논리연구 9-1(2006) pp.63~95

뉴컴의 역설과 합리적 선택*

이 종 권

노직은 뉴컴의 문제를 뉴컴이 묘사한 상황에서 합리적인 행위를 선택함에 있어 지배의 원리와 기대 효용 극대화의 원리의 적용의 결과가 충돌한다는 것으로 설명하고 있다. 슬레징거와 이병욱은 지배의 원리에 의한 선택이 합당함을 논증하고 있다. 그들의 논증은 원래의 뉴컴 상황과 변형된 뉴컴 상황의 유비에 근거하고 있다. 이 글에서는 그러한 유비가 성립하지 않거나 성립할 경우 뉴컴 문제를 해결하는 데 흡족하지 않음을 보이려 하고 있다. 그리고 뉴컴 상황에서 지배의 원리를 사용해야 하는 독자적인 이유를 제시하고 있다.

【주요어】 뉴컴의 역설, 결정 이론, 합리성

* 이 논문은 2003년도 중앙대학교 학술연구비 지원에 의한 것임.

1

인간에게 있어 합리성은 여러 경우에 요구되지만 그 가운데 대표적인 경우가 주어진 상황에서 어떤 행위자에게 열려진, 즉 그가 할 수 있는 여러 행동 내지는 대안 가운데 하나를 선택해야 하는 상황이다. 그러면 어떤 대안을 선택하는 것이 합리적이라고 생각되는가? 칸트에 있어 합리성의 기준을 제공하는 것은 정언 명령(categorical imperative)이다. 즉 정언 명령에 위반하는 것은 비합리적인 행동이며 그러한 한에서 비윤리적이다. 우리가 합리적으로 행동할 수 있는 것은 정언 명령을 확보함으로서만 가능한 데 그것을 확보할 수 있는 것이 바로 우리의 이성(reason)이다. 이에 반해 경험주의자들에 있어 합리성은 인간이 욕망을 지니고 있다는 사실에서 비롯된다. 인간은 욕망의 대상을 소비함으로써 효용(utility)을 얻는다. 경험주의자들에 의하면 어떤 행위자에게 열려진 대안들 가운데 그러한 대안을 선택한 결과로 최대의 효용을 얻을 수 있는 행동이 합리적이며 또한 윤리적이다. 그런데 어떤 행위가 어떤 결과를 가져올 것인가 하는 것은 상황에 의해 결정된다. 반면 그 결과가 행위자에게 얼마만큼의 효용을 갖는가 하는 것은 행위자 자신에 의해 결정될 문제이며 이러한 의미에서 효용은 행위자에 주관적인 것이다. 이러한 상황에서 합리적인 선택을 하기 위해서는 각 행위를 함으로써 얻어지는 효용을 정확하게 계산할 필요가 있으며 경험주의자들에 있어 이성은 바로 이러한 목적에 소용된다. 따라서 이성은 '정념'(passion)의 노예가 되어야 하며 그러한 한에서 도구적이다.

한 예로 여기에 A와 B 두 개의 상자가 있다고 하자. A에는 100만원이 들어 있고 B에는 아무 것도 들어 있지 않으며 이것을

행위자도 알고 있다. 그 행위자에게 열려진 선택은 A 상자를 열거나 B 상자를 여는 것이다. 그러한 행위의 결과로 그 행위자는 각 상자에 들어 있는 돈을 가질 수가 있다. A 상자를 여는 행위를 a_1이라고 하고 B 상자를 여는 행위를 a_2라고 하자. 그리고 행위자가 x원에 부여하는 효용을 $u(x)$라고 하자. 이 경우 이 행위자가 a_1을 선택함으로써 얻는 효용은 u(100만)이고 a_2를 선택함으로써 얻는 효용은 $u(0)$이다. 만일 u(100만)$>u(0)$이라고 한다면 a_1의 행위를 선택하는 것이, 다시 말해 상자 A를 열어 그 안에 있는 돈을 갖는 것이 합리적이다.

일반적으로 어떤 행위자에게 열려 있는 대안적인 행위가 a_1, …, a_n이라고 하고 그 행위를 함으로써 얻어지는 결과(outcome)를 각각 o_1, …, o_n이라고 하자. 이 경우 $u(o_1)$, …, $u(o_n)$ 가운데 최대치를 $u(o_i)$라고 하면, 행위자에게 열려 있는 n 가지 행위 가운데 a_i를 선택하는 것이 합리적이다. 여기서 행위자가 a_1, …, a_n을 선택했을 경우 각각 실제로 o_1, …, o_n과 같은 결과가 야기되어야 하는가 혹은 실제는 어떻든 간에 행위자가 그러한 결과가 야기된다고 믿는 것으로 충분한가 하는 것을 문제 삼을 수 있다. 예를 들어 행위자에게 열려진 행위가 a_1, a_2 둘뿐이며 실제로는 그 행위로부터 야기되는 결과는 o_1, o_2인데 행위자는 반대로 o_2, o_1인 것으로 잘못 알고 있다고 하자. 또한 각 결과에 대해 행위자가 부여하는 효용 혹은 선호도에 의하면 $u(o_1) > u(o_2)$라고 하자. 이 경우 행위자가 a_1을 선택하는 것이 합리적인가 혹은 a_2를 선택하는 것이 합리적인가? 만일 a_1을 선택하는 것이 합리적이라고 생각한다면 그것은 합리적 선택이 이 세계와 행위자의 선호 체계에 의존한다는 이야기가 된다. 반면에 a_2를 선택하는 것이 합리적이라고 생각한다면 그것은 합리적 선택이 이 세계가 어떻게 되어 있는가에 관한 행위자의 믿음과 그의 선호 체계에 의존한다는 결론이 될 것이다. 여기서

는 일단 행위자가 이 세계가 어떻게 되어 있는가에 관해 착각을 하지 않는 것으로 간주한다. 다시 말해 행위자가 a_1, …, a_n을 선택했을 경우 각각 실제로 o_1, …, o_n과 같은 결과가 야기될 경우 또 오직 그 경우에 한해 행위자도 그런 것으로 믿고 있다고 간주하기로 한다.

위의 단락에서와 같은 선택 상황은 행위자에게 열려진 각 행위에 대해 그 결과가 확실하게 알려진 경우이다. 따라서 이러한 선택을 확실성 하에서의(under certainty) 선택이라고 말할 수 있다. 행위자가 선택할 수 있는 행위 a_1, …, a_n 각각에 대해 o_1, …, o_n인 결과가 얻어진다는 것은 그 각각의 쌍에 대해 인과적인 관계가 성립한다는 의미이다. 그와 같은 인과 관계가 성립하는 것은 물론 이 세계가 어떤 식으로 되어 있는가에 좌우되는 문제이다. 그가 a_1을 선택한 이상 그는 더 이상 다른 행위는 선택할 수 없다. 그러나 이 세계에서 앞서와 같은 인과 관계가 성립한다면 a_1을 선택할 당시 그가 만일 예를 들어 a_2를 선택했다면 o_2라는 결과를 얻어졌을 것이라는 반사실적 명제가 성립한다. 그러한 결과가 나오게 되는 것은 이 세계에 의해 결정되는 문제이지만 그러한 결과 때문에 $u(o_2)$라는 효용을 얻게 되는 것은 행위자가 그 결과에 얼마만큼의 가치를 부여하는가에 좌우된다.

행위자가 a_i를 선택했다고 하자. 그러면 그는 실제로 $u(o_i)$라는 효용을 얻을 것이다. 여기서 $u(o_1)$, …, $u(o_n)$에서 $u(o_i)$를 제외한 나머지 가운데 가장 큰 것을 $u(o_j)$라고 하자. 이것은 행위자가 a_i를 선택했을 당시, 그 행위가 아닌 a_j를 선택했을 경우, $u(o_j)$ 만큼의 효용 내지는 만족을 얻었을 것이라는 것을 의미한다. 그러한 효용을 행위 a_i를 선택한 데 따른 기회비용(opportunity cost)이라고 생각할 수 있다. 어떤 행위를 함으로써 얻어지는 효용이 그 기회비용보다 크다는 것은 그 행위를 함으로써 최대의 효용을 얻을 수

있다는 말과 마찬가지이다. 그러한 행위를 지배적인 행위라고 부르기로 하자. 지금까지의 논의는 확실성 하에서의 선택을 하는 상황에서는 지배적인 행위, 다시 말해 최대의 효용을 가져다주는 혹은 기회비용보다도 큰 효용을 가져다주는 행위를 선택하는 것이 합리적이라는 원칙이 성립함을 보여 준다. 이러한 원칙을 노직에 따라서 지배의 원리(*Dominance Principle*)라고 부르기로 하자.[1)]

위의 지배의 원리는 행위자에게 열려 있는 대안적인 행위 각각에 대해 그 행위를 선택했을 경우 어떤 결과가 야기되는가를 알 수 있을 만큼 행위자가 세계에 관해 알고 있을 때 적용 가능하다. 그러나 그렇지 못한 경우가 있을 수 있다. 예를 들어 위의 상자에 예에서 A와 B 상자 가운데 하나에 100만원이 들어 있는데 어느 상자에 들어 있는지 행위자가 모른다고 하자. 이 경우 행위자가 아는 한에 있어 세계가 놓여 있을 수 있는 상태는 두 가지이다. 하나는 A 상자에 100만원이 들어 있는 경우이고 다른 하나는 B 상자에 100만원이 들어 있는 경우이다. 그것을 각각 s_1, s_2라고 하자. 이 행위자는 각각의 가능한 상태에서 자신에게 열려진 행위를 선택했을 때 결과가 어떠한지를 알 수 있으며 그것을 다음과 같은 표(matrix)로 나타낼 수 있다.

