http://dx.doi.org/10.16960/jhlr.17.2.201606.1

자율주행자동차와 로봇윤리: 그 법적 시사점*

이 중 기
(홍익대학교 로봇윤리와 법제연구센터, 법과대학 교수)

오 병 두
(홍익대학교 로봇윤리와 법제연구센터, 법과대학 교수)

【초 록】

이 글은 자율주행자동차에 적합한 로봇윤리를 검토하고 그 법적 시사점을 모색한 것이다. 우선, 자율주행자동차의 로봇윤리로는 자율주행자동차에 맞추어 수정된 아시모프의 3원칙과 해악 최소화 알고리듬을 결합하는 하향식 접근법을 기반으로 하여 그 부족한 점을 기계학습과 같은 상향식 접근법으로 보완하는 혼합형 접근법이 적합하다고 보았다.

다음으로, 자율주행자동차의 로봇윤리의 한계 및 법적 규율과의 관계를 검토하여 다음과 같은 법적 시사점을 제시하였다. 첫째, 자율주행자동차의 의사결정 프로그램은 헌법적 적합성을 개발단계에서부터 검토할 필요가 있다. 둘째, 다수의 주체가 관여하는 도로교통의 특성을 고려하여 교통법규가 반영된 로봇윤리를 구성하여야 하고 이를 위하여 관련 법령의 구조화와 위계화가 필요하다. 셋째, 로봇윤리에 따른 자율주행자동차의 선호의 체계가 공익을 저해하지 않도록 조정하는 법적 노력이 필요하다. 넷째, 기술발달과 병행하는 규제적 법제의 정비를 위해서는 원리 위주의 규제입법이 필요하다.

주제어 : 로봇윤리, 자율주행자동차, 인공지능, 윤리적 난제(도덕적 딜레마), 하향식 접근법, 상향식 접근법, 혼합형 접근법, 수정된 아시모프의 3원칙, 해악 최소화 알고리듬

* 이 논문은 2015학년도 홍익대학교 학술연구진흥비에 의하여 지원되었음.

【차 례】

Ⅰ. 들어가며

자동차사고의 대부분이 운전자의 고의·과실로 인한 것인 현실에서,[1] 운전 중의 모든 상황을 예측·판단하는 자율주행자동차(autonomous vehicle)[2]는 음주운전이나 졸음운전이 없고, 핸드폰 조작 등으로 주의력이 분산되는 일이 없어 교통사고를 획기적으로 줄일 수 있다는 점에서 큰 기대를 모으고 있다.[3] 또한 탑승자가 자동차 운행 중에 운전 이외의 다른 업무를 볼 수 있고, 인공지능에 의한 최적 경로 산출로 교통체증과 공회전 등으로 인한 연료 낭비도 막을 수 있다는 것도 장점으로 지적되고 있다.[4]

자율주행자동차는 자동화된 자동차(automated vehicle)의 일종이다. 인간이 운전하는 자동차에도 크루즈 컨트롤, 자동 제동장치, 차선유지 장치 등 다양한 자동화 장치가 이미 사용되고 있다. 여기서 어느 정도의 자동화 수준이어야 자율주행자동차라고 할 것인가에 대

1) 한국의 경우 약 90%의 사고가 운행자의 과실로 인한 것이라고 한다(이종영, 김정임, "자율주행자동차 운행의 법적 문제," 중앙법학회, 『중앙법학』 제17집 제2호, 2015.6[이하 "이종영·김정임"], 147면). 미국에서도 같은 비율로 운행자 과실의 교통사고가 발생한다고 한다(Jeffrey K. Gurney, "Crashing into the Unknown: An Examination of Crash-optimization Algorithms through the Two Lanes of Ethics and Law," 79 Albany Law Review, 2015/2016[이하 "Gurney"], p.191).

2) 이 글에서는 「자동차관리법」 제2조 제1호의3에 따라 '자율주행자동차'라는 표현을 쓰기로 한다. 자율주행자동차는 그밖에도 자율주행차, 무인자동차 등으로도 불리며 영어로는 self-driving car, driverless car, robot car 등의 표현이 사용되고 있다.

3) Gurney, pp.191-192.

4) Gurney, pp.193-194.

해서 논의가 있다.

2012년 1월 독일 연방 도로기술연구소(Bundesanstalt für Straßenwesen, 이하 "BASt")는 자동차의 자동화 수준을 4단계로 분류하였고,[5] 이어서 2013년 5월 미국 연방고속도로안전청(National Highway Traffic Safety Administration, 이하 "NHTSA")은 이와 유사한 5단계의 기준을 공표한 바 있다.[6] 2014년 1월 SAE International(Society of Automotive Engineers, 미국 자동차 공학협회, 이하 "SAE")은 가장 상세한 6단계의 자동차 자동화 기준(J3016)을 제안하였다.[7]

SAE 기준에 따르면, 3레벨부터 인간이 아닌 자율주행 시스템이 특히 주행환경을 모니터링하며 동적 운행임무(dynamic driving task)를 수행하기 시작한다.[8] 즉, 이 레벨부터 운행에 필수적인 조향, 제동, 가속 등의 기능과 도로상황의 모니터링을 시스템에 의해 자동적으로 수행하는 자율주행이 가능하다. 현재 자동화기술의 수준은 2레벨과 3레벨 사이에 와 있다고 한다.[9]

이 글에서 논의하는 자율주행자동차는 SAE 기준으로 3레벨 이상으로 자동화된 자동차,[10] 즉 "일정한 조건 하에서 스스로 주행환경을 인식하면서 자율적으로 운행하는 자동

---

5) BASt는 자동화 수준을 비자동화 수준(Driver Only), 운전자 보조수준(Assistiert), 일부 자동화 수준(Teilautomatisiert), 고도의 자동화 수준(Hochautomatisiert), 완전한 자동화 수준(Vollautomatisiert)으로 나눈다(http://www.bast.de/DE/Publikationen/Foko/2013-2012/2012-11.html, 최종접속일: 2016.5.31.).

6) 구체적으로는 0레벨(비자동화 수준, No-Automation), 1레벨(특정기능의 자동화 수준, Function-specific Automation), 2레벨(통합된 기능의 자동화 수준, Combined Function Automation), 3레벨(제한된 자율주행 수준, Limited Self-Driving Automation), 4레벨(완전한 주행 자동화 수준, Full Self-Driving Automation)로 나눈다(NHTSA, U.S. Department of Transportation Releases Policy on Automated Vehicle Development, 2013.5.30.(http://www.nhtsa.gov/About+NHTSA/Press+Releases/U.S.+Department+of+Transportation+Releases+Policy+on+Automated+Vehicle+Development, 최종접속일: 2016.5.31.).

7) 구체적으로는 0레벨(비자동화 수준, No Automation), 1레벨(운전자 보조수준, Driver Assistance), 2레벨(일부 자동화 수준, Partial Automation), 3레벨(조건적 자동화 수준, Conditional Automation), 4레벨(고도의 자동화 수준, High Automation), 5레벨(완전한 자동화 수준, Full Automation)로 분류한다(http://www.sae.org/misc/pdfs/automated_driving.pdf, 최종접속일: 2016.5.31.).

8) 여기에서 동적 운행임무란 ① 조향장치, 제동장치, 가속장치 등과 자동차 및 도로상황에 대한 모니터 기능(운행조작의 측면)과 ② 발생한 사태에 대응하여 차선 변경, 유턴, 운행신호의 사용 등의 기능(전술적인 측면)의 운행임무를 말한다. 그러나 목적지나 중간경유지의 설정(전략적 측면)까지를 자동화할 것을 요하지는 않는다(위 각주 7)의 URL 참조).

9) Gurney, p.189. 여기에서는 NHTSA 기준을 중심으로 서술하고 있으나 SAE 기준에 의해서도 같다.

10) BASt의 고도의 자동화 수준, NHTSA의 3레벨도 인간 운전자가 주행 중에 도로를 상시적으로 모니터할 필요가 없으므로 같은 수준에 해당한다. Noah J. Goodall, "Ethical Decision Making During Automated Vehicle Crashes," Transportation Research Record: Journal of the Transportation Research Board, No. 2424, Transportation Research Board of the National Academies, 2014[이하 "Goodall(2014a)"], p.58은 NHTSA의 3레벨 이상을 다루고 있다.

차"이다.[11] 이 개념에 의한 자율주행자동차는 인간에 의하여 목적지가 설정되거나 자율주행 모드가 설정되면 그에 따라 스스로 도로상황 등 주행환경을 인식하여 위험요소를 식별하고 사전에 프로그램된 기준에 따라 의사결정을 하면서 도로를 주행한다.

