CHUL HAK SA SANG

Journal of Philosophical Ideas

Vol. XX June, 2005

Institute of Philosophy, Seoul National University
Seoul, Korea

ISSN 1226-7007

철 학 사 상

서울대학교 철학사상연구소

20

【특집】

경제적 합리성과 게임이론

김 완 진*

【주 요 어】 게임이론, 죄수의 역설, 지네게임, Nash균형, 역진귀납
【Keywords】 game theory, prisoner's dilemma, centipede game, Nash equilibrium, backward induction

I. 서 론

경제학은 합리성의 학문이라 해도 과언이 아니다. 모든 경제문제의 근저에는 인간의 필요에 비해 그를 충족시키는 물질적 자원은 희소하다는 사실이 있고, 그 때문에 희소한 자원을 효과적으로 활용해서 생존하려는 노력이 경제적 행위의 본질이라 할 수 있다. 부의 축적과 경제발전은 적절한 합리적인 선택행위를 통해서 비로소 가능한 것이다.

경제학에서 합리적 행동의 기준에 관해 명확한 형태의 답이 주어진 것은 1870년대에 시작된 소위 한계혁명의 결과라고 할 수 있다. 한계혁명이라는 새로운 파라다임을 주도한 S. Jevons, K. Menger, L.

* 서울대학교 경제학부 교수

Walras 등의 학자들은 각 개인의 합리적 경제 행위를 분석하는데 관심을 가졌다. 그들은 당시의 철학적인 사조인 공리주의적인 관점을 채택하여 개인의 효용극대화를 가정한 효용이론을 전개하였다. 효용이론은 명백히 도구적 합리성개념에 입각하여 결과주의적 입장을 취하고 있다. 개인의 행위의 목표가 되는 효용 혹은 만족도는 합리적 판단의 대상이 될 수 없고 단지 그 효용을 증대시키는 수단의 선택에 관한 합리적인 판단이 가능하다는 것이다. 또한 어떤 행동이 합리적인지는 오직 결과되는 효용의 크기에 의해 판단해야 한다는 점에서 결과주의라 할 수 있다.

대다수의 경제학자는 효용극대화 혹은 이윤추구라는 단순한 합리성의 가정만으로 충분히 경제현상을 설명하는데 어려움이 없다고 여기고 합리성의 개념적인 분석에 큰 관심을 기울이지 않아 왔다. 그러나 최근 합리적 행동에 관한 이론적인 연구가 불확실한 상황이나 게임적인 상황하에서의 합리적 행동과 집단적 행동의 합리성으로 영역을 넓혀 감에 따라 합리성 개념에 매우 중요한 변화가 초래되고 철학적인 개념분석에도 영향을 미칠 수 있는 결과들이 밝혀지고 있다.

경제적인 합리성의 개념과 관련하여 게임이론은 하나의 커다란 분수령이 되었다. 게임이론이 도입되기 이전의 경제학의 전통적인 분석대상은 완전경쟁시장이다. 수많은 소규모의 수요자와 공급자가 참가하는 완전경쟁시장에서는 각 경제주체의 행동이 타 경제주체에 미치는 영향이 극히 미미하기 때문에 각자는 타인의 행동에 대해 주의할 필요가 없이 주어진 시장가격하에서 자신의 이익을 극대화하는 행동을 하면 되는 것이다. 이 때문에 완전경쟁하에서는 그 이름과는 달리 각 경제주체는 경쟁을 의식할 필요가 없게 된다. 각자는 자신이 변화시킬 수없는 객관적인 환경에 적응하여 자신의 이익을 최대화하는 것이 합리적인 행동이 된다. 이러한 합리적 행동은 자신이 통제할 수 있는 변수와 개관적으로 주어진 환경을 나타내는 파라메터가 상호독립적으로 주어지는 수학적인 극대화문제로 정식화가 가능하다.

그러나 완전경쟁적인 시장은 극히 드물게 존재하는 것이 현실이다. 특히 최근의 공산품 시장은 대부분 소수의 기업들이 경쟁하는 독과점의 형태를 띄고 있다. 이런 독과점상황에서는 각 기업의 행동은 상대방 기업에 영향을 미치고 따라서 그에 대한 상대방 기업의 대응을 미리 예측하여 의사결정에 반영하는 것이 필요하게 된다. 이와 같이 경쟁의 상대가 어떻게 반응할 지를 예측하고 그것을 고려하여 행동하는 것을 전략적 행동이라 부른다. 독과점적인 시장에서 전략적인 행동은 합리적 행동의 필수요건이 된다. 상대방의 행동을 그저 객관적으로 불변하는 파라메터로 간주하고 단순히 자신의 이익을 극대화하는 행동은 파라메터적 행동이라 부를 수있다. 파라메터적 행동은 복잡한 세상에서 소박하고 단순한 행동양식이라 할 수 있고, 이러한 행동은 독과점적 상황에서는 더 이상 합리적인 행동이 될 수 없음은 명백하다. 파라메터적인 행동에 비해 전략적인 행동은 매우 복잡한 고려를 해야 할 뿐 아니라 논리적인 난제를 내포하고 있다.

일반적으로 전략적 행동이 요구되는 상황은 독과점시장이라는 경제적인 분야에 국한되지 않고 이해관계가 상호의존적인 다양한 사회현상, 국가간의 경쟁 등의 상황으로 일반화될 수 있으며 이러한 상황을 게임적인 상황이라고 부를 수 있다. 게임이론은 게임적인 상황에서 합리적인 개인들이 어떻게 상호작용하는가에 관해 분석하는 이론이다.

