었다. 처벌 대상이 더 많은 재산을 보유하고 있을 때 그 차이와 처벌 강도 사이의 유의한 양의 상관관계는 처벌의 기저에 자신에게 불리한 불평등을 기피하는 성향이 작동하고 있음을 보여준다. 그러나 오직 한 조건, 보복 가능-평등 조건에서는 처벌 대상이 더 많은 토큰을 보유하고 있을 때 그 차이가 개인의 처벌 의사결정에 유의한 영향을 미치지 못하는 것으로 나타났다. 즉 불리한 불평등을 기피하고자 하는 불평등 기피 성향이 보복 가능-평등 조건에서는 처벌로 이어지지 않았던 것이다.

결과 5-3: 보복 가능-평등 조건을 제외한 나머지 세 조건에서 상대가 자신보다 더 많은 토큰을 보유하고 있을수록 더 강하게 처벌하는 경향은 유의하였다.

3.1.6. 반복된 처벌 단계에서의 처벌 행위 분석

지금까지는 불평등과 보복 가능성의 도입이 전체적인 처벌 행위에 미치는 영향을 분석하였다. 이번에는 보복 가능 조건에서만 관찰 가능한 특별한 처벌 양상, 즉 최초 단계에서 처벌이 발생함으로써 처벌 단계가 반복되었을 때 관찰되었던 참여자의 행위에 대해 분석하여 참여자들이 처벌에 대하여 어떠한 반응을 보였는지 살펴보고자 한다. 특히 이제까지의 분석에서는 최초 단계의 처벌과 그 이후의 처벌을 모두 처벌이라는 단일한 범주에 포함시켰으나 여기에서는 보복의 발단이 되는 처벌과 그 처벌에 대한 보복으로서 발생한 처벌을 나누어 분석을 진행하고자 한다.

먼저, 이미 언급하였듯 보복 가능성이 도입되자 기여 단계 직후에 이루어지는 최초 단계에서의 처벌 행위 자체가 드물게 발생하였다. 처벌에 대한 보복이 가능할 때, 최초 단계에서의 처벌은 평등 조건에서 77건(7.13%), 불평등 조건에서 54건(3.75%)에 불과하였다. 보복이 불가능할 때 참여자들에게 주어진 한 차례의 처벌 단계에서 평등 조건에서 482건(25.10%), 불평등 조건에서 255건(13.28%)의 처벌이 발생한 것과 비교하면 이는 매우 낮은 비율임을 알 수 있다. 자신이 속한 조에서 최초 단계의 처벌이 발생한 경우에 한해서 추가적인 처벌 기회가 주어진다는 점을 감안하면, 최초 단계의 처벌 사건 발생률이 이처럼 낮다는 것은 게임 내에서 보복을 할 수 있었던 표본 또한 한정적이라는 것을 의미한다. 또한 보복 가능-평등 조건에서 발생한 77건의 최초 단계 처벌 중 정당한 처벌은 68건, 부당한 처벌은 9건 발생하였다. 보복 가능-불평등 조건에서는 총 54건 중 정당한 처벌이 25건, 부당한 처벌이 29건 발생하였다.

만약 같은 조 내에서 처벌 사건이 발생한다면, 그 조에 속한 참여자들 모두에게는 추가적인 처벌 기회가 주어진다. 이는 직접 처벌을 당한 당사자뿐만 아니라 처벌을 당하지 않은 참여자도 처벌에 참여할 수 있다는 것을 의미한다. 처벌을 당하지 않은 참여자가 반복된 처벌 단계에서 처벌에 참여할 경우, 이 행위를 보복의 범주에 포함시키는 것은 불가능하다. 오히려 이것은 보복이 아닌, 처벌이 지연되어 발생한 것이라고 보는 것이 타당하다. 하지만 지연된 처벌을 당한 참여자가 다음 단계에서 자신을 처벌한 사람을 처벌한다

면 이것은 여전히 보복으로서 유효하다고 볼 수 있다. 따라서 본 연구에서는 이러한 지연된 처벌을 최초 단계의 처벌과 함께 '보복의 발단이 되는' 처벌의 범주에 포함시키고, 여기에서 얼마나 많은 보복 사건이 발생했는지를 살펴보고자 한다.

지연된 처벌은 보복 가능-평등 조건에서 12건, 보복 가능-불평등 조건에서 3건 관찰되었다. 보복 가능-평등 조건에서 정당한 지연된 처벌은 7건, 부당한 지연된 처벌은 5건 발생하였고, 보복 가능-불평등 조건에서는 정당한 처벌이 1건, 부당한 처벌이 2건 발생하였다. 이를 최초 단계의 처벌에 더하면 보복의 발단이 되는 처벌이 보복 가능-평등 조건에서는 총 89건(정당한 처벌 77건, 부당한 처벌 12건), 보복 가능-불평등 조건에서는 총 57건(정당한 처벌 26건, 부당한 처벌 31건) 발생했음을 알 수 있다(표 3-8).

		보복의 발단이 되는 처벌		보복 발생 횟수 (발생률)	
평등 조건	정당	75	89	14 (18.67%)	21 (23.60%)
	부당	14		7 (50.00%)	
불평등 조건	정당	26	57	11 (42.31%)	17 (29.82%)
	부당	31		6 (19.35%)	

〈표 3-8〉 보복 가능 조건에서 처벌과 보복 발생 횟수

표 3-8을 참고하면 보복의 발단이 되는 처벌이 발생했다 하더라도 그에 대하여 보복을 가하는 일은 잘 관찰되지 않았음을 알 수 있다. 보복 가능-평등 조건에서는 최초 단계 처벌과 지연된 처벌을 합한 89건 가운데 21건(23.60%)에서, 보복 가능-불평등 조건에서는 57건 가운데 17건(29.82%)에서 보복이 발생하였다. 또한 보복 가능-평등 조건에서 보복 발생률은 최초 단계의 처벌이 정당할 경우 18.67%(14/75), 부당할 경우 50.00%(7/14)로, 처벌이 부당할 때 더 높은 확률로 보복이 발생하였다. 그러나 보복 가능-불

평등 조건에서 보복 발생률은 처벌이 정당할 경우 42.31%(11/26), 부당할 경우 19.35%(6/31)로 오히려 처벌이 정당할 때 더 많은 보복이 관찰되었다.

보복 가능-평등 조건에서 자신이 받은 처벌이 정당할 때보다 부당할 때 더 많은 보복이 발생하는 이유는 어렵지 않게 짐작해볼 수 있다. 보복이 처벌에 대한 반감의 표현이라고 하면 자신이 더 많은 토큰을 공공재 계정에 기여했음에도 불구하고 처벌을 받을 때 느끼는 반감이 더 클 것(Fehl *et al.* 2012)이기 때문이다. 그러나 보복 가능-불평등 조건에서는 이러한 직관과 반대되는 결과가 관찰되었는데, 이에 대해서는 다음의 두 가지 이유를 추론해볼 수 있다. 첫째로 불평등 기피 성향이 정당한 처벌에 대한 보복으로 이어졌을 수 있다. 보복 가능-불평등 조건에서 정당한 처벌은 주로 1000유형에 의해 수행되었는데(그림 3-2), 그로 인해 처벌의 정당성 여부와 관계없이 불평등 기피 성향에 의해 보복이 발생했다는 것(정당 처벌에 대한 보복 증가)이다. 둘째로 보복 가능-불평등 조건에서 나타난 250-1000 부당처벌의 결과 1000유형의 토큰 보유량이 0에 가깝게 감소했다는 것을 고려해볼 수 있다. 즉 부당처벌이 발생했음에도 불구하고 사용할 수 있는 토큰이 남지 않아서 보복을 하지 못한 것(부당 처벌에 대한 보복 감소)이다.