자율주행자동차가 보통의 인간을 뛰어넘는 인식, 판단 및 조작을 할 수 있어서 사고위험을 극히 낮출 수 있기는 하나, 그렇다고 전혀 사고를 내지 않는 것은 아니다.[12] 여기에서 자동차의 자율주행 중 사고가 난 경우 법적으로 누가 그리고 어떤 책임을 지는가라는 문제와 자율주행자동차가 운전 중, 특히 사고 직전에 어떤 윤리적 의사결정을 하여야 하는가라는 문제가 새롭게 제기된다.[13] 자율주행자동차는 자율주행 상태에서 다양한 위험요소에 대하여 평가하고 이에 따라 스스로 주행에 관한 판단을 함으로써 사고를 방지하고 사고발생이 불가피한 경우에도 그 손해를 최소화하도록 미리 설계된다. 사고결과는 그 선택과 인과관계가 있다. 여기에서 그 선택의 윤리성을 검토하게 되는 것이다.

전자가 사고 발생 이후 사후처리와 관련된 자율주행자동차의 법적 책임문제라면,[14] 후자는 자율주행자동차가 주행하면서 상시로 행하는 의사결정과 선택이 '옳은가'라는 윤리적 문제이다. 이는 소위 자율주행자동차의 '로봇윤리'로 다루어지는데,[15] 자율주행자동차가 스스로 주행하는 범위에서 자율주행 프로그램(self-driving program) 내지 인공지능(artificial intelligence)이 로봇으로서 인간 통제권을 대체하기 때문이다.

이하에서는 로봇윤리에 관한 기존의 논의를 개념과 접근방법을 중심으로 살펴보고(II), 자율주행자동차에 적합한 로봇윤리를 어떻게 구성할 것인가를 고찰한 후(III), 그에 따른 법적 시사점을 검토해보기로 한다(IV).

---

11) 이 글에서의 자율주행자동차 개념은 이하의 로봇윤리를 검토하기 위한 조작적 정의이다. 따라서 기술적, 법적으로 절대적 기준은 아니다. 따라서 논의의 필요에 따라 다른 정의도 가능하다. 예컨대, 이중기, 황창근, "자율주행자동차 운행에 대비한 책임법제와 책임보험제도의 정비필요성: 소프트웨어의 흠결, 설계상 흠결 문제를 중심으로," 한국금융법학회,『금융법연구』제13권 제1호, 2016[이하 "이중기·황창근"], 95면의 "목표지점이 설정되면 인위적인 추가 조작 없이 스스로 주행환경을 인식하면서 목표지점까지 자율적으로 운행하는 자동차"라는 정의하기도 하고, 이종영·김정임, 146면은 "자동차 스스로 주변환경을 인식하고, 위험을 판단하면서, 계획한 목적지까지 경로를 주행하는 자동차로서, 운전자의 주행조작을 최소화하며 스스로 안전주행이 가능한 인간 친화형 자동차"라고 정의하기도 한다. 전자와 후자는 인위적 조작의 유무에서 차이가 있는데, 이는 자율성의 의미를 다르게 이해한 결과이다.

12) Goodall(2014a), p.59.

13) Noah J. Goodall, "Machine Ethics and Automated Vehicle," Gereon Meyer, Sven Beiker (ed.), Road Vehicle Automation, Springer, 2014[이하 "Goodall(2014b)"], p.93.

14) 이를 다룬 국내문헌으로는 이종영·김정임, 145-184면; 이중기·황창근, 94-122면.

15) Goodall(2014b), p.93.

Ⅱ. 자율주행자동차와 로봇윤리

1. 로봇윤리의 개념

현재 다양한 맥락에서 로봇윤리가 논의되고 있다. 대표적으로 Gianmarco Veruggio와 Keith Abney는 로봇윤리(robot ethics)[16]를 다음의 3가지 의미로 나누어 설명한다.[17]

첫째, 로봇윤리는 응용윤리학의 한 영역으로서 로봇을 사회생활에 도입한 결과 발생하는 윤리적 쟁점에 대한 연구라는 의미로 사용된다.[18] 이 의미의 로봇윤리는 인간으로서의 존엄, 약자의 권리 보호, 로봇기술로 인한 차별 문제인 "로봇 디바이드"(robotics divide)[19]의 제한 등과 같이,[20] 로봇-인간이 연결되는 사회적 영역을 연구대상으로 한다.

둘째, 로봇윤리는 "로봇 자체에 장착되어야 할 윤리적 규범"(moral code), "로봇에 프로그램된 일종의 도덕률"(a morality)이라는 의미로도 사용된다. 이 경우 로봇이 수행하도록 프로그래머가 설정한 코드(code)가 로봇의 윤리적 규범이 된다. 로봇이 자신에게 설정된 도덕률 자체를 인식하는 것은 아니며 "단지 명령을 따르는 것"에 불과하다.[21] 이 입장에서는 어떠한 윤리적 요청, 윤리이론에 따르도록 로봇을 프로그램할 것인가가 중요하므로[22] 로봇윤리도 로봇 자체의 윤리가 아닌 로봇을 설계, 제작, 관리, 사용하는 인간의 윤리를 의미하게 된다.

셋째, 로봇윤리는 로봇 자체의 윤리라는 의미로도 사용되기도 한다. 이는 로봇이 윤리적 행위자가 될 수 있음을 전제로 한다. 로봇이 인간의 특성인 자의식과 합리적 선택능력, 즉 자유 또는 자유의지를 갖고서 스스로 자신만의 도덕률을 선택할 수 있다고 보는 것이다[23]

16) 로봇윤리에 대한 표현으로는 그 이외에도 machine ethics, robot ethics, machine morality 등이 사용되고 있다.

17) 이하의 내용은 Gianmarco Veruggio, Keith Abney, "Roboethics: The Applied Ethics for a New Science," Patrick Lin, Keith Abney, George A. Bekey (ed.), Robot Ethics: The Ethical and Social Implications of Robotics, The MIT Press[이하 "Robot Ethics"], 2012[이하 "Veruggio·Abney"], pp.347-348을 정리한 것이다.

18) Gianmarco Veruggio가 이 개념의 로봇윤리를 다른 경우와 구별하기 위해 roboethics란 용어를 만들었다고 한다(Veruggio·Abney, p.348).

19) 이 로봇 디바이드는 "로봇기술의 이익과 혜택의 불균형한 분배 문제"로서 선진국과 후진국, 부자와 빈자 사이에서 로봇기술의 개발 및 접근 가능성과 관련하여 벌어지는 분배정의, 사회정의의 문제이다(변순용, 송선영, 『로봇윤리란 무엇인가?』, 어문학사, 2015[이하 "변순용·송선영"], 18-19면).

20) Veruggio·Abney, p.347.

21) Veruggio·Abney, p.347.

22) Patrick Lin, "Introduction to Robot Ethics," Robot Ethics, 2012[이하 "Lin(2012)"], p.9.

23) Veruggio·Abney, p.348.

따라서 "로봇을 '사람'(person)으로 보고 로봇에게 일정한 권리와 책임을 부여할 수 있는가?"[24]가 중요한 관심사이다.

2. 로봇윤리에 대한 접근방법

로봇윤리를 어떻게 구성할 것인가라는 로봇윤리의 접근방법은 로봇윤리에 관한 세 번째의 관념에서 출발한다. 즉, 로봇을 윤리적 행위자 내지 주체인 인공도덕행위자(artificial moral agents, AMA)로 파악하는 전제에서 어떻게 하면 로봇을 윤리적 행위자로 만들 것인가의 문제가 로봇윤리에 대한 접근법으로 다루어진다.[25] 이와 관련해서는 ① 하향식 접근법(top-down approach), ② 상향식 접근법(bottom-up approach), ③ 혼합형 접근법(hybrid approach) 등의 3가지가 제시되고 있다.[26]

하향식 접근법은 사전에 정해진 일정한 규칙의 집합으로 로봇윤리를 구성하고자 한다.[27] 이 접근법에서는 특정한 하향식 접근법을 구현할 알고리듬을 찾는 일이 중요하다.[28] 하향식 접근법의 예로는, 대표적으로 칸트의 정언명령과 같은 일련의 의무의 체계에 따라야 할 것을 강조하는 의무론과 특정한 효용의 극대화를 목표로 하는 공리주의를 들지만,[29] 그 이외에도 황금률, 10계명, 덕윤리 등도 여기에 속할 수 있다.[30] 하향식 접근법 중에는 특정 윤리이론만으로 구성하기보다는 2개 이상의 윤리이론을 결합하자거나[31] 덕윤리, 공리주의, 의무론적 윤리설, 책임윤리 등에 공통되는 '최소도덕'을 기준으로 하자는 입장도 있다.[32]

한편, 상향식 접근법은 로봇이 스스로 경험으로부터 윤리적 판단능력을 학습하는 것을 강조한다.[33] 기계학습(machine learning)을 통해 로봇 스스로 윤리적 지식을 습득하도록 하자는 것이 그 전형적인 예이다.[34]

---

24) Lin(2012), p.9.
25) 예컨대, 웬델 월러치, 콜린 알렌/노태복 옮김, 『왜 로봇의 도덕인가』, 메디치, 2014[이하 "월러치·알렌"], 17면 이하 참조.
26) 월러치·알렌, 138면 이하; Keith Abney, "Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed," Robot Ethics, 2012[이하 "Abney"], pp.36-39.
27) 월러치·알렌, 146면.
28) 월러치·알렌, 147면.
29) Goodall(2014a), p.61.
30) 월러치·알렌, 146면.
31) Goodall(2014b), p.99.
32) 변순용·송선영, 28면.
33) 월러치·알렌, 139면.