<표 1>

	s_1	s_2
A 상자를 열기	100만원	0
B 상자를 열기	0	100만원

행위자가 x원을 얻었을 때 효용을 $u(x)$라고 했을 때 $x>y$일 경우 또 오직 그 경우에 한해 $u(x)>u(y)$라면 위의 표는 이 세계가

1) Nozick(1969), 111쪽.

실제로 s_1의 상태에 있다면 A 상자를 여기는 것이 지배적인 행위인 반면에 s_2의 상태에 있다면 B 상자를 여는 행위가 지배적임을 보여주고 있다. 따라서 세계가 실제로 어떤 상태에 있는지 모르는 상황에서는 지배의 원리를 적용할 수 없다. 그러나 행위자가 아는 한에 있어 세계가 어떤 상태에 있건 간에 동일한 행위가 지배적인 경우가 있을 수 있다. 이러한 경우에는 물론 지배의 원리가 적용될 것이다. 예를 들어 B 상자에는 1,000원이 들어 있지만 A 상자에는 100만원이 들어 있거나 들어 있지 않다고 하자. B 상자는 투명하기 때문에 그 안에 1,000원이 들어 있는 것을 행위자는 알고 있지만 A 상자 투명하지 못하기 때문에 그 안에 100만원 들어 있는지 여부를 행위자로서는 알고 있지 못하다. 또한 행위자에게 열려 있는 선택은 A 상자 하나만을 열거나 혹은 A, B 두 상자를 모두 여는 것이다. 지금 A 상자 하나 만을 여는 행위를 a_1, 그리고 A, B 두 상자를 모두 여는 행위를 a_2라고 하고 A 상자에 100만원이 들어 있는 상태를 s_1, 들어 있지 않은 상태를 s_2라고 하면 각 상태에 따라 두 행위를 하였을 때 얻어지는 결과를 나타내는 표를 다음과 같이 작성할 수 있을 것이다.

<표 2>

	s_1	s_2
a_1	100만원	0
a_2	100만+1000원	0+1000원

<표 2>는 a_2 행위가 이 세계가 어떤 상태에 있건 간에 a_2가 지배적인 행위임을 보여주고 있다. 따라서 지배의 원리가 적용될 수 있으며 그 원리에 의하면 a_2를 선택하는 것이 합리적이다. 일반적으로 이 세계가 놓일 수 있는 가능한 상태가 s_1, …, s_m인 것으로

행위자가 보고 있다고 하자. 그리고 세계가 s_j 상태에 있다고 했을 때, 행위자가 선택할 수 있는 행위 a_1, …, a_n을 실제로 했을 때 얻어지는 결과를 각각 o_{1j}, …, o_{nj}라고 하자. 이 세계가 어떤 상태에 있다고 하더라도 a_i를 선택한 결과로 얻어지는 효용이 다른 행위를 선택했다고 했을 때의 효용보다 클 경우, 다시 말해 모든 j에 대해 $\mathrm{Max}(u(o_{1j}), \cdots, u(o_{ij}), \cdots, u(o_{nj}))=u(o_{ij})$일 경우 보다 일반화된 지배의 원리에 의하면 a_i를 선택하는 것이 합리적이다.

<표 1>과 같은 선택 상황에 있는 행위자가 A 상자를 열기와 B 상자를 열기 간에 하나를 선택했다면 이 세계가 실제로 s_1과 s_2 상태 간에 하나가 될 개연성이 다른 상태가 될 개연성보다 크다는 생각에서였을 것이다. 지금 행위자가 보기에 이 세계가 실제로 s 상태일 개연성 혹은 확률을 $Pr(s)$라고 하자.[2] 이 경우 문제의 행위자가 A 상자를 여는 행위를 선택했다면 그는 $Pr(s_1)$의 확률로 100만원을 얻을 것이고 $Pr(s_2)$의 확률로 0원을 얻을 것이라고 생각할 것이다. 또한 B 상자를 여는 행위에 대해서도 비슷한 이야기가 성립할 것이다. 따라서 $Pr(s_1)\times u(100만원)+Pr(s_2)\times u(0원)$과 $Pr(s_1)\times u(0원)+Pr(s_2)\times u(100만원)$을 각각 A 상자를 여는 행위와 B 상자를 여는 행위를 선택했을 때 행위자가 얻을 것으로 기대되는 효용, 즉 기대 효용(expected utility)이라고 생각할 수 있다. 만일 $Pr(s_1)>Pr(s_2)$이면 A 상자를 열었을 때의 기대 효용이 더 클 것이고 반대로 $Pr(s_2)>Pr(s_1)$이면 B 상자를 열었을 때 기대 효용이 더 크게 된다. <표 1>과 같은 선택 상황에서는 행위자가 어떤 행위를 선택했다면 이처럼 각 행위에 대한 기대 효용을 계산하여 최대의 기대 효용을 얻을 수 있는 행위를 선택하였을 것이며 또한 그렇게 하는 것이 합리적이라고 생각된다. 그러한 원리가 기

2) 이러한 확률은 이른바 주관적 확률(subjective probability)이라고 할 수 있다.

대 효용 극대화의 원리(Principle of Maximizing Expected Utility)이다.

지금 일반적으로 이 세계가 s_1, …, s_m인 상태에 놓일 수 있는 확률이 각각 $Pr(s_1)$, …, $Pr(s_m)$인 것으로 행위자가 간주하고 있다고 할 때,[3] 행위자가 a_i를 선택했을 때, 얻을 것으로 기대되는 효용 $EU(a_i)$는 다음과 같을 것이다.

$$EU(a_i)=Pr(s_1)\times u(o_{i1})+\cdots+Pr(s_m)\times u(o_{iim})=\sum_{k=1}^{m} Pr(s_k)\times u(o_{ik})$$

기대치 극대화의 원리는 EU의 값을 최대로 하는 행위를 하는 것이 합리적인 선택이라는 것이다. 다시 말해 Max($EU(a_1)$, …, $EU(a_i)$, …, $EU(a_n)$)=$EU(a_i)$일 경우 a_i를 선택하는 것이 합리적이다. 이러한 원리는 이 세계가 정확하게 어떤 상태인지를 알 수 없기 때문에 행위자로서는 위험을 안고(under risk) 선택할 수밖에 없는 상황에서 적용될 수 있다.

지배의 원리와 기대치 극대화의 원리는 서로 다른 원리이다. 그러나 반드시 서로 충돌하지는 않는다. 예를 들어 <표 2>와 같은 상황에서는 지배의 원리를 적용하건 기대치 극대화의 원리를 적용하건 동일한 행위가 합리적이라는 결론을 얻을 수 있다. 일반적으로 이 세계가 놓일 수 있는 어떤 가능한 상태에서도 a_i가 지배적이라면 임의의 j에 대해 $u(o_{11})$, …, $u(o_{nj})$ 가운데 $u(o_{ij})$가 항상 최대가 된다는 것을 의미한다. 이 경우 $Pr(s_1)$, …, $Pr(s_m)$ 분포가 어떻게 되건 상관없이 $EU(a_1)$, …, $EU(a_n)$ 가운데 $EU(a_i)$의 값이 최대가 될 것이다. 따라서 지배의 원리에 의해 합리적인 선택이 바

3) 물론 여기서 $Pr(s_1)+\cdots+Pr(s_m)=1$이다. 다시 말해 행위자는 이 세계가 s_1, …, s_m 이외에 다른 상태일 개연성은 없는 것으로 믿고 있다.

로 극대화의 원리에 의해서도 합리적인 선택이 된다.

위와 같은 결과는 이 세계가 s_1, …, s_m인 가능성이 각각 행위자가 어떤 행위를 하는가와 무관할 경우에 얻어진다. 그러나 이 세계가 어떤 상태에 놓일 확률이 행위자가 어떤 행위를 선택하는가 하는 것과 무관하지 않다고 하면 어떤 행위가 합리적인가에 대해 지배와 원리와 기대 효용 극대화의 원리가 서로 상반된 결론을 내릴 수 있다. 이 세계가 s_1, …, s_m인 가능성이 각각 행위자가 어떤 행위를 하는가와 무관하지 않게 되는 상황은 예를 들어 그 행위자가 앞으로 어떤 행위를 선택을 하는 것이 이 세계가 어떤 상태가 될 것인가에 대해 얼마만큼의 증거가 된다고 생각할 경우에 벌어질 수 있다. 그러한 정도를 행위자가 그 행위를 하는 조건에서 이 세계가 그 상태가 될 조건적 확률로 나타낼 수 있을 것이다. 행위자가 a_i를 선택한 상황에서 이 세계가 상태 s_j일 조건적 확률을 $Pr\ (s_j/a_i)$로 나타낸다면 행위자가 a_i를 선택했을 때 얻어질 것으로 기대되는 효용 혹은 조건적 기대 효용 $CEU(a_i)$는 다음과 같을 것이다.

$$CEU(a_i) = Pr(s_1/a_i) \times u(o_{i1}) + \cdots + Pr(s_m/a_i) \times u(o_{im}) = \sum_{k=1}^{m} Pr(s_k/a_i) \times u(o_{ik})$$

기대 효용의 개념을 이처럼 일반화한다면 조건 효용 극대화의 원리도 조건적 기대 효용 극대화의 원리(Principle of Maximizing Conditional Expected Utility)로 보다 일반화되어야 할 것이다. 이 경우 예를 들어 a_i가 이 세계가 실제로 놓일 수 있을 것으로 행위자가 믿고 있는 어떤 상태에서도 지배적인 반면에 $CEU(a_i)$의 값이 $CEU(a_1)$, …, $CEU(a_n)$ 가운데 최대가 되지 않는 경우가 일어 날 수 있다. 위의 <표 2>와 같은 선택 상황에서 u(100만원)=100, u(0원)=0, u(100만+1,000원)=101, u(1,000원)=1이라고 하자.

