【북심포지엄】 과학철학 21권 3호 (2018) pp. 213-234

자율성에 대한 개념적 분석:

인공지능의 자율성을 위하여† ‡

강 철*

인공지능에 관한 논의에서 핵심적인 철학적 쟁점들 중 하나는 자율성 개념일 것이다. ‘자율’주행차로 지칭되는 인공지능이 우리 주변에 이미 와있고, ‘자율적 인공지능’이란 말도 빈번히 쓰이고 있다. 필자는 ‘자율’이라는 말이 그 말의 본래적인 의미에 맞게 사용되어야 한다고 생각한다. 그렇게 하기 위해서, 자율성 개념을 이것을 구성하는 필수적인 요소들로 분석해 낼 필요가 있다. 특히, 자율성의 핵심적인 요소로 기속성(commitment)을 제시하고, ‘기속성으로서의 자율성’을 고찰한다. 요컨대, 이 논문은 언어적 차원에서 자율성에 대한 개념적 분석을 시도한다. 이러한 분석이 가지는 의의는, 인공지능을 만드는 엔지니어들이 공학적 차원에서, 말하자면 ‘자율이라는 말의 의미에 부합하는 자율성’을, 실현시키고자 할 때에 무엇을 충족시켜야 하는가라는 문제의식을 가지게 하는데 있다. 본 논문은 『인공지능의 존재론』에 실린 자율성에 관한 논문들을 검토한다.

【주요어】 자율성, 기속성, 인공지능, 개념적 분석

† 본 논문은 필자가 2018년 과학철학회 정기학술대회의 북심포지엄에 참여해서 『인공지능의 존재론』에 실린 논문들을 논평할 기회를 가지게 되었는데, 그때 가진 문제의식을 발전시킨 논문이다.

‡ 논문이 발전될 수 있도록 심사를 해 준 분들께 감사드린다. 제시하신 의견을 최대한 반영하려고 하였다.

* 서울시립대학교, justice1423@hanmail.net.

1. 서론

누가 자율적인 존재자인가? 지금까지 알려진, 자율적인 판단과 행위를 할 수 있는 존재자, 그 유일한 존재자는 우리 인간일 것이다. 그런데 '유일하다'는 데 이의를 제기하는 사람들도 많을 것 같다. 하지만 자율성의 '대표적인' 또는 '전형적인' 존재자가 인간이라는 것을 부정하는 사람은 드물 것 같다. 왜냐하면 자율성이란 무엇인지, 그 자율성에 대한 우리의 근본적이고 근원적인 이해는 우리 자신에 대한 일상적인 경험으로부터 자라나온 것일 수뿐이 없기 때문이다. 따라서 자율성이라는 인간적일 수뿐이 없는 개념을 분석하고자 할 때에, 우리는 자기 자신을 향해야 한다고 주장할 수 있다. 자기 자신의 경험을 향해야 하고, 그리고 그 경험의 녹아있는 언어를 논의해야 한다고 주장할 수 있다.

이러한 주장은 나름 합당해 보이다. 그런데 인간이 아닌 존재자의 자율성을 논의할 경우에는 어떻게 해야 하는가? 동물의 자율성은? 그리고 이글의 주제인 인공지능의 자율성은? 인간이 인간 자신을 경험하듯, 동물 자신이나 인공지능 자신을 경험할 수는 없는 것이다. 더군다나, 병들고, 늙고, 죽으며, 자손을 낳는 등등의 존재조건들을 공유하지 않는 존재자에게 자율성이란 것이 도대체 어떻게 경험될지를 우리는 알 수 없을 것 같다.

이와 같은 문제들에 직면할 때에, 개념(가령, 여기서는 '자율성'이라는 개념)의 사용과 관련해서 우리는 입장을 선택해야 할 것이다. 한 가지 입장은, 인간의 존재조건들을 기반으로 했던 개념들(다시 말해, 기존 개념들)의 의미 경계를 더 명료하게 하고 그 의미에 충실하게 그 개념들을 계속해서 사용하려는 입장이다. 예컨대 인간의 존재조건들을 기반으로 했던 그리고 인간 존재자에게 가능했던 자율성의 의미를 보존한 채, 그 자율성을 동물이나 인공지능이 실현할 수 있는지를 묻는 것이다. 중요한 점은 이 입장이 자율적 인공지능의 실현에 반대할 필요는 없는데, 인간의 존재조건에 기반한 자율성 개념에 충실한 인공지능을 설계하면 되기 때문이다. 또 다른 입장은, 기존 개념들을 상당히 수정하고 변용하여 동물이나 인공지능과 같은 다른 존재자들에게 확대

해서 적용하려는 입장이다.[1] 예컨대 자율성 개념을 자연화 또는 알고리듬화해서, 자율성을 물리적으로 또는 기능적으로 설명하고, 더 나아가 인공지능 로봇을 통해 구현하는 것이다. 실상 이 입장을 따를 경우 자율적 인공지능을 만들 때에, 인간의 존재조건을 기반으로 한 자율성 개념에 원리상 제한을 받을 필요는 없다. 인간의 존재조건을 기반으로 한 자율성만이 실재할 수 있는 유일한 자율성은 아니라고 주장할 수 있기 때문이다. 본 논문은 첫 번째 입장에 입각해서 논의를 전개할 것이다.

인공지능을 연구해왔던 국내의 과학철학자들뿐만 아니라 공학자, 윤리학자, 동양철학자들이 역량을 결집하여『인공지능의 존재론』[2]이라는 책을 올해 출간하였다. 인공지능의 철학적 문제점들을 다양한 관점에서 천착한 단행본으로서 학계의 담론을 활성화시킬 것으로 예상된다.[3] 그런데 책 제목에 '존재론'이라는 명칭이 붙어있기는 하지만, 필자가 보기에 다수의 논문들은 실상은 윤리학의 주제인 '자율성'을 인공지능과 직・간접적으로 관련시키면서 논의를 전개하고 있다. 인공지능에 관한 논의에서 자율성이 그만큼 논쟁적이고 핵심적인 문제라는 점을 반증하는 것이라고 할 것이다.

본 논문의 내용을 요약하자면, 2장 "자율성에 대한 개념적 분석"에서 필자는 자율성에 대한 개념적 분석을 시도한다. 본 장에서는 자율성이라는 말의 의미를 중심으로 자율성을 구성하는 필수적인 요소들이

---

1) 가령 신상규는 다음과 같이 주장한다. "AI 기술을 포함한 첨단 과학기술의 발전에 의해 추동되고 있는 지금의 변화가 우리 삶의 양식 혹은 문법을 뒤바꾸는 변화이므로, 우리의 일상적 세계관을 구성하는 근본 개념이나 판단들 또한 갱신될 필요가 있음을 주장한 바 있다. 이는 우리의 일상적 개념들에 녹아들어 있는 여러 습관적 의미를 해체하고, 새로운 삶의 양식에 맞추어 그 의미를 재구성하는 작업을 요구하는 것이다"(신상규 2017).

