協力理論과 經濟行爲
— 악셀로드의 協力의 進化를 中心으로 —[1]

鄭 基 俊

경제학에서는 경제적 合理性을 중시하고 競爭이 강조된다. 그리고 경쟁은 協力과 대립되는 개념으로까지 인식되어 경제학에서는 협력의 문제가 경쟁의 문제보다 덜 중시되는 경향이 있다. 이 글에서는 競爭理論이 利己性을 기초로 하여 전개되는 것과 마찬가지로, 協力理論도 이기성을 기초로 하여 전개될 수 있다는 악셀로드의 이론을 중심으로 하여, 경제학을 보다 協力中心的인 이론으로 전개할 수 있는 길을 모색한다.

1. 序 論

사회과학 중에서 經濟學이 독보적인 科學性을 가진 학문으로 발전할 수 있게 된 것은 經濟의 行爲主體를 抽象化하는 데 성공했기 때문이라고 할 수 있다. 즉, 경제주체로 하여금 순수한 경제논리 내지 경제적 합리성에 따라서 행동하게 함으로써, 경제행위의 순수이론체계를 구축할 수 있었던 것이다. 즉 경제주체를 이기성의 화신인 호모에코노미쿠스로 봄으로써 오직 利己心 즉 經濟的 合理性에 따라서 행동한다고 봄으로써 일관성 있는 행동의 분석이 가능하게 되는 것이다.

경제학에서는 이처럼 경제적 합리성을 중시하고 경제행위를 분석하기 때문에, 行爲를 설명하는 가장 중요한 개념은 "競爭"이 된다. 그리고 어떤 의미에서 경쟁을 "美化"하게 되고, 경쟁의 결과가 사회적 "공동선"을 가져온다는 아담 스미스의 명제가 중시된다.

이처럼 競爭이 중시되다 보니, 人間行爲 가운데서 또 하나의 중요한 측면인 "協力"의 문제가 아무래도 경제학에서는 덜 중시되는 경향이 있다. 그리고 협력은 경쟁과 대립되는 개념이라고까지 오해되기도 한다. 그러나 경제학에서도 협력의 문제가 경쟁의 문제 못지 않게 중요한 문제임은 말할 필요가 없겠다. 그리고 협력의 문제가 경쟁의 문제와 같은 전제 위에서 설명될 수 있다면 그것은 금상첨화일 것이다.

이 글은 경제문제로서의 이러한 協力의 問題를 다룸에 있어서 정치학자인 악셀로드가

(1) 이 연구는 재단법인 서울대학교 발전기금 연구비의 지원을 받아 이루어졌다.

제시한 "協力理論"〔Axelrod(1984, 1997)〕으로부터 힌트를 얻을 수 있다는 믿음에서, 우선 그 이론의 내용을 설명하고, 이 이론을 경제행위의 설명에 어떻게 적용할 수 있는가를 보기로 한다.

2. 악셀로드의 問題 提起

2.1. 協力의 問題

악셀로드가 제기하는 문제는, "어떤 조건하에서 利己主義者들로 이루어진 세계에서, 公權力의 개입 없이도 協力이라는 행동이 나타날 수 있을까?"이다. 정치학자인 그의 입장에서 보더라도, 이 문제는 장구한 세월에 걸쳐서 사람들을 괴롭혀 온 문제다. 그리고 그럴만한 이유가 있는 문제다. 사람은 천사가 아니다. 사람은 자기를 먼저 배려하고, 自己利益을 먼저 생각한다. 그러나 우리는 삶의 과정에서 協力行爲가 진실로 일어나고 있음을 보며, 우리 문명을 뒷받침하고 있는 것이 바로 그 협력이라는 것을 안다. 그러면 사람 개개인이 이기적으로 자기의 이익을 추구하는 과정에서 어떻게 "協力"이 생겨날 수 있을까?

사람들에게 이 問題를 제시하면 여러 가지 答이 제시되리라고 생각된다. 그리고 각자가 그 문제에 대한 답이라고 생각하고 있는 것이 무엇이냐에 따라서, 그 사람의 대인관계에서 상대방에 대한 태도가 달라질 것이라고 생각된다. 즉, 그 답은 상대방에 대해서 협력할 준비가 얼마나 되어 있는가에 대해서 결정적 영향을 미칠 것이다.

2.2. 否定的인 答의 例

이 문제에 대한 부정적인 답의 예로 토마스 홉스의 답을 보자. 홉스는 주장하기를, 政府가 존재하기 전에, 사회의 자연상태는 個人의 利己性이 지배하는 상태였다는 것이다. 즉 사람들은 매우 잔인한 조건하에서 서로 경쟁했기 때문에, 인간생활은 "외롭고, 가난하고, 사악하고, 단명했다"는 것이다〔Hobbes(1651)〕. 홉스의 견해에 따르면, 사람들 사이의 협력은 정부 없이는 생겨날 수 없으며, 따라서 强力한 政府는 필수적인 존재다. 홉스 이래로, 큰 정부냐 작은 정부냐에 대한 논의가 제기될 때마다, 논의의 초점이 되는 것은, 公權力이 존재하지 않아도, 질서 있는 協力行爲가 생겨날 것이라고 기대할 수 있느냐 없느냐의 문제다.

2.3. 協力의 必要條件

우리가 저녁식사에 초대한 친구가 답례로 초대하지 않는 경우에, 우리는 그 친구를 몇 번까지 더 초대할까? 일상생활에서 우리는 이런 문제를 自問해 볼 때가 있다. 어떤 조직

내에서 한 사람이 다른 사람에게 호의를 베푸는 것은 상대편으로부터의 반대호의를 기대하기 때문이라고 생각될 수 있으며, 신문기자가 특종을 하고 나서 취재원을 호의적으로 은폐하는 것도 차후의 취재를 고려해서이다. 複占産業에서 한 기업이 제품가격을 인상할 수 있는 것은, 상대 기업 역시 따라서 가격을 인상할 것이라고 기대하기 때문이다. 이 예들은 모두 相互主義가 문제가 되는 경우이다.