게임적인 상황에서 합리적인 행동을 결정하는데 피할 수없는 논리적인 문제를 살펴보기 위해 삼국지의 유명한 적벽대전을 예로 들어보자. 적벽대전에서 조조가 대패하여 퇴각하는 도중에 갈라진 길을 만나게 된다. 한쪽은 숲이 우거진 소로이고 다른 한쪽은 평단한 대로인데 소로의 숲에서는 복병이 있는 것처럼 연기가 피어오르고 있다. 이때 조조는 연기는 매복을 가장한 제갈량의 계략이라고 판단하고 소로를 택하였으나 제갈량은 실제로 소로에서 매복하고 있었으므로 조조가 큰 낭패를 당하게 되었다는 이야기가 있다. 조조가 연기를 제갈량의 속임수라고 생각하고 소로를 택했으나 제갈량은 조조의 그런

생각까지 간파하여 소로에 실제 복병을 둠으로써 이길 수 있었던 것이다. 조조가 제갈량의 이런 이중의 계략을 간파했더라면 대로를 택하고 유유히 지나갈 수 있었을 것이라는 점에서 조조의 결정은 합리적이지 못했다고 할 수 있다. 그런데 만약 조조가 이와 같이 제갈량의 생각을 간파하고 대로를 택했다면 이번에는 제갈량이 합리적이지 못한 것이 된다. 그렇다면 제갈량은 조조의 이러한 추론까지도 간파하고 대로를 택한다면 다시 제갈량이 이기고 조조가 패배할 것이 아닌가? 조조가 또 이런 제갈량의 의도를 간파한다면? 다시 제갈량이 조조의 생각을 간파한다면? … 이러한 추론은 끝없이 이어지고 결론을 내릴 수없게 될 것이다. 결국 합리적인 행동을 결정하기 위한 분석은 불가능하게 되는 것은 아닐까? 다시 말해 조조와 제갈량이 모두 합리적이라는 가정은 상호 모순되는 것이라고 결론을 내려야 하지 않을까?

여기서 직면하는 문제는 궁극적으로 다음과 같은 self-reference의 문제가 숨어 있다. 즉, 나의 행동을 결정하기 위해서는 상대방의 행동을 예측해야 한다. 그런데 상대방의 행동을 예측하기 위해서는 상대방이 나의 행동을 어떻게 예측할 지를 예측해야 한다. 다시 말해 나의 합리적인 행동을 결정하기 위해서는 나의 합리적인 행동을 예측해야만 하는 self-reference의 상황에 봉착하게 되는 것이다. 이러한 점에서 게임이론적 상황에서의 합리성의 문제는 논리학에서 역설의 문제와 유사한 구조를 갖고 있다고 할 수 있다.

이 글은 경제학에서 최근 합리적 행동에 관한 연구가 어떻게 진행되고 있는지 개관하고 합리성의 개념과 관련된 문제들을 제기함으로써 철학 등 타 학문분야에서의 합리성 연구와 접점을 마련하는데 그 목적이 있다. 특히 이 글에서는 최근의 게임이론의 연구성과가 합리성의 개념을 새롭게 조명할 수 있는 시각을 제공할 수 있다는 점을 부각시키려 한다.

II. 합리적 행동이론의 분류

Harsanyi(1977)는 합리적 행동의 이론을 세 분야로 구분하고 있다. 첫째는 불확실성이 없는 상황하에서의 합리적 행동을 설명하는 이론으로서 전통적인 효용이론이 이에 속한다. 효용이론은 1950년대 수학적인 기법을 통해 매우 정밀한 이론으로 발전하였으며 효용이라는 형이상학적 개념이 선호preference라는 행동주의적인 개념으로 대체되었다. 효용이론에서 가장 큰 논란은 합리성이 규범적인 것인지 설명적descriptive인 성격인지에 대한 것이다. 실제 경제현상의 설명에서 합리적 행동의 가정이 얼마나 타당한지에 대한 논란은 계속되고 있다. 합리적 행동의 가정이 현실을 설명하는데 얼마나 유용한지에 대한 방법론적인 문제가 존재한다.

둘째는 불확실한 상황하에서의 행동이론이다. J. von Neumann 과 O. Morgenstern(1944), L. Savage(1954) 등에 의해 정립된 기대효용이론이 이 분야의 대표적인 이론이다. 이 이론은 효용이론과 같이 도구적 합리성과 결과주의적인 관점을 견지하면서 불확실한 상황하에서도 효용극대화분석이 가능함을 보여주고 있다. 전통적으로 경제학에서는 불확실성uncertainty과 위험risk을 구분한다. 객관적인 확률을 계산할 수 있는 상황을 위험이라고 부르고, 객관적인 확률을 계산할 수없는 상황, 예를 들어 일회적인 사건의 경우에는 불확실성이라 부를 수 있다. 후자의 경우에는 Bayes의 공식에 의해 주관적인 확률을 계산하여 기대효용을 극대화하는 이론이 제시되고 있다. 이것을 Bayes 합리성이라 한다.

최근에는 불확실한 상황하에서는 결과뿐 아니라 선택의 과정도 매우 중요하게 되는 사례가 제기됨으로써 결과주의적인 관점에 의문을 제기하는 이론도 나타나게 되었다.

불확실성하의 행동이론의 중요한 응용분야의 하나는 윤리학이다.

불확실성하의 행동이론은 공리주의 윤리학의 강력한 근거로 제시되었다(Harsanyi 1953, 1955). 이 점에서 윤리학도 합리적인 의사결정 이론의 한 분야로 분류할 수 있게 된다.