2) 이중원 외 (2018).

3)『인공지능 시대에 인간되기』한국과학철학회 2018년 정기학술대회, 일시: 2018년 7월 11-12일, 장소: 한국교원대학교 융합과학관, 주체: 한국연구재단 일반공동연구지원사업 <포스트휴먼 leo의 인공지능 철학>, 한국과학철학회, 주관: 한국과학철학회, 후원: NRF 한국연구재단.

무엇인지를 열거적으로 설명할 것이다. 보다 많은 구성요소들이 논의되어야 할 테지만, 지면관계상 필자가 보기에 중요하다고 생각되는 요소들에 한정해서 논의를 전개할 것이다. “기속성으로서의 자율성”에서는 ‘강요 없는 강제’를 뜻하는 기속(commitment)에 근거해서, 자율성을 구성하는 가장 핵심적인 개념인 기속에 근거해서 자율성을 해명할 것이다. 3장 “자율성 원칙에 대한 자율성 존중 원칙의 실천적 우선성”에서는 자율성이란 무엇인지를 논의하는 자율성 원칙(the principle of autonomy)에 대해서 자율적인 존재자로서의 상대방을 대우하는 방식에 관한 자율성 존중 원칙(the principle of respect for autonomy)의 실천적 우선성을 주장한다. 그리고 4장 “『인공지능의 존재론』에 실린 자율성에 관한 논문들”에서는 본 책의 논문들 세 편을 필자의 개념적 분석에 의거해서 검토한다.

2. 자율성에 대한 개념적 분석

자율성이란 무엇인가? 자율성 개념의 핵심적인 요소는 무엇인가? 자율성(自律性)이란 사전적 정의에 따르자면, “자기 스스로의 원칙에 따라 어떤 일을 하거나 자기 스스로 자신을 통제하여 절제하는 성질이나 특성”[4]을 말한다. 여기서 ‘통제한다’는 것은 자신의 어떤 욕구에 끌려 다니거나 굴복하거나 하지 않는다는 것을, 또는 자신의 습관에 지배를 받지 않는다는 것을 뜻할 것이다. 따라서 통제한다 함은, 끌려 다니거나 지배를 받거나 하지 않기 위한 심적이거나 행위적인 어떤 작용을 말할 것이다. 그리고 작용이라는 이 측면에서, 자율성이란 실천적 개념인 것이다.

그렇다면 자율성이란 순전히 실천적인 개념일 뿐인가? 다시 말해, 어떤 개체가, 그것이 동물이든 인공지능이든 간에, 인간의 자율성과 같은 그런 자율성을 가졌다고 인정받기 위해서는 스스로 자신을 통제하

4) 다음을 참조할 것. 『표준국어대사전』 https://dict.naver.com/.

는 특성을, 곧 실천적인 특성을 가지는 것으로 충분한가? 필자는 충분하지 않다고 주장한다. 왜냐하면 자신이 누구인지 모르면서 자신을 스스로 통제할 수는 없을 것이기 때문이다.[5]

자율적 행위를 하기 위해서는 자기 자신에 대해 과연 어느 수준에서 어떤 정도의 앎을 가져야 하는가? 가령, 의사를 결정할 능력이 있는 성인의 의식 수준에서 자신이 하려는 행위의 의미와 그 행위가 미칠 영향이나 인과관계에 대한 참된 앎, 그 정도까지의 앎을 요구하지는 않을 것 같다. 아무튼, 자기 자신에 대해 어느 수준의 어떤 정도의 앎을 가져야 하는지는 탐구해볼 가치가 있는 문제라고 본다. 그러나 분명한 점은 자신을 스스로 통제하는 것이 실천적으로 가능하기 위해서는 적어도 자신을 알아채는 정도의 능력은 있어야만 한다는 것이다. 그리고 자신을 알아채는 그 '앎'이란 인식적 개념이다. 따라서 자율성이란 첫째, 순전히 실천적인 개념이 아니라, **인식적 개념을 전제로 하는 실천적인 개념**인 것이다.

둘째, 자율성에 반대되는 개념은 타율성이라고 말해진다. 타율성이란 사전적 정의에 따르자면, "자신의 의지와 관계없이 정하여진 원칙이나 규율에 따라 움직이는 성질"[6]을 말한다. 즉, "자신의 의지와 관계없이 정해진 것들" 곧, "자기가 아닌 것들"로부터 통제나 간섭을 받는다는 의미이다. 따라서 이런 의미에서, 타율성이란 의존적 내지는 관계적 개념인 것이다. 그렇다면 자율성은 순전히 비의존적이거나 비관계적인 개념인가? 그렇지 않다. 왜냐하면 자신의 의지와 관계없이 정해진 것들로부터는 독립해 있지만, "스스로 자기를 통제한다"는 의미에서는 자신과 모종의 관계를 맺어야만 하기 때문이다. 따라서 자율성이란 둘째, 자기 자신과의 관계 속에서 성립한다는 의미에서 **자기관계적인 개념**인 것이다.

---

5) 한 방송국의 다큐멘터리에 의하자면, '아기는 언제부터 자신을 알아볼까'와 관련해서 보통은 20개월 이후가 되어야 자신을 알아본다고 한다. 또한 만 5세 정도 되어야 다른 사람의 생각을 해석할 수 있는 능력을 가지게 된다고 한다. EBS 아기성장보고서제작팀 (2009) 참조.

6) 다음을 참조할 것. 『표준국어대사전』 https://dict.naver.com/.

다른 한편, 자율성을 자신과 모종의 관계를 맺는 개념, 곧 자기관계적인 개념이라고 할 때에 그 관계의 성격은 무엇인가? 다시 말해, “스스로 자기를 통제한다”고 했을 때에 우리는 자기 자신과 어떻게 관계하고 있는 것인가? 무엇보다 먼저 떠오르는 생각은, 어떤 강박으로 인해서 어떤 행위를 한 것이라면 자율적으로 한 것은 아니라는 점이다. 가령 내가 길거리에 담배꽁초를 버리지 않는데, 그 까닭은 누군가 나를 주시하고 있다는 망상에 사로잡혀서 그렇게 한 것이라면, 버리지 않는 그 행위는 자율적 행위로 평가받지 않을 것이란 점이다. 따라서 어떤 결정이나 행위가 자율적이라는 평가를 받기 위해서는 그러한 것들을 하게 만드는 가령, 내부의 어떤 심리적인 강박이 없어야 할 것이다. 뿐만 아니라 외부의 어떤 물리적인 압력 때문에 어떤 행위를 한 것이라면 이 역시도 자율적으로 한 것은 아니라는 점이다. 가령, 쓰레기 무단투기 감시카메라가 설치되어 있기 때문에 쓰레기를 함부로 버리지 않는다면, 그 행위 역시도 자율적 행위라고 인정받지 못 할 것이란 점이다. 따라서 이러한 내적이거나 외적인 강요가 없다고 한다면, 나의 결정이나 행위는 자율적일 수 있을 것 같다.