마지막으로 두 조건에서 발생한 전체 보복 사건 중 28건(73.68%)이 단 한 차례의 보복으로 종료되었다. 이는 보복 사건이 발생하더라도 그것이 보복의 연쇄로 이어지는 일은 드물었다는 것을 의미한다. 이를 제외한 나머지 보복 사건들에서 보복은 2회(6건), 3회(1건), 4회(2건), 5회(1건) 이어진 후 종료되었다.

3.2. 기여 분석

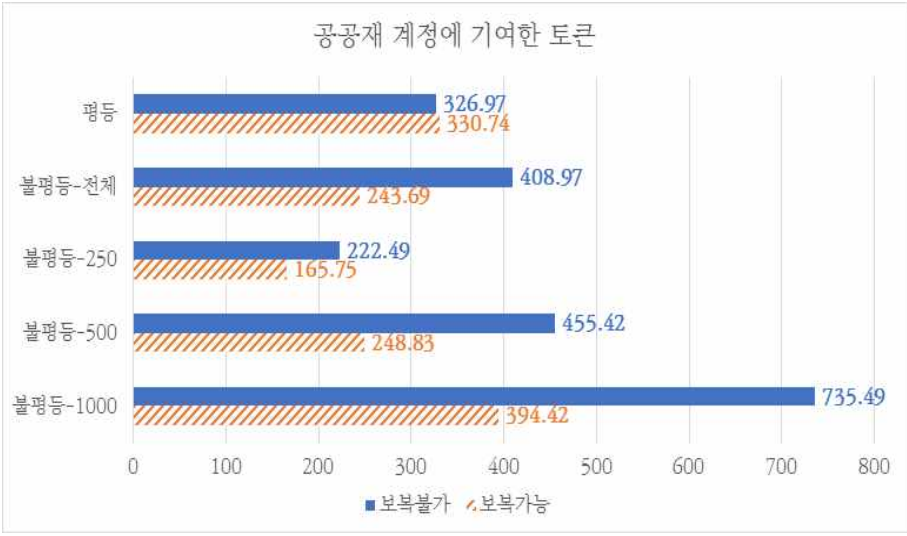
3.2.1. 집단 수준에서의 기여 비교분석

참여자들이 매 라운드 공공재 계정에 기여한 토큰의 수(기여량)는 상이한 조건에서 나타나는 인간의 협력 수준에 대한 지표로 활용된다(그림 3-3). 먼저 보복이 불가능할 때, 매 라운드 평균 기여량은 평등 조건에서 326.97개,

불평등 조건에서 408.97개로, 불평등할 때 더 높은 수준의 협력이 이루어졌다(Mann-Whitney U test: Z=-3.517, p<0.01, N=40). 반대로 보복이 가능할 때 매 라운드 평균 기여량은 평등 조건에서 330.74개, 불평등 조건에서 243.69개로 오히려 불평등할 때 협력 수준이 하락하는 것으로 나타났다(Mann-Whitney U test: Z=-3.801, p<0.01, N=35).

결과 6-1: 보복이 불가능할 때, 참여자들은 평등 조건에서보다 불평등 조건에서 더 많은 토큰을 공공재 계정에 기여하였다.

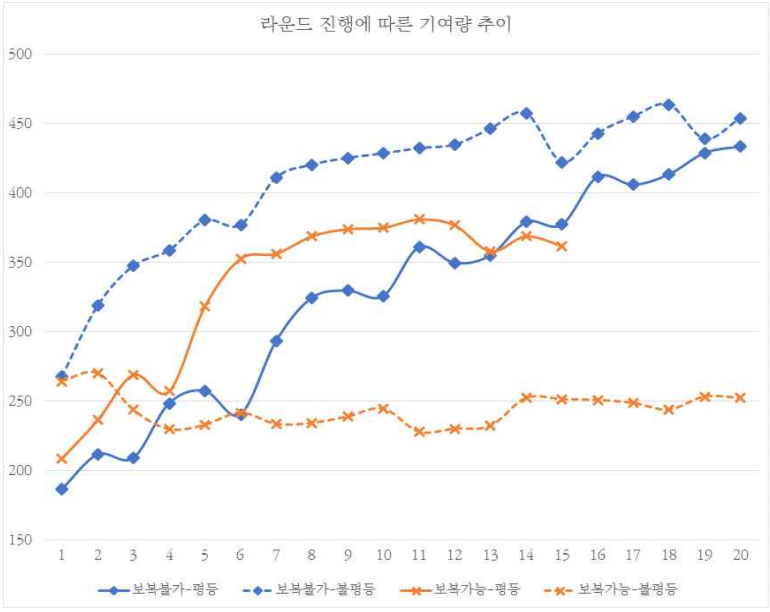
결과 6-2: 보복이 가능할 때, 참여자들은 불평등 조건에서보다 평등 조건에서 더 많은 토큰을 공공재 계정에 기여하였다.



〈그림 3-3〉 공공재 계정에 기여한 토큰

비슷한 양상을 라운드 진행에 따른 평균 기여량 추이 그래프(그림 3-4)에서도 확인할 수 있다. 보복 불가능-평등 조건(◆ 실선) 보복 불가능-불평등 조건(◆ 점선), 그리고 보복 가능-평등 조건(X 실선)에서 참여자들의 기여량은 라운드 진행에 따라 상승하였지만, 보복 가능-불평등 조건(X 점선)에서는 오히려 완만하게 감소하였다.

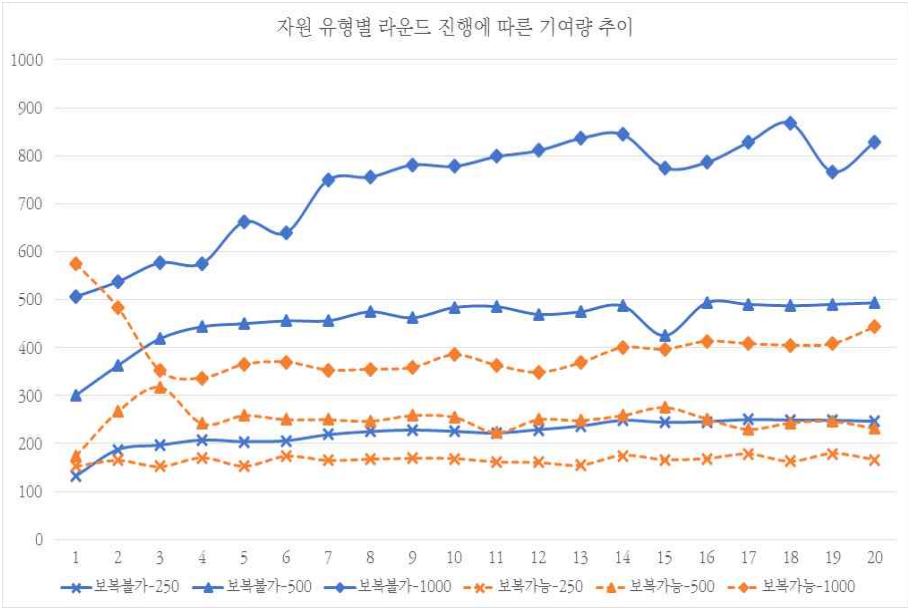
이러한 결과를 종합하면, 불평등의 심화가 참여자들의 기여량에 미치는 영향이 보복 불가능 조건과 보복 가능 조건에서 상반되게 나타났다는 것을 알 수 있다. 그 이유는 앞선 3장 1절에서 분석한 참여자들의 반사회적 처벌 양상과 연관 지어 다음과 같이 추론해볼 수 있다. 보복이 불가능할 때 높은 강도로 빈번하게 발생한 250유형과 500유형의 반사회적 처벌이 1000유형의 기여량 상승을 유도했다는 것이다. 공공재 게임의 원리에 의하면 1000유형은 500토큰을 기여하는 것만으로 무임승차를 피할 수 있다. 자신을 제외하고는 같은 조에 속한 그 누구도 500토큰 이상을 기여할 수 없기 때문이다. 반대로 이야기하자면, 이는 1000유형이 500토큰 이상을 기여할 경우 같은 조에 속한 250유형과 500유형은 필연적으로 무임승차의 이익을 누리게 되고, 1000유형은 상대적으로 적은 이익을 얻게 된다는 것을 의미한다.