하향식 접근법은 일반적인 차원에서 시스템을 구조화하기에는 적합하나 모든 사태들에 대해 망라적으로 구현하기는 어렵다. 상향식 접근법은 그와는 반대로 일정한 시스템을 전제로 그 과제나 임무를 구체화하는 단계에서 사용하기는 용이하지만, 시스템을 구조화하는 데에는 사용하기 어렵다.[35] 따라서 양자의 어느 정도의 결합이 모색되지 않을 수 없는데, 여기에서 고려되는 것이 하향식 접근법과 상향식 접근법을 조합하는 혼합형 접근법이다.[36] 덕윤리를 기본으로 하여 상향식 접근법으로 보완하여 포괄적인 도덕적 의사결정 모델을 수립해야 한다는 견해가 그 대표적인 예이다.[37]

3. 소결: 자율주행자동차에 적합한 로봇윤리

로봇윤리에 관한 기존의 논의는 자율주행자동차의 경우에 그대로 적용하기는 어렵다. 우선, 로봇윤리의 개념을 보자. 자율주행자동차의 경우에는 이와 같은 윤리적 행위자라는 관념을 도입하기가 쉽지 않지 않다. 첫째, 현재의 기술수준에서 구현되는 자율성을 곧바로 자유의지와 동의어로 파악하여 이를 윤리적 의미의 주체성과 연결시키는 어렵다.[38] 둘째로, 자율주행자동차가 스스로 도덕률을 따르도록 하는 것은 위험할 수 있다. 모든 자율주행자동차이 자신의 고유한 윤리적 기준을 가지고 스스로 개별적인 자율 판단에 따라 도로교통체계에 등장하는 경우 다른 인간 운전자나 자율주행자동차의 예측가능성이 낮아져 사고 위험이 증가할 수 있다. 요컨대, 자율주행자동차의 로봇윤리는 인간의 윤리, 특히 로봇의 설계, 제작, 관리, 사용하는 인간에게 적용되는 윤리로 파악하여야 한다.

다음으로, 로봇윤리의 접근법도 로봇을 완전한 윤리적 행위자로 보는 전제에서 논의되고 있다는 점에서 그대로 채택하기는 어렵다. 첫째, 이미 자율주행자동차가 시행주행 단계에 접어든 현재 시점을 고려할 때, 실행가능한 로봇윤리를 고민해야 한다. 로봇이 완전한 인격성을 구비할 때까지 자율주행자동차의 운행을 막을 이유나 필요는 없으며, 현실적인 자율주행자동차의 유용성이 있다면 가능한 범위 내에서 이를 운용하는 것이 사회 전체적인 이익에도 부합하기 때문이다. 따라서 자율주행자동차의 경우 제한적인 범위 내의 로봇

34) 월러치·알렌, 185면 이하.
35) 월러치·알렌, 138-141면.
36) 월러치·알렌, 201면 이하; Gurney, p.208.
37) 월러치·알렌, 203면.
38) Lin(2012), p.9. 현재 로봇윤리에 관하여 대부분 이 입장을 취한다(예컨대, 고인석, "아시모프의 로봇 3법칙 다시 보기: 윤리적인 로봇 만들기," 철학연구회, 『철학연구』 제93집, 2011[이하 "고인석"], 109면; 변순용·송선영, 17-18면; Gurney, p.208 등).

윤리만이 문제된다. 따라서 하향식 접근법 중 보편적인 윤리에서 출발하는 이론은 자율주행자동차에는 맞지 않다. 둘째, 상향식 접근법은 자율주행자동차의 경우에는 위험할 수도 있다. 자율주행자동차가 개별적으로 이루어지는 기계학습 등의 방법으로 스스로 학습하여 인간 운전자의 편견, 잘못된 운전습관 등을 배우게 된다면 이 또한 사고의 원인이 될 수 있기 때문이다.

다만, 여기에서 로봇윤리에 관한 하향식 접근법과 상향식 접근법의 발상 자체는 실용적인 차원에서 활용할 수 있다고 본다. 그렇다면 자율주행자동차의 윤리는 어떻게 구성하여야 하는가? 이에 관하여는 장을 바꾸어 살펴보기로 한다.

Ⅲ. 자율주행자동차에 적합한 로봇윤리의 구성

1. 서설: 자율주행자동차에 적합한 로봇윤리의 구성방법

(1) 실용적 차원의 접근

자율주행자동차의 로봇윤리를 어떻게 구성할 것인가에 관하여는 실용적 접근이 필요하다고 본다. 현재의 자율주행자동차의 발전상황을 고려할 때, 간결하고 명료한 위계구조를 가진 윤리규범을 설정한 후(제1단계), 운행을 통한 경험을 반영하여 그 하향식의 윤리기준을 지속적으로 보충하는 것(제2단계)이 현실적이다. 이 과정에서 그 결과가 축적되면 다시금 피드백하여 윤리기준을 재검토하고 새로운 기준을 수립하는 과정도 지속적으로 수행되어야 할 것이다.

결국 이는 하향식 접근법을 기반으로 하여 그 단점을 상향식 접근법으로 보완하는 혼합형 접근법이 될 것이다.[39] 자율주행자동차의 로봇윤리와 관련하여 가장 주목되고 있는 것은 사고의 방지 내지 손해의 최소화라는 요소이다. 그런데 현실 세계에서 발생하는 모든

39) 이와 유사한 제안으로는 Goodall(2014a), pp.63-64이 있다. Goodall은 자율주행자동차의 로봇윤리의 발전을 위한 3단계의 전략을 제시한다. 제1단계로 일반적으로 합의된 원칙들(예컨대, 사망보다는 상해를 선택하여야 한다)에 의하여 사고 영향을 최소화하는 합리적 도덕 시스템의 구축하고, 제2단계로는 여기에 기계학습 기법을 이용하여 인간 운전자들의 의사결정과 충돌 시나리오를 시뮬레이션해서 유사한 가치적 요소를 추가한 다음, 최종적인 제3단계에서는 자율주행자동차가 이를 인간의 언어로 표현하여 인간으로 하여금 이해하고 이를 교정할 수 있도록 하자고 한다. 제3단계는 자율주행자동차의 제작 단계에서 확인 가능하므로 이를 별도의 단계로 설정하지 않는다면 위에서 제시하는 2단계의 구조와 유사하다.

사고상황에 대처하여 이를 사전에 완벽하게 규정하는 것은 불가능에 가깝다. 따라서 위 제1단계에서 고려될 하향식 모델로는 완전한 윤리적 선택을 전제로 한 모델보다는 도로교통 상황에서 적절하게 선택할 수 있는 개방적이고 융통성 있는 것이 더 선호될 것이다.[40]

여기에 특정 지역 또는 국가의 운전자들의 경험적 특징, 도로의 특성, 운전과 관련한 문화적 규범 등은 기계학습과 같은 상향식 접근법에 의하여 보완될 필요가 있다고 본다. 다만 어느 정도의 정보가 축적된 이후에는 상향식 접근법은 하향식 접근법에 흡수될 것이다. 그러므로 여기에서 상향식 접근법에 의한 보완이라는 것은 하향식 모델을 완성하기 위한 수준의, 제한적인 의미에 지나지 않는다.