그런데 정말 그러한가? 내적이거나 외적인 강요의 부재는 자율성을 보장하기 위한 단지 필요조건일 뿐이며, 충분조건은 아니라고 필자는 주장한다. 더 자세히 말해서, ‘강요는 없어야 한다. 그렇지만 강제는 있어야’ 그 결정이나 행위가 자율적일 수 있다는 것이다. 여기서 필자가 말하는바, ‘강제’란 ‘**자기강제**’이다. 자기강제란, 내가 ‘진술한 말’이나 ‘수행한 행위’나 ‘선언한 원칙’ 그 자체가 가지는 당위적인 효력을, 보다 포괄적으로 말해서 규범적인 효력을 일컫는다. 내가 진술한, 수행한, 선언한 것들이란 나에게 속하는 사실적인 것들이다. 이 사실적인 것들이 내 자신에 대해서 규범적 효력을 갖는다는 것을 전제해야 자율성이 가능한 것이다. 요컨대 ‘강요 없는 강제’는 첫째 자율성을 가능하게 하기 위한 전제조건이다. 바꿔 말해서, 내가 이미 행한 어떤 말이나 행위가, 어떤 것을 해야 한다(should)는 속성을 내재하지 않는다고 한다면, 그 말이나 행위는 자율성이라는 범주에 귀속되지 않는 것이다. 가령, 연극배우가 극중에서 상대방에게 돈을 갚기로 하는 어떤 약속을

했다고 해보자. 그 약속행위는 자율성의 범주에 귀속되는 행위가 애초부터 아니라고 본다. 왜냐하면 그 배우는 그 약속행위에 구속될 의사가 없으며 따라서 갚으려는 동기를 갖지 않기 때문이다. 따라서 극중에서 행한 그 행위는 약속행위로서의 자율적 행위는 아닌 것이다.

뿐만 아니라 '강요 없는 강제'는 둘째, 그 자체가 자율성을 구성하는 핵심요소이다. 이를 사실적인 것들의 규범적 효력(the normative force of the factual)이라고 부를 수 있다. 이 점이 시사하는 바는, 강요 없는 강제로서의 자율성을 인정한다면, 흔히 주장되는 사실-규범의 이분법(fact-value dichotomy)이나 존재로부터 당위를 도출할 수 없다는 소위 흄의 법칙(Hume's Law)은 부정된다는 것이다.

그런데 나의 사실적인 것들이 나에 대해 규범적인 효력을 가질 수 있는 까닭은 무엇인가? 물론, 내가 사실적으로 행한 나의 '모든' 말이나 행위나 원칙 등이 규범적 효력을 가지는 것은 아니다. 따라서 정확히 묻는다면, 나의 '어떤' 사실적인 것들은 내 자신과의 관계에서 왜 규범적인 효력을 가지는가? 그 까닭은 '믿게 할 이유들(reasons for belief)'이나 '행위를 하게 할 이유들(reasons for action)'에 근거를 해서, 즉 동기로서 인식적 이유들이나 실천적 이유들에 근거해서 자신에게 속하는 사실적인 것들을 내가 스스로 선택했기 때문이다.[7] 즉 나에게 속하는 사실적인 것들이 임의로, 무작위로, 우연적으로 그렇게 된 것이 아니라, 내가 이유들에 의거하는 방식으로 즉 정당화가 가능한 방식으로 선택을 했다는 그 사실 때문에 나에게 규범적인 효력을 가지는 것이다.[8]

---

7) 믿게 할 또는 행위를 하게 할 이유들이라는 인식적 또는 실천적 동기에 관해서는 Derek (2011) 참조. 또한 Star (ed.) (2018) 참조.

8) 한 가지 유의할 점은, 자신의 선택이 자신에게 규범적 효력을 가지기 위해서, 자신에 속하는 것들을 선택할 1) 그 당시에, 2) 자신의 인식적 또는 실천적 동기를 의식할 필요도, 명시적으로 표현할 필요도 없다는 점이다. 즉, 자신이 의식하거나 명시적으로 표현을 해야만 자신의 그 인식적 또는 실천적 이유들의 존재가 인정되는 것은 아니라는 점이다. 의식하지 못 했거나 외부적으로 명시하지 못 했다고 하더라도 인정될 수 있다는 점이다. 즉 "선택을 한 당시나 그 이후에라도, 만약 당신의 동기가 무엇이었냐는 물음

그런데 “정당화가 가능한 방식으로 선택을 했다”는 이 말 속에 자율성이 이미 전제되어 있는 것 아닌가? 다시 말해 자율성을 ‘강요 없는 강제’로 설명할 때에, 이 설명 속에 이미 자율성을 포함시키고 있는 것 아닌가? 따라서 자율성에 대한 개념적 정의를 ‘강요 없는 강제’에 의거해서 수행하려는 필자의 설명전략은 실패한 것 아닌가? 그렇지 않다고 필자는 주장한다. 실패한 것이 아니라는 점에 관해서는, “기속성으로서의 자율성”에서 최상의 설명에로의 추론(the Inference to the Best Explanation)을 통해서 상론이 필요해 보인다.

앞서 제시한 사전적 정의에 의거하자면, 자율성이란 “자기 스스로의 원칙에 따라 어떤 일을 하거나 자기 스스로 자신을 통제하여 절제하는 성질이나 특성”을 말한다. 여기서 “자기 스스로의 원칙에 따라 자기 자신을 통제한다”는 말은 다름 아니라, 자신의 사실적인 것들이 규범적 효력을 가진다는, 필자의 용어로는 ‘기속성(羈束性, commitment)’을 가진다는 것을 뜻한다. 앞서 논의했던 “강요 없는 강제”라는 특성을 필자는 ‘기속성’이라고 부르고자 한다.[9] 그렇다면 셋째, 자율성이란 자

을 받을 경우에 그 이유들의 존재를 인정받기에 충분한 대답을 당신이 제공할 수 있는 것만으로도, 자신의 선택이 규범적 효력을 갖기에 충분하다는 점이다.