〈그림 3-4〉 라운드 진행에 따른 기여량 추이

보복 불가능-불평등 구조에서 1000유형의 평균 기여량은 735.49토큰으로, 무임승차를 피할 수 있는 최저 수준인 500토큰을 통계적으로 유의한 수준에서 상회하고 있다(1 sample t-test: $t=14.928$, $p<0.01$, $N=160$). 특히

첫 라운드 1000유형의 평균 기여량은 506.25토큰으로 무임승차를 피할 수 있는 최솟값에 맞춰져 있지만 라운드 진행에 따라 상승하여 마지막 라운드에 이르러서는 828.63토큰에 달하는 것을 확인(그림 3-5, ◆ 실선)할 수 있다. 즉 1000유형은 250유형과 500유형의 반사회적 처벌이 자신에게 집중되자 이를 피하기 위하여 자신에게 상대적으로 적은 이익이 돌아옴에도 불구하고 기여량을 끌어올렸다는 것이다. 따라서 보복 불가능-불평등 조건에서 1000유형을 대상으로 발생한 반사회적 처벌은 다음의 두 가지 맥락에서 불평등을 완화하는 효과를 갖는다고 할 수 있다. 하나는 앞에서 언급하였듯 처벌의 1:3 효율로 인해 즉각적으로 발생하는 조정 효과이고, 다른 하나는 1000유형의 기여량 증가를 강제함으로 인해 기대되는 무임승차 수익 효과이다.



〈그림 3-5〉 자원 유형별 라운드 진행에 따른 기여량 추이

반대로, 보복 가능-불평등 조건에서 1000유형을 대상으로 하는 부당한 처벌은 보복 불가능-불평등 조건에 비하여 크게 감소하였다. 이는 처벌에 대한 보복이 가능해지자, 250유형과 500유형은 더 이상 반사회적 처벌을 통해 1000유형에게 기여량 증가의 압력을 넣을 수 없게 된 것으로 해석할 수 있

다. 그 결과 보복 가능-불평등 조건에서 1000유형의 평균 기여량이 무임승차를 피하기 위한 최저 수준에 미치지 못하는 394.42토큰을 기록하였고(1 sample t-test: $t=-9.776$, $p<0.01$, $N=120$), 라운드 진행에 따른 평균 기여량 역시 완만하게 감소하는 추이를 보이는 것(그림 3-5, ◆ 점선)으로 나타났다.

3.2.2. 과도한 처벌로 인한 협력 이익 파괴

위에서는 불평등의 심화가 보복 불가능 조건과 보복 가능 조건에서 매 라운드 참여자의 평균 기여량에 미치는 영향을 분석하였다. 이번에는 반대로 보복 가능성의 도입이 평등 조건과 불평등 조건에서 기여량에 미치는 영향을 분석해보고자 한다. 먼저 평등 조건에서 평균 기여량은 보복이 불가능할 때 326.97토큰, 보복이 가능할 때 330.74토큰으로 거의 차이를 보이지 않았으며 통계적으로 전혀 유의하지 않았다(Mann-Whitney u test: $Z=0.000$, $p=1.00$, $N=35$). 하지만 불평등 조건에서는 평균 기여량이 보복이 불가능할 때 408.97토큰, 보복이 가능할 때 243.69토큰으로 큰 차이를 보였고 이 차이는 통계적으로도 유의하였다(Mann-Whitney u test: $Z=-5.383$, $p<0.01$, $N=40$).

결과 7-1: 자원의 분포가 평등할 때, 참여자들이 공공재 계정에 기여한 토큰의 수는 보복 불가능 조건과 보복 가능 조건에서 유의한 차이를 보이지 않았다.

결과 7-2: 자원의 분포가 불평등할 때, 참여자들은 보복 불가능 조건보다 보복 가능 조건에서 더 적은 토큰을 공공재 계정에 기여하였다.

불평등 조건에서 보복 가능성의 도입에 따라 이와 같은 차이가 나타나는 이유에 대해서는 앞에서 이미 다룬 바 있다. 여기서는 평등 조건에서 보복 가능 여부에 따른 기여량의 차이가 나타나지 않았다는 사실에 주목하여 논의를 진행하고자 한다. 그림 3-1에 따르면, 참여자들이 평균적으로 처벌에 사

용한 토큰의 수는 보복 불가능-평등 조건에서 102.68개, 보복 가능-평등 조건에서 18.59개로 보복 가능성의 도입에 따라 처벌이 대폭 감소한 것을 확인할 수 있다. 보복 불가능-평등 조건에서 326.97토큰을 기여하고 처벌에 102.68토큰을 사용했다는 것은 참여자들이 매 라운드 평균 83.75토큰 손해를 입었다는 것[4]을 의미한다. 즉 참여자들 사이의 협력(공공재 계정으로의 기여)으로 인해 생산된 이익은 분명 존재하지만, 이것이 과도한 처벌 사용으로 인해 완전히 파괴된 것이다. 반대로 보복 가능-평등 조건에서는 처벌이 거의 사용되지 않았음(18.59)에도 보복 불가능 조건과 유사한 수준의 협력이 유지됨(330.74)으로써 협력으로 인한 이익이 보존(256.38)[5]되는 양상을 보였다.

결과 7-3: 협력을 통해 발생한 이익은 보복 불가능-평등 조건에서 과도한 처벌 사용으로 인해 완전히 소멸되었으나, 보복 가능-평등 조건에서는 처벌이 거의 발생하지 않아 보존되었다.

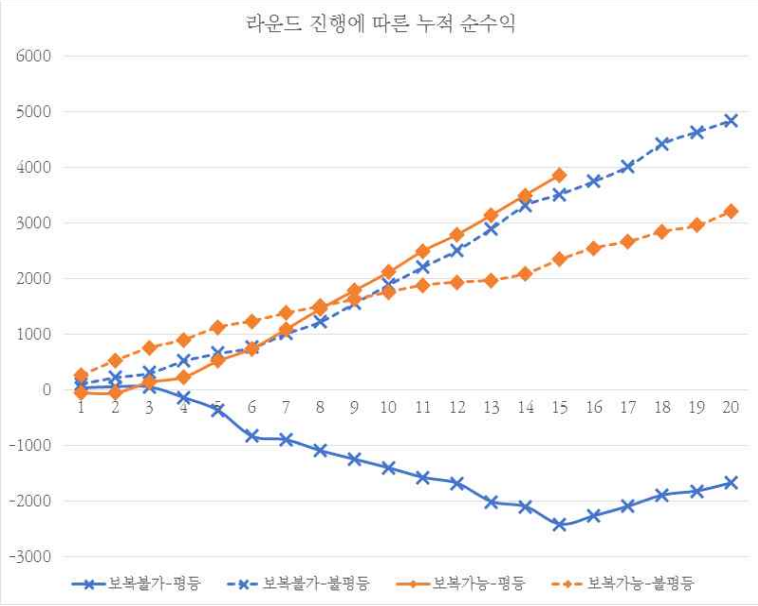
과도한 처벌로 인한 협력 이익의 파괴는 라운드 진행에 따른 참여자들의 평균 누적 순수익 그래프(그림 3-6)를 참고하면 더욱 분명하게 알 수 있다. 여기서 누적 순수익이란 참여자들이 협력과 처벌을 통해 자신들에게 기본적으로 지급된 토큰 외에 얼마나 많은 이익 혹은 손실을 얻었는지를 라운드 진행에 따라 누적하여 기록한 것이다. 보복 불가능-평등 조건(X 실선)에서 누적 순수익은 점차 감소하여 마지막 라운드에서의 누적 순수익 평균이 -1675.28토큰에 이르렀지만, 보복 가능-평등(◆ 실선)에서는 꾸준히 증가하여 마지막 라운드의 누적 순수익 평균이 3845.67[6]토큰을 기록하였다.