사람이 처하는 상황 중에는 상대방과 완전히 이해가 상반하는 상황도 있지만 그렇지 않은 상황도 많이 있다. 즉, 零合게임(zero-sum game)의 상황이 아닌 경우가 있다. 상호 유익한 행동의 기회가 있는 경우에, 그 이익을 실현하기 위한 행동 관행이 생겨날 수 있다. 이 관행 중에서 가장 중요한 것은 相互主義慣行이다. 그러면 이러한 상호주의에 기초를 둔 協力의 慣行은 어떤 조건하에서, 어떻게 생겨날 수 있을까?

3. 協力理論

3.1. 악셀로드의 協力理論의 構造

악셀로드는 협력에 관한 이론 즉, 協力理論을 전개함에 있어서 두 단계의 接近方法을 취한다. 첫째 단계는 個人의 行動을 분석하는 것이고, 둘째 단계는 시스템 全體의 行動을 분석하는 것이다. 다시 말하면, 개인의 행동의 동기에 대한 일정한 가정을 한 다음, 그로부터 개인의 행동을 분석하고, 다음에 시스템 전체의 행동결과를 도출하는 것이다.

여기서 협력의 이론을 전개한다는 것은, 협력이 발생하기 위한 조건들을 찾아낸다는 것이며, 이렇게 함으로써 협력의 증진을 위한 방안도 도출할 수 있게 한다는 것이다.

3.2. 私的利益追求假定의 含意

악셀로드는 利己心을 개인의 행동의 동기로 삼는다. 즉 협력이론 전개의 첫째 단계에서 私的利益 내지 自己利益을 추구하는 개인들의 행동을 연구한다. 그리고 이들의 상호협력을 강제하는 공권력같은 것은 존재하지 않는 것으로 가정한다. 사적이익의 추구의 가정이야말로 악셀로드의 協力理論이 經濟理論과 결합될 수 있는 중요한 연결 고리이다.

악셀로드는 私的利益의 追求를 가정하는 이유를 다음과 같이 설명하고 있다. 즉 이렇게 가정해야만, 협력의 기초가 상대방에 대한 배려 또는 집단 전체의 이익이 아니어서 우리가 설명에 곤혹을 느끼게 되는 경우를 고찰할 수 있기 때문이라는 것이다. 탁월한 合理化다. 이것은 도킨스가 동물행동을 분석함에 있어서 대전제를 遺傳子利己性에 두는 관점과도 유사하다[Dawkins(1989), 정기준(1996)].

그러나 여기서 강조해 두고 싶은 것은, 이 사적이익추구의 가정이 겉으로 보이는 것처

럼 제약적인 가정이 아니라는 사실이다. 가령 형제 간에 서로를 배려하고 있는 경우, 우리는 이것이 사적이익추구의 가정에 배치된다고 볼 필요가 없다는 것이다. 우리는 언니의 이익 속에 (수많은 다른 인자들과 함께) 동생의 이익에 대한 배려가 들어 있다고 생각할 수 있는 것이다. (이것을 경제학적으로 설명하면, 언니의 效用函數의 說明變數 속에 동생의 利益이 들어가는 경우에 해당한다.) 그러나 이것은, 언니와 동생 사이에 아무런 갈등도 없다는 뜻은 아니다. 마찬가지로 두 나라는 여러 가지로 상대방의 이익을 배려하여 행동한다는 의미에서 友邦國이라고 부를 수 있다. 그러나 우방국이라고 해서 그 두 나라가 언제나 相互利益을 위한 協力만 할 수는 없다. 우리는 우방국 간의 문제도 自己利益追求의 가정 속에서 다룰 수 있다. 그러나 만일 우리가 상대방에 대한 배려의 가정으로부터 출발한다면, 상대방과 언제 협력하고 언제 협력하지 않을 것이냐 하는 문제를 다루기가 매우 거북하게 된다.

3.3. 協力問題의 例

협력문제의 예로 貿易障壁問題를 보자. 두 나라가 무역을 하고 있는 경우, 경제이론에 의하면 自由貿易의 相互利益 때문에, 두 나라 모두 무역장벽을 없애는 것이 이익이 될 것이다. 그러나 둘 가운데 한 나라만이 일방적으로 자신의 장벽을 없앤다면, 그 나라는 장벽을 없애기 전보다 불리해질 것이다. 실제로 한 나라만이 장벽을 없애면, 상대방 나라는 자신의 장벽을 그대로 유지하는 것이 이익이 된다. 그러므로 여기서 우리가 직면하는 문제는, 각국이 자신의 무역장벽을 유지하고자 하는 인센티브를 가지게 되며, 이처럼 모두 무역장벽을 유지하는 결과는, 양국이 서로 협력하여 장벽을 철폐했더라면 얻을 수 있는 성과보다 못하다는 것이다.

이 예에서 우리가 직면하는 문제점은, 각국이 이기적으로 추구한 이익의 결과가 양국이 협력했더라면 얻을 수 있을 이익보다 악화된 결과라는 점이다. 이처럼 非協力的 自己利益追求의 결과가 協力의 結果보다 못하게 되는 딜레마 상황의 구체적인 예들은 우리 주변에서 얼마든지 볼 수 있을 것이다. 공통적으로 이와 같은 성질을 가지는 문제들에 대해서 우리가 좀 더 깊이 이해할 수 있으려면, 그리고 그 딜레마의 해결책을 찾을 수 있는 길을 모색하려면, 우리는 어떤 방법을 생각해 볼 수 있을 것인가? 그것은 바로 그 문제들의 공통점을 추출하여 "模型" 내지 모델을 만드는 것이다. 악셀로드는 이 경우에 적합한 모형으로 "共犯者딜레마 게임(Prisoner's Dilemma Game)"이라는 모형을 채택했다.

3.4. 協力理論을 위한 模型: 共犯者딜레마 게임

공범자딜레마 게임은 위와 같은 상황들에 공통적인 문제의 특징을 그대로 유지하면서, 그러나 매우 단순한 구조를 가지는 게임이다. 그러므로 이 게임은 위의 상황을 "代表"할

수 있다.