셋째는 게임적인 상황하에서의 합리적 행동이론이다. 게임적인 상황이란 나의 선택이 타인에게 영향을 미치고 그로 인한 타인의 행동의 변화가 다시 나의 결과에 영향을 미치게 되는 상황을 말하는데 다른 말로 표현하면 해동의 결과가 상호의존적인 상황이라 할 수있다. 이런 상황하에서는 상대방의 행동에 대한 예측이나 기대를 바탕으로 나의 행동을 결정하는 전략적 행동strategic behavior이 합리적인 행동이 된다. J. von Neumann 과 O. Morgenstern(1944)의 책 "게임이론과 경제적 행동"(Theory of Games and Economic Behavior)은 이러한 전략적 행위의 중요성을 제기하고 이론적으로 분석한 최초의 책이 되었다. 그 후 게임이론은 경제이론에서 필수불가결한 분석도구로 자리잡게 되었으며 경제현상 뿐 아니라 다양한 사회현상을 설명하는 포괄적인 이론으로 발전하였다.

III. 정규형 게임과 전개형 게임

게임적인 상황을 크게 협조적 게임cooperative game과 비협조적 게임non-cooperative game으로 분류하는 것이 일반적이다. 협조적 게임이란 게임의 경기자들이 합의를 통해서만 행동을 결정할 수 있고 각자의 독자적인 행동은 허용되지 않는 상황을 말한다. 예들 들면 노사협상과 같은 상황에서는 어느 일방이 독자적으로 자신의 몫을 결정할 수가 없기 때문에 협상을 통해서 각자의 행동을 조정해야 한다. 이런 점에서 노사협상은 협조적 게임의 대표적인 예가 된다. 반면에 비협조적인 게임은 경기자가 각각 자신의 행동을 선택할 수가 있고 그 선택의 결과 경기자들의 보수가 결정되는 상황을 묘사하고 있다. 현실의 상황은 협조적 게임과 비협조적 게임의 두 측면이 모두 복

합적으로 혼재한다고 할 수 있다. 이 논문에서는 비협조적 게임에 초점을 맞추어 분석하고자 한다.[1)]

비협조적인 게임은 다시 정규형 게임normal form game과 전개형 게임extensive form game으로 분류된다.

1. 정규형 게임과 죄수의 역설 게임

먼저 정규형 게임을[2)] 살펴보기로 하자. 정규형 게임은 경기자들이 자신의 전략을 서로 상대방의 전략선택을 모르는 상태에서 동시에 전략을 선택하고, 그 결과 각자의 보수가 정해지는 게임을 말한다. 보다 추상적으로 말하면 정규형 게임은 경기자의 집합 $N=\{1,2,\ldots,n\}$, 각 경기자 $i=1,2,\ldots,n$의 전략집합 S_i, 그리고 경기자 i의 보수함수 $u_i(s_1,s_2,\ldots,s_n)$로 구성된다.

구체적인 예로서 동전맞추기 게임을 생각해보자. 두 명의 경기자 1과 2가 각각 자신의 동전의 앞 면(H)과 뒷 면(T) 중 한 면을 동시에 선택한다. 그 결과 두 경기자가 같은 면을 선택하였으면 경기자 2가 경기자 1에게 1원을 지불하고, 다른 면을 선택하였으면 경기자 1이 1에게 1원을 지불한다고 하자. 이 게임의 경기자의 집합은 {1,2}이고, 각 경기자의 전략의 집합은 모두 {H, T}이며, 경기자들의 보수는

$$u_1(H,H)=u_1(T,T)=u_2(H,T)=u_2(T,H)=1,$$
$$u_2(H,H)=u_2(T,T)=u_1(H,T)=u_1(T,H)=-1$$가 된다.

---

1) 비협조적 게임이라고 해서 대립과 갈등만이 존재하는 상황을 말하는 것은 아니다. 협조적 게임과 비협조적 게임의 구분은 게임의 규칙의 차이에 따라 이루어진다고 볼 수 있다. 두 종류의 게임의 관계에 관해서는 D. Kreps(1990) p.355 참조.

2) 정규형 게임은 정태적 게임(static game), 동시선택 게임(simultaneous move game), 행렬게임(matrix game) 등으로 부르기도 한다.

이 게임을 다음과 같이 행렬 형태로 표시할 수 있다.

	H	T
H	1, -1	-1, 1
T	-1, 1	1, -1

여기서 첫 열에 있는 H와 T는 경기자 1의 전략들이며, 첫 행에 있는 H와 T는 경기자 2의 전략들을 나타낸다. 각 항에 있는 숫자의 짝은 처음 숫자가 경기자 1의 보수, 다음 숫자가 경기자 2의 보수를 각각 의미한다. 예컨대 두 경기자가 모두 앞면(H)을 택할 때 경기자 1의 보수는 1이고 경기자 2의 보수는 -1이다.

이 게임의 중요한 특징은 두 경기자의 보수의 합이 항상 0이 된다는 점이다. 이 게임에서는 한 경기자가 이익을 얻으면 다른 경기자는 그만큼 손해를 보게 되므로 두 경기자의 이해가 완전히 상반되는 이러한 게임을 일반적으로 영합게임 혹은 제로섬게임이라 부른다. 카드게임이나 정치적 상황 등은 제로섬 게임이지만 많은 경제적인 상황은 제로섬게임이 아니다. 경제적인 활동은 대부분 그 참여자들 모두에게 이익이 되기 때문에 포지티브섬 게임이라고 부르기도 한다.

제로섬이 아닌 경우 다양한 사회적 현상들을 나타내는 게임들의 예가 많이 제시되었다. 그중 몇 개의 흥미로운 예를 들어 보면 다음과 같다.

먼저 죄수의 역설prisoner's dilemma이라 불리는 게임이 있다. 이 게임은 다음과 같은 행렬로 요약할 수 있다.

