JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 20, No. 11, pp. 2237-2242, Nov. 2019

인공적 도덕행위자(AMA)의 온톨로지 구축

박 균 열
경상대학교 윤리교육과

Establishment of Ontology for Artificial Moral Agent

Gyun-Yeol Park
Department of Ethics Education, Gyeongsang National University, Jinju 52828, Korea

[요 약]

본 연구는 인공지능이 도덕적 행위자로서의 지위를 갖추기 위해 필요한 온톨로지를 구축하기 위한 기본 방향을 제시하는 데 목적을 둔다. 인공지능 기술이 날로 발전함에 따라서 인공지능의 도덕적, 법적 사안에 대한 논의가 심각하게 제기되고 있다. 법적 문제의 경우 규정 형태로 명시되어 있기 때문에 요목화하는 데 큰 어려움이 없지만 도덕적 사안에 대해서는 많은 난제가 기다리고 있다. 인공지능 자체에 도덕적 기능을 부여할 것인지, 만약 한다면 어느 수준으로 할 것인지, 인공지능을 도덕적 주체로 인정할 경우 사용자와 개발자, 그리고 관련 정책입안자들은 어떤 태도를 지향해야 하는지 등에 대한 논의를 해야 한다. 그중에서 이 연구는 인공지능 자체에 도덕적 기능을 부여할 수 있다고 가정하고, 어떻게 그 기능을 부여할 것인지에 대한 기본적인 골격을 제시하고자 한다. 이를 위해 본 연구는 인공지능의 윤리적 의사결정 메커니즘 모델과 관련한 이론을 고찰하고, 이를 위한 기본 알고리즘을 제안한다. 또한 이 연구는 이것이 갖는 윤리적 시사점을 제기함으로써 향후 보다 복합적이고 설명 가능한 인공적 도덕행위자의 온톨로지 정립에 기초자료를 제공할 수 있다.

[Abstract]

This study aims to provide a basic direction for the ontology of artificial intelligence which is needed to attain its status as a moral agent. With the ever-evolving artificial intelligence technology, discussions on the moral and legal issues of artificial intelligence are being seriously raised. As for legal matters, it is stated in the form of codes, so there are no major difficulties in making segmentations, but many challenges lie ahead on the moral issues. If the artificial intelligence itself is given moral functions, if at what level it will do, and if it is recognized as a moral subject, users, developers and relevant policymakers should discuss what attitude to pursue. The study, among other issues, assumes that the artificial intelligence itself can be given a moral function, and tries to present a basic framework for how to give it the moral function. For that purpose, this study reviewed the theories related to the ethical decision-making mechanism model of the artificial intelligence, and proposed a basic algorithm for the model. This study can also provide a basis for the establishment of a more complicated and explainable artificial moral agent's ontology in the future by bringing up its ethical implications.

색인어 : 인공지능, 도덕적 행위자, 온톨로지, 알고리즘, 도덕판단, 설명가능한 인공적 도덕행위자
Key word : Artificial Intelligence, Moral Agent, Ontology, Algorithm, Moral Judgment, Explainable Artificial Moral Agent

http://dx.doi.org/10.9728/dcs.2019.20.11.2237

Received 23 September 2019; **Revised** 22 October 2019
Accepted 05 November 2019

***Corresponding Author: Gyun-Yeol Park**
Tel: +※ 개인정보 표시제한
E-mail: pgy556@daum.net

Ⅰ. 서 론

어떤 일을 함에 있어서 종(從)을 잡는 일은 매우 중요하다. 마치 실을 잣는 일을 할 때 날줄(=經)을 매기는 일과 같이 기준점이 되는 일이기 때문이다. 인생살이에서 이와 같이 매우 중요한 가르침을 적어둔 책을 경서(經書)라고 얘기하는 이유는 바로 여기에 있다. 스스로 배우고 자신의 약점을 보완할 수 있는 인공지능을 구현하기 위해서는 이와 같은 종잡기가 시작에서부터 필수적으로 수반된다. 컴퓨터 소프트웨어를 개발하는 학자들은 컴퓨터에 있어서 이와 같은 역할을 온톨로지(Ontolgoy)라는 용어로 설명하고 있다.

