〈이타적 인간의 출현 – 최정규지음 〉 정리

– 조항현 h2jo23@gmail.com

0) 소개

– 외딴 마을에 팔이 굽혀지지 않아서 혼자서는 밥을 먹을 수 없는 사람들이 산다고 해보자. 서로 밥을 먹여주면 행복할 것이다. 하지만 밥을 먹여주는 행위는 상대방을 배부르게 하지만 자신에게는 비용이 따르는 일이다. 무엇보다 내가 누군가에게 밥을 먹여주었다고 그가 나에게 밥을 먹여줄 거라는 보장이 없다. 계산적으로 생각해보면, 나는 밥을 얻어먹기만 하고 남에게 밥을 먹여주지 않는 것이 가장 큰 이득이 된다. 하지만 모두가 나와 똑같이 생각한다면 그 섬의 누구도 남에게 밥을 먹여주지 않을 것이며, 결국 모두 굶어죽을 수도 있다. 개인의 입장에서는 남에게 먹여주지 않는 게 최선이지만, 마을 전체로 보면 모두가 서로 먹여주는 게 최선이다. 그래서 딜레마다.

– 죄수의 딜레마(Prisoner's Dilemma)는 두 경기자가 협조와 배반이라는 두 전략 중 하나를 선택할 때, 상대가 무엇을 선택하든 자신은 배반함으로써 더 높은 보수를 얻지만, 어떤 경우든 두 경기자가 모두 협조할 때보다 총 보수는 낮은 상황을 가리킨다. 여기서 '협조'는 상대방에게 이득을 주지만 자신은 비용이 드는 이타적 행위다. '배반'은 자신의 비용을 들이지 않고 상대방의 이타적 행위로 인한 이득을 얻을 수 있는 이기적 행위이며 '무임승차'라고도 한다.

		경기자 2	
		협조(C)	배반(D)
경기자 1	협조(C)	1, 1	−1, 2
	배반(D)	2, −1	0, 0

팔이 굽혀지지 않는 마을의 상황을 위와 같은 보수행렬(payoff matrix)로 나타낼 수 있다. 상대에게 밥을 먹여주는 행위는 자신에게 −1이지만 상대에게는 +2라고 하자. 두 경기자가 만나서 서로 밥을 먹여주면 각각 +1의 이득이 생긴다. 경기자 1이 경기자 2에게 밥을 먹여주기만 하고 끝난다면 경기자 1은 −1, 경기자 2는 +2를 얻을 것이다. 둘 다 밥을 먹여주지 않는다면 둘 다 0의 보수를 얻는다.
경기자 2가 협조를 한다고 할 때, 경기자 1은 배반하는 게 최선이다. 경기자 2가 배반한다고 할 때에도 경기자 1은 배반하는 게 최선이다. 경기자 2도 똑같이 생각할 것이므로 둘 다 배반을 선택한다. 둘의 보수의 합은 0이다. 하지만 둘 다 협조할 때 둘 모두의 보수는 +2이므로 '사회 전체적' 으로 둘 다 협조하는 게 이득이다.

위 보수행렬을 일반적으로 써보자.

		경기자 2	
		C	D
경기자 1	C	b−c, b−c	−c, b
	D	b, −c	0, 0

(죄수의 딜레마가 되기 위해 b 〉 c 라는 조건이 필요하다.)

이렇게 경기자 개인에겐 배반이 최선이지만 사회적으로 서로 협조하는 게 최선인 상황이라 '딜레마'다. 이런 경우가 실제도로 많이 있고, 책 전체에 걸쳐서 다양하고 재미있는 실제 예들이 많이 제시된다. 이런 딜레마 상황에서 모두가 자신의 이익만을 추구한다면 세상에는 이타적 행위가 사라지겠지만 우리는 여전히 이타적 행위(자)들을 많이 발견한다. 왜 그런지, 어떻게 그럴 수 있는지를 이 책을 통해 탐구할 수 있다.

1) 혈연선택 가설

경기자들은 유전자를 많이 공유할수록 서로 더 도우려 한다는 가설이다. 자신과 같은 유전자를 지닌 다른 개체를 도움으로써 자신의 유전자를 더 많이 번식시킬 수 있기 때문이다. 인간의 행위뿐 아니라 집단 생활을 하는 곤충이나 동물들에게서도 나타나는 협조 행위를 설명할 수 있는 가설이다.
만일 상대 경기자가 나와 r의 비율로 같은 유전자를 갖고 있다면 내가 상대를 도움으로써 나에게도 그만큼 이득이 된다. 나의 협조에 의해 상대의 유전자가 b만큼 이득을 얻었다면 그중 나와 같은 유전자에게는 rb만큼 이득이므로 모든 -c에서 rb만큼 보상이 이루어진다.