그러면 $CEU(a_1)$, $CEU(a_2)$는 각기 다음과 같이 계산될 것이다.

$$CEU(a_1)=Pr(s_1/a_1)\times 100+Pr(s_2/a_1)\times 0$$
$$CEU(a_2)=Pr(s_1/a_2)\times 101+Pr(s_2/a_2)\times 1$$

만일 $Pr(s_1/a_1)=Pr(s_2/a_2)$이고 $Pr(s_2/a_1)=Pr(s_1/a_2)$인데 전자의 값이 후자에 비해 상당히 클 경우, a_2가 항상 지배적인 선택이 됨에도 불구하고 $CEU(a_1)>CEU(a_2)$이어서 조건적 기대 효용 극대화의 원리에 의해 a_1이, 즉 하나의 상자만을 여는 선택이 합리적이게 되는 경우가 있을 수 있게 된다. 이것이 바로 뉴컴의 역설이 제시하는 문제 상황이다.

뉴컴은 이 행위자의 행위를 예측할 수 있는 강력한 예측력을 지닌 인물을 도입함으로써 $Pr(s_1/a_1)=Pr(s_2/a_2)\gg Pr(s_2/a_1)=Pr(s_1/a_2)$인 상황을 만들고 있다. 이 예측인이 행위자가 a_1이나 a_2를 선택하기 이전에 그가 a_1을 선택할 것으로 예측했다면 A 상자에 100만원을 넣어두었고 그렇지 않았다면 아무 것도 넣어두지 않았다고 하자. 또한 이 사실을 그 행위자가 알고 있다고 하고 또 그의 예측의 정확성에 대해 행위자가 신뢰하고 있는 정도를 r이라고 하자. 또한 s_1을 그가 A 상자에 100만원을 넣어 둔 상대라고, 그리고 s_2를 아무것도 넣어 두지 않은 상태라고 하자. 그러면 $Pr(s_1/a_1)=Pr(s_2/a_2)=r$이고 $Pr(s_2/a_1)=Pr(s_1/a_2)=1-r$일 것이며 r이 1에 가까우면 $Pr(s_1/a_1)=Pr(s_2/a_2)\gg Pr(s_2/a_1)=Pr(s_1/a_2)$가 성립할 것이다. 따라서 $CEU(a_1)=r\times 100$, $CEU(a_2)=(1-r)\times 100+r\times 1$로서 $CEU(a_1)$이 $CEU(a_2)$보다 훨씬 크게 되어 조건적 기대 효용 극대화의 원리에 의해 a_1, 즉 A 상자만을 여는 선택이 두 상자 모두를 여는 선택에 비해 합리적이라는, 지배의 원리와는 상반된 결론이 도출된다. 합리적 선택에 관한 두 가지 원리가 상반된 결과를 제시하는 상황에서는 그렇다면 어떤 원리가 진정으로 합리적 행위를 제공하는가?

2

프리스트는 Priest[2002]에서 뉴컴의 문제 상황이 합리적인 선택을 하려는 행위자를 딜레마에 빠지게 한다고 주장하고 있다. 그가 제시하는 뉴컴의 문제는 그 행위자가 앞의 <표 2>에서와 같은 두 개의 상자에서 A 상자 하나만을 여는 것(a_1)과 두 상자를 모두 여는 것(a_2) 가운데 하나를 선택해야 상황에서 어떤 행위를 선택할지를 미리 알 수 있는 완벽한 예측인(perfect predictor)의 존재를 전제하고 있다. 이 행위자가 a_1을 선택할 것으로 그 예측인이 미리 알았을 경우 그는 A 상자에 100만원을 넣어 두었으며 a_2를 선택할 것으로 미리 알았을 경우 A 상자에 아무 것도 넣어두지 않았다. 프리스트에 의하면 다음과 같은 원리 R을 받아들이면 행위자는 딜레마 상황에 놓이게 된다.

$$(R)\qquad \frac{\begin{array}{l}P(a_1) \rightarrow G(o_1)\\ P(a_2) \rightarrow G(o_2)\\ u(o_1) > u(o_2)\end{array}}{O(a_1)}$$

R에서 $P(a)$는 행위자가 a의 행위를 선택했다는 것을 의미하며 $G(o)$는 그가 o를 얻게 된다는 것을 뜻한다. 그리고 $O(a)$는 행위자가 a를 하는 것이 합리적이라는 것을 말한다.[4] 프리스트에 의하면

4) 프리스트의 원래의 공식은 다음과 같다.

$$\frac{\begin{array}{l}C(\gamma, \delta)\\ M\gamma \rightarrow Gc_\gamma\\ M\delta \rightarrow Gc_\delta\\ c_\gamma > c_\delta\end{array}}{O(M\gamma)}$$

여기서 $C(\gamma, \delta)$는 행위자가 γ를 참으로 하거나 δ를 참으로 하는 것 가운

행위자를 딜레마 상황으로 몰아넣는 두 뿔(horn)은 다음과 같다.[5]

> 뿔 1: 예측인이 A 상자에 아무 것도 넣어 놓지 않은 경우. o_1=0원이고 o_2=1,000원. 그러므로 $P(a_1)\rightarrow G$(0원), $P(a_2)\rightarrow G$(1,000원)이고 또한 u(1,000원)>u(0원)이라고 할 수 있으므로 원리 R에 의해 $O(a_2)$. 예측인이 A 상자에 100만원을 넣어 놓은 경우. o_1=100만원이고 o_2=100만+1,000원. 그러므로 $P(a_1)\rightarrow G$(100만원), $P(a_2)\rightarrow G$(100만+1,000원)이고 또한 u(100만+1,000원)>u(100만원)이라고 할 수 있으므로 또 다시 원리 R에 의해 $O(a_2)$. 어느 경우에나 $O(a_2)$이므로 결국 a_2를 선택하는 것이 합리적이다.
>
> 뿔 2: 만일 행위자가 a_1을 선택한다면 다시 말해 A 상자만을 여는 것을 선택한다면 완벽한 예측인은 이 사실을 알 것이고 A 상자 안에 100만원을 넣었을 것이다. 그러므로 $P(a_1)\rightarrow G$(100만원). 그러나 행위자가 a_2를 선택한다면 즉 두 상자 모두 여는 것을 선택한다면 완벽한 예측인은 또 다시 이 사실을 알 것이고 A 상자 안에 아무 것도 넣지 않았을 것이다. 그러므로 $P(a_2)\rightarrow G$(1,000원). 또한 u(100만원)>u(1,000원)이라고 할 수 있으므로 또 다시 원리 R에 의해 $O(a_1)$. 즉 a_1을 선택하는 것이 합리적이다.

위와 같은 프리스트의 논증이 타당하다면 합리적 선택에 관한 매우 그럴 듯한 원리 R은 두 가지 선택 모두가 합리적이라는 불합리한 결론을 낳으며 따라서 합리적인 선택을 하려는 행위자는 딜레마에 처하게 된다. 그러한 결론은 원리 R 이외에도 완벽한 예측인이 존재한다는 가정에 의존하고 있다. 그러므로 프리스트의 비판자는 그러한 딜레마 상황을 초래한다는 프리스트의 논증을 원리 R이 잘못되었거나 혹은 프리스트가 가정하는 완벽한 예측인이 존재

데 하나를 선택해야 한다는 것을 의미하며 My는 행위자가 y를 참으로 한다는 것을 그리고 Gx는 행위자가 x를 얻는다는 것을 의미한다. 또한 →는 직설적인 조건(indicative conditional)을 뜻한다(Priest[2002]의 13쪽 참조). 프리스트의 공식에서는 전제에 $u(c_y)>u(c_\delta)$가 아닌 $c_y>c_\delta$이 등장하고 있지만 효용 함수 u가 $x>y$인 경우 또 오직 그 경우에 한해 $u(x)>u(y)$를 만족하면 그 둘 가운데 어떤 것을 전제로 포함해도 무방할 것이다.

5) Priest(2002), 13쪽.

할 수 없다는 것에 대한 귀류 논증으로 삼을 수 있다고 주장할 수도 있을 것이다.