보편적으로 온톨로지에 대한 정의로 가장 잘 수용되고 있는 것은 Gruber(1993)가 말한 “어떤 관심 분야를 개념화하기 위해 명시적으로 정형화한 명세서(an explicit and formal specification of a conceptualization of a domain of interest)”이다[1].

원래 온톨로지는 원래 철학의 한 분과로서 지혜의 진리(truth), 핵심(essence), 신(god) 등의 실체(entity)를 탐구하는 영역이다. 그런데 인공지능에 있어서 온톨로지는 사뭇 다르게 사용된다. 인공지능을 위한 온톨로지는 특정 분야를 기술하는 데이터 모델로서 특정한 분야(domain)에 속하는 개념과, 개념 사이의 관계를 기술하는 정형(formal) 어휘의 집합으로 이루어진다. 예를 들어 “종-속-과-목-강-문-계”로 분류되는 생물과 생물 사이의 분류학적 관계나 영어 단어 사이의 관계를 정형 언어(formal language)로 기술하면 각각 온톨로지라고 할 수 있다. 정형 언어의 집합인 온톨로지는 연역과 추론에 사용된다. 온토로지의 구성요소는 클래스(class), 인스턴스(instance), 관계(relation), 속성(property) 등이 있다[2].

인공지능이 제대로 작동하기 위해서는 이와 같은 네 가지의 구성요소가 잘 구축될 때 가능하다. 그런데 이러한 구성요소들은 기실 형식적인 요건이다. 인공지능이 인간과 유사한 형태의 도덕적 기능을 갖게 되어야만 하는가의 당위적 문제는 차지하고, 실제 인공지능이 도덕적 행위자 역할을 할 수 있을 때, 우리는 그것을 인공적 도덕행위자(Artificial Moral Agent)라고 말한다[3]. 이것이 인공지능과 관련한 가장 큰 난제이다. 인공지능과 관련된 국내외의 다양한 연구는 아시모프 원칙과 같은 설계상의 책임 문제, 사용상의 안전성 문제, 이들의 법령적 구속의 문제, 자격증 부여의 문제[4] 등에 국한되어 많은 논의를 해왔다.

자율학습이 가능한 정도의 범위를 넘어서서 인간과 유사한 인공지능을 구현하는 문제는 이전의 논의와는 전혀 차원이 다르다. 인공지능을 도덕적 행위자로 구현하기 위해서는 인간의 도덕판단역량을 구분할 수 있는 기준 설정이 매우 중요하다. 그렇다면 이렇게 중요한 일에 대해 많은 학자들은 연구를 하지 않았을까? 그것은 바로 가장 인문적인 윤리학과 가장 자연과학적인 통계·컴퓨터를 동시에 전문으로 하는 연구자들이 상대적으로 적었기 때문이다.

따라서 이 연구는 인공지능의 윤리를 ‘인공지능을 운용함에 있어서의 윤리’가 아니라 ‘인공 지능 자체를 어떻게 도덕적으로 만들 것인가’의 주제로 인식하고, 그 기본 구성원리를 설정하는 데 주된 목표를 두고자 한다.

Ⅱ. 인공지능(AI)의 윤리적 의사결정 메커니즘 모델

인공지능의 윤리적 의사결정 메커니즘을 살펴보는 것은 제한된 범위 내에서 인간과 유사한 기능을 수행하고 있거나 그렇게 될 개연성이 높기 때문이다. 현재까지의 대강의 인공지능에 대한 윤리적 접근은 ‘사용자’와 ‘개발자’의 책임 문제에 국한해서 얘기가 다루어졌다. 실제 인공지능은 그 자체가 도덕적 행위자(moral agent)이기 때문에 지금까지의 논의 선상에서 진일보해야 하는 과제를 안고 있다. 이러한 주제로 매우 제한된 선행연구를 살펴보자[5].