		경기자 2	
		C	D
경기자 1	C	b-c+rb, b-c+rb	-c+rb, b
	D	b, -c+rb	0, 0

여기서 r 〉 c/b이면 C가 D보다 우월한 전략이 된다. 이 가설이 맞는다고 해도 유전자를 나누지 않은 타인에 대한 이타적 행위를 설명해주지는 못한다.

2) 반복-상호성 가설

왜 사람들은 서로 협조하는가에 대해 '배반에 대한 보복' 을 생각할 수 있다. 1회성 관계에서는 배반해도 보복당할 위험이 없지만 관계가 지속되고 게임이 되풀이될수록 보복이 가능해지므로 장기적으로는 협조가 서로에게 이득인 경우가 많다. 다시 말해서, 게임이 반복되어야 배반에 대한 보복이 가능해지고 그럴 때에만 '서로 협조'하는 상황이 많아진다. '배반에 대한 보복'은 조건부 협조전략(눈에는 눈, 이에는 이)으로 나타낼 수 있는데, 무조건 협조전략은 무조건 배반전략에게 착취당하지만 조건부 협조전략은 무조건 배반전략을 응징함으로써 협조의 가능성을 높여준다.

이 가설을 다루기 위해 반복 게임을 도입한다. 게임의 횟수가 정해져 있는 경우를 보자. 경기자들은 마지막 회 게임에서는 모두 배반하는 것이 최선이라는 것을 안다. 더 이상 보복당할 가능성이 없기 때문에 마음 놓고 배반한다. 마지막 회에서 서로 배반할 것을 알기때문에 그 전 회에서도 서로 협조할 이유가 없어진다. 이런 식으로 역추론(backward induction)을 하다보면 첫 회에서도 서로 배반하는 게 최선이다. 이런 식으로는 이타적 행위가 왜 나타나는가를 설명할 수 없다. 그래서 역추론이 불가능하도록, 게임의 횟수가 정해져 있지 않고 확률적으로 계속 되는 경우를 생각한다.

여기서 경기자의 전략은 게임을 할 때마다 택하는 협조 또는 배반의 집합으로 표현된다. 예를 들어, 1회 게임에서는 C, 2회 게임에서는 D, 3회 게임에서는 D, 4회부터 게임이 끝날 때까지는 C라는 전략이라면, {C, D, D, C, C, C, ...}로 나타낸다. 하지만 사람들이 게임을 할 때 이런 자세한 전략표를 미리 준비할 것 같지는 않다. 그보다는 나름대로 전략을 선택하는 규칙에 따라 게임을 할 가능성이 높다.

가장 단순하게 만들 수 있는 전략은 '무조건 협조전략(all C)' 또는 '무조건 배반전략(all D)'이다. 게임의 횟수나 상대의 전략이나 자신이 그때까지 얻은 보수와 상관없이 무조건 협조를 하거나 무조건 배반을 하는 전략이다. 무조건 협조전략과 무조건 배반전략이 반복 게임을 한다면 후자가 늘 전자를 착취할 것이다.

이제 '배반에 대한 보복' 을 하는 '눈에는 눈, 이에는 이(tit for tat; TFT)' 전략을 생각한다. TFT 전략은 첫 게임에서는 C를 하고 다음 회부터는 바로 전 회에서 상대방이 낸 전략을 그대로 따라하는 것이다. 이전 회에서 상대가 C였다면 이번 회에서 나는 C를 하고, 이전 회에서 상대가 D였다면 이번 회에서 나는 D를 한다. 즉 상대가 협조하면 다음 회에서 나도 협조로 화답하지만, 상대가 배반하면 다음 회에서 나도 배반함으로써 상대를 보복한다.

한 번 게임하고 다음 번 게임을 할 확률 δ 를 도입한다. 즉 게임의 횟수가 1로 끝날 때도 있고 더 길어질 때도 있다.