---

4) 326.97(공공재 생산으로 인한 이익)-102.68(자신이 처벌에 사용한 토큰)-102.68*3(자신이 받은 처벌로 인한 피해)=-83.75

5) 330.74-18.59-18.59*3=256.38

6) 256.38(라운드 순수익 평균)*15(총 라운드 수)=3845.7(총 누적 순수익 평균)



〈그림 3-6〉 라운드 진행에 따른 누적 순수익

집단 내에서 협력을 통해 발생한 이익이 사적 처벌의 과도한 사용으로 인해 완전히 소멸되고 더 나아가 손해로 이어진다는 것은 사적 처벌이 갖는 대표적인 한계 중 하나로 기존 연구들(노왁 & 하이필드 2012, Dreber *et al.* 2008, Egas & Riedl 2008, Nikiforakis & Normann 2008)에서 여러 차례 지적된 바 있다. 보복 불가능-평등 조건에서 평균 누적 순수익이 0 이하로 떨어졌다는 본 연구의 결과 또한 이를 재현한 것이라 할 수 있다.

이러한 결과는 무임승차자에 대한 처벌을 통해 협력이 진화했다는 기존의 관점에 심각한 도전을 제기한다. 다른 사람들과 집단을 이루고 협력을 하면서 살아가는 것이 평균적으로 손해라면, 집단을 이루고 살아가는 개체에 비해 집단을 이루지 않고 혼자 살아가는 개체가 더 높은 적합도를 갖게 된다. 본 연구에서 실시한 공공재 게임의 구조에 빗대어 말하자면, 그 어떤 조에도 소속되지 않은 채 공공재 계정에 기여하지도 않고, 다른 사람을 처벌하거나 다른 사람에게 처벌받는 일 없이 모든 라운드를 끝마친 참여자(독행자, loner)는 10000(=500*20)토큰을 얻지만, 다른 참여자들과 조를 이루며 공공재 계정에 토큰을 기여하고 처벌을 주고받기도 하는 참여자(집단생활자,

communer)는 (보복 불가능-평등 조건에서) 8324.72(=10000-1675.28)토큰밖에 얻지 못한다는 것이다. 집단생활자보다 독행자가 더 높은 적합도 이익을 누리게 되는 환경에서 인간은 오늘날과 같이 고도의 협력 없이는 유지가 불가능한 사회를 결코 건설할 수 없었을 것이다.

협력이 진화할 수 있으려면 무엇보다도 집단을 이루며 살아가는 사람이 독행자보다 더 큰 이익을 얻을 수 있어야 한다. 다르게 표현하자면, 협력을 통해 발생한 이익이 처벌로 인해 파괴되지 않고 보존되어야 한다는 것이다. 즉 처벌은 억제되면서도, 협력은 유지되어야 한다. 그림 3-6을 보면 알 수 있듯, 보복 불가능-평등 조건을 제외한 나머지 세 조건에서는 협력 이익이 보존되어 누적 순수익이 라운드 진행에 따라 상승하였다. 그러나 처벌에 대한 보복이 불가능한 구조가 비현실적이라는 점을 감안하여 보복 불가능-불평등 조건에서의 누적 순수익 곡선을 제외하면 실제 현실에서 협력이 이루어지는 양상은 보복 가능 조건의 두 곡선과 유사하였다고 예측해볼 수 있다.

이러한 결과를 바탕으로, 처벌과 그에 대한 보복이 자유롭게 발생 가능한, 보다 현실적인 환경에서 인간들을 협력하게 만드는 것은 실제로 발생한 처벌 그 자체가 아닌, 무임승차로 인해 처벌과 보복의 연쇄가 시작될 수 있다는 두려움이라는 것을 유추해볼 수 있다. 보복의 위험으로 인해 참여자들은 서로를 함부로 처벌하지 못하지만, 그런 상황에서도 무임승차 행위가 처벌당할 가능성을 완전히 배제할 수는 없다. 물론 보복이 가능한 환경에서 무임승차자를 처벌하는 일은 쉽지 않다. 자신이 처벌한 무임승차자로부터 보복당할 수 있기 때문이다. 그러나 반대로 무임승차자가 처벌자에게 보복을 가하는 일 또한 쉽지 않다. 마찬가지로, 그 보복이 또 다른 보복을 낳을 수 있기 때문이다. 이러한 논리는 끝없이 반복할 수 있다. 따라서 최초 단계에서 처벌이 발생할 경우 처벌자와 피처벌자 모두 보복의 부담으로부터 자유로울 수 없게 되며, 최악의 경우 양자 모두가 공멸하는 결과가 발생하게 된다. 이러한 상황에서 가장 좋은 전략은 애초에 다른 사람에게 자신을 처벌할 빌미를 주지 않는 것, 즉 협력하는 것이다.

한편 보복 가능-불평등 조건에서는 처벌에 매 라운드 평균적으로 20.88토큰이 사용되어 처벌의 사용 자체는 보복 가능-평등 조건과 거의 차이를 보

이지 않았지만 그럼에도 불구하고 누적 순수익은 15라운드 기준 2338.17토큰으로 보복 가능-평등 조건(3845.67토큰)에 비해 상당히 낮은 수준에 머물고 있다. 이러한 차이가 나타나는 이유는, 위에서 살펴보았듯 보복 가능 조건에서 구성원간 불평등이 심화되자 평균 기여량이 330.74토큰에서 243.69토큰으로 감소하였기 때문이다. 즉 처벌은 비슷한 수준으로 사용되었으나 처벌을 두려워하지 않는 1000유형의 기여량이 낮은 수준을 유지하여 평균적인 협력 수준이 하락하였고, 결과적으로 순이익이 감소했다는 것이다. 이를 고려하면 처벌에 대한 보복이 가능한 현실적인 환경에서는 구성원들이 비교적 평등할 때 더 많은 협력 이익을 누린다는 것을 알 수 있다.

4. 논의

연구를 설계하는 단계에서, 연구자는 무임승차자에 대한 처벌을 통해 협력이 진화했다는 전제에서부터 출발하여, 협력적인 사람에 대한 반사회적 처벌은 어떠한 이유에서 발생하고 또 무엇을 통해 이를 억제할 수 있는지 알아보고자 하였다. 관련된 선행 연구들을 검토한 결과, 문제가 되는 반사회적 처벌의 기저에 불평등 기피 성향이 존재하며 처벌에 대한 보복 가능성을 도입할 경우 이를 억제할 수 있으리라는 가설을 세울 수 있었고, 이러한 가설을 구조에 반영한 공공재 게임을 설계하여 이를 확인해보고자 하였다.

실험 결과 연구자가 예상한 대로 불평등 기피 성향은 반사회적 처벌과 밀접하게 연관되어 있었으며, 처벌에 대한 보복이 가능해지자 반사회적 처벌은 확연하게 감소하였다. 그러나 연구자가 미처 예상하지 못했던 결과들도 관찰되었다. 먼저, 보복 가능성의 도입으로 인해 반사회적 처벌뿐만 아니라 더 적게 기여한 사람을 대상으로 하는 정당한 처벌 또한 크게 감소하였다. 다음으로 보복 불가능-평등 조건에서 강도 높은 처벌의 빈번한 발생으로 인해 참여자들은 협력의 이익을 얻지 못하고 오히려 손해를 보았으나, 보복 가능-평등 조건에서는 처벌이 거의 발생하지 않았음에도 협력이 유지되어 참여자

들이 협력의 이익을 온전히 보존할 수 있었다. 이러한 결과들을 종합해 보면, 처벌과 그에 대한 보복이 자유롭게 이루어질 수 있는 환경에서 협력의 진화를 이끈 것은 실제 처벌이 아니라 서로에게 처벌당할 수도 있다는 두려움, 혹은 그러한 위험에 대한 인식이라는 결론이 도출된다. 결국 무임승차자에 대한 처벌을 통해 협력이 발생하고 유지될 수 있었을 것이라는 전제가 실험을 통해 오히려 반증된 것이다.

