잠자는 미녀의 역설

송 하석

1. 들어가는 말

잠자는 미녀(이후 SB)의 역설을 간단히 소개하면 다음과 같다. 일요일 저녁 SB를 재운다. 월요일 아침에 SB를 깨우고, 잠시 후 기억을 지우는 약을 먹이고 다시 재운다. 그리고 정상적인 동전을 던져서 앞면이 나오면 화요일에는 깨우지 않고, 뒷면이 나오면 화요일 아침에 다시 깨운다. SB는 이러한 실험내용을 알고 있고, 그는 매우 탁월한 확률 계산가 이다. 이런 상황에서 월요일에 SB가 깨워졌을 때, 그가 동전을 던져서 앞면이 나올 것이라고 믿을 믿음의 정도는 얼마인가?

이에 대해서 1/2이라는 주장과 3/1이라는 주장이 그럴 듯한 근거를 가지고 제시되고 있다. 1/2이라고 주장하는 대표적인 철학자는 루이스(D. Lewis), 미첨(C. Meacham), 화이트(R. White), 그리고 프란체스치(P. Francheschi)이다. 반면에 1/3이라고 주장하는 사람들은 또 둘로 나눌 수 있는데, 엘가(A. Elga), 먼튼(B. Monton) 등은 새로운 증거가 없지만 SB가 실험을 시작하기 전과 월요일에 깨워졌을 때 동전이 앞면일 나올 확률에 대한 믿음의 정도가 변하여 1/3이라고 주장하고, 도르(C. Dorr), 와인트로브(R. Weintraub), 호간(T. Hogan) 등은 월요일에 깨워짐은 확률에 대한 믿음의 정도를 변하게 할 새로운 정보이기 때문에 1/3이라고 주장한다.

이 글은 SB의 문제에 대한 올바른 대답은 1/3이라고 주장하고, 왜 루이스 등이 주장하는 1/2이 옳지 않는지 논증할 것이다. 특히 최근 프란체스치가 1/2을 옹호하면서 제시한 입장에 대해서 비판하고 1/3의 입장을 옹호할 것이다. 이를 위해서 다음 절에서 SB 문제에 대한 논의를 촉발시킨 엘가와 루이스의 논변을 간단히 살펴보고, 그 차이에 대해서 설명할 것이다. 그리고 3절에서는 프란체스치가 루이스의 1/2의 입장을 옹호하기 위해서 제시한 논변을 소개하고 그에 대하여 비판할 것이다.

2. 엘가와 루이스의 논변

먼저 엘가의 1/3 옹호 논변을 살펴보자. 그는 SB가 월요일에 깨워졌을 때 그가 있을 수 있는 경우를 다음 세 가지로 구별한다.

H1: 동전이 앞면이 나오고 월요일에 깨워진 경우.
T1: 동전이 뒷면이 나오고 월요일에 깨워진 경우.
T2: 동전이 뒷면이 나오고 화요일에 깨워진 경우.

엘가의 논변은 두 단계로 이루어지는데, 첫 번째 단계는 SB가 T1에 대해서 갖는 믿음의 정도와 T2에 대해서 갖는 믿음의 정도가 같음을 증명하는 단계이고, 두 번째 단계는 SB가 H1에 대해서 갖는 믿음의 정도와 T1에 대해서 갖는 믿음의 정도가 같다는 것을 증명하는

단계이다. 다시 말해서 엘가의 논변은 다음을 증명하는 두 과정으로 이루어진다.

1) P(T1) = P(T2)
2) P(H1) = P(T1)

SB가 동전 던지기 결과가 뒷면임을 안다면, 그는 자신이 T1에 있거나 T2에 있음을 알 것이다. 이 경우 T1에 있음과 T2에 있음은 주관적으로 아무런 차이가 없다. 즉 SB는 자기가 T1에 있든 T2에 있든 정확하게 동일한 명제를 참이라고 생각할 것이다. 그러므로 다음이 성립한다.

$P(T1 | T1 \vee T2) = P(T2 | T1 \vee T2)$

따라서 다음이 성립한다.