TFT와 all D가 붙는 경우를 보자. 첫 회에서 TFT는 C, all D는 D를 하므로 TFT의 보수는 -c, all D는 +b이다. 다음 회부터는 서로 D만 하므로 보수는 계속 0이어서 첫 회의 결과가 최종 결과가 된다. TFT와 TFT가 붙는 경우, 서로 C만 계속 하므로 이때 둘 다 각 회마다 b-c의 보수를 얻는데, 게임이 반복될 확률에 의존하여 다음과 같은 보수를 얻는다.

$$(b-c)+(b-c)\delta+(b-c)\delta^2+\cdots=\frac{b-c}{1-\delta}$$

		경기자 2	
		TFT	all D
경기자 1	TFT	$\frac{b-c}{1-\delta}$, $\frac{b-c}{1-\delta}$	-c, b
	all D	b, -c	0, 0

δ 〉 c/b이면 둘 다 TFT를 택하는 경우와 둘 다 all D를 택하는 경우가 모두 균형이 된다. 여기서 '균형' 이란 일단 두 경기자가 어떤 전략을 택했으면 다른 전략으로 바꿀 유인이 없다는 것을 뜻한다. 반대로 δ 가 c/b보다 작으면 다시 죄수의 딜레마가 되어 무조건 배반전략이 최선이다. 즉 게임이 반복될 확률이 어느 정도 크면 TFT가 all D에게 보복함으로써 TFT가 우세해지는 경우가 나타나고 이로 인해 all D가 도태되어 협조 전략이 살아남을 수 있는 환경을 만들어준다.

혈연선택 가설과 반복-상호성 가설 모두 원래 죄수의 딜레마였던 상황에 새로운 요소를 도입하여 더 이상 죄수의 딜레마가 아니게 함으로써 협조 전략이 살아남을 수 있다는 것을 보여준다. 또한 혈연선택의 경우 '협조'는 더 이상 이타적 행위가 아니다. 그 행위로 인한 비용이 발생하지 않았기 때문이다. 반복-상호성의 경우에도 둘 다 TFT를 선택하는 유인이 더 높은 보수라는 면에서 TFT를 선택한 경기자들도 이해타산의 동기에 의해 움직인다고 할 수 있다. TFT가 all D를 응징하는 건 '정의'를 위해서가 아니라 '더 높은 보수'를 주기 때문이라는 설명이다.

이게 가장 널리 받아들여지는 가설이지만, 이 가설로도 설명되지 않는 현상이 나타난다고 한다. 반복되지 않을 상황(즉 다시 만나지 않을 상황)에서도 사람들은 서로 협조하거나 자신의 비용을 들여서라도 배반자에게 보복하기 때문이다. 이런 사람들을 상호적 인간(Homo reciprocan)이라 부른다.

배반에 대해 보복함으로써 얻는 이득이 보복을 한 사람에게(도) 돌아가는 경우와 그렇지 않은 경우를 나눌 수 있다. 전자의 경우 '이기적 보복', 후자는 '이타적 보복'이라 할 수 있다. 보복이라는 행위(또는 전략)를 선택함으로써 더 높은 보수를 얻었다면 그 행위는 이기적인 동기에 의한 것이라고 해석하자는 것이다. 문제는 '이타적 보복'을 어떻게 이해할지다. 그래서 이득을 극대화하려는 보수대응적 인간(반복-상호성 가설의 행위자)과 상대방의 행동에 대해 반응하는, 예를 들어 상대방이 규범을 따르면 보상하고 규범을 어기면 보복하는, 행위대응적 인간(상호적 인간)을 구분한다.

3) 유유상종 가설

말 그대로 이타적인 사람들은 이타적인 사람들끼리, 이기적인 사람들은 이기적인 사람들끼리 모여 산다는 가설이다. 이타적인 사람이 이기적인 사람한테 착취당할 기회를 줄임으로써 이타적인 사람들이 살아남을 수 있었다는 말이다.
이타적인 사람끼리 만나서 게임을 할 확률과 이기적인 사람끼리 만나서 게임을 할 확률을 s라고 하고 전체 인구 중 이타적인 사람의 비율을 p라고 하자. 이타적인 사람의 평균보수를 π^C라고 하고 이기적인 사람의 평균보수를 π^D 라고 쓰자.

$$\pi^C = [s+(1-s)p] \cdot (b-c) + (1-s)(1-p) \cdot (-c)$$
$$\pi^D = [s+(1-s)(1-p)] \cdot 0 + (1-s)p \cdot b$$

이타적인 사람의 평균보수가 이기적인 사람의 평균보수보다 크려면 p에 상관없이 s 〉 c/b이어야 한다. 즉 끼리끼리 모이는 경향이 어느 정도 이상 된다면 이타적인 사람끼리 게임할 때 더 큰 이득을 얻을 수 있다. 더 큰 이득을 얻는다면 그만큼 살아남을 가능성이 높아진다.
하지만 너무 똑같은 사람들하고만 살면 다양성으로부터 얻는 이득을 포기해야 하므로 끼리끼리에도 비용이 따른다고 볼 수 있다. 하지만 이타적이되 다양한 사람들이 서로 돕는 경우도 있으므로 '끼리끼리에 따른 비용' 에 대한 논의가 더 정교해질 필요가 있다.