이러한 결론은 지금까지 공공재 게임을 사용한 실험 연구들에서 별다른 의심 없이 받아들여졌던 처벌의 원리, 자신이 직접 대가를 치르고 타인에게 피해를 입히는 것에 대해 비판적인 관점에서 생각할 기회를 제공한다. 처벌은 그 정의상 부당한 행위에 대하여 고통을 주는 것을 의미한다. 협력의 맥락에서 보자면 여기서 부당한 행위란 곧 협력하지 않고 협력으로 발생한 이익을 누리는 것, 즉 무임승차하는 것이다. 그렇다면 이러한 처벌의 정의가 공공재 게임의 구조에서 구현되기 위해서는 그 대상을 무임승차자로 엄격하게 한정할 수 있어야 한다.

그러나 공공재 게임의 참여자가 자신보다, 혹은 다른 조원들보다 적게 기여한 사람만을 선택할 수 있도록 제한을 두는 일은 방법론적으로 바람직하지 못하다. 그렇게 할 경우 참여자에게 남들보다 적게 기여하는 것은 잘못된 일이라는 인식을 사전에 심어줌으로써 참여자로 하여금 연구자가 원하는 방향으로 행동하도록 유도하는 효과가 발생할 수 있기 때문이다. 같은 맥락에서 연구자는 참여자 앞에서 '처벌'이라는 표현을 직접 사용하지 않아야 한다. 이는 참여자에게 자신이 누군가를 처벌한다는 인식을 미리 심어주지 않은 상태에서 다른 사람의 자원을 차감하는 행위가 과연 무임승차자에 대한 처벌의 양상을 보이는지 객관적으로 확인하기 위함이다.

이러한 이유들로 인해 일반적으로 공공재 게임에서 참여자는 같은 조에 속한 사람이면 누구라도 선택하여 그가 가진 자원을 '차감'할 수 있다는 안내를 받게 된다. 따라서 참여자가 선택한 대상이 무임승차자일 경우에만 자원을 차감하는 행위는 처벌이 될 수 있다. 즉 다른 사람의 자원을 차감하는 행위 그 자체만으로는 처벌의 조건을 만족하지 못한다는 것이다. 이를 고려하면 참여자의 차감 행위가 자신보다 더 많이 기여한 사람을 향했을 때, 이것

을 '반사회적 처벌'로 분류하는 일반적인 관행에는 사실 논리적 오류가 있는 셈이다. 공공재 게임에서 발생한 차감 행위가 더 적은 토큰을 기여한 사람을 대상으로 하지 않는다면 그 행위는 애초에 처벌이 될 수 없기 때문이다.

따라서 공공재 게임에서의 차감, 자신이 비용을 치러 가며 다른 사람이 가진 것을 없애는 행위는 그 성질에 비추어봤을 때 타인에 대한 공격성의 발현, 혹은 폭력적 행위를 모형화한 것이라 할 수 있다. 무임승차자에 대한 정당한 처벌이든 반사회적 처벌이든 근본적으로는 폭력의 범주에 포함된다는 것이다. 이를 처벌이라고 부르는 완곡어법으로 인해 공공재 게임에서 발생하는 차감 행위는 실제에 비해 덜 폭력적이고 덜 공격적인 것으로 간주되어 왔다. 물론 자신이 가진 자원을 차감당한다고 해서 그 사람이 실제로 상처를 입거나 피를 흘리는 것은 아니므로, 이러한 차감 행위를 폭력이라고 생각하기 어려운 것은 사실이다. 그러나 이는 공공재 게임에서의 차감이 폭력의 모형화한 것이기 때문이지, 그 본질이 덜 폭력적이라서 그런 것은 아니다.

그렇다면 첫 문단에서 언급한, 무임승차자에 대한 처벌을 통해 협력이 진화하였다는 기존의 관점에는 결국 폭력의 행사가 협력의 진화에 기여했다는 의미가 내포되어 있음을 알 수 있다. 협력 진화의 초기 단서를 찾을 수 있을 것으로 기대되는 소규모 부족사회에서, 대규모 현대 사회에서보다 더 높은 빈도로 폭력 사건이 관찰되었다는 민족지적 연구(Chagnon 1988, 1992, Knauft 1991, Lee 1979, Marlowe 2010)들은 이러한 관점을 지지하는 것처럼 보인다(Guala 2012: 9에서 재인용). 그러나 Guala(2012)는 이러한 폭력 사건은 대부분 배우자를 둘러싼 성적 경쟁의 과열로 인해 발생하며, 무임승차 행위로 인해 발생하는 경우는 거의 없다는 것을 지적한다. 소규모 부족사회의 무임승차자는 무시당하거나 뒷소문의 주인공이 될 수는 있지만(Dunbar 2004), 직접적으로 처벌받는 일은 드물다는 것이다. 또한 칼라하리 사막의 쿵 족 사회에서 사적으로 폭력을 행사한 사람은 집단 내에서 인정을 받기보다 기피되고, 폭력에 대한 두려움이 집단 구성원들 사이에 만연한 것으로 조사되었다(Lee 1979). 이러한 결과들은 폭력을 통해 유지되는 협력의 사례가 실험실을 벗어난 현실 사회에서는 잘 관찰되지 않는다는 것을 의미한다.

공공재 게임에서와는 달리 현실 사회에서 협력의 수단으로서 폭력이 선호되지 않는 이유는, 폭력의 사용에는 언제나 보복의 위험이 뒤따르기 때문이다. 다시 쿵 족의 사례를 들면, 자신의 친족을 살해한 사람을 죽이는 일, 즉 보복에 대해서 쿵 족은 관대한 편(Lee 1979)이다. 본 연구의 공공재 게임에서도 보복 가능성을 도입하자 참여자들의 공격성 발현(처벌 사용)은 상당히 억제되었다. 또한 폭력을 사용할 경우, 사용자와 대상 모두 손실을 입게 된다. 따라서 설령 그것이 무임승차자를 벌하기 위한 것이었다 해도, 폭력의 사용은 협력을 통해 생산된 이익을 완전히 파괴할 수 있다는 점에서 문제가 된다. 실제로 본 연구의 실험 결과, 보복 불가능-평등 조건에서 참여자들의 협력 수준이 결코 낮지 않았음에도 처벌의 과도한 사용으로 인해 평균 순이익은 0 이하로 떨어지는 것으로 나타났다. 폭력적 처벌은 사회성을 촉진하기보다는 오히려 저해한다는 것이다(Guala 2012).

이처럼 소규모 부족사회에 대한 민족지적 연구들, 그리고 공공재 게임에 보복 가능성을 도입한 실험 연구들을 종합적으로 검토하여 Guala(2012)는 협력을 유지하는 데 있어 '값비싼' 처벌의 기여가 실제에 비해 과장되어왔음을 비판하고, 더 나아가 험담(gossip)이나 따돌림(ostracism), 중앙화된 집단적 처벌 (centralized coalitional punishment) 등과 같은 '비싸지 않은' 제재 수단으로도 협력은 충분히 유지될 수 있다고 주장하였다. 그러나 그것만으로 협력에 있어서 분산화된 사적 처벌(decentralized peer punishment)의 기여를 완전히 부정하는 것은 성급한 결론일 수 있다. 그 대신 집단 내에서 구성원들이 서로에게 처벌과 보복, 즉 폭력을 사용할 수 있다는 위험에 대한 인식이 협력에 미치는 영향에 주목해볼 필요가 있다.