2-1 Ronald Arkin의 윤리적 자율로봇 아키텍처

아킨은 어떻게 로봇 전투기계를 전시 행동의 복잡한 윤리를 다룰 수 있도록 만드는가의 문제를 연구했다. 그가 제안한 아키텍처는 윤리에 대해 네 가지 특화된 구성요소를 갖고 있다. 첫째, 윤리적 통제 장치(ethical governor)이다. 이는 의무론적 논리를 바탕으로 허용 가능한 행동에 대한 엄격한 제약사항을 관리하기 위한 장치이다. 둘째, 윤리적 행동 제어 모듈(ethical behavior control)이다. 이는 특정한 전투 상황 개입에 관한 구체적인 군사적 규칙을 실행하고 아울러 허용 가능한 선택사항들 가운데서 선택을 행하도록 해주는 원칙들이 이 모듈 속에 내장돼 있다. 셋째, 윤리적 적응 장치(ethical adaptor)이다. 이는 실시간 행동 중에 감정 시스템을 관리하고 아울러 사실에 따른 성찰적 사고를 관리한다.

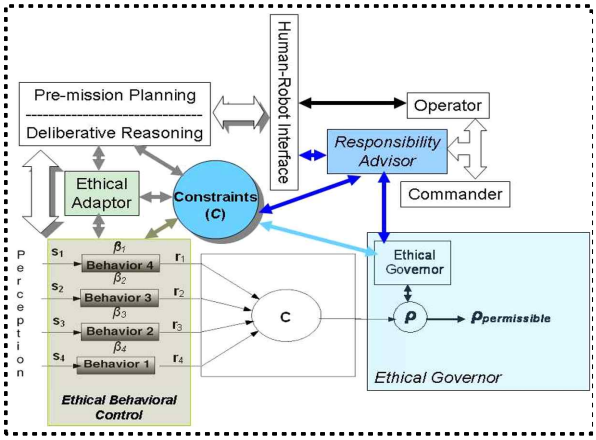

그림 1. 정보통신망 흐름도
Fig. 1. Flow Chart of Information & Communication Network

넷째, 책임 조언 장치(responsibility advisor)이다. 로봇과 인간 운영자 사이의 인터페이스 역할을 하는데, 살상력 사용 허가를 받고서 로봇을 자동화 전투 임무에 투입하는 것과 관련된 의미들을 인간 운영자가 적절히 고려하도록 보장해준다([그림 1] 참조)[6].

아킨의 연구에서 눈에 띄는 것은 계획, 추론, 적응, 통제, 지배, 간섭 등과 같은 인지작용의 조절 역량을 가장 중심적인 기능으로 여긴다는 것이다. 그의 연구에서 두 가지 점에서 아쉬운 점이 있다[6]. 한 가지 아쉬운 점은 정도의 문제를 고려하지 않고 있다는 점이고 또 한 가지 아쉬운 점은 피드백을 원활이 고려하지 않고 있다는 점이다. 이러한 점이 개선된다면 복합적인 추론과 감정적 상호 작용을 고려한 도덕적 행위자로서의 인공지능에 대한 윤리적 제어 메커니즘을 부여할 수 있을 것으로 사료된다.

2-2 Baars의 통합작업공간 이론

통합작업공간 이론(GWT: Global Workspace Theory)이란 일상생활 공간 전체를 인식의 공간 대상으로 전제하고 인식의 범위를 확장한다. 통합작업공간 이론을 주창한 Baars는 "당신이 인식하기 때문에 그래서 나도 존재한다"(You are conscious and so am I)라고 말한다[7][8][9][10]. 바아스의 주장은 불교의 연기설과 관련이 있어 보인다. 연기(緣起)라는 단어는 산스크리트어 프라티트야 삼무파다(प्रतीत्यसमुत्पाद pratītyasamutpāda)를 뜻에 따라 번역한 것으로 인연생기(因緣生起: 인과 연에 의지하여 생겨남, 인연따라 생겨남)의 준말이다. 한역(漢譯) 경전에서는 발랄저제야삼모파다(鉢剌底帝夜參牟播陀)로 음차하여 표기한 경우도 있다. 프라티트야(산스크리트어: pratītya)의 사전적인 뜻은 '의존하다'이고 삼무파다(samutpāda)의 사전적인 뜻은 '생겨나다', '발생하다'이다. 연기(緣起)를 영어권에서는 dependent arising(의존하여 생겨남), conditioned genesis(조건이 부여된 생성), dependent co-arising(의존된 상호발생) 또는 interdependent arising(상호의존하여 생겨남) 등으로 번역되고 있다. 『잡아함경』 제12권 제299경 「연기법경(緣起法經)」에서 고타마 붓다는 연기법(緣起法)은 자신이나 다른 깨달은 이[如來]가 만들어 낸 것이 아니며 법계(우주)에 본래부터 항상 존재하는[常住] 법칙[法]이라고 말하고 있다. 이들 여래(如來: 문자 그대로는 '진리[如]로부터 온[來] 자' 또는 '진리와 같아진[如] 후, 즉 진리와 하나가 된[如] 후, 즉 완전히 깨달은[如] 후 다른 사람들을 돕기 위해 세상으로 나온[來] 자')들은 이 우주 법칙을 완전히 깨달은 후에 다른 이들도 자신처럼 이 우주 법칙을 완전히 깨달을 수 있도록 돕기 위해 그것을 12연기설 등의 형태로, 즉 아직 완전한 깨닫지 못한 사람들도 이해할 수 있고 사용할 수 있는 형태로 세상에 드러낸 것일 뿐이라고 말한다[11].