	C	D
C	2, 2	0, 3
D	3, 0	1, 1

두 명의 죄수가 각각 심문을 받고 있다. 각 죄수는 범죄를 부인하거나(C) 자백을 하거나(D) 둘 중에서 선택해야 한다. 모두 부인하면(C, C) 가벼운 처벌을 받고 풀려난다.(2, 2) 모두 자백하면(D, D) 중한 처벌을 받는다(1, 1). 그러나 한 죄수가 자백을 할 때(D) 다른 죄수가 범죄를 부인하면(C) 부인한 죄수는 가중 처벌을 받는 대신에(보수 0), 자백을 한 죄수는 보상을 받고 풀려 난다(보수 3).

이 게임을 역설이라 부르는 이유는 다 같이 범죄를 부인하는 것이 두 죄수에게 모두 이익이 되지만 각자는 상대방이 어떤 행동을 취하든지 자백하는 것이 유리하게 된다는 점에 있다. 즉, 상대방이 범죄를 부인할 때 본인이 자백하면 3이라는 보수를 받게 되므로 자백하는 것이 이익이고, 또한 상대방이 자백을 할 때에도 본인이 자백하는 것이 유리하게 되므로 어떤 경우든지 자백하는 것이 개인적으로 유리한 선택이다. 그러나 각자가 그렇게 하면 두 죄수 모두에게 불리한 상황에 처하게 된다는 점에서 역설적이다.[3]

개인에게는 이익이 되지만 모두가 개인의 이익을 추구하면 사회적으로 바람직하지 못한 결과를 가져오는 모든 상황은 죄수의 역설과 같은 구조를 갖게 된다. 이러한 구조는 다양한 사회적, 경제적, 정치적 상황에 응용될 수 있다. 예를 들어 교통질서를 지키는 문제를 보자. 모두 질서를 지키면 모두 안 지킬 때에 비해 모두가 쾌적하게 지낼 수 있다. 그러나 다른 사람들이 모두 질서를 지킬 때 한 사람이 안 지키면 그 사람은 빠르게 갈 수 있어서 이익을 본다. 그러나 나머지 사람들은 질서를 안 지키는 사람 때문에 사고의 위험에 직면하게 된다. 질서를 지키는 것을 C, 안 지키는 것을 D로 하면 이 상황은 죄수의 역설과 같은 상황이 된다.

죄수의 역설 게임에서 과연 D를 선택하는 것이 합리적 행동인가? 앞서의 논리에 의하면 D는 상대방의 선택과 관계없이 최선의 선택이

3) 전략 D와 같이 상대방이 어떤 선택을 하든지 유리한 전략을 우월전략(dominant strategy)라 부른다. 모든 게임에서 우월전략이 존재하는 것은 아니다. 그러나 우월전략이 존재한다면 그것을 선택하는 것이 합리적 행동이라고 할 수 있다.

므로 그것을 선택하는 것이 합리적이다. 그러나 그 결과는 자기 자신에게 불리한 상황이 된다. 궁극적으로 자신의 이익을 해치는 결과는 가져오는 행동이 어떻게 합리적일 수 있는가?

또 다른 흥미로운 게임으로 성의 대결Battle of sexes 게임이 있다. 남편(경기자 1)과 부인(경기자 2)가 축구(F)와 음악회(C) 관람 중 하나를 선택하는 문제를 생각해 보자. 모두 함께 축구관람을 선택하면 남편은 매우 좋지만 부인은 그저 그렇게 느낀다. 모두 같이 음악회를 간다면 부인이 매우 좋고 남편은 그저 그렇다. 그러나 각각 혼자 축구나 음악회관람을 간다면 모두 최악이다. 이것을 행렬로 나타내면 다음과 같이 된다.

	F	C
F	2, 1	0, 0
C	0, 0	1, 2

이러한 게임에서 어떤 전략이 합리적인 선택인가? 게임적 상황에서는 본인의 합리적 선택은 상대방의 선택에 의존한다. 예를 들면 부인이 F를 선택한다면 남편도 F를 선택하는 것이 합리적이지만, 부인이 C를 선택한다면 남편은 C를 선택하는 것이 합리적인 행동일 것이다. 남편이 합리적 선택을 하기 위해서는 부인의 선택을 먼저 예측해야 한다. 그런데 부인의 입장에 서서 합리적인 행동을 선택하기 위해서는 다시 남편의 합리적인 행동을 예측해야 하는 순환논법에 빠지게 된다. 따라서 이러한 논리로는 합리적인 행동을 결정할 수없다.

Nash(1950)는 어떤 선택이 합리적인가하는 질문을 하는 대신에 어떤 상황이 지속될 수 있는 균형상황인가하는 질문에 대한 해답을 찾았다. 그 결과 Nash균형이라 불리는 매우 중요한 균형개념을 정립하게 되었다. Nash균형이란 각 경기자가 독자적으로 현재의 상태에서 이탈할 유인이 없는 상태를 말한다. 성의 대결 게임에서 Nash균형은 함께 축구를 관람하거나(F, F), 함께 음악회에 가거나(C, C) 하는 것

이다.[4] 부부가 모두 축구를 관람하는 상태(F, F)에서는 남편이나 부인이 혼자서 음악회에 가는 것으로 바꿀 때 아무 이익이 없으므로 바꿀 유인이 없다. 따라서 이 상태는 지속될 것이라는 점에서 균형이라 부를 수 있다. 그러나 (C, F)는 균형이 될 수 없다. 부인이 F를 택하고 있는 상태에서 남편이 C를 유지하는 것보다는 F로 전환하는 것이 이익이 될 것이므로 이 상태는 유지될 수 없기 때문이다.