처벌에 관한 고전적인 공공재 게임 연구(Fehr & Gächter 2000; 2002)에서 처벌이 불가능할 때 참여자들의 기여량은 게임 초반, 지급된 토큰의 절반에 이르는 높은 수준을 기록했지만 라운드가 진행됨에 따라 점차 감소하여 게임 종료를 앞둔 시점에서는 0에 수렴하는 것으로 나타났다. 그러나 본 연구의 보복 가능-평등 조건에서 참여자들의 처벌 사용은 억제되었으나 기여량은 라운드의 진행에 따라 점차 증가하였고, 평균적인 수준에서도 보복이 불가능했을 때(따라서 처벌이 활발하게 사용되었을 때)와 큰 차이를 보이지 않

았다. 원천적으로 폭력을 사용할 수 없는 것(처벌이 불가능한 공공재 게임)과 폭력을 사용할 수 있음에도 불구하고 보복의 부담으로 인해 사용하지 않는 것(처벌과 보복이 가능한 공공재 게임)은 참여자의 협력 의사결정에 완전히 상반된 영향을 미친 것이다.

인간이 서로 협력하여 오늘날과 같은 사회를 건설할 수 있었던 것은, 인간이 폭력을 사용할 줄 모르는 아주 온순한 양과 같은 존재였기 때문이 아니다. 폭력의 사용을 완전히 배제한 채 협력 여부를 결정하는 온순한 양들의 공공재 게임에서, 참여자들의 협력 수준은 장기적으로 완전한 무임승차에 수렴하였다(Fehr & Gächter 2000; 2002). 인간들은 얼마든지 서로에게 상처를 입히는 괴물이 될 수 있다. 특히 그것이 자신의 이익과 관련된 일이라면, 상대에게 상처를 입힘으로써 상대를 끌어내리거나 자신이 더 우월한 지위를 획득할 수 있다면 주저하지 않고 폭력을 사용하려고 드는 성향이 인간에게는 내재되어 있다. 처벌이 1회만 가능한 공공재 게임에서 빈번하게 나타났던 높은 강도의 처벌을 통해 이를 확인할 수 있다.

그러나 다행히도 오늘날 인류는 그런 괴물들이 다른 모든 괴물들을 상대로 투쟁을 벌이는 홉스의 자연 상태와 같은 세상을 살고 있지 않다. 홉스는 그 이유를 훨씬 더 큰 괴물, 리바이어던과 같은 국가의 탄생에서 찾고 있지만 본 연구의 결과는 그보다 앞선 단계에서 대답을 제시하고 있다. 인간이 누군가에게 함부로 괴물(처벌)이 될 수 없는 까닭은, 그 누군가도 자신에게 괴물(보복)이 될 수 있기 때문이다. 앞에서 이미 언급했지만, 쿵 족 사회에 폭력에 대한 두려움이 만연하다(Lee 1979)는 것도 같은 맥락에서 이해할 수 있다. 폭력이 없는 세상에는 폭력에 대한 두려움도 있을 수 없다. 사람과 사람 사이에는 폭력의 씨앗이 심어져 있고, 그것이 꽃을 피우면 양자 모두 돌이킬 수 없는 상처를 입게 된다는 것을 이해할 때에만 폭력은 두려움의 대상이 될 수 있기 때문이다.

보복이 두려워 함부로 처벌하지 못하는, 아슬아슬한 비폭력의 균형을 이루고 있는 괴물들의 세계는 애초에 처벌할 줄 모르는 양들의 세계와 표면적으로 동일하지만 그 균형이 조금이라도 깨질 경우 치명적인 파국으로 치닫게 될 수 있다는 점에서 결정적인 차이를 보인다. 따라서 보복 가능 조건에서

공공재 게임의 참여자들은 함부로 무임승차자를 처벌할 수 없겠지만, 섣불리 무임승차할 수도 없고, 무임승차하지 않은 사람에게 처벌을 사용할 수도 없다. 결국 서로가 서로에게 괴물이 될 수 있기 때문에, 또 그것을 너무나도 잘 이해하고 있기에 인간은 역설적으로 오늘날과 같이 협력으로 가득한 사회를 건설할 수 있었던 것이다.

다만 처벌과 보복에 대한 두려움이 언제나 인간을 협력하게 만드는 것은 아니다. 본 연구의 보복 가능 조건에서 불평등이 심화되자 참여자들의 평균 기여량은 오히려 감소하는 것으로 나타났다. 양자 관계에서 한 쪽이 압도적으로 더 많은 자원을 가지고 있으면, 그 사람은 더 이상 상대가 자신에게 괴물이 될 것을 두려워하지 않게 된다. 그 사람이 자신에게 줄 수 있는 고통보다 훨씬 더 큰 고통을 그에게 돌려줄 수 있기 때문이다. 따라서 적게 가진 사람은 자신이 가진 것이 적기 때문에, 많이 가진 사람은 어차피 적게 가진 사람이 어차피 자신에게 해를 입힐 수 없기 때문에 적은 양의 토큰을 공공재 계정에 기여하고, 그로 인해 평균 기여량은 평등할 때에 비해 감소하게 된다.

이를 보다 현실적인 맥락에서 설명하기 위하여, 다시 한 번 쿵 족의 사례를 들어보고자 한다. 칼라하리 사막에서 현지조사를 수행하던 리처드 리는 크리스마스를 맞아 쿵 족에게 살찐 소를 대접하고자 했다. 그는 구할 수 있는 가장 큰 소를 구해서 대접했지만 부족 사람들은 그 소가 늙고 말라서 먹을 것이 없다며 그를 조롱했다. 그러나 실제로 그 소는 모든 부족 사람들이 이틀 밤낮으로 먹고도 남을 정도로 큰 소였다. 리는 이러한 일화를 통해 사냥꾼이 아무리 큰 짐승을 사냥하더라도 그 성과를 비하하고, 사냥꾼을 모욕하는 것이 쿵 족의 문화임을 깨닫는다(한국문화인류학회 2011: 77에서 재인용). 그 모욕이 갖는 의미를 묻자, 쿵 족 정보제공자는 다음과 같이 대답한다.

“어떤 사람이 너무 많은 짐승을 잡게 되면 그는 자기가 무슨 추장이나 그에 버금가는 대단한 사람이 된 걸로 착각하게 되죠. 그리고 다른 사람들을 자기 하인이나 자기보다 못한 사람으로 여기게 돼요. 그렇게 되는 것을 그냥

보고만 있어서는 안 돼요. 반드시 막아야 해요. (중략) 이런 식으로 그의 마음에 교만함이 차지 않게 하여 그를 겸손하게 만들어주는 거지요(한국문화인류학회 2011: 77에서 재인용)."

이 말을 보복 가능 조건에서 진행되는 공공재 게임의 맥락으로 옮겨 보면 다음과 같을 것이다. 모든 세상이 그러하듯, 쿵 족 사회에도 개인들 사이의 격차는 존재하며 그로 인한 불평등은 필연적이다. 또한 자신이 입은 피해를 되돌려줄 수 있는 현실적인 환경에서, 더 많이 가진 사람에게 자신보다 적게 가진 사람의 처벌은 그리 큰 위협이 되지 않는다. 처벌이 두렵지 않다면, 협력 수준 또한 자연스럽게 감소할 것이다. 자신이 불평등 조건에서 유리한 위치에 놓여 있다는 것을 깨닫고 더 이상 협력하지 않는 것, 그것이 바로 위의 정보제공자가 말하는 '교만'인 셈이다. 본 연구의 보복 가능-평등 조건과 보복 가능-불평등 조건의 누적 순수익 곡선을 비교해 보면 알 수 있듯, 더 많이 가진 사람의 이러한 교만은 협력을 통해 발생하는 집단적 이익이 감소하는 결과를 낳는다. 따라서 쿵 족 사람들이 더 많은 자원을 보유한 것으로 간주할 수 있는 뛰어난 사냥꾼을 모욕하는 것은 불평등의 심화로 인해 협력 수준이 하락하는 것을 경계하는, 반대로 말하자면 구성원들 사이의 관계를 평등하게 만들고자 하는 심리적 적응의 일종이라고도 볼 수 있을 것이다.