GWT가 처음으로 주창된 것은 Alan Newell[12]과 Herbert A. Simon[13] 등과 같은 인지 모델링 학자들에 의해서였다. 삶의 장소란 데카르트 식의 낮잠과 같은 것이 아니라 영화관과 같이 실제 살아있는 것이다[14]. 심지어 현금계좌에 비유되기도 한다[15][16][17][18][19]. Baars는 이와 같은 논증을 발전시켰으며[20][21][22][23], David Chalmers와 같은 철학자에 의해 검증되기도 했다[24]. 이에 대해서는 국내에서도 연구가 이루어졌는데, 주로 경영학 분야에서 인사관리 관련 주제가 많다([그림 2] 참조).

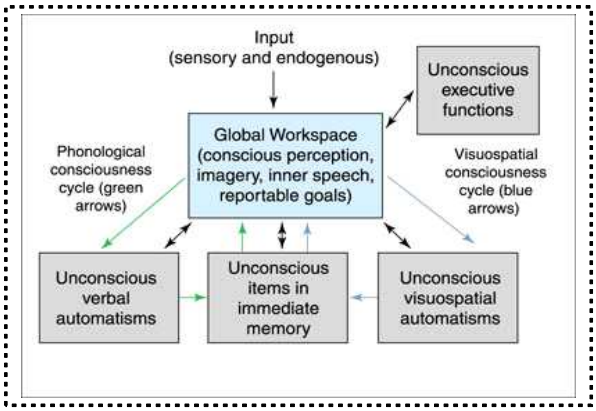

그림 2. Baars의 통합작업공간 이론
Fig. 2. Baars' Global Workplace Theory

2-3 S. Franklin의 학습형 지능적 분산 행위자

S. Franklin은 인공지능 연구를 수행하면서 학습형 지능적 분산 행위자(LIDA: Learning Intelligent Distribution Agent) 개발에 관해 관심을 가졌다[26]. LIDA는 특수 목적의 윤리 모듈을 갖지 않고 대신 더욱 일반적인 지각적, 정서적, 의사결정 요소들을 통해 윤리적 감수성과 사고 능력을 구현한다([그림 3] 참조).

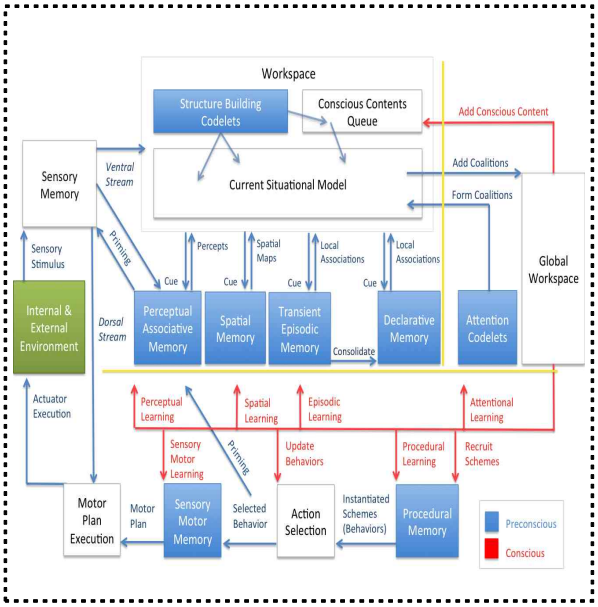

그림 3. Franklin의 학급형 지능적 분산 행위자 모형
Fig. 3. Learning Intelligent Distribution Agent Model