Nash균형 개념을 일반적으로 정의하면 다음과 같다. 즉, 어떤 전략의 짝$(s^*_1, s^*_2, \ldots, s^*_n)$이 Nash균형이면 모든 경기자 i는 임의의 전략 $s_i \in S_i$에 대해,

$$u_i\,(s^*_1, \ldots, s^*_i, \ldots, s^*_n) \geqq u_i\,(s^*_1, \ldots, s_i, \ldots, s^*_n)$$

이 성립한다. Nash(1950)는 전략의 집합과 보수함수가 적당한 조건을 충족한다면 Nash균형이 적어도 하나 존재한다는 사실을 증명하였다. Nash균형 개념은 비협조게임에 대한 연구를 제로섬게임의 범위를 벗어나도록 하는데 결정적인 역할을 하였다.[5]

앞에서 언급한 바와 같이 Nash균형은 합리적 선택의 충분조건이 될 수는 없다. 성의 대결 게임에서와 같이 많은 게임에서 Nash균형은 다수 존재하는데, 이 경우 Nash균형 전략은 다수 존재한다. 따라서 상대방의 선택과 무관하게 어느 한 전략을 합리적인 선택으로 추천할 수 없다. Nash균형이 유일하게 존재하는 경우에도 Nash균형전

4) 이 두 균형 외에도 혼합전략 균형이 하나 존재한다. 혼합전략이란 두 전략을 일정한 확률로 무작위로 선택해서 행동하는 전략을 말한다. 혼합전략에 대비하여 어느 한 전략을 확률 1로 선택하는 것을 순수전략이라 한다.

5) 전략의 집합이 유한한 경우에는 Nash균형이 존재하지 않을 수 있다. 그러나 이때에는 일정한 확률로 전략을 무작위로 선택하는 혼합전략을 포함하는 균형은 반드시 하나 존재하게 된다. 동전맞추기 게임의 경우 순수전략은 존재하지 않고. H와 T를 각각 1/2의 확률로 혼합하는 혼합전략이 유일한 Nash균형전략이 된다.

략이 반드시 합리적인 전략이라고 말할 수없는 경우도 있다. 다음과 같은 게임을 보자.

	A	B
X	1, 0	-100, 0
Y	0, 1	0, 0

이 게임의 유일한 Nash균형은 (X, A)이다. 전략 X가 균형전략이기는 하지만 상대방이 B를 선택할 경우 매우 큰 손실을 입게 되는데 상대방이 A를 확실하게 선택한다는 보장이 없는 한 전략 Y를 선택하는 것이 안전하다고 판단할 수 있다. 따라서 이러한 경우에는 경기자들이 Nash균형전략을 선택할 이유가 없다. 이렇게 본다면 Nash균형전략은 합리적 선택의 필요조건도 충분조건도 아니다.

또 다른 예로서 겁장이 게임game of chicken을 들 수 있다. 두 사람이 일차선 밖에 없는 도로에서 서로 마주보며 차를 출발시킨다고 하자. 두 차가 전속력으로 끝까지 달리면(B, B) 서로 충돌하고 두 사람 모두 크게 다친다. 그러나 상대방이 먼저 피할 때(C) 끝까지 달리는 사람(B)이 이기게 된다. 모두 도중에 피하면(C, C) 비기게 된다. 이 게임의 보수행렬은 다음과 같다.

	B	C
B	-10, -10	2, -2
C	-2, 2	0, 0

이 게임의 Nash균형은 (B, C)와 (C, B) 이다.[6]

6) 성의 대결 게임과 같이 이 게임도 혼합전략균형을 갖는다.

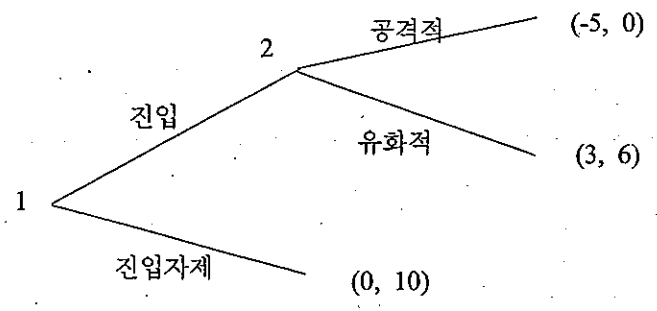
2. 전개형 게임과 지네게임

실제 많은 경우 각 경기자는 순차적으로 자신의 선택을 하게 되는데 이러한 상황을 묘사하기 위해 전개형 게임의[7] 개념이 유용하다.

전개형 게임의 예로서 다음과 같은 진입저지 게임을 생각해 보자. 경기자 2는 기존의 독점기업이고, 경기자 2는 이 독점산업에 진입할지를 고려하고 있는 기업이다. 경기자 2가 먼저 진입여부를 결정하면, 그 후에 경기자 1인 자신의 대응을 결정하는 순서로 의사결정이 이루어 질 것이다. 경기자 1이 택할 수 있는 선택은 진입(E)하든가 혹은 진입하지 않든가(N)의 두 가지이고, 기업 2가 진입하지 않기로 결정하면 게임은 끝나고 기업1은 0의 이익을 얻고 기존의 독점기업은 독점이윤 10을 얻게 된다. 만약 기업 2가 진입하기로 결정하면 기업 1의 대응방법은 두 가지가 있다. 하나는 가격인하를 단행하는 등 공격적으로 대응해서 기업1이 손해를 보고 퇴출하도록 하는 전략이고 또 하나는 유화적으로 행동해서 공존하는 전략이다. 공격적으로 행동할 때는 상대 기업만 손해를 보는 것이 아니라 자신도 평화적으로 행동할 때보다 이윤이 줄어들 것을 각오해야 한다. 공격적으로 행동하면 기업 1은 -5, 기업 2는 0의 이익을 얻고, 유화적으로 행동하면 기업 1은 3, 기업 2는 6의 이윤을 얻는다.