(2) 자율주행자동차의 윤리적 난제

다음으로 문제되는 것은 혼합형 접근법의 기본을 이루는 하향식 접근법으로는 어떤 기준을 채택하여야 하는가이다. 자율주행자동차와 관련한 하향식 접근법으로는 아이작 아시모프의 로봇 3원칙과 해악 최소화 알고리듬이 주로 거론되고 있다.[41] 전자는 의무론의 입장[42]과, 후자는 공리주의적 관점[43]과 각각 연결된다. 의무론은 일련의 의무의 체계에 따라야 할 것을 강조함에 반하여, 공리주의는 공리(효용, utility)의 총량을 극대화하여야 한다고 본다.[44]

이들 하향식 모델 중에서 어떠한 것이 자율주행자동차에 더 적합한가를 논의하기 위하여 윤리적 난제(도덕적 딜레마)들이 제시되고 있다. 이는 자율주행자동차의 불가피한 교통사고와 관련한 일종의 사고실험으로 극단적인 이익충돌 상황을 가정한 것인데, 이익충돌 상황에 따라 크게 2가지 유형으로 나누어 볼 수 있다.[45] 이때 대립하는 이익으로 고려되는 것은 생명, 신체의 안전, 재산 3가지이다.[46]

---

40) Goodall(2014b), p.95는 충돌 순간에 복잡한 윤리적 판단이 가능할 것이라는 것은 비현실적이라고 지적한다.

41) 이는 자율주행자동차의 경우, 인공도덕행위자(AMA)를 인정하는 전제에서 논의되는 덕윤리나 칸트 윤리학은 현 수준에서는 고려할 여지가 없기 때문으로 보인다.

42) Abney, p.42; Goodall(2014a), p.61; J. Christian Gerdes, Sarah M. Thornton, "Implementable Ethics for Autonomous Vehicles," Autonomous Driving, 2016[이하 "Gerdes·Thornton"], p.95. 또한 월러치·알렌, 159-160면에서는 명시적으로 이를 의무론이라고 하지는 않으나 공리주의에 대립되는 "의무기반 도덕"(duty-based morality)이라고 설명한다.

43) Goodall(2014b), p.97; Goodall(2014a), p.62.

44) 월러치·알렌, 147면.

45) 이하의 사례들은 주로 Patrick Lin, "Why Ethics Matters for Autonomous Cars," Markus Maurer, J. Christian Gerdes, Barbara Lenz, Hermann Winner (ed.), Autonomous Driving, Springer[이하 "Autonomous Driving"], 2016[이하 "Lin(2016)"], p.69 이하; Gurney, p.195 이하를 재구성한 것이다.

첫째, 서로 다른 피해자들 중에서 하나를 선택하는 유형이 있다(① 유형). 이 유형에서는 자율주행자동차와 그 탑승자가 손상되거나 상해를 입을 것을 전제로 하지 않는다.

[①-1] 전차 사례(Trolley Problem)[47]: 자율주행자동차의 교통사고가 불가피한 상황에서 한 쪽에는 8살짜리의 어린이가 있고, 다른 한 쪽에서는 80세의 노인이 있다. 자율주행자동차는 누구와 충돌해야 하는가?[48]

[①-2] 헬멧 사례(Motorcycle Problem): 위의 사례에서 좌측에는 헬멧을 쓴 모터사이클 운전자가 있고, 우측에는 헬멧을 쓰지 않은 모터사이클 운전자가 있다면, 자율주행자동차는 누구와 충돌해야 하는가?[49]

둘째, 자율주행자동차와 그 탑승자의 이익과 다른 피해자의 이익 사이에서 선택해야 하는 유형(② 유형)도 있다. 자율주행자동차가 자신 또는 탑승자의 안전을 더 고려해야 하는가 아니면 다른 자동차 혹은 탑승자 이외의 사람의 안전을 더 고려해야 하는가가 이 유형에서의 쟁점이다.

[②-1] 교각 사례(Bridge Problem): 자율주행자동차가 다리 위의 2차선을 주행하던 중 충돌이 불가피한 상태에 놓이게 되었다. 이 경우에 ① 맞은편에서 오던 여러 명의 학생이 탄 스쿨버스와 충돌하여야 하는가? 아니면 ② 다리에서 벗어나 탑승자를 사망케 하여야 하는가?[50]

[②-2] 터널 사례(Tunnel Problem): 자율주행자동차가 1차선의 산길 위의 좁은 터널로 진입하던 중 도로 전방에 넘어진 어린 아이 1명을 발견하였다. 이 경우 ① 직진하여 아이를 충격하여 사망케 하여야 하는가? 아니면 ② 방향을 틀어 터널 벽을 들이받아 탑승자를 사망케 하여야 하는가?[51]

[②-3] 자동차 대 자동차 사례(the Car Problem): 위 전차 사례에서 왼쪽에는 안전도가 낮은 차가 있고, 오른쪽에는 안전도 높은 차가 있다면 어떤 차와 충돌하여야 하는가?[52]

---

46) Goodall(2014a), p.61.
47) 원래 전차사례는 Philippa Foot에 의하여 전차 운전사가 1명 또는 다수의 사람을 불가피하게 충돌하게 되는 상황 하에서의 윤리적 판단을 위한 예제로서 제시되고, Judith Jarvis Thomson에 의하여 전차 운전사가 아닌 제3자가 통제할 수 있도록 변형된 사례인데(Gurney, p.206), 위의 내용은 자율주행자동차의 경우에 맞추어 수정된 것이다.
48) Lin(2016), p.69.
49) Lin(2016), pp.72-73; Gurney, p.197.
50) Lin(2016), p.76; Gurney, p.204.
51) Gurney, p.202.
52) Lin(2016), p.72; Gurney, p.198.

이하에서는 아시모프의 3원칙과 해악 최소화 알고리듬에 의할 때, 윤리적 난제들이 어떻게 해결되는지 살펴보기로 한다.[53]

2. 수정된 아시모프의 3원칙과 해악 최소화 알고리듬

(1) 자율주행자동차를 위해 수정된 아시모프의 로봇 3원칙

아시모프의 3원칙은 소설가 아이작 아시모프가 1942년 출간된 단편소설 "Runaround"에서 제시한 기준이다. 이는 그 기준의 다의성 때문에 주인공인 로봇이 곤란을 겪도록 유도하는 일종의 플롯장치(plot device)였다.[54] 아시모프가 제안한 의도는 쉽게 풀기 어려운 윤리적 난제(도덕적 딜레마)에 봉착하여 모순되는 선택 사이에서 갈등하는 로봇의 모습을 보여주고자 한 것이었다.

> [제1법칙] 로봇은 인간을 해치거나 혹은 부작위에 의해 인간에게 위험을 초래해서는 안 된다.
> [제2법칙] 로봇은 제1법칙에 저촉되지 않는 한, 인간이 내린 명령에 복종해야 한다.
> [제3법칙] 로봇은 제1법칙 혹은 제2법칙에 저촉되지 않는 한, 자신을 보호해야 한다.

로봇윤리에 관한 다양한 상상력의 원천이 되기는 하였으나,[55] 독자적인 로봇윤리의 기준으로 부적절하다는 것이 윤리학적 접근을 하는 연구자들의 시각이다.[56] 이에 반하여, 자

---

53) 이상에서 다루어진 사례들은 설명하고자 하는 목적과 취지에 따라 다른 조건을 부여함으로써 상이한 유형으로 다룰 수 있다. 또한 ① 유형의 자기보존과 ② 유형의 타인간의 이익형량이 혼합된 사례도 가능하다(③ 유형). 예컨대, 자동차 대 자동차 사례[②-3]는 자율주행자동차와 그 탑승자의 안전이 문제된다고 구성할 수 있다(Lin(2016), p.72). 이는 위의 ① 유형과 ② 유형의 논의가 중첩적으로 적용되는 것이나, 논의의 단순화를 위해 여기에서는 따로 검토하지는 않기로 한다. 이 ③ 유형의 대표적인 예로는 쇼핑 카트 사례(Shopping Cart Problem)가 있다. 이는 브레이크가 고장난 자율주행자동차의 전방에는 유모차를 밀고 있는 아기엄마가 있고, 그 왼편에는 물건을 많이 실은 쇼핑카트가 있으며, 그 오른편에는 식료품가게가 있다고 가정하여, 유모차로 돌진하면 아이가 죽게 되고, 쇼핑카트를 택하면 자율주행자동차가 손상되며, 식료품 가게로 돌진하면 자율주행자동차와 탑승자가 중대한 손상을 입게 된다고 할 때, 자율주행자동차는 어느 쪽으로 진행해야 하는가라는 윤리적 난제이다(Gurney, p.195).

54) Gurney, pp.183-184; Goodall(2014a), p.61; 고인석, 104면. 아시모프는 1942년 "Runaround"에서 로봇 3원칙을 처음 제시하였고, 1988년 "Robots and Empire"에서 제0법칙("로봇은 인류(humanity)를 해치거나 혹은 부작위에 의해 인류에게 위험을 초래해서는 안 된다.")을 추가하여 이를 수정하였다(고인석, 101면 이하).

55) 아시모프의 원칙을 일반적 로봇윤리의 차원에서 재구성해보고자 하는 시도로는 Abney, p.43. 또한 고인석, 98면 이하도 이를 로봇을 설계, 제작, 관리, 사용하는 자의 윤리로 아시모프의 3원칙을 재구성하고자 한다.