9) “commitment”라는 용어는 영미 분석철학 분야에서 특히, 콰인의 ontological commitment라는 개념과 관련하여 다양하게 번역되고 있다. 가령, 존재론적 “개입”, “연루”, “언질” 등으로 번역되고 있다. 유명한 문장인 “현재 프랑스왕은 대머리이다”라는 문장이 상식을 가진 사람들이 믿기에 거짓이기 위해서는 프랑스왕이 존재해야 할 것이 요구되는 것처럼 보인다. 즉 그 말을 하는 순간, 우리는 현재 프랑스왕이 실재한다는 것에 존재론적으로 개입하거나 연루되는 것처럼 보인다. 윤리학적 맥락에서는 “헌신”이나 “전념”라는 용어가 사용되기도 한다. 그리고 경영학에서는 “몰입”이라는 용어로 번역되고 있다. 그런데 이러한 모든 번역어를 관통하는 추상적 관념은 “강요 없는 강제”라고 필자는 생각한다. 그리고 “강요 없는 강제”라는 관념을 사용하고 적용해도 되지만, 용어의 단순성과 편리성을 위해 “기속”이라는 말로 대체하고자 한다.

기속(羈束)이라는 말은 우리의 법학 분야에서 사용되는 용어인데, 가령, 하

기 자신과의 관계에서 강요 없는 강제를 뜻한다는 의미에서 **자기기속적인 개념**인 것이다.

넷째, 이제 자율성 개념이 어떻게 '진정한 자기다움'을 즉, 자기진정성(authenticity) 개념을 함의하는지에 대한 논의를 다음의 물음을 통해서 시작해 보자. 자신이 진술한 말이나 수행한 행위나 선언한 원칙 등이 어떻게 해서 우리에게 강요 없는 강제로 다가올 수 있는 것인가? 다시 말해 그것들에 우리가 자발적으로 구속, 즉 기속되는 까닭은 무엇인가? 그 까닭은 나의 그러한 사실적인 것들이 **나를 진정으로 표현**해 주고 있기 때문이다. 만약 그렇지 않다면, 그러한 것들에 자발적으로 구속될 까닭은 없는 것이다. 단지 심리적인 어떤 강박이나 외부의 어떤 위력으로 인해서 마지 못 해 그러한 말이나 행위나 원칙들에 구속될 뿐이다. 요컨대 내가 어떤 결정이나 행위를 자율적으로 한다고 함은 그 결정이나 행위를 지은 저자(author)가 바로 나 자신임을 천명하는 것, 즉 자기진정성을 표현하고 있는 것이다. 나의 내적인 욕구나 강박과 같은 개인에 속하는 것들(subpersonal)로 인해서 또는 외적인 위력이나 권위로 인해서 그러한 결정이나 행위가 야기된 것이 아니라, 실상은 나 자신(person)이 그것들을 만든 저자임을, 따라서 그것들이 내 것임을 밝히는 것이다. 자율성이란 진정한 자기를 표현한다는 의미에서 **자기진정성 개념**인 것이다.

자율성에 대한 개념적 분석을 통해서 자기진정성 개념이 도출되었고, 이 자기진정성으로부터 자기인정이라는 개념이 또한 도출된다. 나의 어떤 결정이나 행위가 자율적이라는 말은 자율적으로 내린 결정이나 행위가 나의 미래 행위를 규제하고 지도하는 규범적 효력을 가진다는 말이며, 강요 없는 강제로서 그렇게 한다는 말이다. 또한 그 말은

위법원의 판결이 상위법원의 판결에 기속된다고 말한다. 풀어서 말하자면, 법관은 판결에 있어서 양심에 따른 독립성과 재량을 가지고 있다. 이점에서 법관의 판결은 자발적인 것이다. 그러나 그럼에도 불구하고 상위법원의 판결이나 선례(판례)에 구속되는 것이다. 그리고 구속되는 것이 법관의 독립성을 해치는 것은 아니다. 이점에서 "자발적 구속"이라는 표현을 사용할 수도 있을 것인데, 이와 같은 현상을 지칭하는 말이 "기속"이다.

그 결정이나 행위에 대해 만약 누군가 의문을 제기한다면, 그 의문에 반응하거나 응답할 적임자가 자신임을 인정하는 것이다. 즉 자신의 것들에 대해서는 '자신이' 대답을 하는 것이 도덕적으로 정당하며(legitimacy), 그것들에 대해 '자신이' 도덕적인 권위(authority)를 가진다는 것을 말한다. 이 점에서 자율성이란 **자기인정 개념**인 것이다. 자율성으로서의 자기인정이 사회적 차원에서 벌어지는 쌍방 간의 인정투쟁을 성립시켜 주는 토대인 것이다.

지금까지 우리는 자율성에 대한 개념적 분석을 통해서, 자율성 개념을 구성하는 핵심요소로서 "인식적 개념을 전제로 하는 실천적 개념", "자기관계 개념", "자기기속 개념", "자기진정성 개념", "자기인정 개념"이라는 다섯 가지 요소를 살펴보았다. 이러한 개념들은 인공지능의 자율성을 구현하는 데 있어서 어떤 지도 원리로 작동할 수 있다고 본다. 이 이외에도 자율성에 속하는 개념들을 생각할 수 있을 것 같다. 이러한 개념들 모두를 충족시키는 인공지능도 생각해 볼 수 있고, 그 중 일부를 충족시키는 인공지능도 생각해 볼 수 있을 것이다. 요컨대 자율성에 대한 개념적 분석에 의거한 의미론적 논의는 다양한 종류의 자율성을 지닌 인공지능을 설계하는 데 활용될 수 있을 것이다.

3. 자율성 원칙에 대한 자율성 존중 원칙의 실천적 우선성

필자가 말하는 바, '자율성 원칙'이란 자율성이란 무엇인지, 그 핵심적인 요소들을 구성하거나 설명하는 원칙을 말한다. '자율성 존중의 원칙'이란 자율적인 존재자를 어떻게 대우해야 하는가, 적합한 대우의 방식은 무엇인가 등을 논의하는 원칙을 말한다.

이제, 자기진정성과 자기인정, 이 두 개념이 가지는 의의에 대해서 논의하고자 한다. 이 두 개념이 중요한 이유는 자율적인 어떤 대상을 우리가 과학기술을 통해 만들려고 할 때에, 고려해야 할 보다 더 핵심적인 사안은 '자율성이란 무엇인가?'라는 인식론적 문제가 아니라, '우

리가 만들려는 그 대상을 우리는 어떻게 대우해야 하는가?'라는 **실천적 문제**라는 데 있다. 자율성이란 그 본성이 규명되어야 할 단순히 인식적 개념이 아니라, 우리로 하여금 어떤 방식의 대우를 요구하는 실천적 개념인 것이다. 이 점을 이제 자율성이라는 가치에 대한 대우방식을 통해서 논의해 보고자 한다.