5. 결론

기존의 협력과 처벌 연구에서 반사회적 처벌의 문제는 여러 차례 지적된 바 있다. 본 연구에서는 이러한 반사회적 처벌을 인간의 불평등 기피 성향과 연결할 수 있으리라 가정하였고, 이를 확인하기 위해 자원이 불평등하게 분포된 공공재 게임을 시행하였다. 또한 처벌이 1회만 가능한 기존의 게임 구조가 현실을 반영하지 못한다는 지적을 고려해 참여자들에게 복수의 처벌 기회를 부여하여 처벌에 대한 보복이 가능하게끔 만들었고, 보복 가능 조건에서는 불평등 기피 성향에 기반을 둔 반사회적 처벌이 억제될 것이라고 판단하였다.

실험 결과 이러한 예측은 상당 부분 타당한 것으로 밝혀졌다. 먼저 보복이 불가능한 조건에서 불평등이 도입되자, 보유한 자원이 적은 250유형과 500유형의 참여자는 더 많은 자원을 보유한 1000유형의 참여자를 대상으로 강도 높은 반사회적 처벌을 가하는 것으로 나타났다. 1000유형이 250유형과 500유형에 비해 더 많은 토큰을 기여했음에도 불구하고 불평등 기피 성향으로 인해 1000유형과 자신 사이의 격차를 줄이고자 하는 목적에서 처벌이 발생한 것이다.

한편 동일한 불평등 조건에서 참여자들에게 복수의 처벌 기회를 부여함으로써 처벌에 대한 보복이 가능해지자, 1000유형을 대상으로 하는 500유형의 반사회적 처벌은 큰 폭으로 감소하였다. 그러나 1000유형을 대상으로 하는 250유형의 반사회적 처벌은 그다지 감소하지 않은 것으로 나타났는데, 이는 1총 12명의 250유형 중 단 두 명이 1000유형에게 강도 높은 반사회적 처벌을 사용하였기 때문이었다. 현재 라운드에서 획득한 토큰만을 처벌에 사용할 수 있는 게임의 구조를 이용하여 1000유형이 자신에게 보복을 할 수 없게끔 1000유형의 토큰보유량을 0으로 만드는 처벌을 가한 것이다. 나머지 10명 중 9명은 1000유형을 대상으로 단 한 차례의 부당한 처벌도 가하지 않았다. 따라서 처벌 강도가 아닌 처벌이 발생한 빈도를 대상으로 비교하였을 때에는 불평등 조건에서 처벌에 대한 보복이 가능해지자 1000유형을 대상으로 하는 250유형의 반사회적 처벌의 발생 빈도가 통계적으로 유의한 수준에서 크게 감소하는 것으로 나타났다.

실험을 기획하는 단계에서 연구자가 미처 고려하지 못했거나 예상과 어긋난 결과들을 관찰할 수 있었다. 하나는 보복 가능성의 도입으로 인해 반사회적 처벌뿐만 아니라 정당한 처벌 역시 크게 감소하였다는 것이다. 이러한 결과는 사회적 상호작용이 벌어지는 현실적인 환경에서는 서로가 서로를 처벌하는 사건이 실제로는 잘 발생하지 않는다는 것을 의미한다. 즉 집단 구성원 간의 사적 처벌이 협력의 진화에 기여한 바가 그다지 크지 않을 수 있다는 것이다.

미처 예상하지 못한 또 다른 결과는 불평등 조건에서 참여자들의 협력 수준이 보복 가능 여부에 따라 크게 다르게 나타났다는 것이다. 보복이 불가능

한 조건에서 250/500유형이 1000유형에게 집중적으로 부당한 처벌을 가함에 따라 1000유형은 자신이 얻는 이익이 상대적으로 크지 않음에도 불구하고 기여량을 높은 수준으로 끌어올렸으나, 보복이 가능한 조건에서는 250/500유형의 부당한 처벌이 억제됨에 따라 1000유형의 기여량은 보복 불가능 조건과 비교했을 때 상당히 낮은 수준을 유지했다.

마지막으로, 참여자간 불평등이 존재할 때에는 보복 가능성의 도입이 협력 수준의 하락으로 이어지지만, 참여자들이 평등할 때에는 보복 가능 조건에서도 보복 불가능 조건과 유사한 수준의 협력이 유지된다는 결과도 주목할 만하다. 더 나아가 보복 불가능-평등 조건에서는 높은 강도의 처벌이 발생하여 협력으로 인해 발생한 집단 이익이 완전히 제거되었지만, 보복 가능-평등 조건에서는 처벌이 거의 발생하지 않았음에도 유사한 수준의 협동성이 유지되었으며, 결과적으로 협력으로 인해 발생한 집단 이익이 보존되는 현상을 관찰할 수 있었다.

처벌(1회)이 가능한 공공재 게임에서 처벌의 과도한 사용으로 인한 협력 이익의 소멸은 이미 다른 연구들(노왁 & 하이필드 2012, Dreber *et al.* 2008, Egas & Riedl 2008, Nikiforakis & Normann 2008)에서 여러 차례 지적된 바 있다. 본 연구의 공공재 게임에서 보복이 불가능할 때 관찰된 높은 처벌 강도는 이러한 연구 결과를 재현하고 있다. 그러나 보복이 가능한 평등 구조의 공공재 게임에서 처벌은 거의 발생하지 않았음에도 보복이 불가능할 때와 유사한 수준의 협동성이 유지되었다는 사실은 인간 협력의 진화에 있어 처벌이 수행한 기능을 새로운 각도에서 바라보게 해준다. 이는 처벌과 그에 대한 보복이 자유롭게 발생 가능한, 보다 현실적인 환경에서 인간들을 협력하게 만드는 것은 실제로 발생한 처벌 그 자체가 아닌 서로에게 처벌당할 수도 있다는 두려움이라는 것이다.

한편 구성원들이 평등할수록 보복 가능성이 처벌을 억제하면서도 협력을 유지하는 효과가 더 크게 나타나는 것을 관찰할 수 있었다. 보복이 가능한 처벌 환경에서 참여자들 사이의 불평등이 도입되자, 불평등 기피 성향에서 비롯된 부당한 처벌은 증가하였고, 공공재 계정으로의 기여량은 반대로 감소하였기 때문이다. 보복이 불가능한 조건에서 참여자들 사이에 불평등이 도입

되자 평균 기여량이 증가했던 것과는 정반대의 결과가 관찰된 것이다.

참고문헌

Balafoutas, Loukas, Kristoffel Grechenig, and Nikos Nikiforakis, 2014, “Third-party punishment and counter-punishment in one-shot interactions,” *Economics Letters* 122(2): 308-310.

Bone, Jonathan E., Brian Wallace, Redouan Bshary, and Nichola J. Raihani, 2015, “The effect of power asymmetries on cooperation and punishment in a prisoner’s dilemma game,” *PloS one* 10(1): e0117183.

Bone, Jonathan E., and Nichola J. Raihani, 2015, "Human punishment is motivated by both a desire for revenge and a desire for equality," *Evolution and Human Behavior* 36(4): 323-330.

Boyd, Robert, Herbert Gintis, and Samuel Bowles, 2010, "Coordinated punishment of defectors sustains cooperation and can proliferate when rare," *Science* 328(5978): 617-620.