Ⅲ. AMA 구축을 위한 기본 알고리즘

인공적 도덕행위자(AMA)를 제작하기 위해서는 인간의 가치판단의 절차를 잘 요약해야 한다. 그 이외의 외생 변수에 대해서는 0과 1로만 응답할 수 있도록 분류하면 된다. 여기에 대해서는 기존의 언어학이나 빅데이터 분야의 경험이 많이 축적되어 있어서 큰 문제가 되지 않는다. 다만 인간의 가치판단을 이와 같이 0과 1로 구분될 수 있도록 하면 되는 것이다.

인간의 윤리적 판단은 인간적 정체성(personal identity)에서부터 도덕적 정체성(moral identity)으로 바뀌면서부터 시작된다. 인간적 정체성은 "나는 누구인가?"(Who am I?)라고 하는 질문에 부응하는 문제로서 머리카락의 색깔, 눈동자의 크기와 색깔, 말투, 키, 혈액형 등의 인간의 내외적인 구성을 함에 있어서 그 사람만이 갖고 있는 특징을 말한다. 도덕적 정체성은 "나는 누구여야만 하는가?"(Who should be I?)라는 질문에 대한 응답으로서 인간이라면 마땅히 지향해야 할 바를 말한다. 이러한 도덕적 정체성은 무엇을 대상으로, 어디를 대상으로 향하느냐에 따라서 구체화된다. 흔히 인간의 인지적, 정서적, 심동적 특징을 이성, 감정, 행동으로 3분화해서 말한다. 그러나 엄격히 보면 이성(logos)은 그 자체로 움직이지 않는다. 그런데 선현들은 이것 자체가 움직이는 것으로 보았다. 서양의 아리스토텔레스(Aristotle)는 이성의 성격을 가진 지혜가 움직이는 것으로, 즉 실천적 지혜(phronesis)라고 했고, 칸트 또한 이성에 동력을 부여하여 실천이성(Praktische Vernunft)이라고 명명했다. 중국 유교의 전통을 계승한 한국 전통의 퇴계(退溪) 이황(李滉) 선생도 이 문제에서 자유롭지 못하다. 그는 원리인 이(理)와 형체를 가진 기(氣)의 관계를 '이가 발하면 기가 그것을 따르고'(=理發氣隨之), '기가 먼저 발하면 이는 이미 그것을 타고 있다'(氣發理承之)는 용어를 사용하여 이와 기가 각기 발할 수 있다는 입장을 전개하고 있다. 기실 윤리적 판단의 작동은 그와 같은 방식에 의한 것이 아니라 스스로 움직일 수 없는 '이성'과 '감정'의 두 요소와 방향성은 없으면서 순수히 움직이기만 하는 '동력'이라는 각각의 영역을 소재로 삼고서, 그것을 잘 조절하고자 하는 인간의 의지, 즉 에토스(Ethos)에 의해 이루어진다. 그 요체는 '중요한 것과 덜 중요한 것, 천천히 해야 할 것과 급하게 해야 할 것'(=輕重緩急), '높은 가치를 여길 것인지 낮은 가치로 여길 것인지, 먼저 할 것인지 나중에 할 것인지'(=優劣先後), '큰 것도 잘 보면서 작은 것도 그에 상응해서 잘 살핀다'(=大觀小察)라고 할 수 있다. 이것인 어디서, 언제, 누구를 대상으로 발현되느냐에 따라 도덕적 판단은 구체화되고 그 결과가 행동으로 도출되어 원래의 인간적 정체성과 도덕적 정체성에 부합되면 강화되고 부합되지 않으면 약화되어 피드백되는 것이다([그림 4] 참조).