$P(T1) = P(T2)$

또 SB가 오늘이 월요일임을 안다면, 그는 자신이 H1의 상태에 있거나 T1의 상태에 있음을 알 것이다. 이 경우 그가 H1의 상태에 있음을 믿을 믿음의 정도는 정상적인 동전을 던져서 앞면이 나올 것이라고 믿을 믿음의 정도와 같다. 즉

$P(H1 | H1 \vee T1) = P(H)$

그런데 P(H) = 1/2이므로,

$P(H1 | H1 \vee T1) = 1/2$이다.

이로부터 다음을 증명할 수 있다.

$P(H1) = P(T1)$

따라서 P(H1) = P(T1) = P(T2)이고, P(H1) = 1/3이다.

엘가는 이상의 논변을 통해서 SB가 실험에 참여하기 전에 동전이 앞면이 나올 확률은 1/2이라고 믿지만, 그가 월요일에 깨워질 때, 그 믿음이 변해서 1/3이 된다고 주장한다. 엘가는 SB가 월요일에 깨워졌을 때 그의 믿음이 변하는 이유는 그에게 새로운 정보가 주어졌기 때문이 아니라 자신의 시간적 위치(temporal location)가 동전이 앞면임이라는 사건(이하 H)의 참과 무관한 것으로 여겨지는 상황에서 자신의 시간적 위치가 그 사건의 참과 관련된 것으로 여겨지는 상황으로 변했기 때문이라고 설명한다. 엘가에 따르면, 새로운 정보란 정보를 받아들이는 주체가 지금까지 가진 증거에 의해서 배제하지 않았던 가능세계를 배제하는 증거이다. 그리고 SB는 월요일에 깨워질 것이라는 사실에 대해서 이미 알고 있었고 따라서 그가 깨워졌다는 사실은 지금까지 배제하지 않았던 가능세계를 배제할 만한 새로운 증거가 아니라는 것이다. 그러나 SB가 깨워짐은 자신의 시간적 위치가 어떤 명제의 참과

관련된 것으로 여기지 않다가 새롭게 그 명제의 참과 관련된 것으로 여기게 하는 사건이다. 엘가에 따르면, 행위자가 자신의 시간적 위치를 어떤 명제의 참과 관련된 것으로 간주한다는 것은 그 행위자의 믿음이 자신이 어떤 시각 t에 있음과 양립가능하고, 자신이 t에 있다는 조건 하에서 그 명제에 대한 자신의 믿음의 정도가 그러한 조건이 없는 경우와 다른 그러한 시각 t가 있다는 말이다. 다시 말해서 SB가 월요일에 깨워졌을 때, 그러나 그는 그 날이 월요일인지 화요일인지 알 수 없을 때, SB가 실험이 시작되기 전에 H의 참에 대해서 가졌던 믿음과 다른 믿음을 갖게 되는데, 그 이유는 SB가 새로운 정보를 갖게 되어서가 아니라 자신의 시간적 위치가 H의 참과 무관한 것으로 여겨지는 상황에서 자신의 시간적 위치가 H의 참과 관련된 것으로 여겨지는 상황으로 변했기 때문이다.[1)]

이에 대해서 루이스는 엘가의 첫 번째 논변을 받아들이지만, 두 번째 논변이 옳지 않다고 주장한다. 그는 다음과 같은 몇 가지 확률 함수를 제시한다.

P: 월요일에 깨워진 직후 SB가 갖는 믿음의 정도를 나타내는 함수

P+: 월요일 깨워져서 오늘이 월요일임을 알려준 후, SB가 갖는 믿음의 정도를 나타내는 함수

P-: 실험에 관한 설명을 듣고 재워지기 전에 SB가 갖는 믿음의 정도를 나타내는 함수

여기서 루이스는 엘가의 처음 논변의 주장, P(T1) = P(T2)을 받아들인다. 그리고 다음과 같은 엘가도 동의하는 주장을 제시한다.