이 "共犯者딜레마 게임"이라는 模型에는 두 競技者(two players)가 있다. 그리고 각 경기자는 두 選擇肢(two choices)를 가지고 있다. 즉 協力(cooperate, C)이란 선택지와 非協力(defect, D)이란 선택지다. 각 경기자는 상대방이 어떤 선택을 할지를 모르는 상황에서 선택을 해야 한다. 상대방이 어떤 선택을 하든, 비협력을 선택할 때는 협력을 선택할 때보다 더 높은 이득을 얻는다. 딜레마는, 경기자 둘 다 비협력을 선택할 때에, 둘 다 협력을 선택했더라면 얻을 수 있었을 이득보다, 경기자 둘 다 적은 이득을 얻는다는 데 있다. 이 단순한 게임이, 우리의 協力理論의 유일한 기초가 된다. 즉 우리가 앞으로 행할 모든 분석의 유일한 대상이 되는 모형이다.

4. 共犯者딜레마 게임

4.1. 이 게임의 作動原理

이 게임의 作動原理는 다음과 같다. 選擇의 結果를 2×2行列로 표현할 때, 한 경기자(이를 "나"라고 하자)는 행에서 協力 C 또는 非協力 D을 선택한다. 다른 경기자(이를 "상대방"이라 하자)는 동시에, 열에서 協力 C 또는 非協力 D을 선택한다. 이 두 선택으로 나타나는 결과 전체는 다음과 같이, 4개의 가능한 結果(results)로 이루어지는 2×2행렬로 나타낼 수 있다. 즉,

$$\text{결과행렬}: \begin{bmatrix} (C,C) & (C,D) \\ (D,C) & (D,D) \end{bmatrix}$$

두 경기자가 모두 협력을 선택한다면, 둘 다 꽤 좋은 결과를 얻는다. 즉 둘 다 "相互協力의 褒賞값(reward)" R을 얻는다. 우리는 그 구체적인 값을 3점이라고 놓는다. 나는 협력을 선택하고, 상대방은 비협력을 선택한다면, 나는 "봉잡이의 값(sucker's payoff)" S를 얻는다. 나는 비협력을 선택하고, 상대방은 협력을 선택한다면, 나는 "非協力에의 誘惑값(temptation)" T를 얻는다. 우리는 이 값을 5점으로 놓는다. 경기자 둘 다 비협력을 선택한다면, 둘 다 "相互非協力의 懲罰값(punishment)" P를 얻는다. 우리는 이 값을 1점으로 놓는다.

이와 같이 결과에 따라서 내가 얻는 점수는 나의 "利得값(payoffs)"이다. 이는 결과의 "함수"이므로 다음과 같이 利得값 함수 V로 나타낼 수 있다.

$$V(C, C) = R = 3$$
$$V(C, D) = S = 0$$
$$V(D, C) = T = 5$$
$$V(D, D) = P = 1$$

상대방의 이득값도 나의 이득값과 "對稱"이라고 가정하면, 이 기호표현은 상대방의 이득값을 나타내는 것으로 해석해도 무방하다. 다만 이득값 함수의 앞의 선택을 상대방의 선택으로, 뒤의 선택을 나의 선택으로 보면 된다.

4.2. 내게 有利한 選擇

이런 게임상황에서 나는 어떤 선택을 하는 것이 유리한가? 결과는 상대방의 선택에도 의존하므로 열선택자인 상대방이 어떤 선택을 하는가에 따라 결과를 따져보자. 상대방이 協力 C를 선택한다면 나는 나의 선택으로 결과행렬의 첫째 열의 두 결과 중의 하나를 얻을 수 있다. 즉 나도 協力 C를 선택함으로써 結果 (C, C)를 얻을 수 있고, 이득값으로

$$V(C, C) = R = 3$$

즉 포상값 3점을 얻을 수 있다. 그러나 나는 非協力 D를 선택함으로써 結果(D, C)를 얻을 수 있고, 이득값으로

$$V(D, C) = T = 5$$

즉 유혹값 5점을 얻을 수도 있다. 따라서 相對方이 協力할 것이라고 생각될 때는, 나는 協力을 선택하는 것보다 非協力을 선택하는 것이 유리하다.

$$V(D, C) > V(C, C)$$

이기 때문이다. 상대방이 非協力 D를 선택할 것이라고 내가 생각하고 있는 경우에도 마찬가지다. 왜냐하면

$$V(D, D) > V(C, D)$$

이기 때문이다. 이 말의 의미는, 상대방이 협력할 것이라고 생각될 때도, 나는 비협력하는 것이 유리하고, 또, 상대방이 비협력할 것이라고 생각될 때도, 나는 역시 비협력을 선택하는 것이 유리하다는 것이 된다. 상대방이 어떻게 나오든지 관계없이, 나는 非協力하는 것이 有利한 것이다.

똑같은 논리가 相對方에게도 적용된다. 그러므로, 내가 어떻게 나올 것이라고 상대방이 생각하는지에 관계없이, 상대방도 역시 나에게 非協力할 수밖에 없게 된다.

이처럼 쌍방의 "合理的인 選擇"은 쌍방이 모두 非協力을 선택하는 것이다. 그러나 그렇게 되면 둘 다 相互非協力의 징벌값 1점밖에 얻지 못하게 되는데, 이 점수는 쌍방이 모두 협력을 선택하였더라면 얻을 수 있었을 相互協力의 포상값 3점보다 나쁜 점수인 것이다. 즉,

$$V(D, D) < V(C, C)$$

각자는 合理的으로 행동했지만, 그 결과는, 가능한 최선의 결과보다 둘 다에게 모두 불리하다. 이것이 바로 딜레마인 것이다.

4.3. 딜레마 게임이 되기 위한 利得값들 사이의 關係

共犯者딜레마 게임은, 단순히, 매우 흔하고 매우 흥미있는 상황들의 집합을 抽象的으로 模型化한 것이다. 이 상황이란, 위의 예에서처럼, 각자가 선택한 최선의 결과는 相互非協力으로 나타나지만, 모두가 相互協力했더라면 모두에게 더 좋은 결과가 나타났을, 그런 상황인 것이다.

이런 상황은 위의 세 부등식으로 대표되는 상황이다. 이 부등식을 이득값들의 기호로 나타내면 다음과 같다.

$$T > R, \quad P > S, \quad R > P$$

그리고 이를 하나로 통합하면 다음과 같다.