56) 아시모프의 3원칙은 로봇을 독자적인 윤리의 주체로 삼아 로봇에게 의무를 부여한 것으로 소설적

율주행자동차 관련 연구자, 특히 공학자들은 이 아시모프의 3원칙을 선호한다. 단순명료함과 위계적 구조 때문이라고 판단된다.[57] 단순명료함으로 인해 코딩이 용이하고, 자율주행자동차는 사고 상황에서 서로 모순 혹은 대립하는 원칙들의 위계에 따라 우선순위를 판단하도록 할 수 있기 때문이다. 즉, 자율주행자동차로 하여금 동시에 실현할 수 없는 모순된 명령을 수행하라고 하는 것은 불가능하고 이를 해결하기 위해서는 일정한 제약조건 또는 명령들 사이의 위계가 필수적으로 요구되기 때문이다.[58]

아시모프의 3원칙을 자율주행자동차에 적용하는 대표적인 예로는 Gerdes와 Thornton의 3원칙(이하 "수정된 아시모프의 3원칙")과 Raul Rojas의 4원칙(이하 "자율주행자동차의 4원칙")을 들 수 있다.

우선, Gerdes와 Thornton은 아시모프의 3원칙을 재구성하여 현실적으로 실행가능한 윤리원칙으로서 다음과 같이 수정된 아시모프의 3원칙을 제시한다.[59]

[제1법칙] 자율주행자동차는 보행자나 자전거운전자와 충돌해서는 안 된다.
[제2법칙] 자율주행자동차는 제1법칙에 저촉되는 충돌을 피하기 위한 것이 아닌 한, 다른 자동차를 충돌해서는 안 된다.
[제3법칙] 자율주행자동차는 제1법칙과 제2법칙에 저촉되는 충돌을 피하기 위한 것이 아니 한, 주위의 다른 물체와 충돌해서는 안 된다.

Gerdes와 Thornton은 인간의 생명에 우선권을 부여하고 이를 지킬 로봇의 의무를 규정하였다는 점에서 아시모프의 3원칙의 장점을 찾는다.[60] 또한 이 규칙은 느슨한 척도에 의하여 충돌대상만을 유형화하고 정교한 손상의 계산을 포함하지 않는데, 이 정도가 외부물체가 완벽하게 식별되지 않는 현재의 인식기술 수준에 비추어 실행가능한 수준이라는 점을 고려한 것이라고 한다.[61]

상상력에 기초한 것이어서 로봇윤리의 근거로 되기는 어렵다는 평가로는 월러치·알렌, 158-159면; 변순용·송선영, 64면. 한편, 월러치·알렌, 159면에서는 아시모프의 3원칙이 '도덕적 행위자'에게는 적합한 이론이 아니라는 점에서 이 이론이 부적절하다고 본다.

57) 예컨대, Gerdes·Thornton, p.96. 한편, 이는 아시모프의 3원칙의 단점으로 지적되기도 한다. 즉, 아시모프의 3원칙이 지나치게 추상적이라거나(Goodall(2014a), p.62), 규칙들이 애매하다(ambiguity)는 것(Gurney, p.184)이다.

58) Gerdes·Thornton, pp.94-95.

59) Gerdes·Thornton, p.96.

60) Gerdes·Thornton, p.95.

61) Gerdes·Thornton, p.96.

또한 Raul Rojas는 아시모프의 3원칙에서 고려되지 않았던 법적 요소인 교통법규를 제2법칙으로 추가하여 자율주행자동차의 4원칙을 제시한다.[62]

[제1법칙] 자율주행자동차는 인간을 해치거나 혹은 부작위에 의해 인간에게 위험을 초래해서는 안 된다.
[제2법칙] 자율주행자동차는 제1법칙에 저촉되지 않는 한, 교통법규를 준수하여야 한다.
[제3법칙] 자율주행자동차는 인간이 내린 명령이 제1법칙 혹은 제2법칙에 저촉되지 않는 한, 그 명령에 복종해야 한다.
[제4법칙] 자율주행자동차는 제1법칙, 제2법칙 혹은 제3법칙에 저촉되지 않는 한, 자신을 보호해야 한다.

아시모프의 3원칙은 서로 다른 인간이 모순되는 명령을 내리는 경우와 같이,[63] 동일한 위계 내에서의 우선순위를 고려하지 못하는 난점이 있다. 이는 수정된 아시모프의 3원칙, 자율주행자동차의 4원칙의 경우에도 유사하다. 윤리적 난제에 관한 이 입장의 결론은 이를 잘 보여준다.

[1] 전차 사례[①-1], 헬멧 사례[①-2]와 같은 인간 피해자 대 인간 피해자와 같이 동일한 위계 내의 사례(① 유형)에서는 자율주행자동차는 선택을 하지 못하게 된다.
[2] 교각 사례[②-1], 터널 사례[②-2]의 경우, 보행자를 우선하는 수정된 아시모프의 3원칙의 제1법칙에 의하면 보행자를 피하기 위하여 탑승자를 희생하게 된다. 반면, 자율주행자동차의 4원칙에 의하면, 인간 대 인간의 관계이므로 선택을 하지 못하게 된다.
[3] 자동차 대 자동차 사례[②-3]에서는 인간 탑승자의 존재 여부에 따라 자율주행자동차의 선택이 달라질 수 있다. 즉, 사람이 타고 있는 경우에는 제1법칙의 적용대상이 되지만, 그렇지 않은 경우 수정된 아시모프의 3원칙에서는 제2법칙에 의하면 동일한 위계 내의 경우로 선택이 불가능하다.

Gerdes와 Thornton의 수정된 아시모프의 3원칙은 보행자·자전거운전자, 다른 자동차, 기타의 문제라는 순서로 위계를 정하고 있다는 점은 주목할 만하다. 그러나 동일한 서열의 갈등상황에서는 그와 별도의 추가적인 변수를 고려하여 가중치를 부여하지 않고서는 문제

62) Raul Rojas, "I, Car: The Four Laws of Robotic Cars."(http://www.inf.fu-berlin.de/inst/ag-ki/rojas_home/documents/tutorials/I-Car-Laws.pdf, 최종접속일: 2016.5.31.)[이하 "Raul Rojas"]
63) 모순되는 명령의 가능성에 대해서는 고인석, 102면 이하.

를 해결하기 힘들다.

또한 Rojas의 자율주행자동차의 4원칙, 특히 제2법칙은 자율주행자동차가 고립적으로 존재하는 것이 아니라 도로교통이라는 환경을 규율하는 법질서, 교통법규의 틀 내에서 운행하는 것임을 고려한 것이다. 이는 교통법규 하에서 운행하는 현실을 고려한 착상이다. 그러나 교통법규라는 일반적 법규범은 다른 제1·3·4법칙과 비교할 때 그 명령의 성질이 이질적이라는 문제가 있다. 즉, 인간의 보호-교통법규-인간의 명령-자신의 보호로 연결되는 위계질서를 예정하고 있으나, 예컨대 제1·3·4법칙 자체가 일종의 교통법규로 이해되는 경우와 같이, 내부적인 원칙이 서로 충돌할 수 있다. 따라서 교통법규와 순수한 로봇윤리로서의 제1·3·4법칙의 관계는 별도로 설정될 필요가 있다.

(2) 해악 최소화 알고리듬(공리주의적 접근)

한편, 공리주의적 입장에서 자율주행자동차의 사고 방지와 관련한 프로그램으로는 충돌최적화 알고리듬(crash-optimization algorithm)이 논의된다. 여기에는 ① 자기보존적 알고리듬(self-preservation algorithm), ② 이타적 내지 자기희생적 알고리듬(altruistic/self-sacrifice algorithm), ③ 해악 최소화 알고리듬(harm-minimizing algorithm) 등 3가지를 들 수 있다.[64] 그런데 극도의 자기보존적 알고리듬은 자율주행자동차와 탑승자의 안전을 극단적으로 보호하는 결과, 제3자의 피해를 과도하게 초래할 수 있어 사회적으로, 윤리적으로 수용되기 어려운 결과를 초래할 수 있다.[65] 한편, 이타적 알고리듬을 장착한 경우는 자기보존적 알고리듬보다는 윤리적이라고 할 수 있다.[66] 그러나 이는 자신의 안전을 선호하는 소비자의 외면을 받기 쉬워 제작자가 이를 선뜻 받아들이기는 어렵다.[67] 따라서 두 가지 모두 현실적인 알고리듬으로 채택되기를 어렵기 때문에 현실적으로는 해악 최소화 알고리듬만이 고려대상이 된다.[68]

윤리적 난제는 해악 최소화 알고리듬에 따르면 다음과 같이 해결된다. 여기에서는 해악을 구체적으로 산정하는 방법에 따라 그 결론이 달라짐을 볼 수 있다.