다양한 가치들은 그 가치의 본성에 맞추어 반응할 것을 우리에게 요구한다. '존중'과 '증진'이라는 대표적인 두 가지 방식을 검토해 보자.[10] 첫째, 자율성이라는 가치는 그 본질에 있어서 '존중(respecting)'의 방식으로 우리가 그 가치를 대우해야 할 것을 요구한다. 존중의 방식으로 대우한다는 것은 어떻게 대우해야 하는가를 결정할 주도권(initiative)이 대우를 하는 행위자(the agent)인 우리에게 있는 것이 아니라, 대우를 받는 피행위자(the patient)인 그에게 있다는 것이다 (피행위자는 사람일 수도 있고, 그 사람의 자율성과 같은 속성일 수도 있다). 그리고 대우에 대한 궁극적인 도덕적 정당성(legitimacy)은 대우의 피행위자를 통해서 확보된다. 이러한 측면에서 자율성이라는 개념은 쌍방 간의 작용(가령, 나와 너로 구성되는 '대화'와 같은 대면할 수 있는 방식) 속에 이루어지는 역동적인 윤리, 필자가 말하는 바 2인칭 윤리에 속하는 것이다.

예를 들어, 철수의 자율적인 결정이나 행위가 철수 자신을 진정으로 표현하고 있고(자기진정성), 그 결정이나 행위에 관한 도덕적 정당성과 권위가 철수 자신에게 있다(자기인정)고 해보자. 이제, 우리는 철수의 결정이나 행위를 어떻게 대우해야 하는가? 결론부터 말하자면, 존중의 방식으로 대우해야 한다는 것이다. 그리고 존중의 방식이 구체적으로 어떠해야 하는지는 '우리의 목적이나 이해관심'에 의해서가 아니라 '그 결정이나 행위를 통해서 철수가 이루고자 하는 목적이나 이해관심'에 의해서 (그 목적이나 이해관심이 합당하다는 전제 하에) 규정되어야 하는 것이다. 존중의 방식이란 행위자인 우리가 말하고 싶은 대로가 아니라, 피행위자로부터 들려오는 대로 규정되어야 하는 것이다. 그리고 대우방식의 궁극적인 도덕적 정당성은 말함이 아니라 들음에

10) 가치에 대한 다양한 반응방식에 대한 논의로는 Swanton (2005) 참조.

의해서 보장되는 것이다.

반면에 둘째, 가치에 대한 또 하나의 대우방식으로서 '증진'의 경우는 어떠한가? 가령, 많은 사람들의 '생명', 곧 '목숨'이 위급한 상황에 처해 있다고 해보자. 이때 우리는 더 많은 사람들을 구하려고 하며, 이 점에서 우리는 '생명'이라는 가치를 증진(promoting)의 방식으로 대우하고 있는 것이다. 다시 말해, 모든 인간 생명이 평등하다는 것을 전제로, 한 명을 구조하는 것보다는 열 명을 구조해야 하는 것이다. 다시 말해, 다다익선(多多益善)인 것이다. 왜냐하면 열 명을 구조하는 행위 속에, 한 명을 구조하는 행위가 이미 포함되어 있기 때문이다. 이 점에서 '생명'이란 같은 값(同價)의 다른 생명으로 대체될 수 있는 가치인 것이다.

우리의 문제로 돌아와서, 자율성을 존중하는 것과 자율성을 증진하는 것은 전혀 다른 사안이다. 우리가 어떤 자의 자율성을 존중한다고 한다면, 그의 자율성 자체를 내재적 가치로 인정하며, 생명과 달리 또 다른 자율성으로 대체될 수 없음을 보장하는 것이다. 다시 말해, 철수의 자율성의 고유성을 인정하고 대체될 수 없는 것으로 소중히 여기는 것이, 바로 그것을 존중하는 것이다. 다섯 개의 자율성을 증진하기 위해 한 개의 자율성을 침해하는 것을 자율성 존중은 허용하지 않는 것이다.

요약하자면, '자율성'이라는 가치에 대한 규범적으로 정당화되는 반응은 '존중'의 방식이다. 자율성 개념이 자기진정성과 자기인정이라는 이 두 개념을 본성적인 요소로 하는 한, **자율성은 자율성 존중을 함축한다**. 자율성의 본성 그 자체가 타인으로부터 또는 외부로부터의 인정과 존중을 요구하는 것이기 때문에 바로 이 점에서 자율성이란 **정치적 개념**인 것이다. 따라서 우리는 '어떤 존재자를 자율적인 존재자로 대우할 준비가 되어 있는지'를 진지하게 물어야 하는 것이다. 그런데 인공지능의 자율성에 관한 많은 논문들은, 무엇이 자율적인지를 규명하려는 인식론적인 작업에 치우쳐 있는 것은 아닌가라는 인상을 준다. 즉, 자율성의 의미를 분석하는 어떤 종류의 '자율성 원칙'을 먼저 제시한 후에, 그 의미에 의거해서 자율성 존중을 논의하는 '자율성 존중의

원칙'을 제시해야 한다고 생각하는 것 같다. 그러나 앞서 필자의 '자율성에 대한 개념적 분석'을 통해 드러난 자율성의 본성에 의하자면, 자율성이란 실상 정치적 개념이다. 자율성 자체가 행위자로서의 우리가 상대방에 대해 존중의 방식으로 반응할 것을 요구한다. 그렇다면 자율성에 대한 피행위자의 관념과 행위자인 우리의 관념이 다르다고 하더라도, 피행위자의 결정과 행위가 합당한 한에서, 우리는 피행위자를 존중의 방식으로 대우해야 하는 것이다. 따라서 실천적 맥락에서는 자율성 존중의 원칙이 자율성 원칙에 우선하는 것이다. 이제『인공지능의 존재론』에 실린, 자율성을 논의하는 논문들을 살펴보고자 한다.

4.『인공지능의 존재론』의 자율성에 관한 논문들에 대한 검토

1)「제 2장 의식적 인공지능에서」

이 논문에서 이영의 교수는 전통적인 철학에서 말하는 "실체로서의 자아"를 환상이라고 비판하면서 자아를 "내러티브의 산물" 또는 "언어적 과정의 산물"로 따라서 "우리의 구성물"로 보고자 한다.[11] "인간은 내러티브망 속에서 태어나며 내러티브적 질문을 통해 삶의 목적을 설정한다. 이런 점에서 인간은 내러티브적 주체이고 호모 나랜스(Homo Narrans)이며 인간의 행위는 내러티브망과 독립해 이해될 수 없다." 그러면서 저자는 "인공지능이 인간과 같은 내러티브적 자아를 가질 수 있는가"라는 흥미로운 문제를 제기한다. 저자는 "의식적인 인공지능"이 가능하다는 점을 주장하기 위해서 "내러티브적 자아"와 "인공적 밈"을 제시한다.