(4.1) $$T > R > P > S$$

즉, 경기자가 얻을 수 있는 최선의 이득은 T, 즉 상대방이 협력할 때 자기가 비협력하려는 유혹의 이득값이다. 최악의 이득은 S, 즉 상대방이 비협력할 때, 자기는 협력하여 봉

잡히는 이득값이다. 나머지 두 이득값의 크기 순서에 관해서는, 상호협력의 포상값인 R이 상호비협력의 징벌값인 P보다 큰 것이다. 그리하여 네 이득값들의 크기 순서는 큰 것부터 작은 것으로, T, R, P, S의 순이다. 이것은 어떤 게임이 딜레마 게임이 되기 위한 "必要條件"이라고 말할 수 있다.

우리가 共犯者딜레마 게임을 정의할 때는 여기에 한 가지 조건을 추가하여 정의한다. 그 追加條件이란, 경기자들이 착취와 피착취를 번갈아 함으로써 그 딜레마를 벗어날 수 없다는 의미를 가지는 조건이다. 즉, 搾取의 利得값 T와 被搾取의 利得값 S의 산술평균이 相互協力의 利得값 R보다 좋지는 못하다는 조건, 즉

$$2R > T + S \tag{4.2}$$

이다. 공범자딜레마 게임은 이 조건과, 앞의 이득값들 사이의 순서 조건에 의해서 정의되는 것이다.

4.4. 反復 게임과 協力可能性

이상에서 본 바와 같이 공범자딜레마 게임의 상황에서, 그 게임이 단발 게임인 경우에는 게임 당사자 간에 협력의 가능성이 없다. 그리고 게임이론의 일반 논리에 의하면, 단발 게임의 경우에 협력가능성이 없으면, 回數가 확정된 確定게임 역시 협력가능성이 없다. 협력가능성은 不確定 게임뿐이다. 횟수가 정해져 있지 않아서 게임 당사자가 그 게임이 언제 끝날지 모르는 共犯者딜레마 게임을 우리는 "反復 共犯者딜레마 게임(iterated prisoner's dilemma game)"이라고 부르기로 한다. 그렇다면 우리의 문제는, 반복 공범자딜레마 게임에서, 협력이 생겨날 必要充分條件은 과연 정확히 무엇이냐를 발견하는 문제로 된다.

4.5. 共犯者딜레마 問題의 性質

(1) 認識과 記憶: 우리의 공범자딜레마 게임에서는 경기자가 단 둘뿐이다. 한 경기자는 다수의 상대방과 경기할 수 있다. 그러나 상대방이 다수라 할지라도, 한번에 접촉하는 상대는 하나뿐이라고 가정한다. 그 경기자는 상대방이 누군지 인식하고, 또 둘 사이에 그때까지 어떤 관계에 있었는지를 기억한다. 이 인식과 기억의 그 능력은, 특정 상대와의 접촉의 "역사"를 감안하여 이후의 전략을 세울 수 있게 해준다.

(2) 問題의 根本形態: 공범자딜레마의 문제를 해결하기 위해서는 추가적인 행동을 허용함으로써 문제의 성격을 근본적으로 바꿀 수도 있다. 그러나 우리는 원래의 문제를 그대로 인정한 상태에서, 즉 문제를 있는 그대로, 어떤 변경도 시도함이 없이, 다룬다.

(3) 威脅 또는 約束의 不可能性: 경기자들은 강제력을 가지고 상대방을 위협하거나 상대방에게 약속을 할 어떠한 실용적 메카니즘도 존재하지 않는다. 경기자들은 어떤 특정한 전략을 쓰겠다고 상대방에게 약속할 아무런 방도도 없다. 때문에, 각 경기자는 상대방의 모든 가능한 전략에 대비해야 하며, 또 자기 자신도 얼마든지 다양한 전략을 쓸 수 있다.

(4) 追加情報의 不在: 상대방이 언제 어떤 수를 쓸 것인가를 알 수 있는 방법은 아무것도 없다. 경기자들이 이용할 수 있는 상대방에 대한 정보는, 그들 사이에 지금까지 있었던 역사뿐이다.

(5) 게임 中止의 不可能性: 각 경기자는 특정 상대방을 배제할 수 있는 아무런 방법도 없다. 즉, 경기를 자기 뜻대로 그만둘 수 있는 방법은 없다. 각 경기자는 매 번 협력 또는 비협력을 선택할 수 있을 뿐이다.

(6) 利得값의 確定性: 경기자는 이득값을 변경할 수 있는 아무런 방법도 없다. 주어진 이득값은, 각 경기자가 게임에 임하기 전에 이미 정해진 것이다.

(7) 意思疏通의 유일한 手段은 行動뿐: 게임에서 말은 의사소통의 수단이 아니다. 경기자들은 행동을 통해서 서로 의사소통할 수 있을 뿐이다. 그 밖의 의사소통 수단은 없다.

5. 協力과 未來와 未來利得값의 割引

5.1. 協力可能性의 源泉

반복 공범자딜레마 게임에서 상호협력이 생겨날 수 있는 가능성은 어디에 있는 것일까? 그것은 경기 당사자들이 다시 만날 수 있을지 모른다고 생각하는 바로 그 사실이다. 즉 이런 不確定 게임에서는 게임과정 내내 당사자들은 모두 어느 번째나 이번이 끝이라는 생각을 하지 않으며, 이 경우에 協力의 可能性이 있다. (끝이라고 생각할 때는 비협력만이 합리적인 선택임은 이미 설명한 바 있다.) 이 협력의 발생 가능성은, 지금 행하는 선택이 지금의 결과를 결정할 뿐 아니라, 경기자들의 미래의 선택에 영향을 미칠 수 있다는 것을 의미한다. 그러므로 거꾸로, 未來는 현재에 그림자를 드리울 수 있고, 또 그리하여 현재의 전략적 상황에 영향을 미칠 수 있는 것이다.

그러나 未來는 現在보다 그 중요성이 덜하다. 그 이유는 두 가지다. 첫째로, 경기자들은 利得값을 얻는 시점이 현재가 아니라 미래일 때, 그 가치를 보다 낮게 평가하는 경향이 있기 때문이다. 둘째로, 경기자들이 현재는 만나고 있지만 미래에는 다시 만나지 못하게 될 가능성이 어느 정도는 언제나 있기 때문이다. 즉, 경기자 중 하나가 이사가 버리거나, 직장을 바꾸거나, 죽거나, 파산하거나 하여, 게임이 끝나버릴지 모르기 때문이다.