프리스트의 논증 뿔 1은 예측인이 있건 없건 타당하다. 어떤 경위로서든지 B 상자에는 1,000원이 들어 있는 것이 확실하고 A 상자에는 100만원이 들어 있거나 들어 있지 않다면 뿔 1과 같은 논증에 의해 두 상자를 모두 여는 선택을 하는 것이 합리적이라는 결론은 내릴 수 있다. 그러나 완벽한 예측인이 존재하지 않는다면 뿔 2와 같은 논증은 성립하지 않는다. 따라서 뿔 1에 의해 두 상자를 모두 여는 것이 합리적이라는 해답을 내림으로써 뉴컴의 문제는 해결된다. 이러한 길을 슐레징거(G. Schlesinger)는 걷고 있다.[6] 우리의 행위자가 어떤 선택을 하기 이전에 그와 유사한 선택 상황에 있던 다른 많은 행위자의 선택을 그때마다 정확하게 예측했던 예측인이 존재한다면 그러한 예측인은 프리스트가 말하는 완벽한 예측인으로 생각될 수도 있을 것이다. 또한 그러한 예측인이 존재한다는 가정은 논리적으로도 모순되지 않는다. 그러나 슐제징거에 의하면 그러한 예측인도 뿔 2와 같은 논증을 타당하게 만들 정도로 완벽한 예측인이라고 생각할 수 없다. 왜냐하면 문제의 예측인이 프리스트가 말하는 완벽한 예측인이라는 추론은 일종의 귀납 논증에 의거하고 있는데 그 논증보다도 뿔 1의 결론을 뒷받침하는 더 강력한 연역 논증이 존재하기 때문이다.[7]

슐레징거는 더 나아가 한 행위자가 자유로운 선택을 할 수 있다는 것과 그의 선택을 완벽하게 예측할 수 있다는 것은 상호 충돌한다고 생각한다. 우리의 행위자는 자유로운 선택을 할 수 있으며 따라서 그의 행위를 완벽하게 예측할 수는 없다. 따라서 프리스트의 뿔 2의 논증을 가능하게 하는 완벽한 예측인은 존재할 수 없다.

---

6) Schlesinger(1974).

7) Schlesinger(1974), 212쪽.

행위자가 자유로운 선택을 한다는 것과 그의 행위를 완벽하게 예측할 수 있다는 것이 상호 양립불가능하다는 주장은 중세 신의 전지함을 문제 삼을 때에도 제기되었던 비판이다. 그러한 비판에 의하면 만일 신이 전지한 존재로서 우리의 행위자가 어떤 선택을 할지를, 예를 들어 a_1을 선택하리라는 것을 미리 알고 계시다면 그 행위자는 실제로 a_1을 할 뿐더러 그 행위 이외에 다른 행위를 선택하는 것은 불가능하다. 따라서 그는 실제로 하나의 상자를 여는 선택을 하였을 뿐더러 두 개의 상자를 모두 여는 행위를 선택할 수도 없었다. 그러나 이병욱은 인간에 관한 한 이러한 논증을 받아들이지 않는다. 그는 Yi[2003]에서 이렇게 말하고 있다. "예를 들어 부시의 일과를 충분히 알고 있는 사람들은 (예컨대 부시 자신은) 부시가 내일 정오 이전에 잠자리에서 일어나리라는 것을 정확하게 예측하고 또 미리 알고 있을지 모른다. 그러나 이 사실이 그가 (현재) 그러한 행위 이외에 다른 선택이 없다는 것을 의미하지는 않는다."

인간의 행위가 아닌 자연적 사건에 대한 정확한 예측은 현재 입수하고 있는 자료와 예측하려는 사건 간에 인과 관계에 관한 지식에 의존하여 이루어진다. 현재 지구가 어떤 위치에 있다는 것을 알고 있다면 물리 법칙에 의존하여 내일 지구가 어떤 위치에 있으리라는 것을 미리 알 수 있으며 또한 정확하게 예측할 수 있다. 이 경우 그 예측이 정확하다는 것은 지구가 내일 그 위치 이외에 다른 위치에 있는 것이 불가능하다는 것을 함축한다. 인간의 경우에도 예를 들어 부시의 선택을 정확하게 예측할 수 있을 정도로 부시의 일과에 관한 충분히 알기 위해서는 현재 상황에 대한 지식과 더불어 내일 부시의 선택과 현재의 상황을 연결하는 인과 법칙을 아는 것이 필요하다면, 그리고 그러한 법칙에 관한 지식을 토대로 내일 부시의 선택을 예측했다면 부시는 내일 예를 들어 정오 이전

에 다른 선택을 하는 것이 불가능하다는 결론을 내려야 할 것이다. 반대로 부시가 실제로 한 선택 이외에 다른 선택을 하는 것이 불가능하다면 부시의 선택에 관한 예측은 불가능하다고 해야 할 것이다.

예측인이 행위자가 a_1을 선택하리라는 것을 미리 알았음에도 불구하고 행위자가 a_2를 선택할 수도 있다고 하자. 그리고 예측인의 예측이 행위자의 선택을 인과적으로 초래할 수 있는 어떤 원인, 예를 들어 C를 앎으로서 이루어졌다고 하자.[8] 행위자가 a_1을 선택하리라는 예측인의 지식을 결과한 원인 C는 또한 그가 100만원을 A 상자에 넣는 행위를 인과적으로 초래할 것이며 또한 그러한 결과와 행위자의 선택이 원인이 되어 행위자가 100만원을 얻게 되는 결과가 초래될 것이다. 그런데 행위자는 a_1을 선택할 당시 a_1이 아닌 a_2를, 다시 말해 두 상자를 모두 여는 행위를 선택할 수도 있었다. 그렇다면 그 결과는 어떻게 되었을까? 뿔 1에서와 같은 추론에 의하면 그 결과는 100만+1,000원을 얻는 것이다. 즉 뿔 1에 의하면 행위자가 a_1을 선택하여 100만원을 얻은 상태에서 "만일 그가 a_1이 아닌 a_2를 선택했다면 100만+1,000원을 얻을 수 있었을 것이다."는 반사실적 진술이 성립한다. 그러나 완벽한 예측인의 존재 때문에 "만일 그가 a_1이 아닌 a_2를 선택했다면 1,000원만을 얻을 수 있었을 것이다."는 반사실적 진술이 성립한다는 결론도 귀결된다. 따라서 뿔 1에서처럼 a_2를 선택하는 것이 합리적이라는 결론과 동시에 뿔 2에처럼 a_1을 선택하는 것이 합리적이라는 모순된 결론이 나오게 되는 것이다. 그러나 행위자가 a_1을 선택하여 100만원을 얻은 상태에서 위의 두 반사실적 명제가 동시에 성립한다고 생각할 수 없다. 그러므로 프리스트가 가정하는 것과 같은 예측인의 존

8) 그 원인 C는 말하자면 예측인의 예측과 행위자의 선택의 공동 원인이 되는 셈이다.

재는 거부되어야 한다.

프리스트가 가정하는 것과 같은 행위자의 선택을 미리 완벽하게 알 수 있는 존재는 아니더라도 행위자가 절대 신뢰하는 예측인의 존재는 생각할 수 있다. 예를 들어 어떤 사람이 과거 많은 사람의 선택을 정확하게 예측하는 데 성공했다고 하자. 이러한 관찰을 토대로 행위자는 그의 예측에 대해 절대적인 믿음을 가질 수 있다. 그러나 물론 이것이 그 예측인이 그 행위자의 선택에 관해 미리 정확하게 예측할 수 있다는 것을 함축하지 않는다. 행위자가 절대적으로 신뢰하는 예측인이 그가 A 상자만을 열 경우 그 상자에 100만원을 넣어 두었으며 두 상자를 모두 열 경우 아무 것도 넣어 두지 않았다는 것을 행위자 자신이 알고 있는 새로운 상황에서도 뿔 1과 같은 논증은 성립할 것이다. 이 경우에도 프리스트의 원리 R은 여전히 타당하게 적용된다. 그렇다면 다른 뿔은 어떻게 되는가?

행위자는 자신이 A 상자만을 여는 행위를 선택하면 그 상자에 100만원이 들어 있을 것이라는 것을 예측인의 예측을 신뢰하는 정도로 믿을 것이다. 따라서 그러한 정도로 u(100만원)에 해당하는 효용을 얻게 될 것으로 기대할 것이다. 반대로 두 상자를 모두 연다면 A 상자에 100만원이 들어 있지 않고 오직 B 상자에만 1,000원이 들어 있을 것이라고 같은 정도로 믿을 것이다. 이 경우에 얻을 것으로 기대되는 효용은 u(1,000원)에 불과하다. 따라서 그 두 효용을 비교하여 a_1을 선택하는 것이 합리적이라는 결론을 내릴 것이다. 그러나 이 경우에 비교되는 것은 기대효용이다. 따라서 이 경우에는 R이 아닌 다음의 원리가 적용된다고 생각해야 한다.

$$(U) \qquad \frac{CEU(a_i) > CEU(a_j)}{O(a_i)}$$

행위자는 예측인을 절대적으로 신뢰하고 있다. 그러므로 자신이 a_1을 선택할 경우 A 상자에 확실하게 100만원이 들어 있을 것이라고 생각할 것이다. 마찬가지로 a_2를 선택할 경우 확실하게 들어 있지 않을 것이라고 생각할 것이다. 따라서 A 상자에 100만원이 들어 있는 상태를 s_1이라고 하고 들어 있지 않은 상태를 s_2라고 하면 $Pr(s_1/a_1)=Pr(s_2/a_2)=1$이고 $Pr(s_2/a_1)=Pr(s_1/a_2)=0$. 그리고 $o_{11}=$ 100만원, $o_{21}=$100만+1,000원, $o_{12}=$0원, $o_{22}=$1,000원이므로

$$CEU(a_1)=Pr(s_1/a_1)\times u(o_{11})+Pr(s_2/a_1)\times u(o_{12})=u(o_{11})=u(100\text{만원})$$
$$CEU(a_2)=Pr(s_1/a_2)\times u(o_{21})+Pr(s_2/a_2)\times u(o_{22})=u(o_{22})=u(1{,}000\text{원})$$

따라서 $CEU(a_1)>CEU(a_2)$이므로 a_1, 즉 A 한 개의 상자만을 여는 것이 합리적이라는 결론이 귀결된다. 그러나 이것은 원리 R이 아닌 U를 적용한 결과인 것이다.