$P-(H) = P-(T) = 1/2$

$P+(H) = P(H|H1 \vee T1)$

$P(H|H1 \vee T1) = P(H1) / P(H1) + P(T1)$

$P(H) = P(H1)$

그런데 루이스는 월요일에 SB가 깨워졌다는 사실은 믿음을 바꿀만한 새로운 정보가 아니라는 점에 주목하여 다음을 주장한다.

$P-(H) = P(H)$

즉 SB가 실험을 하기 전에 정상적인 동전을 던져서 앞면이 나올 것이라고 믿을 믿음의 정도와 월요일에 깨워졌을 때 정상적인 동전을 던져서 앞면이 나올 것이라고 믿음 믿음의 정도가 같다는 것이다. 그런데 P(H) = 1/2이고, 따라서 다음이 성립한다.

$P(H1) = 1/2$

루이스와 엘가의 가장 근본적인 차이는 P(H) = 1/2 인가, P+(H)= 1/2인가에 있다. 엘가는 P+(H)가 1/2이라고 주장하고, 반면에 루이스는 P(H)가 1/2이라고 주장한다. P+는 SB가 지금이 월요일임을 아는 경우의 확률함수를 나타낸다. 그런데 SB가 지금이 월요일임을 아는 경우에, 아직 동전 던지기가 이루어지지 않았다고 믿는다면 동전 던지기는 미래의

1) Elga (2000), p. 145.

사건이므로 당연히 동전의 앞면이 나올 확률은 1/2이라고 믿을 것이고, 이미 동전 던지기가 이루어졌다고 해도, 자신이 지금 월요일에 깨워졌음을 아는 경우 자신이 월요일에 깨워지는 것은 동전 던지기의 결과와 상관이 없다는 것을 알고 있기 때문에 여전히 H에 대한 그의 믿음의 정도는 1/2이라는 것이 엘가의 설명이다. 그러나 루이스는 SB가 자신이 깨워졌다는 것은 동전 던지기에 대한 H에 대한 그의 원래의 믿음에 아무런 영향을 주지 못하기 때문에 P-(H)와 P(H)는 동일하다고 주장한다.

누구의 주장이 더 설득력 있는 살펴보기 위해서, SB가 월요일에 깨워져서 그가 그날이 월요일임을 알 때, 그가 H에 대해서 갖는 믿음의 정도에 대해서 생각해보자. 이에 대해서 루이스는 다음과 같이 주장한다.

$$\begin{aligned} P+(H) &= P(H \mid H1 \vee T1) \\ &= P(H1) \,/\, P(H1) + P(T1) \\ &= 2/3 \end{aligned}$$

루이스는 SB가 월요일에 깨워졌고, 그날이 월요일임을 아는 사실은 미래에 일어날 동전 던지기에서 앞면이 나올 확률에 대한 그의 믿음을 바꾸게 하는 정보임을 설명해야 한다. 루이스는 이에 대해서 "나는 그 원리[미래의 사건에 대한 확률은 그 사건이 발생할 가능성과 같다는 원리]는 단서를 필요로 한다"고 말하면서 그런데 그 단서는 "우리가 [P-(H)=1/2을] 얻기 위해서 그 원리를 사용할 때는 만족하지만, 엘가가 [P+(H)=1/2을] 얻기 위해서 그 원리를 사용할 때믄 만족하지 않는다"고 말한다.[2] 이어서 그는 "[SB]가 월요일에 깨워져 있는 동안 그 날이 월요일임을 듣게 될 때, 그는 미래에 관한 증거, 자신이 지금 T2의 경우에 있지 않다는 증거를 얻게 되는 셈이고," 그것은 그날이 월요일이라고 듣기 전에는 갖지 않은 새로운 증거라고 주장한다. 그리고 루이스는 엘가도 다음이 성립한다는 것을 인정한다고 주장한다.