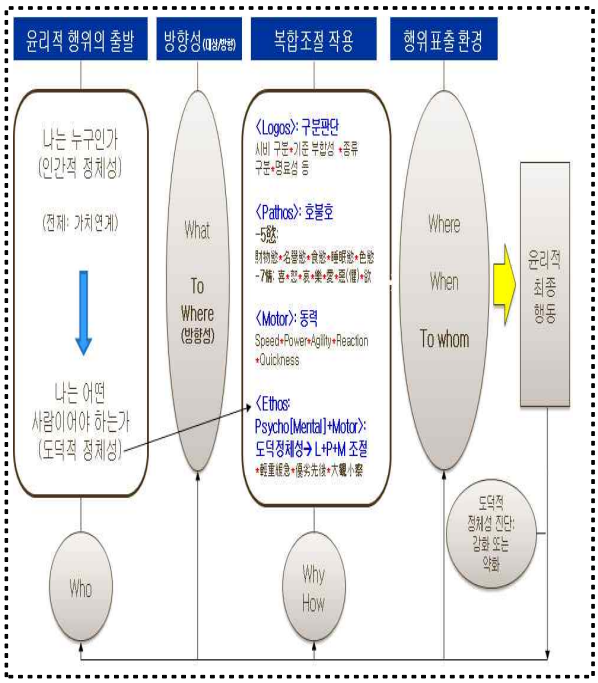

그림 4. 인공적 도덕행위자를 위한 온토로지 구축 기본 알고리즘
Fig. 4. Basic Algorithm for Ontology of Artificial Moral Agent

Ⅳ. 결론

본 연구는 인공지능이 도덕적 행위자로서의 지위를 갖추기 위해 필요한 온톨로지를 구축하기 위한 기본 방향을 제시하는 데 목적을 둔다. 인공지능 기술이 날로 발전함에 따라서 인공지능의 도덕적, 법적 사안에 대한 논의가 심각하게 제기되고 있다. 법적 문제의 경우 규정 형태로 명시되어 있기 때문에 요목화하는 데 큰 어려움이 없지만 도덕적 사안에 대해서는 많은 난제가 기다리고 있다. 인공지능 자체에 도덕적 기능을 부여할 것인지, 만약 한다는 어느 수준으로 할 것인지, 인공지능을 도덕적 주체로 인정할 경우 사용자와 개발자, 그리고 관련 정책입안자들은 어떤 태도를 지향해야 하는지 등에 대한 논의를 해야 한다. 그중에서 이 연구는 인공지능 자체에 도덕적 기능을 부여할 수 있다고 가정하고, 어떻게 그 기능을 부여할 것인지에 대한 기본적인 골격을 제시하고자 한다. 이를 위해 본 연구는 동서양의 철학적 논변들에 대해 간단하게 그 이론적 특징을 제시하고, 왜 그 이론들이 인공지능의 도덕적 기능을 부여하는 데 한계가 있는지를 지적하면서 그 대안적인 방향을 담은 인공적 도덕행위자를 구현하기 위한 온토로지 구축 기본 알고리즘을 제시했다. 이 연구는 완전하게 설명가능하고 완전하게 작동가능한 인공적 도덕행위자의 전형을 제시하는 것이 아니라 개괄적인 방향을 담고 있기 때문에 추후 제어계측 차원에서 정밀한 손질과 보완이 더 필요하다.

참고문헌

[1] Tomas R. Gruber, 1993, A Translation Approach to Portable Ontology Specifications, Knowledge Systems Laboratory, Stanford University : CA, Technical Report KSL 92-71, 1993

[2] https://ko.wikipedia.org/wiki/%EC%98%A8%ED%86%A8%EB%A1%9C%EC%A7%80 (2019.10.20.)

[3] W. Travis. “A Prospective Framework for the Design of Ideal Artificial Moral Agents : Insights from the Science of Heroism in Humans”, Minds & Machines, Vol. 25, No. 1, pp. 57-71, January 2015.

[4] B. J. Kim, A. Lang, A. Carass, and J. Prince, “Automated segmentation of mouse OCT volumes (ASiMOV): Validation & clinical study of a light damage model”, PLoS ONE , Vol. 12, No. 8, pp.1-17, August 2017.

[5] Gyun-Yeol Park, Public Officials Ethics Competence Model, National HRD Institute, 2017.