프리스트가 말한 것과 같은 합리적 선택을 둘러싼 딜레마 상황은 행위자의 선택을 실제로 완벽하게 미리 알고 있는 예측인을 설정하고 원리 R만을 사용함으로써 야기된다. 그렇지만 그러한 절대적 예측인이 존재할 수 있는지에 대해서는 의문의 여지가 있다. 그러나 그러한 예측인이 아닌 단지 행위자가 절대적으로 신뢰하는 예측인을 설정하고 프리스트 원리 이외에도 다른 원리를 U를 채택함에 의해서도 딜레마 상황은 야기된다. 이것은 본질적으로 앞의 절에서 설명한 것과 같은 딜레마이다. 왜냐하면 U는 조건적 효용 극대화의 원리에 해당하며 또한 프리스트의 원리 R은 바로 지배의 원리로부터의 귀결이라고 말할 수 있기 때문이다. 그렇다면 이러한 딜레마 상황에서 어떤 원리를 적용한 결과가 정말로 합리적인 선택이 될 수 있는가 하는 문제는 여전히 남게 된다.

3

뉴컴의 문제에서 지배의 원리와 조건적 효용의 극대화의 원리가 서로 충돌하는 딜레마적인 상황은 $u(o_{21})>u(o_{11})$이고 $u(o_{22})>u(o_{12})$이어서 이 세계가 실제로 어떤 상태인 것으로 확인되더라도 a_2의 행위가 지배적인 것으로 드러나는 반면에 $CEU(a_1)>CEU(a_2)$일 정도로 $Pr(s_1/a_1)=Pr(s_2/a_2)$이 $Pr(s_2/a_1)=Pr(s_1/a_2)$에 비해 충분히 크고 $u(o_{21})$와 $u(o_{11})$, 그리고 $u(o_{22})$와 $u(o_{12})$의 값의 차이가 적절할 경우에 항상 발생한다. 이러한 경우에 노직은 행위자의 선택이 앞으로 어떤 상태가 벌어질지에 대해 인과적인 영향을 미치지 않을 경우, 그의 표현에 의하면 "행위 내지는 행위를 하겠다는 결정이 어떤 상태가 벌어질지에 대해 어떤 영향을 미치거나, 그러한 결과를 초래하는 데 보탬이 되거나 혹은 좌우하지 않을 경우" 지배의 원리를 적용해야 할 것으로 생각한다고 말하고 있다.[9] 그러나 곧 이어 $Pr(s_1/a_1)=Pr(s_2/a_2)=1$이고 $Pr(s_2/a_1)=Pr(s_1/a_2)=0$인 극단적인 경우 즉 예측인의 예측이 100% 정확할 확률이 1인 경우에는 지배의 원리가 아닌 조건적 기대 효용 극대화의 원리가 적용되어야 한다고 말하고 있다. 그에 의하면 행위자는 다음과 같이 추론하리라는 것이다.

> 전제 1: 만일 내가 두 상자를 모두 연다면 1,000원만을 얻으리라는 것을 알고 있다. 반면에,

9) Nozick(1969), 123쪽. a가 s를 인과적으로 초래한다는 것을 $a\Rightarrow s$로 나타내면 노직의 말은 곧 $Pr(a_1\Rightarrow s_1)=Pr(a_2\Rightarrow s_1)$이고 $Pr(a_1\Rightarrow s_2)=Pr(a_2\Rightarrow s_2)$일 경우 지배의 원리가 적용되어야 한다는 것이다. 그런데 1절에서 $CEU(a_i)$에 관한 정의에서 조건부 확률 $Pr(s_i/a_j)$를 모두 인과적 확률 $Pr(a_j\Rightarrow s_i)$로 바꾼다면 지배의 원리의 결론과 기대효용 극대화의 원리의 결론은 서로 일치하게 될 것이다.

전제 2: 만일 내가 A 상자만을 연다면 100만원을 얻으리라는 것을 알고 있다.
따라서 A 상자만을 열어야 한다.[10]

위의 논증은 본질적으로 프리스트의 뿔 2이다. 노직은 왜 예측인의 예측이 정확할 확률이 1인 극단적인 상황에서는 위와 같은 논증이 지배의 원리에 우선할 수 있는지에 대한 이유를 제시하고 있지 않다. 만일 그러한 논증을 제시하는 데 성공한다고 하더라도 노직 스스로도 지적하는 것처럼 왜 $Pr(s_1/a_1)=Pr(s_2/a_2)=1$이고 $Pr(s_2/a_1)=Pr(s_1/a_2)=0$인 극단적인 상황에서는 지배의 원리가 무력화되고 그 확률이 1보다 적은 경우, 따라서 예측인의 예측인 거짓일 확률이 조금이라도 있는 경우에는 조건적 기대 효용 극대화의 원리 대신 지배의 원리가 적용되어야 하는지에 대한 설명이 필요하게 될 것이다.

슐제진저는 위에서 언급한 것처럼 인간의 선택에 관해 100% 신뢰 가능한 예측을 한다는 것은 불가능하다고 본다. 만일 그러한 예측인이 존재한다면 우리의 행위자는 A 상자만을 선택해야 한다는 노직의 주장에 슐래진저는 동의한다. 그러나 그러한 주장을 뒷받침할 수 있는 논증보다 더 강력한, 두 개의 상자를 여는 선택을 뒷받침하는 논증이 존재한다고 슐레징거는 생각한다. 그러한 논증과 충돌하므로 100% 신뢰할 수 있는 예측인의 존재와 더불어 자유 의지에 의한 인간의 행위에 대한 완벽한 예측 가능성도 부정해야 한다. 두 개의 상자를 여는 선택을 뒷받침하는 논증을 전개하기 위해 슐레징거는 뉴컴의 문제 상황을 약간 바꾸어 두 상자에 무엇이 들어 있는지를 완전히 알고 있는, 그러면서도 동시에 나에게 가장 유리한 선택을 하도록 도와주려는 선의로 충만한 조언자가 존재하는 상황을 가정하고 있다. 그 조언자는 A 상자에 100만원이 들어 있

10) Nozick(1969), 128-8쪽.

는지 아니면 비어 있는지를 볼 수 있다. 그러나 자기가 본 것이 얼마이건 간에 그는 행위자에게 두 상자를 모두 여는 것이 행위자에게 이익이 된다는 것을 알 것이고 따라서 그렇게 조언할 것이다. 그러므로 그러한 조언에 따르는 것이 행위자에게 가장 이익이 될 것이고 그렇기 때문에 행위자는 두 상자를 모두 여는 선택을 해야 한다는 것이다. 슐레징거에 의하면 그러한 조언자가 실제로 존재했는가는 문제가 되지 않는다. 중요한 것은 그 조언자가 존재했다면 어떤 조언을 했을까 하는 것이다. 그에 의하면 그러한 조언자가 존재했다면 제공했을 조언에 반대되는 행동을 하는 것이 나에게 가장 이익이 된다고 주장하는 것은 모순이다.[11]

지금 뉴컴의 문제 상황을 N이라고 하고 슐레징거의 상황을 S라고 하자. 슐레징거의 논증은 다음과 같이 진행되는 것 같다.

> 전제 1: S의 상황에서 조언자가 조언하는 선택이 그 상황에서의 행위자에게 합리적일뿐더러 N의 상황에서의 행위자에게도 합리적인 선택이다.
> 전제 2: S의 상황에서 조언자는 a_2를 선택하도록 조언할 것이다.
> 결론: N 상황에서 a_2를 선택하는 것이 행위자에게 합리적이다.

우선 슐레징거의 전제 1이 참인지 의심스럽다. 앞서 <표 1>에서의 선택 상황에서 실제로 A 상자에 100만원이 들어 있는데 행위자는 A 상자는 비어 있고 B 상자에 100만원이 들어 있을 확률이 0.8이라고 믿고 있다고 하자(여기서 A와 B 상자에 100만원이 들어 있을 확률이 행위자의 선택과는 독립적이라고 가정하기로 하자). 이 경우 위의 S 상황에서와 같은 조언자가 있다고 하면 그는 A 상자를 열도록 조언할 것이다. 그러므로 S 상황에서는 행위자가 그 조언자의 조언을 따르는 것이 합리적이다. 그러나 실제 상황에서는 B 상자를 여는 선택의 기대 효용이 더 크다. 그러므로 행위

11) Schlesinger(1973), 211쪽.