$$P+(H) = P(H) + 1/6$$

이로부터 루이스는 SB가 월요일에 깨워져서 그 날이 월요일임을 듣게 되는 것은 그의 믿음을 바꾸게 하는 증거라고 말한다. 그런데 루이스가 설명해야 하는 것은 SB가 월요일에 깨워졌을 그 날이 월요일임을 모르는 상태에서 H에 대해서 갖는 믿음의 정도와 그 날이 월요일임을 알게 되었을 때 그가 H에 대해서 갖는 믿음의 정도가 변한다는 것이 아다. 그가 설명해야 하는 것은 SB가 월요일에 깨워져서 그 날이 월요일임을 들음으로써 미래의 동전 던지기에서 앞면이 나올 확률을 1/2이라는 일반적인 믿음을 버리고 2/3이라는 믿음을 갖게 되는 이유이다. 다시 말해서 그는 P-(H)=1/2이지만 P+(H)=2/3인 이유가 무엇인지를 설명해야지, P+(H)에 대한 믿음의 정도와 P(H)에 대한 믿음의 정도가 다른 이유를 설명해야 하는 것은 아니다. 왜냐하면 루이스도 지적하듯이 엘가도 후자는 받아들이기 때문이다. 그런데 위에서 루이스는 P+(H)에 대한 믿음의 정도와 P(H)에 대한 믿음의 정도가 다른 이유를 설명하고 있을 뿐, P-(H)와 P+(H)에 대한 믿음의 정도가 다른 이유를 설명하고 있지는 않다.

---

2) Lewis (2001), p. 175.

3. 프란체스치의 3분의 1 주장에 대한 비판과 응답

프란체스치는 "잠자는 미녀와 세계 환원의 문제"라는 논문에서 SB 문제의 옳은 답이 1/2이고, 왜 1/3의 주장이 옳지 않은지를 설명한다. 이를 위해서 그는 다음과 같은 경우를 생각해보도록 제안한다.

정상적인 동전을 던져서 앞면이 나오면 빨간색 공 하나를 항아리에 넣고, 뒷면이 나오면 빨간색 공 하나와 초록색 공 하나를 항아리에 넣는다. 이렇게 만들어진 항아리에서 공을 꺼냈을 경우 빨간색 공을 꺼낼 확률은 얼마인가?

이에 대해서 다음과 같은 대답이 가능할 것이다.

$$P(R) = (1/2 \times 1) + (1/2 \times 1/2) = 3/4$$

즉, 만약 정상적인 동전을 던져서 앞면이 나온다면, 항아리에는 빨간색 공만 하나 들어갈 것이므로 빨간색 공을 꺼낼 확률은 1이고, 뒷면이 나온다면 항아리에 빨간색 공과 초록색 공이 하나씩 들어갈 것이므로 빨간색 공을 꺼낼 확률은 1/2이다. 따라서 이 항아리에서 빨간색 공을 꺼낼 확률은 위와 같은 계산에 의해서 3/4이다. 그러나 다른 대답도 가능해 보인다. 위와 같은 시행을 n번 반복해서 항아리를 채웠다고 하자. 그러면 항아리에 빨간색 공의 개수와 초록색 공의 개수는 다음과 같이 될 것이다.

$$N(R) \fallingdotseq (1/2 \times 1 \times n) + (1/2 \times 1 \times n) = n$$
$$N(G) \fallingdotseq (1/2 \times 0 \times n) + (1/2 \times 1 \times n) = n/2$$

그러므로 항아리 안의 전체 공의 개수는 3n/2이고, 그 중 n개가 빨간색 공이므로, 빨간색 공을 꺼낼 확률은 2/3이다. 즉,

$$P(R) = 2/3$$

프란체스치는 이 문제에 대한 옳은 대답은 3/4이지 2/3가 아니라고 주장한다. 그는 두 번째 풀이가 옳지 않은 이유를 하나의 독립적 사건과 하나의 사건의 일부분일 뿐인 것 사이의 차이를 파악하지 못했기 때문이라고 말한다. 즉 동전을 던져서 앞면이 나오고 항아리에 빨간색 공을 하나 넣은 것은 하나의 독립적인 사건이고 동전을 던져서 뒷면이 나오고 항아리에 빨간색 공과 초록색 공을 넣는 것도 하나의 독립적 사건이다. 다시 말해서 동전의 뒷면이 나와서 항아리에 빨간색 공을 넣는 것과 동전의 뒷면이 나와서 초록색 공을 넣은 것은 하나의 사건의 일부분을 구성하는 것이지 두 개의 별도의 사건이 아니다. 이것은 분리할 수 없는(indissociable) 사건의 일부이다. 따라서 빈도를 계산할 때, 이 둘을 따로 분리가능한 별도의 사건인 것처럼 생각해서는 안 되는데, 두 번째 풀이는 이 점에서 잘못을 범하고 있다는 것이다.[3]