Clutton-Brock, T. H., and G. A. Parker, 1995, “Punishment in animal societies,” *Nature* 373(6511): 209.

Dawes, C. T., J. H. Fowler, T. Johnson, R. McElreath, and O. Smirnov, 2007, “Egalitarian motives in humans,” *nature* 446(7137): 794-796.

Denant-Boemont, Laurent, David Masclet, and Charles N. Noussair, 2007, "Punishment, counterpunishment and sanction enforcement in a social dilemma experiment." *Economic theory* 33(1): 145-167.

Dreber, Anna, and David G. Rand, 2012, "Retaliation and antisocial punishment are overlooked in many theoretical

models as well as behavioral experiments," *Behavioral and brain sciences* 35(1): 24-24.

Dreber, Anna, David G. Rand, Drew Fudenberg, and Martin A. Nowak, 2008, "Winners don't punish," *Nature* 452(7185): 348-351.

Dunbar, Robin IM, 2004, "Gossip in evolutionary perspective," *Review of general psychology* 8(2): 100.

Egas, Martijn, and Arno Riedl, 2008, "The economics of altruistic punishment and the maintenance of cooperation," *Proceedings of the Royal Society of London B: Biological Sciences* 275(1637): 871-878.

Elster, Jon, 1990, "Norms of revenge," *Ethics* 100(4): 862-885.

Engelmann, Dirk, and Nikos Nikiforakis, 2015, "In the long-run we are all dead: On the benefits of peer punishment in rich environments." *Social Choice and Welfare* 45(3): 561-577.

Fehl, Katrin, Ralf D. Sommerfeld, Dirk Semmann, Hans-Jürgen Krambeck, and Manfred Milinski, 2012, "I dare you to punish me—vendettas in games of cooperation," *PloS one* 7(9): e45093.

Fehr, Ernst, and Simon Gächter, 2000, "Cooperation and punishment in public goods experiments," *The American Economic Review* 90(4): 980-994.

Fehr, Ernst, and Simon Gächter, 2002, "Altruistic punishment in humans," *nature* 415(6868): 137-140.

Fischbacher, Urs, 2007, "z-Tree: Zurich toolbox for ready-made economic experiments." *Experimental economics* 10(2): 171-178.

Guala, Francesco, 2012, "Reciprocity: Weak or Strong? What punishment experiments do (and do not) demonstrate," 2012, *Behavioral and Brain Science* 35: 1-59.

Henrich, Joseph, Richard Mcelreath, Abigail Barr, Jean Ensminger,

Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, Michael Gurven, Edwins Gwako, Natalie Henrich, Carolyn Lesorogol, Frank Marlowe, David Tracer, and John Ziker, 2006, “Costly punishment across human societies,” *Science* 312(5781): 1767-1770.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter, 2008, "Antisocial punishment across societies," *Science* 319(5868): 1362-1367.

Hilbe, Christian, and Arne Traulsen, 2012, "Emergence of responsible sanctions without second order free riders, antisocial punishment or spite," *Scientific reports* 2: 458.

Janssen, Marco A., and Clint Bushman, 2008, "Evolution of cooperation and altruistic punishment when retaliation is possible," *Journal of theoretical biology* 254(3): 541-545.

Johnson, Tim, Christopher T. Dawes, James H. Fowler, Richard Mcelreath, and Oleg Smirnov, 2009, “The role of egalitarian motives in altruistic punishment,” *Economics Letters* 102(3): 192-194.

Lee, Richard B., 1979, *The! Kung San: men, women and work in a foraging society*, CUP Archive.

Leimgruber, Kristin L., Alexandra G. Rosati, and Laurie R. Santos, 2016, "Capuchin monkeys punish those who have more," *Evolution and Human Behavior* 37(3): 236-244.

Marczyk, Jesse, 2017, "Human punishment is not primarily motivated by inequality," *PloS one* 12(2): e0171298.

Nikiforakis, Nikos, 2008, “Punishment and counter-punishment in public good games: Can we really govern ourselves?,” *Journal of Public Economics* 92(1): 91-112.

Nikiforakis, Nikos, Charles N. Noussair, and Tom Wilkening, 2012,

"Normative conflict and feuds: The limits of self-enforcement," *Journal of Public Economics* 96(9): 797-807.

Nikiforakis, Nikos, and Hans-Theo Normann, 2008 "A comparative statics analysis of punishment in public-good experiments," *Experimental Economics* 11(4): 358-369.

Przepiorka, Wojtek, and Andreas Diekmann, 2013, "Individual heterogeneity and costly punishment: a volunteer's dilemma," *Proceedings of the Royal Society of London B: Biological Sciences* 280(1759): 20130247.

Prediger, Sebastian, 2011, "How does income inequality affect cooperation and punishment in public good settings?," *Joint discussion paper series in economics.*

Tan, Fangfang, 2008, "Punishment in a linear public good game with productivity heterogeneity," *De Economist* 156(3): 269-293.

Raihani, Nichola J., and Katherine McAuliffe, 2012, "Human punishment is motivated by inequity aversion, not a desire for reciprocity," *Biology letters* 8(5): 802-804.

Rand, David G., and Martin A. Nowak, 2011, "The evolution of anti-social punishment in optional public goods games," *Nature communications* 2: 434.

Wolff, Irenaeus, 2012, "Retaliation and the role for punishment in the evolution of cooperation," *Journal of theoretical biology* 315: 128-138.

김상인, 2006, "구성원간 불평등이 집단 내 협력에 미치는 영향 - 공공재 게임을 통한 진화심리학적 연구," 서울대학교 대학원 인류학과 석사학위논문.

노왁, 마틴, 로저 하이필드 (허준석 역), 2012, 『초협력자』, 서울: 사이언스 북스.
한국문화인류학회, 2011, 『낯선 곳에서 나를 만나다』, 서울: 일조각.

Abstract

The effects of inequality and retaliation on peer punishment and cooperation in public goods games

Hwang Joon
Department of Anthropology
The Graduate School
Seoul National University

A number of experimental studies have shown that costly peer punishment promotes cooperation among individuals in a public goods game (PGG). In the standard PGG settings, all the players are assumed to be equal in terms of their endowments. However, from the evolutionary perspective, assuming perfect equality is unrealistic because in every social species, hierarchy is evident among individuals. Furthermore, it is also problematic that the researchers only allowed for a single punishment stage in standard PGGs, because in nature the punisher can be retaliated by the punished. In this research, a series of PGGs was conducted in 2*2 (equal/unequal endowments * single/repeated punishment stage(s)) conditions to examine the effects of inequality and retaliation on the player's behavior.

In the inequality condition with only a single punishment stage, low-endowment players actively punished high-endowment players in order to restore equality. Observed punishments against players

with higher endowment were mainly anti-social, because in many cases high-endowment players made higher contributions. However, anti-social punishments motivated by inequality aversion were suppressed when retaliation against punishment became possible in PGG. Low-endowment players could not accept the risk of being retaliated by the high-endowment players.

High-endowment players in a single punishment condition continuously raised their contributions in order to avoid the punishments. As a result, in a single punishment condition, average contribution levels were higher in the environments where endowments were differentially distributed. By contrast, when the punishment stages were repeatedly given, high-endowment players did not raise their contributions, because low-endowment players did not punish them by the fear of retaliation. As a result, in repeated punishment condition, average contribution levels were lower when income inequality was introduced in PGG.

In the equality condition, endowments spent on the punishment were higher when the retaliation against punishment was prohibited. Meanwhile, players contributed similar amount of endowments in both punishment conditions. Therefore players in the repeated punishment condition could preserve the benefits produced through cooperation. This result shows us that in the realistic environment where the punishers can be retaliated, people can cooperate without wasting their resources on the punishment.

keywords : Cooperation, Punishment, Inequality, Retaliation, Public Goods Game

Student Number : 2015-22542