5.2. 未來의 割引과 割引因子

이런 이유들 때문에, 다음 번의 이득값은 이번의 이득값에 비하여 언제나 낮게 평가된다. 즉 未來는 割引된다. 이 사실을 합리적으로 고려하는 하나의 자연스러운 방법은, 현재와 미래에 걸친 이득값들을 누적계산할 때, 다음번은 이번보다 얼마만큼 덜한 가치를 가지는 것으로 하여, 시간에 걸쳐서 미래의 이득값들을 할인하여 누적계산하는 것이다. 이것은 현재의 가중치를 1로 하고, 1期 후의 가중치는 1보다 작은 값으로 하면 된다. 그 가중치를 w로 나타내고 이를 割引因子라고 부르기로 한다. 경제학에서 시간에 따라 할인할 때 利子率 r을 이용하는 경우가 많은데 이 경우 할인인자는 $1/(1+r)$이고 이것이 바로 w에 해당한다.

5.3. 게임의 累積利得값 計算의 例

이 할인인자 w는 전체 이득값의 列의 누적계산에 다음과 같이 사용된다. 상호비협력의 이득값 V(D, D)를 앞에서의 수치 예대로 1점이라고 가정하고 이 이득값의 열이 현재에서 미래에 걸쳐서 이어진다고 하면 그 累積利得값은, 累積利得값 함수 AV를 쓰면, 다음과 같다.

$$\mathrm{AV(D,\,D)} \equiv 1 + w + w^2 + w^3 + \cdots = \frac{1}{(1-w)}$$

相互協力의 利得값 V(C, C)의 이득값인 R의 열이 현재에서 미래에 걸쳐서 이어진다고 하면 그 누적이득값은 다음과 같다.

$$\mathrm{AV(C,\,C)} \equiv R \times (1 + w + w^2 + w^3 + \cdots) = \frac{R}{(1-w)}$$

예컨대 $w = 0.9$라면 위의 두 누적이득값은

$$\mathrm{AV(D,\,D)} = \frac{1}{(1-0.9)} = 10$$

$$\mathrm{AV(C,\,C)} = 10R$$

이다.

이제는 게임의 累積利得값을 보기로 하자. 경기자 중 한 사람은 항상 비협력하는 전략, ALL D을 따르고, 다른 한 사람은 TFT 전략을 따른다고 가정하자. TFT 戰略이란, 첫 번에는 무조건 협력하고, 그 다음 번부터는 상대방의 직전 번의 선택을 따라하는 전략을 말한다. (이를 영어로 tit for tat 戰略이라 하는데, 이를 줄여서, TFT 戰略이라 한다.) 이 경기에서 선택의 열은 다음과 같이 진행된다.

ALL D: D D D D D D ···
TFT: C D D D D D ···

그러므로 ALL D를 따르는 경기자의 누적이득값은

$$AV(ALL\ D, TFT) = V(D, C) + wV(D, D) + w^2V(D, D) + \cdots$$
$$= T + wP + w^2P \cdots = T + \frac{wP}{(1-w)}$$

로 되며, TFT를 따르는 경기자의 누적이득값은

$$AV(TFT, ALL\ D) = V(C, D) + wV(D, D) + w^2V(D, D) + \cdots$$
$$= S + wP + w^2P \cdots = S + \frac{wP}{(1-w)}$$

로 된다.

6. 戰略

6.1. 意思決定規則으로서의 戰略

위의 ALL D와 TFT는 둘 다 전략이다. 그리고 일반적으로 전략이란 意思決定規則이다. 자기가 직면하게 될 어떤 상황에서, 경기자가 어떤 선택을 할 것인가를 정해놓은 규칙이다. 그 상황 자체는 그때까지 진행된 그 게임의 역사에 의존한다. 그러므로 전략의 내용은 어떤 역사적 상황에서는 협력을 선택하고, 다른 상황에서는 비협력을 선택한다는 등의 규칙으로 구성되어 있다. 더욱이, 어떤 전략은 確率槪念을 사용할 수도 있다. 그런

전략의 예에는 매번 협력과 비협력을 같은 확률로 무작위로 선택하는 完全無作爲戰略이 있을 수 있다. 어떤 전략은, 그때까지 진행된 역사를 매우 복잡하게 고려하여, 다음번의 수를 결정할 수도 있다. 또 전략은 여러 다른 전략들을 나름대로 결합한 것일 수도 있다.

6.2. 最善의 戰略

우리가 이 단계에서 묻고 싶어지는 질문은, 경기자에게 가장 이로운 "최선의 전략은 무엇인가?"라는 질문일 것이다. 다른 말로 하면, "어떤 전략이 경기자에게 가장 높은 점수를 올리게 해줄 것인가?" 라는 질문이다. 이것은 타당한 질문임에 틀림없다. 그러나 뒤에서 밝혀지게 되지만, 상대방이 사용하는 전략과 무관하게 어떤 경기자에게 가장 좋은 전략은 존재하지 않는다. 이 의미에서, 우리의 반복 공범자딜레마 게임은, 바둑이나 장기와 같은 게임과는 완전히 다르다. 바둑의 고수는, 상대방이 자기에게 가장 위협적으로 나올 것이라고 가정하는 것이 안전하다. 이 가정하에서 그 고수는 바둑게임을 구상해 나간다. 이 때 개념적으로는 그 고수의 최선의 전략이 있을 수 있다. 바둑은 경기자들의 이해가 완전히 상반되는 게임이기 때문이다. 그러나 공범자딜레마 게임으로 표현되는 상황은 이와는 전혀 다르다. 경기 당사자들의 이해가 완전히 상반되는 상황이 아니기 때문이다. 두 경기자 모두 相互協力의 褒賞값으로 R을 얻어 둘 다 좋아질 수도 있고, 相互非協力의 懲罰값 P를 얻어 둘 다 나빠질 수도 있다. 상대방이 언제나 나에게 가장 위협적으로 나올 것이라는 가정을 할 수 있는 상황이라면, 우리는 상대방이 나에게 결코 협력하지 않을 것이라고 생각해도 좋을 것이다. 그렇게 되면 나 스스로도 비협력을 선택하여, 결국 끊임없는 징벌값 P를 얻게 되고 말 것이다. 바둑 게임과는 달리, 반복 공범자딜레마 게임에서는, 상대방이 늘 나를 무너뜨리려 한다는 가정은 옳지 않다.