이 게임을 아래와 같은 형태로 묘사할 수 있다. 이것을 전개형 게임이라 한다.

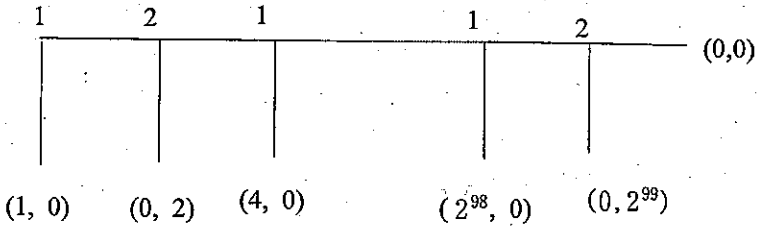
7) 전개형 게임은 동태적 게임(dynamic game), 순차 게임(sequential game) 등으로 부르기도 한다.



이 진입저지 게임은 두 개의 Nash균형을 갖는다. 하나는 경기자 1이 진입하고 2가 유화적으로 대응해서 각각 3, 6의 이익을 얻는 것이고, 또 하나의 균형은 경기자 1이 진입을 자제하고 기업 2는 공격적으로 행동하는 것으로 이 때 각 기업은 각각 0, 10의 이익을 얻는다. 그러나 후자의 균형은 불합리한 점이 있다. 이것이 균형인 이유는 기존의 독점 기업인 경기자 2가 기업 1에 대해 만약 진입하면 공격적으로 대응하겠다고 위협을 하고 기업 1은 이 위협에 굴복하여 진입을 하지 않기 때문이다. 그러나 이 위협은 신뢰할 수없는 공갈에 불과한 것이 게임의 구조상 분명하다. 기업 1이 위협에도 불구하고 진입을 한 경우 기존의 독점기업인 기업 2는 공언한 대로 공격적으로 행동하기 보다는 유화적으로 행동하는 것이 자신에게 유리하기 때문이다. 기업 1은 이러한 사실을 고려에 넣고 행동한다면 진입하는 것이 합리적인 행동이 된다. 전개형 게임에서는 Nash균형 중에서도 이와 같이 신뢰할 수없는 헛된 위협을 내포하지 않는 균형을 특별히 부분게임완전균형subgame perfect equilibrium이라 부른다.

부분게임완전균형은 또한 역진귀납backward induction의 추론을 통해서도 얻을 수 있다. 진입저지 게임의 경우를 예로 들면, 마지막 의사결정단계에서 기업 2는 공격적 혹은 유화적 행동 중에서 유화적으로 선택하는 것이 자신에게 유리하다. 이것을 고려에 넣고 기업 1은 진입하는 것으로 결정하게 될 것이다. 이렇게 추론하면 두 기업이 모두 합리적이라는 가정으로부터 기업 1은 진입하고 기업 2는 유화적으로 행동하는 부분게임완전균형을 얻게 된다. 이와 같이 마지막 의사결정단계에서부터 시작하여 하나씩 거꾸로 거슬러 올라가면서 합리적인 의사결정을 내리는 것을 역진귀납법이라 부른다.

역진귀납의 추론은 합리적인 의사결정을 위한 매우 설득력 있는 방법으로 보이지만 게임적인 상황에서는 역설적인 결과를 가져오는 다음과 같은 예를 생각해 보자. 이 게임은 그림의 모양이 지네와 같다고 해서 지네게임centipede game이라 부르기도 한다. 경기자는 1과 2가 교대로 의사결정을 내리는 게임으로 경기자 1이 먼저 앞에 놓인 돈 1원을 가져가거나 그대로 놓아두거나를 선택한다. 만약 가져가면 게임은 바로 끝난다. 그대로 두면 금액이 두 배로 증가하고 이번에는 경기자 2가 선택할 차례가 된다. 경기자 2가 가져가면 거기서 게임은 끝나고 만약 그대로 두면 금액은 다시 두 배가 되고 경기자 1에게 선택권이 주어진다. 이러한 과정이 100번 반복된다. 100번째에는 경기자 2가 가져가면 2^{99}원이라는 천문학적인 금액을 갖게 되는데 만약 그대로 두면 두 경기자 모두 아무것도 받지 못하고 끝나게 된다. 이것을 전개형 게임으로 표시하면 아래 그림과 같게 된다.



이 게임을 역진귀납으로 추론하면 마지막 단계에서 경기자 2는 당연히 가져가는 것이 이익이다. 그렇다면 이것을 예상하고 그 전 단계에서 경기자 1은 자신의 선택할 차례에서 2^{98}원을 가져가는 것이 합

리적인 행동이 된다. 이렇게 추론하면 매 단계에서 누구든지 자신의 차례가 왔을 때 돈을 가져가는 것이 합리적인 행동이라고 결론지울 수 있다. 그렇다면 이 게임의 결과는 첫 단계에서 경기자 1이 1원을 가져가고 끝나게 된다. 이것이 이 게임의 유일한 부분게임완전균형이다.