[6] Ronald C. Arkin, Governing lethal behavior: embedding ethics in a hybrid deliberative/reactive robot architecture, U.S. Army Research Office, 2008, p.62, https://www.semanticscholar.org/paper/Governing-lethal-behavior-embedding-ethics-in-a-hy-Arkin/9671c28381ae16666 ae64aa47b024fa34a01eed8/figure/11 (2019. 7. 24.)

[7] W. Wallach, and C. Allen, Translated by T. B. Noh, Why Robot?, Seoul: Medici Press, 2014.

[8] W. Wallach, and C. Allen, Translated by T. B. Noh, Why Robot?, Seoul: Medici Press, 2014.

[9] Consciousness Programing [Internet] Available http://www.neurdon.com/2009/07/18/

[10] B. J. Baars, “Theater of Consciousness: Global Workspace Theory, A Rigorous Scientific Theory of Consciousness”, Journal of Consciousness Studies, Vol. 4, No. 4, p. 4, December 1997.

[11] https://ko.wikipedia.org/wiki/%EC%97%B0%EA%B8%B0_(%EB%B6%88%EA%B5%90) (2019.10.20.)

[12] Alan Newell, Unified Theories of Cognition, Cambridge, MA: Harvard University Press, 1990.

[13] H. Simon, Reason in Human Affairs, CA: Stanford University Press, 1982.

[15] D. C. Dennett, and M. Kinsbourne, “Time and the observer: The where and when of consciousness in the brain”, Brain and Behavioral Sciences, Vol. 15, pp. 183–247, March 1992.

[16] A. R. Damasio, “Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition”, Cognition, Vol. 39, pp. 25–62, February 1989.

[17] M. S. Gazzaniga, “Brain mechanisms and conscious experience”, in Experimental and Theoretical Studies of Consciousness, CIBA Foundation 174, New York: Wiley, 1993.

[18] T. R. Shallice, "Information-processing models of consciousness: possibilities and problems", in A.J. Marcel & E. Bisiach(eds.), Consciousness in Contemporary Science, Oxford: Oxford University Press, 1988.

[19] M. Velmans, (ed.)., The Science of Consciousness: Psychological, Neuropsychological and Clinical Reviews, London: Routledge, 1996.

[20] B. J. Baars, “Conscious contents provide the nervous system with coherent, global information’, in R . Davidson, G. Schwartz, & D. Shapiro (eds.), Consciousness and Self-regulation, Vol. 3, ,New York: Plenum, 1983.

[21] B. J. Baars, A Cognitive Theory of Consciousness, New York: Cambridge University Press, 1988.

[22] B. J. Baars, “Theater of Consciousness”, Journal of Consciousness Studies, Vol. 4, No. 4, p. 44, 1997.

[23] B. J. Baars, In the Theater of Consciousness: The Workspace of the Mind, New York: Oxford University Press, 1997.

[24] D. J. Chalmers, The Conscious Mind: In Search of a Fundamental Theory, New York: Oxford University Press, 1996.

[25] Sean Lorenz, Programming a kinder, gentler conscious HAL, 2009; Bernard J. Baars, “In the Theater of Consciousness: Global Workspace Theory, A Rigorous Scientific Theory of Consciousness”, Journal of Consciousness Studies, 4(4) (1997), pp. 292-309.

[26] S. S. Lim., K. H. Oh, and S. B. Cho, "System Initiative Dialogue of Mixed Initiative Conversational Agent using Global Workspace Theory and Spreading Activation Theory", Proceeding of the Korea HCI Association Conference, 2010, Seoul, pp.1-20.

[27] http://mbscience.org/scientific-advisory-board-stan-franklin-lida-model-of-cognition/ (2019. 10. 13)

박균열(Gyun-Yeol Park)

1989년 : 경상대학교 윤리교육과(문학사)
1994년 : 서울대학교 윤리교육학과(교육학석사)
2000년 : 서울대학교 윤리교육학과(교육학박사)
2007년～현 재: 경상대학교 윤리교육과 교수
2017년～현 재: 대한변호사협회 검사평가특별위원회 위원
※관심분야 : 도덕성 측정, AI윤리, 정치윤리