6.3. 協力可能性과 未來와 割引因子 w

내가 상대방을 다시는 만날 것 같지 않다거나, 내가 미래의 이득에 대해서 아무런 가치도 느끼지 못하고 있다면, 즉 w가 충분히 작다면, 나는 미래를 고려할 필요가 없게 되고, 나의 최선의 선택은 이번에 비협력하는 것이다. 따라서 w가 충분히 작다면, 나의 最善의 戰略은, 상대방의 전략과 관계없이, 항상 非協力을 선택하는 전략, 즉 ALL D이다.

그러면 割引因子가 충분히 큰 경우는 어떠한가? 악셀로드는 그 답을 그의 첫 명제로 다음과 같이 제시하고 있다.

命題 1: 우리의 할인인자 w가 충분히 크다면, 상대방이 사용하는 전략과 무관하게 가장 좋은 전략이란 존재하지 않는다.

이 命題는 그리 만족스러워 보이지 않는다. 즉 미래가 중요하다면, 단 하나의 가장 좋은 전략은 존재하지 않고, 상대방의 전략이 무엇이냐에 따라서 나의 전략이 달라진다는 말이 된다. 그러나 미래가 중요하지 않다면 그때의 최선의 전략이 ALL D임을 상기하면, 이 명제의 중요성을 인식할 수 있다. 즉 協力可能性의 必要條件을 이야기해 주는 것이 이 명제인 것이다. 이 명제의 전제가 충족되지 않으면 우리는 협력의 가능성조차도 이야기할 수 없다는 것을 이 명제는 말해 주고 있는 것이다.

이 명제의 타당성을 확인하는 일은 그리 어렵지 않다. 이 명제의 전제가 충족되는 상황에서, 즉 미래가 충분히 중요한 상황에서, 상대방이 항상 非協力하는 戰略, 즉 ALL D를 사용한다고 가정하자. 이 경우 나의 최선의 전략은 역시 ALL D이다. 내가 한번이라도 협력하게 되면 그 때에 나는 최악의 이득값 S를 얻게 되고 이 불이익은 결코 회복할 방법이 없기 때문이다. 이번에는 역시 이 명제의 전제가 충족되는 상황에서, 상대방이 "永久報復"의 戰略을 사용한다고 가정하자. 상대방의 永久報復戰略이란, 내가 비협력하기 전에는 상대방은 나에게 계속 협력하다가, 내가 한번 비협력하면 그 다음 번부터는 끝까지 나에게 비협력하는 전략이다. 이 경우에 나의 最適戰略은 절대로 비협력하지 않는 것이다. 미래가 충분히 중요한 경우에는 내가 먼저 비협력함으로써 얻게 되는 유혹의 이득값 T의 달콤한 맛은, 그 뒤에 징벌값 P밖에 얻지 못함으로 말미암은 장기적 불이익의 쓴 맛을 당할 수 없다. 이것은, 割引因子 w가 충분히 큰 경우에는 언제나 그러하다.

그러므로 w가 충분히 크면, 상대방의 전략과 무관한 나의 최선의 전략이란 존재하지 않는다. 즉, 최선의 전략의 내용 속에 협력의 요소가 들어올 가능성이 있다. 그러나 w가 충분히 크다는 것이 협력의 충분조건은 아니다.

6.4. 토나멘트를 통한 最適 戰略의 探索

그러면 反復 共犯者딜레마에 직면할 때 무엇이 좋은 戰略일까를 탐색하려면 어떤 방법이 있을까? 악셀로드는 이것이 수학의 최적해를 구하는 식의 방법으로는 불가능함을 인식하고, 전혀 다른 방법을 쓰고 있다. 컴퓨터를 이용한 토나멘트의 방법을 쓰는 것이다.

즉 이 연구에서 악셀로드는 게임이론 전문가들을 대상으로 그들이 생각하기에 가장 좋은 전략을 컴퓨터 프로그램 형식으로 제출해 줄 것을 요청하였다. 그리고 제출된 전략들을 서로 짝지워 컴퓨터 상에서 토나멘트 게임을 하게 하여 가장 좋은 전략을 찾아내고자 하였다. 그런데 놀랍게도, 이 토나멘트에서 승자는 제출된 모든 戰略들 가운데서 가장 간단한 전략인 TFT 戰略이었다. 그것은 앞에서 이미 설명한 대로 첫 번에는 協力하고, 그 다음 번부터는 상대방의 직전 번의 수를 따라하는 전략이다.

악셀로드는 첫 번째 토나멘트의 결과를 분석하여 몇 가지 잠정적 교훈을 얻었다. 즉

게임에서 좋은 성적을 올리는 전략들의 공통적인 성질은 무엇인가에 관심을 가진 것이다. 그 결과 그는 좋은 戰略의 性質로서 가장 중요한 것이 好意性(niceness)임을 알아냈다. 그 다음으로 報復性(provocability), 寬容性(forgiveness)도 중요한 성질임을 알아냈다. 好意性이란 자기가 먼저 비협력하지 않는 성질로 정의되며, 報復性이란 상대방의 비협력에 대해서 비협력으로 보복하는 성질, 寬容性이란, 상대방의 한 번의 비협력을 오래 기억하지 않고, 상대방이 비협력 뒤에 협력으로 용서를 빌어오면, 이를 받아들여 협력으로 용서해 주는 성질로 정의된다.

이러한 결과의 "發見"은 어디까지나 暫定的인 것이다. 그리하여 악셀로드는 첫 번째 토나멘트의 結果를 더욱 확실히 하기 위하여, 그 결과를 일반에게 공개하고, 두 번째 토나멘트를 준비하였다. 이번에는 첫 번째 토나멘트의 결과를 알고 있는 게임이론의 프로와 아마추어가 모두 포함된 훨씬 많은 수의 전략 엔트리들을 가지고 토나멘트가 수행되었다. 그런데 놀랍게도 이번 결과도 TFT의 승리였다. 그리고 전번 토나멘트에서 얻은 교훈들이 거의 다 재확인되었다.