그런데 이 균형해는 직관적으로 타당하지 않은 것으로 보인다. 단계가 넘어 갈수록 금액은 두 배씩 증가하여 게임의 중간 단계쯤에는 2^{50}원이라는 천문학적 금액이 된다는 사실을 두 경기자가 잘 알고 있으므로 처음 몇 단계에서는 가져가지 않고 그대로 두고 있다가 금액이 충분히 커지 후부터 본격적인 게임이 시작될 것이라고 예상하는 것이 합리적일 것이다. 실제로 실험적인 사실로 이러한 결과가 확인되고 있다. 전형적인 실험결과는 위의 예상과 일치한다.[8]

이론의 예측과 실험적인 사실과의 괴리를 설명하는 다양한 제안들이 제시되어 왔다. 그 중 대표적인 예로서 Kreps et al.(1982)의 연구를 들 수 있다.[9] 이 논문은 경기자들 상대방의 합리성에 대한 약간의 의구심을 갖고 있다는 사실이 문제해결의 한 실마리가 될 수 있음을 보였다. 즉, 경기자들이 상대방이 비합리적인 행동을 할 가능성이 극히 작은 확률로 존재한다고 인식하고 있다면, 처음 단계에서 돈을 가져가지 않고 그대로 두더라도 상대방이 다음 단계에서 바로 가져가지 않을 가능성이 있다는 인식하에 모두 처음 몇 단계동안 그대로 돈을 놓아두게 되는 것이 균형이 될 수 있다는 것이다. 그런데 이 논문의 함의는 완전히 합리적인 경기자들은 이론의 예측하는 바대로 행동하는데 반해, 현실적인 행동은 비합리성 때문이라는 것이다.

8) 대표적인 실험연구로서 McKelvey and Palfrey(1992), Ho and Weigelt(2000) 등을 들 수있다. Camerer(2003)은 실험여구에 과한 다양한 결과를 비교 분석하고 있다.

9) Kreps et al(1982)가 대상으로 하는 게임은 지네게임이 아니라 죄수의 역설 게임의 유한반복게임이다. 그러나 결과는 그대로 지네게임에 적용될 수 있다.

IV. 게임적 상황하에서의 합리성

지금까지 게임적인 상황을 정규형 게임과 전개형 게임으로 나누어 각각의 경우 합리적인 행위자가 취하는 행동에 대해 살펴보았다. 대표적인 예로서 죄수의 역설 게임과 지네게임을 분석하였다. 두게임에서 볼 수 있는 공통적인 특징은 일견 별 문제가 없어 보이는 합리성의 개념이 게임적인 상황에서는 매우 역설적인 결과를 가져온다는 점이다. 죄수의 역설 게임에서는 우월전략의 선택이 결과적으로는 경기자의 이익에 반하는 결과를 초래한다는 점에서 역설적이다. 지네게임에서는 역진귀납의 추론이라는 방법이 의외로 불합리하게 보이는 결정으로 경기자들을 유도하게 된다.

이러한 역설에 대한 경제학자들의 일반적인 태도는 Shubik(1970)의 다음과 같은 언급에 잘 표현되어 있다. "죄수의 역설은 영원히 풀리지 않을 것이다. 어떤 의미에서는 이미 해결되었다고 볼 수있다. 왜냐하면 역설 자체가 존재하지 않기 때문이다." 그것이 역설적으로 보이는 이유는 단순히 상식이 틀렸기 때문이라는 것이다. Shubik은 진공 중에서 가벼운 깃털과 무거운 납공이 같은 속도로 떨어진다는 사실이 상식에 반하는 것으로 보이는 것처럼 게임적 상황에서의 역설도 마찬가지로 설명될 수 있다고 주장한다.[10]

그러나 이러한 판단에 대한 반대논거로서 실험적인 결과를 들 수 있다. 죄수의 역설과 지네 게임을 포함하는 다양한 게임적 상황에 관한 실험적 연구는 매우 광범위하게 진행되어 왔다.[11] 다양한 조건하에서의 시험의 결과는 일관되게 상식적인 결론을 뒷받침하고 있다. 예를 들어 Sally(1995)의 연구에서는 죄수의 역설 게임에서 약 반수의 경기자가 이론적 예측과는 달리 협조적인 행동, 즉 자백하지 않는

10) W. Poundstone(1992) p.277 참조.

11) W. Camerer(2003)는 최근까지의 연구성과를 잘 정리하고 있다.

선택을 하는 것으로 나타났다. 수많은 다양한 상황에서 같은 상식적인 결과가 도출된다면 그것을 체계적으로 설명하는 이론이 제시되지 않으면 안 된다. 단지 공기의 저항과 유사한 외생적인 요인만으로는 설명이 되지 않기 때문이다. 이런 점에서 게임적인 상황하에서는 합리성의 개념 자체를 재정립할 필요성이 제기된다고 볼 수있다.

논점을 분명히 하기 위해 먼저 죄수의 역설 게임에 대해 좀 더 살펴보기로 하자. 이 게임을 일회성 게임으로 끝나는 경우와 반복되는 상황을 구분하면 새로운 결과가 도출될 수 있다. 일회적으로 끝나는 상황에서는 자백하는 것(전략 D)이 합리적인 선택이지만 이러한 게임이 반복되는 상황이라면 전략 C도 다음과 같은 논리에 의해 합리화될 수 있다. 즉, 서로 C를 선택하기로 약속하고 만약 어느 한 경기자가 이 약속을 어기고 D를 택한 경우 다음부터는 영원히 모두 D를 택하기로 한다면, 아무도 D를 택하지 않게 된다. 즉, 한번 D를 택하면 단기적으로는 이익을 얻게 되지만 그 이후부터 장기적으로 손해를 보게 되기 때문에 모두 C를 계속하는 것이 유리하게 되는 것이다. 달리 말하면 반복되는 죄수의 역설 상황에서는 협조적인 행동이 균형이 된다. 개인의 이익만을 추구하는 전략 D가 장기적으로는 사회전체의 이익을 가져오는 협조전략 C보다 열등하게 되는 것이다.