저자의 글은 인간의 어떤 특성들을 인공지능에 유비적으로 투사하려는 글이다. 즉 인간이 밈(meme)을 통해서 문화정보를 전달하듯이, 인공지능도 인공적 밈을 통해서 그럴 수 있다는 것이다. 또한 인간이

11) 이중원 외 (2018), p. 71.

내러티브를 통해서 자아를 형성하듯이, 인공지능 역시도 인간과 인공지능 간에 그리고 인공지능끼리의 내러티브를 통해서 자아를 형성할 수 있다는 것이다.

두 대상 간의 유비적인 관계가 성공하기 위해서는 핵심적인 특성들을 공유해야 할 것이다. 그런데 필자가 보기에, 인간과 인공지능의 '기억'의 특성에 있어서 중요한 차이가 있기 때문에, 결국 인공지능에게는 내러티브와 인공적 밈의 실효성이 없다고 생각한다. 결론부터 말하자면, 기억이 리셋되거나 포맷되거나 교체될 수 있는 존재자에게는(인공지능이 그러한 존재자인데) 내러티브를 통한 자아의 형성이나 밈을 통한 문화의 전승이 인간의 경우와는 다르다는 것이다. 다시 말해 인간에 있어서 내러티브와 밈이 가지는 그런 실효성은 없다는 것이다. 인공지능에게는 우리 인간에게는 '내러티브'를 전개할 자율적인 존재자를 인정하거나 상정할 수 있다. 하지만 인공지능에게는 우리와 같은 성격의 '내러티브'가 존재하지 않으며, 그것을 전개할 존재자를 상정할 수조차 없는 것이다.

인간에게 있어서 문화는 안정적이고 지속적으로 세대에서 세대로 전승된다. 그렇게 되는 이유는 문화의 운반체인 '뇌'에 문화의 특성이 각인되기 때문일 것이다. 그리고 유년시절을 포함해서 인생의 초기에 어떤 양육환경에 노출되느냐는 통상적으로 개인의 삶에 심대한 영향을 미친다. 뇌에는 양육환경을 포함해서 문화가 기입되며, 기입된 그것이 비가역적으로 과거라는 이름으로 저장되는 것이다. 기억상실이 아니라면 저장된 과거에 대한 기억은 개인의 자아와 정체성을 형성하는데 깊은 영향을 미치는 것이다. 그 기억에 인간은 '기속'되는 것이다. 자신이 살아온 기억으로서의 내러티브는 강요 없는 강제로서 우리의 삶을 때로는 인도하고 때로는 제한하는 것이다. 내러티브에로의 기속이 인간을 자율적 존재자로 만들어 주는 것이다.

그러나 인공지능의 CPU나 기억처리장치는 기입의 불가역성이 없다. 즉, 과거를 지우고 새롭게 시작하기 위해 리셋하거나 새로운 내용을 장착하거나 기억의 공간을 늘리거나 줄이거나에 있어서 원리적으로 자유롭다. 이러한 존재인 인공지능에게는 인간의 내러티브처럼 특정한

내러티브가 가지는 구속력은 없다. 인간의 뇌가 문화를 운반할 수 있는 것은 그것이 가진 기입의 상당한 불가역성 때문이다. 그러나 인공지능의 기억장치에는 기입의 용이함이 있을 뿐이다. 초기 양육환경이 가지는 자아 정체성에 대한 심대한 영향력 역시 인공지능에게는 찾아볼 수 없다. 또한 삶과 내러티브의 양방향성도 인공지능에게는 성립하지 않는다.

필자가 말하는 삶과 이야기의 양방향성이란, 살아온 '삶'이 이야기로 되고, 다시 그 '이야기'가 삶을 이끌어 가는 그런 양방향성이다. 우리는 하루를 등가치의 시간으로 생각하지 않는다. 우리는 하루의 시간을 기승전결이 있는 이야기로 마치 짜임새가 내재하는 것처럼 이해한다. 그리고 어제의 이야기가 오늘을 오늘의 이야기가 내일을 살아가는 지침을 준다. 삶과 이야기의 이러한 양방향성이 가능한 것은 우리가 삶에도 그리고 이야기에도 기속되기 때문이다. 그리고 이러한 기속성이 자율성의 필요조건이며 자율성을 가능하게 해주는 것이다.

그러나 인공지능에게는 언제든 새로운 이야기를 시작할 수 있기 때문에 이야기가 삶을 구속하지도 않으며, 살아온 삶이 미래를 구속할 이야기가 되지도 않는다. 인공지능의 과거는 본성상 리셋되거나 포맷될 수 있으며, 따라서 과거로서의 이야기에 대해 인간이 가지는 그런 기속성은 결국 존재하지 않는 것이다. 그리고 과거에 대한, 이야기에 대한, 기억에 대한 기속이 없는 존재자는 자율적인 존재자가 될 수 없다. 그렇다면 "인공지능이 인간과 같은 내러티브적 자아를 가질 수 있는가?"라는 저자의 물음으로 돌아가 보자. 과거가 새겨질 뇌가 없는 따라서 전승의 압력이 없는 그런 존재자에게 과연 문화가 가능할까? 가능하더라도 그런 문화는 우리가 익히 알고 있는 문화는 아닐 것이다.

2) 「제3장 인공지능이 자율성을 가진 존재일 수 있는가에 대하여」

이 논문에서 고인석 교수는 "공학적 개념으로서의 '자율성'과 철학적인 '자율성' 개념의 관계"를 해명하려고 한다.[12] 저자는 철학적 자율성 개념으로서 루소와 칸트의 자율성 개념과 크리스먼이 설명하고 있는

자율성을 검토한다.

루소의 자율성 개념을 설명하면서 저자는 말하기를 “X가 자율성을 지닌 존재인가? 하는 문제는 X라는 존재 자체의 속성들만으로 결정되지 않는다. 그것은 X를 성원(member)으로 하는 사회적 관계의 맥락 속에서만 결정 가능하다”고 한다.