3) Francheschi (2005), p. 2.

프란체스치는 위의 설명을 적용하여 엘가의 논변을 비판한다. T1과 T2는 분리할 수 없는 사건의 일부, 즉 하나의 사건을 구성하는 일부분인 반면, H1은 별도의 하나의 사건이다. 그런 점에서 H1과 T1(또는 T2)를 같은 유형의 사건으로 간주하여 빈도를 계산하는 것은 잘못이다. 그러므로 P(H1) = P(T1∨T2) = 1/2이라는 것이 프란체스치의 주장이다.

그렇다면 프란체스치가 제시한 문제에서 두 번째 풀이는 과연 틀린 것인가? 동전 던지기를 한 번 할 경우 프란체스치의 주장처럼 P(R)의 값은 3/4이다. 그러나 두 번 던질 경우는, 세 번 던질 경우는, 그리고 n 번 던질 경우는 어떤가?

두 번 던질 경우: P(R) = 17/24
세 번 던질 경우: P(R) = 101/160
n 번 던질 경우: $P(R) = (\frac{1}{2})^n\{{}_nC_0\frac{n}{n} + {}_nC_1\frac{n}{n+1} + \ldots + {}_nC_k\frac{n}{n+k} + \ldots + {}_nC_n\frac{n}{n+n}\}$

결국 이러한 시행이 무한히 반복될 경우, P(R)은 2/3에 수렴한다. 그러므로 프란체스치는 자신이 제시한 문제에 대해서 2/3이라고 답하는 것이 옳지 않고, 옳은 답은 3/4이라고 하기 위해서는 그의 물음이 보다 정확해져야 한다. 다시 말해서 동전을 한 번 던져서 위와 같은 방법으로 항아리를 채울 경우, 그 항아리에서 공을 꺼냈을 때 빨간색일 확률이라고 물어야 한다. 위의 계산이 주는 교훈은 동전을 던져서 앞면이 나오고 빨간색 공을 채우는 사건은 하나의 독립적인 사건이지만, 뒷면이 나와서 빨간색 공을 채우는 사건이나 뒷면이 나와서 초록색 공을 채우는 사건은 서로 분리할 수 없는 사건의 일부이고 따라서 빈도 계산을 할 때 달리 해야 한다는 그의 생각이 옳지 않다는 것이다. 항아리에서 공을 꺼내서 그 공이 빨간색 공일 확률을 계산하기 위해서는 여전히 앞에서 어떤 사건이 벌어졌던지 그 사건의 결과 항아리 안에 채워진 공의 전체 수에 대한 빨간색 공의 수라는 것은 여전히 동일하다.

이제 다른 경우를 살펴보자.

정상적인 동전을 던져서 앞면이 나오면 항아리에 두 개의 빨간색 공을 넣고, 뒷면이 나오면 한 개의 초록색 공을 넣는다. 이러한 작업을 n 번 계속했다. 이 항아리에서 공을 꺼냈을 때, 그 공이 초록색일 확률은 얼마인가?