4) 값비싼 신호 보내기 가설

자신이 능력이 있다는 걸 믿게 하려면 능력이 없는 사람이 절대로 할 수 없는 일을 보여주면 된다. 즉 그만큼의 비용이 들고, 그럴 때에만 사람들이 믿어주며 그렇게 함으로써 배우자를 선택할 때 유리해진다. 그럼 그 능력을 어떻게 보여줄까. 용맹하게 사냥을 하여 사람들에게 베풀면 된다. 사실 '능력 과시'는 협조가 아니라 배반을 통해서도 할 수 있으므로 이 가설이 '이타적 행위'에만 적용된다고 볼 수 있을지 의문이다. 다만 협조할 때 사람들이 더 그 신호를 믿어줄 것 같다. (이것도 또다른 게임인가?)

5) 의사소통 가설

게임 참가자들이 의사소통을 함으로써 서로 더 잘 도와주게 된다는 실험 결과가 있다. 협조가 사회적으로 바람직하다는 걸 이해하게 되었다든지(이 설명에 반하는 실험 결과가 있다), 협조에 대한 의무감이 생긴다든지, 참가자 사이의 신뢰가 쌓인다든지, 집단의식이 생긴다든지, 배반에 대한 죄의식이 생긴다든지 등의 설명이 있다.

6) 집단선택 가설

이타적인 사람들이 많은 집단과 이기적인 사람들이 많은 집단 사이에서도 선택(즉 집단선택)이 일어난다면, 전자의 총보수가 후자의 총보수보다 높으므로 전자가 선택되고 후자는 도태된다. 개인선택은 이기적 행위를 선호하지만 집단선택은 이타적 행위를 선호한다고 볼 수 있다. 예를 들어 용맹하고 이타적인 전사가 많은 집단이 전쟁에서 이길 가능성이 높아진다.
그런데 집단선택의 속도는 개인선택의 속도보다 느리다고 많은 학자들이 주장한다고 한다. 하지만 인간 사회에는 개인선택의 속도를 낮추고 집단선택의 효과를 크게 해주는 '제도'가 존재하여 집단선택 가설을 뒷받침할 수 있다고 글쓴이는 강조한다.
이와 관련하여 글쓴이는 자신의 모형연구를 소개한다. 각 개인은 같은 집단의 구성원에게 이타적(A)이거나 이기적(N)이며, 또한 다른 집단의 구성원(외부인)에 대해 적대적(P)이거나 관용적(T)이기도 하다. 그래서 모두 네 가지 조합(AP, AT, NP, NT)이 가능하다. 죄수의 딜레마 게임에서 봤듯이 N이 A보다 우세하고, T가 외부와의 교류로부터 더 많은 보수를 얻는다고 하면 P보다 우세하다. 그런데 AP가 많은 집단일수록 집단 사이의 전쟁에서 이길 가능성이 높아지며, 전쟁에서 진 집단의 AP는 모두 사라

지고 그만큼 이긴 집단의 AP 비율로 보충된다고 하자. 시뮹내기 결과 전체적으로 AP 비중이 높은 상태와 NT 비중이 높은 상태가 번갈아가며 나타나는데, NT가 많은 건 N과 T가 우세하므로 자연스러운데, AP가 많은 상태는 "이타성과 외부인에 대한 적대가 공진화"한 결과라고 한다.