자가 B 상자를 여는 선택을 하는 것이 합리적이다. 설사 B 상자를 열어 본 결과 그 안에는 아무 것도 없고 A 상자에 100만원이 들어 있는 것으로 확인되었다 하더라도 상자를 열었을 당시에는 B 상자를 여는 것이 합리적이었다고 말할 수 있다. 이것은 슐레징거가 묘사하는 것과 같은 가상적 상황에서의 합리적인 선택과 실제 상황에서의 합리적인 선택이 서로 별개일 수 있음을 보여준다. 마찬가지로 뉴컴의 문제에서의 S 상황에서의 행위자에게 합리적인 선택이 반드시 N 상황에서의 행위자에게도 합리적인 선택이 된다고 생각할 이유가 없다.

슐레징거가 묘사한 S 상황에서도 행위자의 선택에 관한 예측에 정확한 예측인이 존재하는지, 그리고 그 사실과 더불어 예측인이 어떤 예측을 하였는지를 조언자가 알고 있는지가 분명하지 않다. 만일 그러한 예측인이 존재하지 않거나 혹은 그 예측인이 예측한 내용을 조언자가 알고 있지 못하다면 조언자가 판단한 합리적인 선택이 실제 상황에서도 합리적인 선택이 될 것이라고 주장할 여지는 더욱 없게 된다. 그런데 만일 그러한 예측인이 존재하고 조언자가 그 사실과 더불어 예측인의 예측 내용을 알고 있었다면 그가 과연 슐레징거의 주장대로 a_2를 선택하는 것이 행위자에게 가장 유리하다는 의미에서 합리적이라고 판단했을까? 그렇지 않다고 불여지가 전혀 없는 것이 아니다.

S의 상황을 조언자가 존재하지 않고 단지 행위자가 두 상자의 내용물을 모두 볼 수 있는 상황으로 바꾼다고 해도 슐레징거의 논증은 그대로 유지될 것 같다. 그러한 변형된 상황에서 행위자에게 가장 유리한 선택이 실제 상황에서의 행위자에게도 가장 유리한데 그 변형된 가상 상황에서 행위자는 a_2를 가장 유리한 것으로 판단하여 그것을 선택할 것이며 따라서 실제 상황에서도 a_2를 선택하는 것이 합리적이라는 것이 슐레징거의 논증인 것으로 생각된다. 그런

데 그러한 변형된 선택 상황은 이병욱이 Yi[2003]에서 도입한 O 상황과 유사하다. 이병욱의 O 상황은 A 상자 안에는 아무 것도 들어 있지 않으며 또한 행위자는 B 상자는 물론 A 상자도 투명하여 그 안의 아무 내용물도 없음을 보고 알 수 있는 상황이다. 이병욱은 O 상황에서도 위의 절에서의 프리스트의 논증 뿔 1과 뿔 2는 그대로 성립하기 때문에 프리스트가 말하는 합리적 딜레마 상황은 발생해야 한다고 주장한다. 그러나 프리스트마저도 O 상황에서는 뿔 2의 논증은 성립하지 않을 것으로 보리라고 이병욱은 진단한다. 왜냐하면 B 상자에는 1,000원이 들어 있으나 A 상자에는 아무 것도 들어 있지 않은 것을 알고 있는 상황에서도 A 상자만을 선택하는 것이 합리적이라고 생각할리는 없기 때문이다. 일단 A 상자에 아무 것도 들어 있지 않다는 것을 알고 있는 한, A 상자 하나만을 선택하는 것이 그 안에 100만원이 들어 있게 하는 데 아무 보탬이 될 리 없다는 것은 명백하다.[12] 프리스트에 의하면 O 상황에서도 원리 R은 적용 가능하며 따라서 그의 이른바 합리적 딜레마 상황이 야기되는데, 사실은 O 상황에서는 두 상자는 여는 것이 합리적이다. 그러므로 프리스트의 원리는 배격되어야 한다.

우리는 여기서 이병욱의 상황 O에 처한 행위자가 구체적으로 어떤 논증을 전개할지를 물어 볼 필요가 있다. 우선 그는 프리스트의 뿔 1과 같은 논증을 전개하여 두 상자 모두를 여는 선택이 합리적이라는 결론에 도달했을 것이다. 뿔 1은 완전한 예측인이 존재하는가와 상관없이 전개할 수 있는 추론이며, 본질적으로 지배의 원리라고 할 수 있는 프리스트의 원리 R이 적용된다. 그러한 추론은 다음과 같이 진행될 것이다.

(1) $P(a_1) \rightarrow G$(0원)
$P(a_2) \rightarrow G$(1,000원)

12) Yi(2003), 239쪽.

$$\frac{u(1{,}000\text{원}) > u(0\text{원})}{O(a_2)}$$

위에서 a_1은 A 상자 하나만을 여는 행위를 의미하며 a_2는 두 상자 모두를 여는 행위를 의미한다.

여기까지는 프리스트의 원리 내지는 지배의 원리를 적용하는 데 아무 문제도 없다. 문제는 프리스트의 원리 혹은 원리 U를 이용하여 뿔 2와 같은 논증을 전개하는 데 있다. 이병욱은 이 경우에 프리스트의 원리나 U를 이용한 추론은 잘못된 것이라고 생각하는 듯하다. 왜 그런가?

만일 O 상황에서는 N 상황에서와 같은 행위자의 선택을 미리 예측할 수 있는, 혹은 행위자가 그의 예측을 완전히 신뢰하는 예측인이 없다고 하면 뿔 2와 같은 논증은 해당이 되지 않는다. 그러나 이것은 O 상황을 N과는 전혀 다른 종류의 선택 상황으로 만들게 된다. N 상황은 행위자가 위험을 안은 선택 상황이다. 그러나 O 상황이 위와 같다면 그것은 확실성 하에서의 선택 상황이 되며 따라서 지배의 원리만이 적용 가능하다. 그러므로 예측인의 존재가 배제된 O 상황에서는 프리스트의 원리의 적용이 잘못된 결과를 야기한다고 말할 수 없다. 또한 이 상황에서 합리적인 행위가 N 상황에서도 그러하리라고 장담할 수 없다.[13]

그러나 이병욱은 O 상황이 프리스트의 원리 R이 뿔 2와 같은 방식으로 적용이 허용되는 상황임에도 불구하고 그와 같은 적용이

13) 예측인의 존재가 배제된 O 상황에서의 행위의 선택과 그 결과를 나타내는 도표(matrix)는 다음과 같다.

a_1	0원
a_2	1,000원

그러나 N 상황에서의 도표는 다음과 같을 것이다

잘못된 결과를 낳는다고 생각하는 듯하다. 그는 아마도 위에서 든 노직의 논증에 대해서도 마찬가지로 생각할 것이다. 프리스트의 뿔 2와 노직의 논증은 다음과 같이 정식화될 것이다.

(2)
$$\frac{\begin{array}{l}P(a_1) \rightarrow G(100\text{만원}) \\ P(a_2) \rightarrow G(1000\text{원}) \\ u(100\text{만원}) > u(1000\text{원})\end{array}}{O(a_1)}$$

이병욱은 이러한 잘못된 추론을 봉쇄하기 위해 프리스트의 원리

	A 상자에 100만원을 넣은 경우(s_1)	아무 것도 넣지 않은 경우(s_2)
a_1	100만원	0원
a_2	100만+1,000원	0 + 1,000원

따라서 논증 뿔 1도 O 상황과 같이 진행되지 않고 양도 논법 형식으로 진행된다. 즉 s_1인 경우,

$$\frac{\begin{array}{l}P(a_1) \rightarrow G(100\text{만원}) \\ P(a_2) \rightarrow G(100\text{만+1000원}) \\ u(100\text{만+1000원}) > u(100\text{만원})\end{array}}{O(a_2)}$$

또한 s_2인 경우,

$$\frac{\begin{array}{l}P(a_1) \rightarrow G(0\text{원}) \\ P(a_2) \rightarrow G(0\text{+1000원}) \\ u(0\text{+1000원}) > u(0\text{원})\end{array}}{O(a_2)}$$

어느 경우에나 $O(a_2)$이므로 결국 $O(a_2)$. 그러나 O 상황에서는 이 가운데 두 번째 추론만이 행해질 것이다. 이 점이 바로 이병욱이 묘사한 O 상황이 사실은 앞의 도표로 표현되는 확실성 하에서의 선택 상황이라는 의심을 갖게 하는 것이다.

R을 수정할 것을 제안한다. $t(a)$를 a를 함으로써 결과 되는 총이익(the total benefit *resulting* (or *deriving*) from choosing a)이라고 할 때 이병욱이 제안하는 새로운 원리 $R*$는 다음과 같이 정식화된다.[14]

$$(R*) \quad \frac{t(a_i) > t(a_j)}{O(a_i)}$$

다음이 성립한다고 하자.

$$(Y) \quad (t(a)=x) \rightarrow (\exists y)((P(x) \rightarrow G(y)) \wedge x=u(y))$$

Y는 x가 a를 함으로써 얻어지는 총이익이라고 할 경우, 그것은 a를 하면 얻어지는 것의 효용이 된다는 것을 의미하고 있다. Y가 성립할 경우 $R*$는 R을 함축한다. Y의 역이 성립할 경우 R도 $R*$를 함축하겠지만 이병욱은 그러한 역이 성립한다는 것을 부정한다.