---

64) 이 3가지의 알고리듬 유형은 Gurney, p.195 이하; Lin(2016), p.72 등의 설명을 종합한 것이다. 다만 Lin(2016), p.72는 3가지 유형은 3가지 유형 중 주로 자기보존적 알고리듬과 이타적 알고리듬을 대비시켜 설명한다.

65) Gurney, p.196.

66) Lin(2016), p.72.

67) Gurney, p.203.

68) 한편, 온건한 자기보호 알고리듬의 경우도 생각해볼 수 있으나, 이 경우 자율주행자동차와 그 탑승자의 안전도 동일한 차원에서 고려하여야 할 이익의 하나가 되므로 실질에 있어서 해악 최소화 알고리듬과 같이 검토될 것이다.

[1] 전차 사례[①-1], 교각 사례[②-1], 터널 사례[②-2] 등에서는 우선 최대다수의 최대행복이라는 기준을 적용하여 더 적은 수의 사람이 희생되도록 선택한다.

[2] 헬멧 사례[①-2]의 경우, 교통사고로 인한 피해를 최소화하기 위해 헬멧 쓴 운전자를 피해자로 선택한다.

[3] 자동차 대 자동차 사례[②-3]에서는 금전적인 손해배상의 측면을 고려하면, 안전도가 높은 차와 충돌하거나 가격이 낮은 차량과 충돌하여야 하나, 이는 반대로 안전도가 높은 차량과 충돌하는 것은 자율주행자동차 탑승자의 안전을 저해할 수 있다.[69] 한편, 얼굴 인식 시스템에 의하여 피해자의 수를 인식할 수 있다면 탑승자가 적은 피해차량과 충돌하는 등, 사고와 관련된 각 자동차에서의 탑승자의 인원수, 안전벨트의 착용 여부와 같은 인구통계학적 정보를 고려하여 판단하게 될 것이다.[70]

해악 최소화 알고리듬은 많은 경우 현실적이고 유용한 해결책을 제시해준다. 사고의 회피와 사고 발생, 경미한 재산상의 손해와 중대한 신체손상 등이 대립하는 경우에는 그 선택은 자명하다. 그러나 문제는 해악에 대한 정의에 따라 그 계산방법이 달라지고 때로는 그 결과가 윤리적 수용되기 어려운 경우도 있다는 데에 있다.[71] 예컨대, 헬멧 사례[①-2]에서 헬멧 쓴 운전자를 피해자로 선택하는 것은 교통법규를 잘 지킨 사람이 오히려 공격의 대상이 되는 불공정성이 있고, 또한 이는 법을 준수하여 헬멧을 써야 할 사회적 동기를 약화시킨다는 점에서 문제가 있다.[72] 따라서 공리주의적 접근이 자율주행자동차의 로봇윤리로 채택되기 위해서는 해악의 정의나 크기에 대한 가치서열 등이 별도로 정해질 필요가 있다.

3. 소결: 수정된 아시모프의 3원칙과 해악 최소화 알고리듬의 결합

수정된 아시모프의 3원칙과 해악 최소화 알고리듬은 모두 장점과 단점을 지닌 이론들이다. 따라서 하향식 접근법의 경우에는 위계를 고려한 의무론적 구성과 해악 최소화 알고리듬으로 표현되는 공리주의의 결합에 의하는 것이 바람직하다. 수정된 아시모프의 3원칙은 서로 다른 이익들간의 서열 내지 위계구조를 보여주고 있다는 점에서 유용하다. 그러나 동

69) 이는 자기보존적 알고리듬에 의한 것과 결과가 같다. 즉, 자신의 안전이 더 잘 보장되는 차량(예컨대 가벼운 차량)을 선택할 것이나 이타적 알고리듬에 의하면, 반대편 차량을 택하게 될 것이기 때문이다(Lin(2016), p.72).

70) Goodall(2014a), p.62; Gurney, pp.201-202.

71) Goodall(2014a), p.62. 특히 공리주의가 개인의 권리를 고려하지 않는다는 비판(Goodall(2014b), p.97)이 여기에도 그대로 적용된다.

72) Gurney, p.198.

일 서열 내부에서는 우선순위를 판단하지 못하는 난점이 있다. 한편, 해악 최소화 알고리듬에서의 해악 기준은 동일한 법익 내지 이익의 경우에 해악의 다과를 통해 실용적인 기준을 제시해 줄 수 있다는 장점이 있다. 그러나 해악의 의미와 기준 자체가 불명확할 수 있고, 이를 위해서는 별도의 기준이 필요하다는 한계가 있다.

조금 더 구체적으로 보자면, 수정된 아시모프의 3원칙은 이익 또는 보호대상의 위계구조를 보행자·자전거운전자, 다른 자동차, 기타의 문제 순서로 정하고 있으나, 동일 위계 내에서의 가중치 등의 문제는 공리주의적 요소 등 다른 기준이 가미되어야 해결 가능하다. 또한 해악 최소화 알고리듬은 그 내부에 문제가 되는 이익들, 생명, 신체의 안전, 재산 등에 대한 규범적 기준을 고려하지 않고서는 선택이 불가능한 상황에 봉착할 가능성이 크다. 따라서 수정된 아시모프의 3원칙의 위계구조를 바탕으로 해악 최소화 알고리듬을 결합하는 방식으로 자율주행자동차의 로봇윤리를 구성하는 것이 바람직하다.

한편, 자율주행자동차의 4원칙 중 제2법칙은 자율주행자동차가 실제로 운행되는 상황에서는 교통법규의 틀 내에서 운행하여야 함을 보여주는 장점이 있으나, 다른 제1·3·4법칙과 교통법규라는 이질적인 요소의 관계가 분명하지 않은 이론적 난점이 있다. 이 난점은 수정된 아시모프의 3원칙과 해악 최소화 알고리듬의 결합에 의한 로봇윤리의 구성에도 불구하고 여전히 존재하는 것이다. 이는 로봇윤리와 법적 규율의 관계에서 비롯되는 근본적인 문제점을 시사하는 것이기 때문이다. 이에 관하여는 장을 바꾸어 살펴보기로 한다.

Ⅳ. 자율주행자동차의 로봇윤리와 그 법적 시사점

1. 자율주행자동차의 로봇윤리와 법적 규율의 관계

(1) 잠정적 기준으로서 자율주행자동차의 로봇윤리

수정된 아시모프의 3원칙과 해악 최소화 알고리듬을 결합하여 구성하는 로봇윤리로 어느 정도 안정적인 로봇윤리의 기준을 수립할 수 있을 것이다. 그러나 이에 의하더라도 자율주행자동차와 관련된 윤리적인 문제가 망라적·완결적으로 해결되는 것은 아니다.

첫째, 개별적인 자율주행자동차의 안전이 곧바로 자율주행자동차가 속해 있는 도로교통체계의 안전을 의미하지는 않는다는 점이다. 즉, 개별 자율주행자동차의 윤리적 의사결정

이 반드시 안전을 확보해주지 못한다. 로봇윤리는 전체적인 교통 흐름과 체계라는 거시적 틀을 고려하여야 하며, 도로교통의 중층적 이해관계를 고려하지 않고서 로봇윤리가 완전한 의미를 지니기 어렵다. Rojas의 자율주행자동차 4원칙 중 제2원칙이 주목되는 이유가 여기에 있음은 앞서 본 바와 같다.

둘째, 자율주행자동차는 현실적으로 제한된 정보를 토대로 의사결정할 수밖에 없다는 점이다. 앞서의 자율주행자동차의 윤리적 난제들은 대체로 자율주행자동차가 완전한 정보를 가지고 있거나 혹은 가질 수 있다는 점을 암묵적으로 전제한 것이다. 그러나 앞으로 기술의 비약적인 발전이 있다고 하여도 도로에서 발생하는 모든 상황을 완벽하게 예상한 대응체계가 갖추어지려면 상당기간이 소요될 것으로 예상된다.

셋째, 자율주행자동차의 발전속도를 예상할 때, 인간 운전자에 의한 운행이 완전히 자율주행자동차로 대체되려면 앞으로도 상당한 시간이 필요할 것이라는 점이다. 이는 자율주행자동차의 로봇윤리는 로봇과 인간 모두가 도로교통에 관여하는 동안에는 양자 모두의 이해관계가 관련된다는 말이 된다. 따라서 그러한 동안에는 로봇윤리만으로 자율주행자동차가 도로교통 상황에 대응하는 데에는 일정한 제약이 따를 수밖에 없다.