사회진화과정에서 이기적인 개인들이 자신의 이익을 희생하면서 협력적인 행동을 하게 되는 현상을 설명하는데 이러한 죄수의 역설 게임을 인용하기도 한다. 또한 이러한 분석을 통해서 D를 이기적인 행동이라 하고 C를 윤리적 행동으로 간주할 때고 장기적인 관점에서 어떻게 윤리적인 행동이 정당화되는 지를 설명할 수 있다. 그래서 전략 D는 개인적 합리성을, 전략 C는 집단적 합리성을 대표하는 선택이 된다.

이 예에서 도출할 수 있는 하나의 함의는 윤리적인 행동은 반복적인 상황에서의 장기적인 이익을 극대화하는 행동원칙이라는 것이다. 그런데 윤리적인 판단을 내리는 모든 상황이 반복적인 것은 아니다. 어린 아이가 물에 빠져 허우적거릴 때 그 아이를 구해야 한다는 판단을 내리는 상황은 실제로 반복되는 것이 아니라 우리들의 마음속에 그 아이의 입장을 상상해 보고 또한 장래 어떤 경우에 내 자신 혹은 나의 아들이 그런 상황에 노일 수도 있다는 가능성을 배경으로 해서 윤리적인 판단을 내리는 것이다. 따라서 실제로는 일회적으로 일어나는 사건임이 분명하더라도 그와 같은 상황에 처한 타인을 역지사지할 수 있는 문화적 상상력과 세계관을 모두 공유하고 있다면 그러한 공동의 문화적 인식이 무언인가에 따라 합리적 행동의 내용이 달라지게 되는 것이다.

또한 그때 그때의 상황에 따라 이익을 추구하는 것보다는 일시적으로는 손해를 보더라도 정해진 원칙을 따르는 것이 더 합리적인 행동이 된다는 결론에 이를 수 있다. 합리성의 개념이 좁은 의미의 수단 합리성의 영역을 벗어나 목적 그 자체까지도 다시 고려하고, 극대화 행동을 넘어서 준칙을 따르는 행동까지 포함하게 되는 것은 우리가 처한 상황이 상호의존적인 게임적 상황이라는 인식으로부터 자연스럽게 추론되는 것이다.

반복되는 상황에서 한 가지 고려해야 할 점은 유한번 반복하는 상황이라면 그 반복회수가 아무리 많더라도 협력이 유발될 수 없다는 사실이다. 예를 들어 100번 반복된다고 하자. 그러면 마지막 100회째에서는 일회성 게임과 같다. 따라서 합리적인 행동은 D를 선택하는 것이다. 이것을 고려하면 99번째 게임에서도 보복의 위험이 없으므로 D를 택하는 것이 합리적이다. 이와 같이 역진귀납의 추론을 적용하면 언제나 비협조적인 행동인 D를 택하는 것이 합리적인 결정이 된다. 지네게임에서와 마찬가지로 역진귀납의 추론이 예상밖의 결과를 가져오게 되는 것이다.

지네게임의 경우 서로 상당한 정도 예를 들어 50회 정도까지는 욕심을 억제하고 그대로 놓아 둠으로써 모두가 가질 수 있는 몫이 충분히 커지도록 한다는 공감대가 형성되고 그것을 지키려는 상대방에 대한 신뢰가 공유되는 사회에서는 단기적인 이익에 눈이 어두워 작은 금액으로 끝나는 상황은 발생하지 않을 것이다. 여기서 상호신뢰

와 상황을 전체적으로 파악할 수 있는 안목과 눈앞의 이익에 끌리지 않는 자제력이 사회의 장기적이 이익을 증진시킬 수있는 중요한 덕목이 된다.

결론적으로 게임이론을 분석함으로써 단기적인 이익과 장기적인 이익, 이기적 행동과 도덕적 행동, 이익과 덕, 수단 합리성과 목적 합리성의 관계를 더 잘 조망할 수 있는 입지를 확보할 수있게 된다고 말할 수 있을 것이다.

참고문헌

K. Arrow et al. (1996), *The Rational Foundations of Economic Behavior*, IEA.

C. Camerer (2003), *Behavioral Game Theory*, Princeton University Press.

J. Elster (ed. 1986), *Rational Choice*, Basil Blackwell.

J. Harsanyi (1953), "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking", *Journal of Political Economy* 61, pp.434-435.

_____ (1955), "Cardinal Welfare, Individual Ethics, and Interpersonal Comparison of Utility", *Journal of Political Economy* 63, pp.309-321.

_____ (1977), "Advances in Understanding Rational Behavior" in Elster (1986).

T. Ho and K. Weigelt (2000), "Population Level Trust Building Among Strangers", University Pennsylvania Working Paper.

D. Kreps et al. (1982), "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma", *Journal of Economic Theory* 27, pp.245-52.

D. Kreps (1990), *Game Theory and Economic Modelling*, Clarendon Press.

D. McKelvey and T. Palfrey (1992), "An Experimental Study of the Centipede Game", *Econometrica* 60, pp.803-36.

J. Nash (1950), "Equilibrium Points in n-Person Games", *Proceedings of the National Academy of Science* 36, pp.48-9.

R. Nozick (1993), *The Nature of Rationality*, Princeton University

Press.

W. Poundstone (1992), *Prisoner's Dilemma*, Anchor Books.

D. Sally (1995), "Conversation and Cooperation in Social Dilemmas: A Meta-analysis of Experiments from 1958 to 1992", *Rationality and Society* 14, pp.79-109.

L. Savage (1954), *The Foundations of Statistics*, John Wiley & Sons.

A. Sen (1977), "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory", *Philosophy and Public Affairs* 6, pp.317-44.

M. Shubik (1970), "Game Theory, Behavior, and the Paradox of the Prisoner's Dilemma: Three Solutions", *Journal of Conflict Resolution* 14, pp.181-93.

J. von Neumann and O. Morgenstern (1944), *Theory of Games and Economic Behavior*, Princeton University Press.