그런데 칸트와 크리스먼이 말하는 자율성에 대한 저자의 평가에 대해서는 의문이 든다. 저자는 말하기를 “칸트의 관점에서 볼 때 인공지능은 그 수준이나 세부 속성과 무관하게 자율적 존재일 수 없는 것으로 보인다. (중략) 인간과 달리 그것들[인공지능이나 지능형 로봇]은 이성을 지닌 존재가 아니고, 도덕 법칙에 스스로 순응할 수 있는 존재가 아니기 때문이다”고 말하다. 필자가 보기에, 도덕 법칙의 본성이 무엇이고 구체적으로 무엇이 도덕법칙이냐에 관해서 우리들 사이에서도 합의가 이루어지지 않고 있으며 앞으로도 그럴 것 같다. 그리고 도덕법칙이 있다고 가정하더라도, 일부의 인간은 그것에 순응하지 않거나 못 할 것이다. 사정이 그러하더라도 칸트의 입장에 서서 저자는 말하기를 “칸트에게는 우주에 존재하는 모든 것들 가운데 이런 특유의 속성[“이성이 부여하는 실천 법칙에 스스로 복종하는 속성”]을 지닌 존재인 인간만이 자율적 존재의 지위에 있다. 따라서 ‘책임’이나 ‘의무’는 오로지 인간에게만 적용 가능한 개념이다”고 말한다. 그러나 이러한 저자의 언급은 칸트의 생각과는 상충하는 것 같다. 왜냐하면 필자가 아는 한, 인간만이 우주에서 지금까지 발견된 유일한 이성적 존재자이기 때문에 인간만이 자율적인 존재이며 책임의 의무의 주체라고 칸트가 말하고 있기 때문이다. 따라서 칸트는 이성적인 존재자로서 ‘외계생명체’가 있을 수 있을 가능성을 열어 두고 있는 것이다. 같은 취지에서 보자면, 저저의 주장과 달리, 칸트는 인공지능이 이성적 존재자인 한 자율적 존재이며 책임이나 의무의 주체가 될 수 있다는데 기꺼이 동의할 것 같다.

한편, 크리스먼이 설명하는 바 자율성 개념을 다음과 같이 옮겨 적고 있다 (크리스먼 스스로가 옹호하는 자율성 개념은 아닌 것 같다).

---

12) Ibid., p. 83.

> Autonomy, it is argued, implies the ability to reflect wholly on oneself, to accept or reject one's values, connections, and self-defining features, and change such elements of one's life at will.[13]

자율성은 **자기 자신 전체를** 반성의 대상으로 삼는 능력, 자신이 인정하고 있는 가치들, 연관성 그리고 자신을 규정하는 속성들을 수용하거나 거부하는 능력들, 그리고 자신의 삶 속에서 그와 같은 요소들을 자기 뜻에 따라 변경하는 능력을 함축한다고 일컬어진다.

필자가 보기에 저자는 밑줄 친 부분을 과도하게 해석하는 것은 아닌가 하는 의문이 든다. 그리고 한글로 번역된 그런 종류의 자율성은 지나치게 높은 기준이다. 우리가 자율적이라고 생각하는 많은 성인들은 그 수준에서 벗어 날 것이다. 즉 자기 자신 '전체'를 반성하지 않더라도 자율적이라고 평가할 수 있기 때문이다. 저자는 밑줄 친 부분 "reflect wholly on oneself"를 "체계 자체의 전복 가능성"이라고도 새기는데, 그럴 경우 자율성의 기준이 너무 높다는 생각이 든다.

그리고 위에서 인용한 영어 부분은 (그 뒤에 샌델의 주장이 뒤따르고 있는 데서도 알 수 있듯이) 샌들과 같은 공리주의자들이 롤즈의 인간관을 비판하기 위해서 그의 정의론에 언급된 내용을 토대로 하고 있는 문장들이다. "reflect wholly on oneself"와 같은 유사한 구절로 롤즈가 말하고자 한 바는 (롤즈는 "the highest-order interest"라는 표현을 사용하는데) "자기 자신을 철저하게 반성할 능력"을 말하는 것이다.

저자는 결론으로 주장하기를 "만일 공학적으로 가능하다고 해도 인공지능에 스스로의 목표를 설정 변경할 수 있다는 의미의 자율성을 부여하는 것은 정당하지 않음을 확인했다."고 한다. "반면에 스스로 목표를 변경하거나 새롭게 설정할 역량을 가진 인공지능에 그런 새로운 목표를 추구할 권능을 인정하는 일은 현재 우리가 알지 못하는 따라서 그것을 적절히 대응할 수 있는지 어떤지 알 수 없는 종류의 위험을 세계에 끌어들인다."[14]

---

13) Christman (2015).

14) 이중원 외 (2018), p. 111.

그런데 철학적 의미의 자율성을 부여하지 않아야 하는 까닭은 저자의 따르자면 우리가 알 수 없는 위험을 야기하기 때문이다. 만약 그렇다면 필자가 보기에 관건은 '인공지능이 철학적 의미에서 자율성을 갖느냐'가 아니라, '우리의 통제 범위를 벗어나느냐'에 있는 것이라고 생각한다. 통제 범위 내에 있다면 칸트적으로 자율적이라고 하더라도 허용될 수 있는 것이다. 주의할 것은 이분법적으로 생각하는 오류이다. 즉, 만약 우리가 인공지능의 자율성을 우리의 통제 범위에서 벗어나면 자율적이고, 통제 범위 내에 있으면 자율적이지 않다는 식으로 인공지능의 자율성을 이분법적으로 생각하는 것은 오류이다. 우리의 통제 범위 내에 있으면서 인공지능은 칸트적으로 자율적일 수 있는 것이다.

3) 「제4장 인공지능과 관계적 자율성에 대하여」

이 논문에서 이중원 교수는 인공지능에 적합한 자율성 개념으로 관계적 자율성을 주장하고 있다. 저자는 말하기를 "다시 말해 인공지능(로봇)의 행위의 정체성 자체가 인공지능(로봇)의 내재적 자율성보다는, 인공지능(로봇)이 인간 사회와 맺는 사회적 관계에 의해 규정되고 있다고 말할 수 있다. 그런 의미에서 인공지능(로봇)은 관계적 자율성을 갖고 있다고 말할 수 있다"[15]고 한다.

그런데 자율성에 있어서 "관계"가 왜 중요한가? 왜 "관계"를 중요하게 여겨야만 하는가? 그것은 자율성 개념이 정치적이기 때문이라고 생각한다. 필자는 앞서 '자율성에 대한 개념적 분석'을 통해서 자율성이란 실상 '정치적 개념'이라고 주장하였다. 인간 사회에서 어떤 행위나 선택을 또는 어떤 자를 자율적이라고 규정하는 것은 그것을 우리가 규범적으로 어떻게 대우해야 하는지를 함축한다. 따라서 무엇이 어떤 대상을 자율적이게 만들어 주는 요소인지를 규명하는 일은 단순히 개념적이거나 철학적인 과제에 그치는 것이 아니라, 궁극적으로는 정치적이고 사회적인 과제인 것이다. 예컨대 자율성이라는 개념적 도구는 억압과 무시에 대한 시정과 존중과 배려에 대한 요구를 정당하게 해주는

15) Ibid., p. 134-35.

권원을 부여해 주는 것이다.