6.5. 토나멘트의 結果에서 얻은 敎訓

이 두 토나멘트의 데이터를 분석한 결과, 악셀로드는 成功的인 戰略이 공통적으로 가지는 네 가지 성질을 찾아낼 수 있었다. 그 네 가지 성질이란:

(1) 상대방이 協力하는 한, 나도 協力함으로써 불필요한 분란을 회피할 것;
(2) 상대방의 이유없는 非協力에 대해서는 報復할 것;
(3) 한 번 報復을 한 다음에는 容恕할 것;
(4) 상대방이 나의 行動類型에 적응할 수 있도록 나의 行動을 透明하게 할 것.

이다.

두 번 모두 일등을 한 TFT 戰略은 분명히 이 네 性質을 모두 갖추고 있다. TFT는 協力으로부터 시작한다. 그리고 상대방이 協力하는 한 자신도 협력한다. 그리고 상대방이 非協力하면 자신도 따라서 비협력한다. 이는 報復 내지 膺懲의 효과가 있다. 그런데 TFT 전략은 상대방의 비협력을 오래 기억하지 않는다. 한번 보복하고 나서는 깨끗이 잊어버린다. 그리하여 비록 비협력을 한 前歷이 있는 상대방이라도, 상대방이 협력으로 돌아서면 즉시 자신도 협력함으로써, 상대방의 과거 비협력의 전력을 容恕한다.

이러한 성질들을 갖춘 나의 TFT 전략은 전혀 복잡한 전략이 아니다. 상대방은 매우 쉽게 나의 전략을 간파할 수 있다. 나는 상대방에게 전혀 숨김이 없는 것이다. 透明한

것이다.

이 두 토나멘트로부터 얻은 결과에서 우리가 얻을 수 있는 교훈은, 적절한 조건만 마련된다면, 利己主義者들의 세계에서 公權力의 개입 없이도, 協力이 진정으로 발생할 가능성이 있다는 것이다. 즉 두 번의 토나멘트에서 일등을 한 TFT 전략은 그것을 실증하는 한 예이다. 두 당사자가 모두 TFT 전략을 사용하는 경우, 그 게임의 진행은 相互協力의 연속으로 나타나게 되기 때문이다. 첫 번에 두 당사자는 협력을 선택한다. 두 번째에서 두 당사자는 상대방의 직전 선택인 협력을 선택한다. 그 다음 번에도 마찬가지다. 이리하여 결과적으로 協力의 連續이 되는 것이다.

이러한 협력의 연속은 TFT 전략만의 특징이 아니다. 악셀로드는 상대방보다 먼저 비협력을 선택하는 일이 결코 없는 전략인 好意性 戰略(nice strategy)의 경우 그 전략을 채택하는 경기자 사이의 게임에서는 언제나 협력의 연속이 나타남을 보이고 있다.

7. 協力의 進化의 理論과 實際

7.1. 充分히 큰 w

두 번에 걸친 토나멘트에서 얻은 결과는 중요하고 유용하지만, 그것은 어디까지나 特定狀況에서 얻은 결과일 뿐이다. 一般的 狀況에 관한 명제를 얻으려면 理論的 接近方法을 써야 한다. 그리하여 악셀로드는 이론적 접근을 시도한다. 이 과정에서 얻어지는 일련의 명제들은 협력의 발생을 위해서는 어떤 조건들이 요구되는가를 보여 줄 뿐만 아니라, 또한 協力의 進化의 시간적 과정을 보여 준다. 그 논증 내용을 요약하면 다음과 같다.

우선 TFT 전략이 집단의 대다수가 채택하고 있는 전략이라면, 그런 TFT 집단에서는 구성원들끼리 협력이 지속되는 바람직한 상황이 벌어질 것이다. 그런데 이런 상황이 과연 안정적으로 지속될 수 있을까? 다른 전략의 침입을 받지 않고 안정적으로 지속되는 성질을 악셀로드는 集團安定性(collective stability)라고 부르면서, 다음 命題를 數學的으로 입증하고 있다.

命題 2: 전략 TFT가 集團安定性을 가지기 위한 必要充分條件은 w가 충분히 큰 값을 가지는 것이다. 그리고 이때 그 臨界값은 이득값 T, R, P, S의 함수이다.

一般的 好意性戰略에 관해서는 다음 명제를 제시한다.

命題 3: 임의의 호의성전략이 集團安定性을 가지기 위한 必要條件은 w가 충분히 큰 값을 가지는 것이다.

이 두 命題는 협력의 진화를 위해서 요구되는 조건은 경기 당사자들이 다시 만날 확률이 충분히 커서, 그들의 利害得失의 몫이 미래 속에 있어야 한다는 것이다. 이 조건이 충족되면, 協力은 다음과 같이 세 단계를 거쳐서 進化할 수 있다.

7.2. 協力의 進化의 세 段階

(1) 第1段階: 이야기는 먼저, 무조건적 비협력의 ALL D의 세계에서조차도 협력이 생겨날 수 있다는 것으로부터 시작한다. 이 협력의 전개는, 서로 접촉할 수 있는 기회를 사실상 가지지 못하는 흩어져 있는 개인들에 의해서만 시도될 때는, 불가능하다. 그러나 협력은 개인들로 이루어진 小集團(cluster)으로부터 진화할 수 있다. 단, 그 소집단을 이루는 개인들은 그들의 협력을 相互主義에 기초를 두고 실행하며, 작지만 그들 상호간에 접촉할 기회를 가진다.

(2) 第2段階: 이야기의 중간은, 많은 상이한 전략들이 각축을 벌이고 있는 세계에서, 相互主義에 기초를 둔 전략이 세력을 확장할 수 있다는 것이다.

(3) 第3段階: 이야기의 마지막은, 협력체계가 상호주의의 기초 위에서 일단 확립되고 나면, 협력성이 떨어지는 전략들의 침입으로부터 자신을 방어할 수 있다는 것이다. 그리하여 協力進化의 과정은 逆轉防止裝置를 내장하고 있는 셈이다.