7) 공간구조 효과

사람들은 다른 사람의 보수와 자신의 보수를 비교하여 보수가 높은 쪽의 전략을 따라간다(전수받는다고 한다). 그런데 '다른 사람'을 사회 전체에서 랜덤하게 고를 수도 있지만 자신의 이웃들 중에서 고를 수도 있다. 이렇게 '이웃'의 효과를 고려하는 걸 일반적으로 공간구조 효과라 부를 수 있다. 특히 2차원 격자 위에 각 행위자들이 있다고 하고 동서남북의 이웃들하고만 게임하고 또 보수를 비교하여 전수받는다고 하면 처음에는 이타적인 전략이 급격히 줄어들지만 우연히 이타적인 전략들이 이웃한 경우 이들은 세를 확장하여 이기적인 전략보다 더 우세해지기도 한다. 사실 공간구조 효과보다는 그냥 '국소성 효과'라고 하는 게 더 정확해 보이며(책에서도 국소성으로 설명한다), 이 효과는 끼리끼리 효과이기도 하고 집단선택 효과이기도 하다.

8) 내멋대로 정리

죄수의 딜레마를 다시 간단히 써보면, "C 〈 D & C+C 〉 D+D"이다. 개인에게는 협조(C)보다 배반(D)이 우월하지만 사회적으로는 둘 다 협조(C+C)하는 게 둘 다 배반(D+D)하는 것보다 더 낫다. (사실 "C+C 〉 C+D"도 포함되어야 한다.) 이 딜레마를 해결하는 방법이 여러 가지 소개되었다. 혈연이나 반복 게임을 통해서 C 〉 D로 만드는 방법도 있고, 국소성으로 인한 집단선택을 통해 C+C인 가능성을 높이는 방법도 있다.
상호적 인간에 대한 여러 가설/이론(유유상종, 집단선택 등)은 "이타적으로 보이는 것에 그치지 않는 진정한 이타적 행위"(306쪽)가 어떻게 유지되거나 진화했는지에 초점을 둔다고 한다. 그런데 유유상종이나 집단선택이 '진정한 이타적 행위'를 설명하는 것인지 의문이 든다. 일단 이후의 논의들은 " 더 높은 보수를 받는 사람의 전략을 더 낮은 보수를 받는 사람들이 전수받는다"는 진화 메커니즘에 바탕을 두고 있다. 그런데 바로 이런 진화 메커니즘은 결국 ' 더 높은 보수'를 지향하지 않으면 작동할 수 없다. 반복-상호성 가설에서 명시적이었던 '이해타산'의 요소가 유유상종, 집단선택 가설에서는 진화 메커니즘으로 (암시적으로) 바뀌었을 뿐 본질은 달라지지 않은 것으로 보인다.
책의 맥락을 오해했을 가능성이 있으므로, 몇 가지 단서를 제시하고 넘어가자. "이제부터 우리가 살펴볼 이야기에서는 게임은 반복되지 않는다고 가정할 것이다. 게임이 일회적임에도 불구하고, 협조적 전략이 혹은 상호적 인간형이 어떻게 진화과정에서 살아남을 수 있었을까?"(163쪽) 바로 다음 쪽부터 유유상종 가설에 대한 이야기가 시작된다. 유유상종 가설의 메커니즘에 대한 설명이 있는 175쪽에는 " 이타적인 사람이 얻게 되는 평균보수가 이기적인 사람의 평균보수보다 크다면 사회에 이타적인 사람들이 늘어날 것이라고 예상할 수 있다."라는 문장이 있다. 그리고 이 조건으로부터 "유유상종의 확률이 1/2을 넘어서면, 이타적 전략이 사회에 퍼져나갈 수 있다."고 한다.

마지막으로 살펴볼 내용은 맺음말에 나온다. 이타적 행위로 인한 심리적 만족감이 물질적 이득을 통한 만족감과 함께 행동에 영향을 미친다고 한다. 이런 접근은 위의 질문과 무관하게 논의될 수 있다. 여튼 만족감 U를 물질에 의한 만족감 M과 타인의 행복으로부터 얻는 만족감 V에 어떤 수 a를 곱해 나타낸다.

$$U = M + aV$$

a가 0보다 크면 타인의 행복에 만족하는 사람이고, a가 0보다 작으면 타인의 불행에 만족하는 사람이다. a가 0이면 타인의 행복은 전혀 상관하지 않고 물질적 요인에만 영향 받는 사람이다. 이 a를 어떻게 규명하고 이해할 거냐가 앞으로 더 탐구되어야 할 주제라고 한다. 우선 M으로 환원되지 않는 aV의 존재를 확인한 후, a가 0이 아닌 이유와 그것의 진화적 기원을 밝히고, 개인 사이의 상호작용, 문화/제도/역사적 요인들이 a에 어떻게 영향을 미치는지를 규명할 필요가 있다고 한다.