지금 O 상황에서는 $t(a_1)=u$(0원), $t(a_2)=u$(1,000원)이라고 할 수 있으므로 명백히 다음과 같은 추론이 성립한다.

$$(3) \quad \frac{t(a_2) > t(a_1)}{O(a_2)}$$

그러므로 Y에 의해 위의 (1)과 같은 추론도 성립한다. 만일 O 상황에서 $t(a_1)=u$(100만원), $t(a_2)=u$(1,000원)이라고도 할 수 있으면 (2)도 성립할 것이다. 그러나 그 상황에서 A 상자만을 엶으로

14) Yi(2003), 240쪽 참조. $R*$에 대한 이병욱의 정식화는 다음과 같다.

> 만일 두 가지 대안 가운데 선택해야 하는 상황이라고 하고 하나의 대안을 선택함으로써 *결과 되는* (혹은 *파생되는*) 총이익이 다른 대안을 선택함으로써 *결과 되는* 총이익보다 크다고 하자. 이 경우 첫 번째 대안을 선택하는 것이 합리적인 면에서 요구된다.(이탤릭체 원저자)

써 얻게 되는 총이익이 u(1,000원)이기도 하다는 것은 그러한 선택을 함으로써 얻게 되는 총이익이 u(0원)이자 동시에 u(1,000원)이기도 하다는 것이므로 모순이다. 그러므로 R*에 의지할 경우 (3)과 같은 추론 이외에는 성립하지 않는다.

그런데 어떤 행위를 선택함으로써 결과 된다는 것이 무슨 의미인가? 만일 그 행위를 선택함으로써 인과적으로 결과 된다는 것을 의미한다면 a_i를 선택함으로써 결과 되는 것은 위의 1절에서 언급한 o_i라고 할 수 있다. 이 경우 $t(a_i)$는 $u(o_i)$에 해당하게 되며 따라서 이병욱의 원리 R*는 지배의 원리에 불과하게 된다. 원리 R*를 고집하는 이병욱의 입장은 곧 O 상황이나 N 상황에서 지배의 원리만을 적용해야 한다는 입장과 통한다. 물론 지배의 원리만을 적용한다면 애초에 뉴컴의 문제 상황은 발생하지 않는다. 뉴컴이 문제는 지배의 원리와 U나 기대 효용 극대화의 원리의 적용이 서로 충돌하는 결론을 야기한다는 사실에서 비롯된다. 그리고 이러한 문제는 행위자가 그의 예측을 절대 신봉하는 예측인이 있는 한, O 상황에서도 발생하는 것으로 보인다.

O 상황에서도 행위자가 그의 예측을 절대 신봉하는 예측인이 있다고 하자. 그리고 그 예측인이 N 상황에서와 같은 예측을 했다고 하자. 물론 O 상황에 있는 행위자가 그 예측인의 예측을 절대 신봉하는 일이 벌어진다는 것은 그럴듯하지 않다. 그러나 만일 그 행위자가 그러한 불합리한 믿음을 가지고 있다면 그는 프리스트와 같은 딜레마 상황에 봉착할 것이다. 왜냐하면 한편으로 이병욱의 원리 R*나 프리스트의 원리 R을 이용하여 두 상자를 모두 여는 행위를 선택하는 것이 합리적이라고 판단할 것이나 다른 한편으로 기대 효용 극대화의 원리 U에 의해 $CEU(a_1)=u$(100만원)과 $CEU(a_2)=u$(1,000원)을 비교하여 a_1, 즉 A 상자 하나만을 여는 것이 합리적이라고 판단할 것이기 때문이다. 같은 이유로 S 상황에

서의 슐레징거의 논증에서 전제 2도 명백하게 참이 된다고 말할 수 없다. 왜냐하면 그 상황에서도 조언자가 지배의 원리와 기대 효용 극대화의 원리의 적용 간에 딜레마 상황에 빠질 수 있기 때문이다.

O 상황에서 행위자가 N 상황에서와 같은 예측을 한 예측인을 완전히 신뢰한다는 것은 있을 법한 일이 아니다. 왜냐하면 그러한 믿음을 갖는다면 곧 A 상자가 비어 있는 것을 보아 알면서도 A 상자를 열 경우 100만원을 얻을 것이라는 믿음을 가지게 될 것이기 때문이다. 그러나 행위자가 그처럼 신뢰하는 예측인이 존재하지 않는다는 것은 O 상황에서 얻은 결론이 N 상황에서도 성립한다는 것을 보장할 수 없을 만큼 두 상황의 유사성이 상실된다는 것을 의미한다. 반대로 그러한 예측인이 존재한다면 두 상황의 유사성은 유지되지만 N 상황에서와 마찬가지로 O 상황에서도 합리적 딜레마가 야기되는 것을 막을 수 없다. 슐레징거는 S 상황에서의 조언자에 있어, 그리고 이병욱은 O 상황에서의 행위자의 입장에서 오직 지배의 원리만이 적용되어야 하며 따라서 딜레마 상황은 해소된다고 보고 있다. 그리고 그러한 상황과 뉴컴의 문제 상황의 유사성 때문에 N 상황에서도 진정으로 합리적인 것은 지배의 원리를 적용하여 얻어진 결과라는 결론을 내리고 있다. 그러나 뉴컴의 문제 상황은 물론이지만 S 상황이나 O 상황에서 왜 기대 효용 극대화의 원리가 배제되어야 하는지에 대한 보다 설득력 있는 설명이 제공되지 않는 그러한 생각은 완전히 정당화하기 어려울 것이다.

4

행위자가 a_1, …, a_n을 선택했을 때 각각 실제로 o_1, …, o_n과 같은 결과가 야기될 때, 그 가운데 최대인 $u(o_i)$에 해당하는 a_i를 선

택하는 게 합리적이라는 것이 지배의 원리이다. 이 지배의 원리가 전제하는 것은 각 o_i가 행위자가 a_i를 선택한 인과적인 결과로 야기되었다는 것이다. 즉 행위자가 a_i를 선택한 사건과 그가 o_i를 얻은 사건 사이에는 인과적인 관계가 성립한다는 것이다. 그렇기 때문에 실제 a_i를 함으로써 o_i를 얻었다고 하더라도 그것과 다른 행위 예를 들어 a_k를 선택했다면 o_k를 얻게 되었을 것이라는 반사실적 명제가 성립한다.[15] 따라서 $u(o_i)$가 최대라면 이미 a_i를 선택한 상태에서도 만일 그 행위 이외에 다른 행위를 선택했다면 그보다 적은 효용을 얻었을 것이라는 이야기가 성립한다. 이러한 의미에서 a_i를 선택함으로써 얻게 되는 효용은 그것의 기회비용보다 크다. 왜냐하면 기회비용이란 그 행위 이외에 다른 행위를 선택했다면 얻을 수 있는 최대의 효용이라고 할 수 있기 때문이다. 그러나 이러한 이야기가 기대 효용에 대해서는 성립하지 않는다.

이 세계의 실상이 s_1, …, s_m 가운데 어떤 것인지가 확실하지 않은 상태에서는 지배의 원리를 적용하지 못할 수 있다. 왜냐하면 만일 이 세계가 s_k인 상태라면 a_i가 지배적인 행위가 되는 반면에 s_l인 상태라면 그와는 다른 a_j가 지배적인 행위가 되는 경우도 있을 수 있기 때문이다. 이러한 상황에서 각 행위 a_i에 대한 상태 s_k의 조건적 확률을 안다면, 각 상태에서 각각의 행위를 함으로써 인과되는 결과를 안다면 각 행위에 대한 조건적 기대 효용을 계산할 수 있다. 그러나 기대 효용은 사전에 그 행위에 대해 행위자가 합리적으로 부여할 수 있는 효용이지 그 행위를 함으로써 인과적으로 얻게 되는 효용은 아니다. 앞서 2절에서의 프리스트의 예에서처

15) 이병욱의 t함수에 대해서도 같은 말을 할 수 있을 것이다. 즉 $t(a)=x$는 행위자가 a를 선택한 인과적 결과로 각각 x만큼의 효용을 얻는다는 것이다. 따라서 $t(a)=x$는 실제로 행위자가 a를 선택하지 않은 상황에서도 만일 그가 a를 선택했다면 x만큼의 효용을 얻었을 것이라는 반사실적 명제를 뒷받침한다.

럼 $Pr(s_1/a_1)=Pr(s_2/a_2)=1$, $Pr(s_2/a_1)=Pr(s_1/a_2)=0$이고 o_{11}=100만원, o_{21}=100만+1,000원, o_{12}=0원, o_{22}=1,000원이라고 하면 $CEU(a_1)=u(o_{11})=u$(100만원)이고 $CEU(a_2)=u(o_{22})=u$(1,000원)이다. 그러나 $CEU(a_1)=u(o_{11})$, $CEU(a_2)=u(o_{22})$라는 것이 행위자가 a_1을 혹은 a_2를 선택한 인과적인 결과로 $u(o_{11})$ 혹은 $u(o_{22})$ 만큼의 효용을 얻게 될 것이라는 것을 의미하지는 않는다. 그러므로 그 두 관계는 행위자가 실제로 a_1이나 a_2를 선택하지 않은 상황에서 그가 만일 a_1을 혹은 a_2를 선택했다면 $u(o_{11})$ 혹은 $u(o_{22})$ 만큼의 효용을 얻었을 것이라는 반사실적 명제를 뒷받침하지도 않는다.