7.3. 實際의 例 1: "나도 살고 너도 살기" 體制

악셀로드는 구체적인 상황을 들어서, 위의 이론적 결과가 얼마나 넓은 응용범위를 가질 수 있는지를 보여 준다. 첫째 예인 "나도 살고 너도 살기" 體制(the "live and let live" system)는 제1차 세계대전의 진지전에서 생겨났던 적군 병사들 간의 "協力"의 예이다. 극심한 분쟁의 와중에서, 최일선의 병사들은 흔히 적을 사살하는 일을 자제하는 경우를 보는 것이다. 단 이 자제가 적방의 병사들로부터 상호주의적으로 받아들여질 때 그러하다. 이 相互主義的 自制를 가능하게 한 것은 참호전의 정돈상태적 성질, 즉, 동일한 小部隊가 장기간에 걸쳐서 서로 대치한 상태를 유지하고 있다는 성질에 기인하였다. 이 대치상태에 있는 소부대의 병사들은, 상호의 暗默的 協力을 달성하기 위하여, 실제로 上部의 명령을 어기고 있었던 것이다. 이 경우를 자세히 들여다 보면, 協力의 發生을 위한 條件들이 갖춰져 있었음을 알 수 있고, 이처럼 협력의 조건이 갖추어져 있을 때에는, 극히 협력이 있을 법하지 않은 상황 속에서도, 협력은 시작될 수 있고, 퍼질 수 있고, 또 안정적으로 유지될 수 있다는 것이다. 특히, 이 "나도 살고 너도 살기" 체제가 보여 주는 것은, 협

력의 전개를 위해서 우정 따위의 감정은 필요없이, 적절한 조건만 갖추어 지면, 서로 증오하는 적대자 사이에서조차도 相互主義에 기초를 둔 協力이 전개될 수 있다는 것이다.

7.4. 實際의 例 2: 生物界

악셀로드와 해밀턴은 生物界의 例를 통하여, 협력은 미래를 내다보는 능력, 즉 豫見力이 없어도 생겨날 수 있음을 보여 주고 있다. 즉 그들은 박테리아에서 고등동물에 이르는 생물계의 광범위한 영역에서 발견되는 行動類型들이 "協力理論"에 의하여 설명될 수 있음을 보여 주고 있다. 생물계에서 나타나는 협력현상을 보면, 협력은 그 당사자들이 친연관계가 없을 때는 물론이고, 자기들의 행동의 결과가 무엇인지를 인식하지 못할 때조차도 생겨날 수 있다. 이것을 가능하게 해주는 메카니즘은, 遺傳과 進化 메카니즘이다. 어떤 개체가 다른 개체로부터 유리한 반응을 획득할 수 있다면 그 개체는 후손들을 둘 확률이 높고, 또 그 후손들은 다른 개체들로부터 유리한 반응을 유도해냈던 선조들의 바로 그 행동유형을 계속 수행할 확률이 높다. 그리하여, 적절한 조건하에서, 相互主義에 기초를 둔 協力은 생물계에서 安定的 行動類型으로 자리매김할 수 있게 된다. 우리는 다윈이 강조한 개체의 이익을 가지고, 실제로, 동일 物種(species)에 속하는 또는 다른 물종에 속하는 개체들 간의 협력의 존재를 설명할 수 있다. 協力理論에서 우리가 얻은 적절한 조건만 갖추어진다면, 협력은 생겨날 수 있고, 번성할 수 있고, 안정적으로 유지될 수 있다는 결론이 다시 한 번 확인되는 것이다.

7.5. 악셀로드의 忠告事項

악셀로드는 토나멘트의 결과와 이론적으로 도출된 命題들에 기초하여, 각자 자신의 선택에 도움이 될 네 개의 忠告事項들을 제시하고 있다. 그 충고사항이란,

(1) 상대방의 성공에 대해서 猜忌心을 품지 말라;
(2) 내가 먼저 非協力하지 말라;
(3) 상대방의 협력과 비협력에 대해서 모두 相互主義로 대하라;
(4) 너무 약게 굴지 말라.

이다. (Do not be envious of the other player's success; do not be the first to defect; reciprocate both cooperation and defection; and do not be too clever.)

악셀로드는 또 社會改革者의 거시적 시각을 가지고, 사회 전반의 協力增進을 위한 改革의 방법도 이야기하고 있는데, 예를 들면, 당사자 간의 관계를 보다 장기적이고 자주 갖게 하는 일; 경기 당사자들로 하여금 상대방을 서로 배려하도록 가르치는 일; 그들에게

상호주의의 중요성을 가르치는 일 등이다. 이들 역시 協力理論에서 도출되는 그 이론의 含意이다.

8. 結 論

악셀로드의 協力理論은 여러 방면으로 확장될 수 있다. 즉, 公權力 없이 利己主義者들 간에 협력이 발생하는 문제의 논의로부터, 사람들이 실제로 서로를 '配慮하는' 경우에 어떤 일이 일어날까에 관한 분석과, 공권력의 개입이 '있는' 경우에 어떤 일이 일어날까에 관한 분석으로 진행될 수 있다. 그러나 기본적 接近方法은 동일하다. 즉, 個人들이 자기 자신의 이익 속에서 어떻게 행동하는가를 알면, 全體 集團에 어떤 일이 생기는지를 알 수 있다는 관점의 접근방법이다. 즉 전형적인 經濟學的 接近方法이다. 이 접근방법은 단일 경기자의 관점을 이해할 수 있게 해줄 뿐만 아니라, 주어진 상황 속에서 상호 협력의 안정성을 증진하려면 무엇을 요하는가를 인식할 수 있게 해준다. 통찰력과 이성을 갖춘 인류가 "協力理論"을 이해한다면, 우리가 바라는 協力의 進化는 그 속도를 더할 수 있고 경제행위에서 "協力"의 위치를 제고할 수 있을 것이다.

서울大學校 經濟學部 名譽敎授
137-846 서울특별시 서초구 방배2동 967-28
전화: 011-9775-6370
E-mail: kjjeong@snu.ac.kr

參 考 文 獻

정기준(1996): "경제인, 유전자 이기성, 그리고 이타적 행동의 기초로서의 근친도와 통합 근친도," 『경제논집』, **35**. **1**, 서울대학교 경제연구소.

Axelrod, R.(1984): *The Evolution of Cooperation*, London, England, Penguin Books.

_____(1997): *Complexity and Cooperation*, Princeton, Princeton University Press.

Dawkins, R.(1989): *The Selfish Gene*, New Edition, Oxford, U.K., Oxford University Press.

Hobbes, T.(1651): *Leviathan*, as London England, Penguin Classics(1985).