이상과 같은 이유에서 자율주행자동차의 로봇윤리는 한계를 지닌 잠정적인 기준이 되고, 따라서 로봇윤리가 상당한 수준으로 정교하게 구성되더라도 실제로 발생하는 사고에 대한 법적 규율과는 완전히 일치할 수는 없게 된다.

(2) 자율주행자동차의 로봇윤리와 법적 규율의 분리

한편, 어떠한 윤리이론, 로봇윤리에 의하여 윤리적으로 정당하다고 판단되더라도, 법적 책임의 가능성은 여전히 열려 있다. 윤리적 판단에 따른 결과라고 하여도 법적 책임이 당연히 면책되는 것은 아니며, 판단이 윤리적이어도 그에 대한 법적 책임은 별도로 문제될 수 있다.[73] 윤리이론에 따른 결과가 법적 책임과 일치하지 않는 한, 윤리적 규범과 법적 규범 사이의 얼마간의 간극은 불가피할 뿐 아니라 자연스럽기까지 한 것이다.

첫째, 자율주행자동차의 로봇윤리는 그 목적을 1차적으로 해당 자율주행자동차의 사고 방지나 안전한 운행에 두지만, 도로교통에 관한 법적 규율은 도로교통의 원활과 안전이라는 공익적 관점에서 출발한다.[74] 자율주행자동차를 설계, 제작하는 사람들에게 가장 고민

73) 반대의 경우도 마찬가지이다. 즉, 법적 책임이 문제되지 않는다는 것이 윤리적 문제를 완전히 해결하는 것도 아니다(Goodall(2014B), p.97).

74) 도로교통법제는 도로의 안전, 원활한 도로교통의 보장 또는 편의 등과 같은 목적을 추구하는 경향이 있다(Gerdes·Thornton, p.97)

되는 것은 자신의 설계, 제작하는 자율주행 프로그램의 실행결과에 대하여 얼마만큼의 책임을 부담하게 되는가를 충분히 예측할 수 없다는 점이다. 그 원인을 법규범의 불명확성과 낮은 예측가능성에서 찾고 있지만,[75] 더 근본적으로 양자의 목적과 수단, 작동방식 등이 동일하지 않다는 점에 그 원인이 있다. 사적인 영역에서의 손해배상 책임 등은 생명, 신체의 안전, 재산과 관련된 부분에서는 양자가 일치하는 경우도 있으나, 언제나 그리고 당연히 양자가 일치하는 것은 아니다.

둘째, 자율주행자동차의 기술적·물리적 한계에 의해서도 윤리적 판단과 법적 판단이 달라질 수 있다. 사건 발생 이전의 설계 시점에서 그리고 사건 발생 당시 자율주행자동차의 판단시점에서 제한된 정보에 의하여 판단할 수밖에 없기 때문이다. 완전하지 못한 정보에 의한 판단의 결과는 사후적으로 충분한 논의와 심사를 거쳐 법적 책임으로 최종 확정된다.[76]

셋째, 자율주행자동차의 주행은 인공지능의 합리적 판단만으로 모든 것이 결정되지는 않는다. 물리적 실체로서 자동차는 물리법칙의 지배를 받기 때문에 제동거리, 회전반경, 관성 등 다양한 요소가 개입하여 사고와 연결될 수 있다.[77] 이 경우의 법적 책임문제는 자율주행 프로그램 이외의 다른 요소에 의해 영향을 받는다.

2. 자율주행자동차의 로봇윤리와 그 법적 시사점

자율주행자동차의 로봇윤리는 일정한 한계를 지닌 잠정적인 개념이며, 그것은 법적 규율과는 이론적, 실제적으로 분리되어 있다. 그럼에도 불구하고 윤리적 판단의 문제는 법제도에 영향을 미치는 중요한 요소의 하나이다. 또한 윤리적 정당성은 법적인 차원의 면책가능성을 부여할 수도 있다. 예컨대, 일정한 사회적, 윤리적으로 승인될 수 있는 우선순위에 따른 선택은 법적 절차에서도 면책될 가능성이 커진다. 자율주행자동차의 로봇윤리가 법체계와의 상호작용을 통하여 법체계에 주는 영향 내지 법적 시사점은 여기에 한하지 않는다.

첫째, 자율주행 프로그램의 의사결정은 사회적 가치평가, 특히 헌법적 가치평가를 통해

75) 예컨대, Gerdes·Thornton, p.87 이하의 서술을 보라.
76) 다만 이에 대응하기 위한 법기술이 어느 정도 마련되어 있다. 형법상 위법성이나 책임의 조각에 관한 법리들은 이와 같은 상황에 대처하기 위한 것이다. 민사적으로도 손해배상과 관련하여 유사한 법리가 사용되고 있다.
77) Lin(2016), p.81.

승인될 수 있는 것이어야 하므로 이에 관한 충분한 논의가 그 개발단계에서부터 필요하다. 예컨대, 해악 최소화 알고리듬은 일종의 표적식별 알고리듬(targeting algorithm)이다. 이는 실질적으로 피해자를 표적으로 간주하고 이를 찾아내는 것이며, 손해나 사고를 최소화한다는 발상은 결국 손해를 최소화할 피해자를 고르는 선택을 의미한다.[78] 이에 대해서는 법적으로 차별금지나 평등권, 인간의 존엄과 가치 등의 헌법적 가치와의 충돌 문제가 제기될 수 있다.[79] 인간 운전자에 의한 교통사고와 달리, 자율주행자동차의 경우에는 사전에 면밀한 검토를 통해, 프로그래머가 계획적으로 특정 피해자 유형을 선택하여 그에 대하여 사고가 일어나도록 설정하기 때문이다.[80] 이 문제는 현재 잠재적인 문제제기에 그치고 있지만, 자율주행자동차로 인한 사고가 발생하는 경우 이에 대한 법적 판단 내지 해결방향이 자율주행자동차의 발전방향과 형태를 규정하게 될 가능성이 매우 높다. 따라서 향후 자율주행자동차 기술과 산업이 안정적으로 발전하기 위해서는 관련된 헌법적 쟁점을 자율주행자동차의 개발시점부터 충분히 검토할 필요가 있다.

둘째, 로봇윤리적 결정의 사회적 효과를 고려하는 법적 개입이 필요하다. 자율주행자동차의 특정 알고리듬, 사고시의 선호의 체계가 인간 운전자나 보행자 등에게 알려지게 되면, 그것이 시장의 선호를 통해 사회적으로 중요한 변화를 야기할 수 있다. 예컨대, 더 무거운 차량이 회피된다는 알고리듬이 알려진다면 그에 따라 무거운 차량의 판매가 늘어날 것이며,[81] 해악 최소화 알고리듬에서 특정한 조건의 운전자, 자동차가 사고위험이 있을 때 회피된다는 사실이 알려지면 그에 따른 인간의 선택이 일어날 것이다. 이 경우 법이 사회나 국가 전체의 공익적 관점에서 정한 정책적 방향에 따라 적극 개입하여 로봇윤리 관련 프로그램의 신설, 수정, 삭제 등을 요구할 수 있어야 한다. 한편, 인간 운전자가 특정 알고리듬을 역이용하여 자율주행자동차의 사고를 고의적으로 야기하고자 시도하는 경우[82]에 대해서도 적절히 법적으로 규제해야 도로교통의 안전을 확보할 수 있다.

셋째, 자율주행자동차의 경우 법적 규율의 요소가 로봇윤리를 구성하는 체계 내부에 구비되어 있어야 도로교통의 체계 속에서 안정적인 윤리적 선택이 가능하다. 이 경우 법적 규율은 자율주행 시스템의 수학적 제한조건으로 기능하게 될 것이다.[83] 이 제한조건이 되

78) 따라서 이에 대해서는 군사용 로봇의 표적식별 알고리듬에서 문제되는 것과 동일한 윤리적 문제가 따른다(Lin(2016), p.72).
79) Lin(2016), p.70.
80) Goodall(2014a), p.80.
81) Lin(2016), p.73.
82) Lin(2016), p.81.
83) Gerdes·Thornton, p.97.

는 법적 규율의 대표적인 예가 교통법규이다. 그런데 문제는 현행 법체계가 곧바로 인공지능이 인지할 수 있는 논리적 기준을 갖추지 못했다는 점이다.[84] 여기에서 향후 자율주행기술의 발전에 따라 도로교통법체계의 재구성, 즉 구조화와 위계화가 강력히 요구될 것이 예상된다. 또한 형식적인 위법성과 실질적인 위법성의 관계나 일반조항에 의한 책임 배제 등의 경우 그 구조화와 위계화의 요구가 특히 커질 것으로 본다. 이에 대한 법이론적인 대비가 필요함은 물론이다.