그동안 자율성은 (영장류에 대해 논쟁이 있을 수 있지만) 일반적으로 인간 종에게만, 다시 말해 인간 유기체에게만 부여되어 왔다. 근원적인 측면에서 생물학적으로 죽어야만 하는 이성적인 존재가 생존기간 동안에 누려야 할 또는 그에게 정당하게 돌아가야 할 몫을 분배받아야 한다는 이 실존적 절박함 속에서 자율성 개념은 사회적으로 중대한 의미를 가진다고 할 것이다. 죽음이라는 한계 상황 속에서 자율성은 근본적인 차원에서 정치적 중요성을 가지는 것이다. 그렇다면 이제 인공지능에게 자율성을 부여하거나 또는 자율성이 공학적으로 구현된 인공지능을 개발하는 것이 무슨 의미인지를 '정치적 차원'에서 고찰해야만 할 것이다.

저자는 인공지능에게 적합한 자율성 개념으로 칸트적인 개체 중심의 자율성을 비판하면서 관계적 자율성을 정당화하고자 한다. 그런데 인공지능은 인간 유기체가 처한 생물학적 한계상황들이나 취약성을 가지고 있지 않다. 인공지능이란 인간 유기체가 처한 죽어야만 한다는 죽음의 사멸성이나 죽음으로 인해 완전히 끝장난다는 죽음의 절멸성을 가지지 않은 존재자이다. 전원이 공급되면 영원히 작동할 수 있으며, 테세우스의 배처럼 부품을 바꿔가며 지속할 수 있는 존재인 것이다. 따라서 인간이 한정된 생존시간 동안에 각자의 권리를 인정받기 위해 호소했던 자율성 개념이 가지는 정치적 의미를 바탕으로 한 관계적 자율성 개념을 과연 인공지능에게 적용하는 것이 적합한지 의문이 든다.

앞서 인용된 저자의 언급에서 보듯이, "인공지능의 행위의 정체성"을 "인공지능의 내재적 자율성보다는, 인공지능이 인간 사회와 맺는 사회적 관계에 의해 규정하려"는 시도를 통해서 인공지능에게 자율성을 부여할 수 있다. 그러나 인간과 인공지능의 물리적인 본성이 다르다면, 인간들끼리 맺는 사회적 관계와 다른 관계가 인공지능과 인간 사이에서 성립할 것이다. 따라서 저자의 시도는 단순히 소위 말하는 "관계적 자율성"을 외삽하는 것으로 인공지능의 자율성을 논의하는 것일지 모른다. 오히려 자율성 개념 자체가 가지는 자기진정성이나 자기인정에 의거해서, 인공지능이 자신의 결정이나 행위에 기속될 수 있기

위해서는 어떻게 해야 하는가에 대한 논의를 통해서 인공지능에 자율성을 부여해 볼 수도 있을 것 같다.

5. 결론

필자는 인공지능에게 자율성이 불가능하다는 주장을 하고자 하지 않는다. 실상은 자율성에 대한 개념적 분석을 통해서, 엔지니어들이 인공지능을 설계할 때에 '자율이라는 말의 의미에 부합하는 그러 자율성'을 위해서 어떤 개념적 요구사항을 충족시켜야 하는지에 대한 통찰을 가지게 하는데 있다. 우리 사회가 인공지능의 자율성을 어느 수준에서 인정할 것인가, 하는 사회적 인식의 문제일 것이다. 공학적 측면에서 인공지능의 자율성 문제를 논의하기에 앞서서, 언어적 또는 의미적 차원에서 자율성이란 무엇인지를 논의하는 것이 필요하다고 생각한다. 그리고 의미적 차원의 분석은 공학적 차원에서 개발된 어떤 인공지능의 자율성 인정 여부에 대한 논쟁이 붙었을 때, 보다 합리적인 토론으로 이끌 수 있는 개념적 도구를 제공해 준다는 데 그 의의가 있을 것이다. 또한 자율성 개념 자체가 정치적 관계를 함축하는 개념이라고 한다면, 어떤 개체를 자율적 개체라고 부르기에 앞서서, 우리가 그 개체를 자율적 존재로서 어떻게 대우해야만 하는가의 문제는 차후적으로 해결해야 할 문제가 아닌 것이다. 왜냐하면 이름은 그 이름에 해당하는 속성을 지칭하는 것처럼 보이기 때문이다. 다시 말해, 자율성은 '존중'이라는 속성을 지칭하는 것이다. 따라서 자율성에 대한 논의는 인식론적 논의와 윤리학적 논의가 중첩되어 있다는 점을 확인할 필요가 있을 것 같다.

참고문헌

고인석 (2018), 「인공지능이 자율성을 가진 존재일 수 있는가?」, 이중원 외, 『인공지능의 존재론』, 한울.

이영의 (2018), 「의식적 인공지능」, 이중원 외, 『인공지능의 존재론』, 한울.

이중원 외 (2018), 『인공지능의 존재론』, 한울.

신상규 (2017), 「인공지능은 자율적 도덕행위자일 수 있는가?」, 『철학』 132권, pp. 265-292.

EBS 아기성장보고서제작팀 (2009), 『아기성장보고서: EBS 특별기획 다큐멘터리』, 위즈덤하우스.

Christman, J. (2015), 'Autonomy in Moral and Political Philosophy', https://plato.stanford.edu/entries/autonomy-moral/ (검색일: 2018. 11. 15.)

Parfit, Derek (2011), *On What Matters*, vol. I, Oxford University Press.

Star, D. (ed.) (2018), *The Oxford Handbook of Reasons and Normativity*, Oxford University Press.

Swanton, C. (2005), *Virtue Ethics: A Pluralistic View*, Oxford University Press.

논문 투고일	2018. 11. 16.
심사 완료일	2018. 11. 20.
게재 확정일	2018. 11. 21.

Conceptual Analysis on Autonomy

Cheul Kang

One of the key philosophical issues in artificial intelligence discussion is the concept of autonomy. The artificial intelligence called ‘autonomous’ driving car has already been around us, and the term ‘autonomous artificial intelligence’ has already been used. In this situation, I think that we should basically ask the meaning of autonomy. I insist that the term ‘autonomy’ should be used in its intrinsic meaning. To do so, we need to analyze the concept of autonomy into its integral components. In particular, I want to present commitment as a key element of autonomy and consider ‘autonomy as commitment’. In short, this paper attempts to do the conceptual analysis of autonomy. The significance of this analysis is that it helps the engineers to have a critical mind concerning what kind of conditions they should satisfy when they want to realize ‘autonomy in accordance with the meaning of autonomy’. This paper examines articles on autonomy in *Ontology of Artificial Intelligence* (2018).

Keywords: autonomy, commitment, artificial intelligence, conceptual analysis