지배의 원리는 어떤 다른 행위를 했더라도 a를 선택했을 때 얻는 효용보다도 더 많은 효용을 얻지 못했을 것이라는 명제가 성립할 경우 a를 선택하는 것이 합리적이라는 직관을 반영하고 있다. 이러한 지배의 원리는 실제로 행위자에게 열려 있는 각 행위를 선택했을 때 인과적으로 어떤 결과가 야기되는지가 알려져 있을 경우 적용가능하다. 그러나 1절의 <표 1>의 상황에서처럼 이 세계가 놓일 수 있는 각 상태에서 각 행위를 선택했을 때 인과적으로 야기되는 결과는 알 수 있다고 하더라도 각 상태마다 최대의 효용을 얻게 되는 행위가 일치하지 않을 경우 지배의 원리는 적용될 수 없다. 이 경우에는 이 세계가 각 상태에 있을 확률 값에 의거하여 최대의 효용을 기대할 수 있는 선택을 하는 것이 합리적이며 기대효용 극대화의 원리는 이러한 취지를 반영하고 있다. 따라서 <표 1>의 상황에서 A 상자를 여는 행위의 기대 효용이 더 크기 때문에 그러한 선택을 했는데 실제로 이 세계가 실제로 s_2 상태에 있는 것으로 확인되면 그 선택은 실제로는 최대의 효용을 얻게 되는 선택이 아니었던 것으로 드러나는 셈이지만, 그러나 그럼에도 불구하고 A 상자를 여는 기대 효용이 크다면 그 선택이 불합리하다고 말할 수는 없다.

기대 효용 극대화의 원리는 이 세계가 실제로 어떤 상태에 있는지 불확실하여 어떤 선택이 최대 효용을 갖게 될지 알 수 없기 때문에 위험을 감수한 선택을 하여야 하는 상황에서 위험을 줄이기 위해 사용되는 원리이다. 그러나 이 세계가 어떤 상태에 있는지 불확실하다고 해도 이 세계가 어떤 상태에 있는 것으로 드러나건 간에 그러한 상태에서 어떤 특정한 행위 이외에 다른 행위를 선택하였을 경우 그보다 적은 효용을 얻을 것이라는 점만은 알 수 있을 정도로 행위의 인과적인 결과에 관한 지식을 행위자가 가지고 있을 경우 지배의 원리를 적용하는 것이 타당하다. 따라서 위의 <표 2>와 같은 상황에서는 a_2를 선택하는 것이 합리적이다. 이러한 결론은 각 행위를 했을 때 이 세계가 여러 상태에 있게 될 조건적 확률이 서로 달라 $CEU(a_1) > CEU(a_2)$인 것으로 드러나는 경우에도 달라질 수 없다고 생각된다. 특히 $Pr(s_1/a_1)=Pr(s_2/a_2)=1$, $Pr(s_2/a_1)=Pr(s_1/a_2)=0$이어서 $CEU(a_1)=u$(100만원), $CEU(a_2)=u$(1,000원)인 경우에도 그러하다. 왜냐하면 $CEU(a_1) > CEU(a_2)$이기 때문에 a_1을 선택한 결과로 실제로 얻게 되는 것은 100만원 혹은 u(100만원)이든가 0원 혹은 u(0원)이지만, 그러나 예를 들어 a_1을 선택한 결과로 100만원을 얻은 상황에서 만일 a_2를 선택했을 때 얻게 되었을 것이라고 말할 수 있는 것은 $CEU(a_2)=u$(1,000원)아니라 u(100만+1,000원)이기 때문이다.

프리스트의 뿔 2와 노직의 논증을 정식화한 위의 (2)는 a_1을 선택하면 그 인과적 결과로 100만원을 얻게 되며 반대로 a_2를 선택하면 1,000원을 얻게 될 것이라는 말처럼 들린다. 그러나 (1)에서와 달리 (2)에서의 화살표는 인과적 결과로 해석될 수 없다. 이 세계가 어떤 상태에 있건 그 상태에서 성립하는 인과 관계에 의하면 a_2를 선택할 경우 a_1을 선택했을 경우보다도 1,000원을 더 얻게 되어 있기 때문이다.

정규 주사위를 던져서 1의 눈이 나올 경우 100원을 얻고 그 밖의 경우에는 아무 것도 얻지 못한다고 하면 그 주사위를 한번 던졌을 때 얻을 것으로 기대되는 값은 100/6원이다. 그러므로 그 주사위를 n회 던졌을 경우, $n\times(100/6)$원을 얻을 것으로 기대된다. n이 클수록 이 값은 실제로 얻게 되는 결과와 일치하게 될 것이다. 뉴컴의 상황에서 $CEU(a_1)=x$, $CEU(a_2)=y$라면 이것은 그 행위자가 반복적으로 뉴컴 상황에 처해서 계속적으로 n번 a_1을[a_2를] 선택할 경우 nx[ny] 만큼의 효용을 얻으리라는 것을 의미한다. 따라서 뉴컴의 상황에서와 같은 선택을 장기적으로 반복할 경우 $CEU(a_1)>CEU(a_2)$이라면 a_1을 선택하는 것이 유리하다고 생각될 수도 있다. 그러나 이 경우에도 a_1을 n번 선택해서 실제로 얻은 결과가 X원이었다면 만약 a_1 대신에 a_2를 선택했더라면 $X+n\times1{,}000$원을 얻을 수 있었다고 말이 성립할 것이다.

뉴컴의 문제와 같은 상황에서 하나의 상자만을 여는 것의 기대효용이 더 크기 때문에 사전에 그러한 선택에 가치를 부여할 수 있다. 그러나 그 선택에 대해 사전에 부여한 가치가 얼마이건 간에 실제로 그 행위를 하였을 경우, 그것이 아닌 다른 대안적 행위, 즉 두 상자를 모두 여는 행위를 선택하였더라면 하나의 상자를 엶으로써 실제로 얻은 가치보다 더 큰 가치를 얻었을 것이라는 결론은 여전히 성립한다. 어떤 행위의 기대 효용이 최대가 된다고 하더라도 그것을 선택한 경우 다른 대안적인 행위를 하였더라면 더 큰 효용을 얻을 수 있었으리라는 의미에서 그 행위를 실제로 선택함으로써 얻는 효용이 기회비용보다 적다는 사실이 알려져 있다면 그 행위를 선택하는 것은 합리적이지 않다고 해야 할 것이다. 이러한 이유에서 우리는 뉴컴의 선택 상황에서 두 개의 상자를 모두 여는 선택이 합리적이라고 생각해야 한다.

참고문헌

Bar-Hillel, M. & Margalit, A.(1972), "Newcomb's Paradox Revisited," *The British Journal for the Philosophy of Science* 23, 295-304쪽.

Bendit, T. M. & Ross, D. J.(1976), "Newcomb's 'Paradox'," *The British Journal for the Philosophy of Science* 27, 161-4쪽.

Burgess, S.(2004), "The Newcomb Problem: An Unquatlified Resolution," *Synthese* 138, 261-87.

Campbell, R. & Sowden, L.(eds.)(1985), *Paradox of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*, The University of British Columbia Press, 1985.

Eells, E.(1982), *Rational Decision and Causality*, Cambridge University Press, 1982, 우정규 옮김, 『합리적 결단과 인과성』, 서광사, 서울, 1994.

Gibbard, A. & Harper, W. L.(1978), "Counterfactuals and Two Kinds of Expected Utility," 후커(C. A. Hooker) 등이 편집한 *Foundations and Applications of Decision Theory*, Vol. I, Reidel, Dordrecht, 125-62쪽에 게재, Campbell & Sowden(1985), 133-58쪽에 전재.

Jeffrey, R. C.(1983), *The Logic of Decision*, 2판, The University of Chicago Press, Chicago and London, 이좌용 옮김, 『결단의 논리』, 성균관 대학교 출판부, 서울, 1998.

Ledwig, M.(2000), *Newcomb's Problem*, University of Constance 박사 학위 논문.

Nozick. R.(1969), "Newcomb's Problem and Two Principles of Choice," 레셔(N. Rescher) 등이 편집한 *Essays in Honor of Carl G. Hempel*, Reidel, Dordrecht, 114-46쪽에 게재, Campbell & Sowden[1985], 107-33쪽에 전재.

Priest, G.(2002), "Rational Dilemmas," *Analysis* 62.1, 11-16쪽.

Schlesinger, G.(1974), "The Unpredictability of Free Choice," *The British Journal for the Philosophy of Science* 25, 209-21쪽.

Yi, B.(2003), "Newcomb's Paradox and Priest's Principle of Rational Decision," *Analysis* 63.3, 237-42쪽.

중앙대학교 철학과
E-mail: leejk@cau.ac.kr