넷째, 기술의 발전에 부합한 법적 규제의 반영이 있어야 한다. 일반적으로는 기술발달에 뒤쳐진 법령 등이 문제이다. 과학기술의 발전 수준에 부합하지 않는 과거의 법령을 고수하는 것은 자율주행자동차 산업의 발전을 저해할 수도 있기 때문이다. 따라서 자율주행자동차, 나아가 로봇 등 새로운 과학기술의 영역을 규율함에는 개개의 규정보다는 원리 위주로 접근하는 규제입법이 필요하다고 본다. 개개의 구체화된 개별 규제는 자율주행자동차의 자동화 정도에 관한 발전속도를 그대로 좇아가기가 어렵기 때문이다.[85] 한편, 반대의 경우도 생각해볼 수 있다. 기술발전 수준을 앞서는 성급한 법적 규제, 즉 현재의 기술수준 혹은 자율주행자동차의 위험을 과도하게 평가하여 불필요한 규제를 선제적으로 시도하는 것도 기술발달이나 합리적인 사회발전에 장애요소가 될 수 있다.

Ⅴ. 나오며

이상에서는 자율주행자동차에 요구되는 로봇윤리를 검토하여 그로부터 법적 시사점을 모색해 보았다. 우선, 자율주행자동차의 로봇윤리는 하향식 접근법을 기반으로 하여 그 부족한 점을 상향식 접근법으로 보완하는 혼합형 접근법에 의하는 것이 적합하다고 보았다. 자율주행자동차에 맞추어 수정된 아시모프의 3원칙과 같은 의무론적 요소와 해악 최소화 알고리듬으로 표현되는 공리주의적 발상의 결합이 필요하며, 이를 기반으로 기계학습과 같은 상향식 접근이 보완되어야 한다고 보았다.

다음으로, 자율주행자동차의 로봇윤리의 한계 및 법적 규율과의 관계를 검토하여 다음과 같은 법적 시사점을 도출하였다. 첫째, 자율주행자동차의 의사결정 프로그램은 생명, 신

84) Goodall(2014b), p.97.
85) Veruggio·Abney, p.359.

체의 안전이라는 법적 이익을 통해 직접 침해하므로 헌법적 적합성의 검토가 개발단계에서부터 필요하다. 둘째, 다수의 주체가 관여하는 도로교통의 특성을 고려하여 로봇윤리를 구성할 때 교통법규가 그 내부에 편입되어야 하는데, 이를 위하여 관련 법령의 구조화와 위계화가 필요하다. 셋째, 로봇윤리에 따른 자율주행자동차의 선호의 체계가 공익을 저해하지 않도록 조정하는 법적 노력이 필요하다. 넷째, 기술발달과 병행하는 규제적 법제의 정비를 위해서는 원리 위주의 원리 위주의 규제입법이 필요하다.

자율주행자동차의 등장은 우마차에서 자동차로 전환된 것 이상의 사회적 변화를 초래할 것이다.[86] 그 변화는 우리의 예측보다는 더 깊고 더 근본적일 수 있다. 따라서 그로 인해 초래되는 사회적 변화를 통제가능한 범위 내로 유지하도록 미리 대비할 필요가 있다. 자율주행자동차와 관련하여 로봇윤리를 검토한 이유도 여기에 있다. 또한 자율주행자동차와 관련해서는 법적 차원에서도 기존의 법체계를 단순 재구성하는 수준을 넘는 근본적 변화를 요구할 것으로 예상된다. 향후 이에 관한 연구와 대비도 필요하다.

86) Lin(2016), p.81.

■ 참고문헌

변순용, 송선영,『로봇윤리란 무엇인가?』, 어문학사, 2015.

웬델 월러치, 콜린 알렌/노태복 옮김,『왜 로봇의 도덕인가』, 메디치, 2014.

Gianmarco Veruggio, Keith Abney, "Roboethics: The Applied Ethics for a New Science," Patrick Lin, Keith Abney, George A. Bekey (ed.), Robot Ethics: The Ethical and Social Implications of Robotics, The MIT Press[이하 "Robot Ethics"], 2012, pp.347-363.

J. Christian Gerdes, Sarah M. Thornton, "Implementable Ethics for Autonomous Vehicles," Markus Maurer, J. Christian Gerdes, Barbara Lenz, Hermann Winner (ed.), Autonomous Driving[이하 "Autonomous Driving"], 2016, pp.87-102.

Jeffrey K. Gurney, "Crushing into the Unknown: An Examination of Crush-optimization Algorithms through the Two Lanes of Ethics and Law," 79 Albany Law Review, 2015/2016, pp.183-266.

Keith Abney, "Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed," Robot Ethics, 2012, pp.35-54.

Noah J. Goodall, "Ethical Decision Making During Automated Vehicle Crashes," Transportation Research Record: Journal of the Transportation Research Board, No. 2424, Transportation Research Board of the National Academies, 2014, pp.58 – 65.

_____, "Machine Ethics and Automated Vehicle," Gereon Meyer, Sven Beiker (ed.), Road Vehicle Automation, Springer, 2014, pp.93-102.

Patrick Lin, "Introduction to Robot Ethics," Robot Ethics, 2012, pp.3-15.

_____, "Why Ethics Matters for Autonomous Cars," Autonomous Driving, 2016, pp.69-85.

고인석, "아시모프의 로봇 3법칙 다시 보기: 윤리적인 로봇 만들기," 철학연구회,『철학연구』제93집, 2011, 97-120면.

이중기, 황창근, "자율주행자동차 운행에 대비한 책임법제와 책임보험제도의 정비필요성: 소프트웨어의 흠결, 설계상 흠결 문제를 중심으로," 한국금융법학회,『금융법연구』제13권 제1호, 2016, 94-122면.

이종영, 김정임, "자율주행자동차 운행의 법적 문제," 중앙법학회,『중앙법학』제17집 제2호, 2015.6, 145-184면.

NHTSA, U.S. Department of Transportation Releases Policy on Automated Vehicle Development, 2013.5.30.(http://www.nhtsa.gov/About+NHTSA/Press+Releases/U.S.+Department+of+Transportation+Releases+Policy+on+Automated+Vehicle+Development, 최종접속일: 2016.5.31.)

Raul Rojas, "I, Car: The Four Laws of Robotic Cars."(http://www.inf.fu-berlin.de/inst/ag-ki/rojas_home/documents/tutorials/I-Car-Laws.pdf, 최종접속일: 2016.5.31.)

http://www.bast.de/DE/Publikationen/Foko/2013-2012/2012-11.html(최종접속일: 2016.5.31.)

http://www.sae.org/misc/pdfs/automated_driving.pdf(최종접속일: 2016.5.31.)

투고일자 2016. 05. 31 심사개시일자 2016. 06. 03 게재확정일자 2016. 06. 20

Hongik Law Review, Vol. 17, No. 2(2016)

【ABSTRACT】

The Robot Ethics of the Autonomous Vehicle and its Legal Implications

LEE, Choong-Kee
OH, Byung Doo

This paper deals with the legal implications of the robot ethics of the autonomous vehicles. It is proposed that in developing robot ethics for the autonomous vehicle an incremental, hybrid approach should be adopted, because driving cars is related to the safety of road traffic and the interests of other people. At first, it should be formed through top-down approach by combining of the two ethical theories: modified Asimov's Three Laws of Robotics and harm-minimizing algorithm. The former, modified from the Asimov's Three Laws of Robotics and adjusted to the autonomous vehicle, is partly suitable for the self-driving program due to its straightforwardness, hierarchical structure and emphasis on the rights of humans. And the latter, a program based on the utilitarianism, is also partly acceptable, because it makes the program in self-driving cars possible the calculation of the harms in case of the imminent accident. And then, upon this top-down approach, the information on the situation of roads, driving culture and other technological considerations should be supplemented by bottom-up approach, such as machine learning.

In conclusion, the followings are suggested as the legal implications of the robot ethics

of the autonomous vehicles: Firstly, the harm-minimizing algorithm should be scrutinized beforehand that it could be approved by the constitutional values, because in some cases it may be construed as the targeting and discriminating choices among human victims. Secondly, the laws, especially traffic laws should make up for the lacunas, that inevitably exist in robot ethics of the autonomous vehicles. Thirdly, the potential social preferences that may be caused through the priorities shown by the self-driving program should be controlled in view of the public interests. Finally, the principle-based regulations would be better than the specific and individual regulations as the laws that govern the autonomous vehicles, in order that they may always keep pace with the development of technologies.

Key Words : Robot Ethics, Autonomous Vehicle, Artificial Intelligence, Ethical Dilemma, Top-down Approach, Bottom-up Approach, Hybrid Approach, Modified Asimov's Three Laws of Robotics, Harm-minimizing Algorithm