이에 대한 일반적인 대답은 아마 1/3일 것이다. 왜냐하면 1/2의 확률로 빨간색 공이 두 개 채워지고, 또 1/2의 확률로 초록색 공이 채워지므로, 항아리 안에는 빨간색 공은 초록색 공의 2배가 들어있을 것이기 때문이다. 그러나 프란체스치의 위의 계산 방법을 따르면 그 대답은 1/2이어야 한다. 가능한 사건은 "동전을 던져서 앞면이 나오고 항아리에 두 개의 빨간색 공이 채워짐"과 "동전을 던져서 뒷면이 나오고 항아리에 한 개의 초록색 공이 채워짐"이라는 두 개의 사건밖에 없을 것이고 각각의 확률은 1/2이기 때문이다. 정확하게 말해서 이 문제도 동전 던지기의 횟수에 따라서 그 답은 달라진다. 즉,

동전 던지기를 한 번 한 경우: P(G) = 1/2
동전 던지기를 두 번 한 경우: P(G) = 5/12
동전 던지기를 n 번 할 경우:

$$P(G) = (\frac{1}{2})^n\{{}_nC_0\frac{0}{2n} + {}_nC_1\frac{1}{2n-1} + ... + {}_nC_k\frac{k}{2n-k} + ... + {}_nC_n\frac{n}{2n-n}\}$$

이러한 시행을 무한히 계속할 경우 P(G)는 1/3에 수렴한다. 이를 통해서 우리는 프란체스티가 자신이 제시한 문제에 대해서 그 답은 3/4이고 하나의 독립된 사건과 하나의 사건을 구성하는 분리할 수 없는 부분을 구별해야 한다는 그의 설명은 설득력이 없음을 알 수 있다. 오히려 그가 제시한 문제는 그 자체로 명확하지 않고 그 문제에 대한 답이 3/4이라고 하기 위해서는 동전 던지기를 한 번했다는 조건이 필요하다.

4. 결론

지금까지 논의를 통해서 SB 문제에 대한 보다 설득력 있는 답은 1/3이라고 주장했다. 그리고 SB 문제에 대해서 1/3이라고 답하는 것이 1/2이라고 답하는 것보다 직관적이다. 뉴의 문제를 약간 수정해 보자. 동전을 던져서 앞면이 나오면 월요일에만 깨우고 끝내지만, 동전을 던져서 뒷면이 나오면 월요일부터 일요일까지 7번 깨운다고 하자. 이렇게 수정되었을 경우, SB가 월요일에 깨워졌을 때, 그가 동전이 앞면이 나올 확률에 대한 믿음의 정도는 무엇일까?

여전히 루이스와 프란체스치는 1/2이라고 답할 것이다. 그러나 이는 직관적으로 받아들이기 어렵다. 수정된 문제에서 SB가 월요일에 깨워졌을 때 그가 그 날이 월요일임을 알지 못한 경우, 그는 그 날이 7일 중의 어느 하루라고 생각하는 것이 자연스럽다. 그리고 SB가 깨워지는 경우는 8가지 중 어느 하나라고 판단할 것이고, 각각의 가능성 사이에 차이가 있다고 생각해야 할 아무런 근거도 없다. 그런 점에서 수정된 설정에서 SB의 합리적인 믿음은 1/8이이어야 할 것이다.

끝으로 다음 경우를 생각해 보자.

세 개의 OX문제가 있다. 그 중 두 개의 답은 O이고, 다른 하나는 X이다. 물론 어떤 것이 X인지 알 수는 없다. 이 세 문제 중 첫 번째 문제의 답이 X일 확률은 얼마인가?[4)]

첫 번째 가능한 풀이는 모든 OX문제의 경우 그 답이 X일 확률은 1/2이고, 전체 문제에 대한 O와 X의 개수는 첫 번째 문제의 답이 X일 확률에 영향을 미치지 않는다. 따라서 답은 1/2이다. 그러나 이 풀이는 직관적으로 받아들이기 어렵다. 세 개의 문제 중 그 답이 X인 것은 하나이고 O인 것은 둘이라는 사실은 OX문제의 일반적인 확률값을 변경하게 할 중요한 정보이다. 따라서 첫 번째 문제의 답이 X일 확률은 1/3이다.

SB 문제에 대하여 그 답이 1/2이라고 주장하는 루이스 등은 김명석의 문제에 1/2이라고 답하는 경우와 유사하다.

4) 이것은 김명석 선생님과 대화 중에서 김명석 선생님께서 제시한 것이다.