



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

박 사 학 위 논 문

머신러닝 알고리즘과 차별



2018년 12월

이 준 일 교수지도

박 사 학 위 논 문

머신러닝 알고리즘과 차별

이 논문을 법학박사 학위논문으로 제출함.

2018년 10월

고려대학교 대학원

법 학 과

남 중 권




남중권의 법학박사 학위논문
심사를 완료함.

2018년 12월

위원장 이 준 일 

위원 윤 영 미 

위원 김 하 열 

위원 차 진 아 

위원 박 성 빈 



차례

제1장 서론	1
제1절 연구의 배경	1
I. 헌법 만능주의와 기술 만능주의	1
II. 알고리즘의 데이터 분석과 감시	3
III. 알고리즘 활용 기술의 발전과 그 파급력	5
제2절 연구의 과제	9
I. 연구 대상과 목표	9
II. 연구 방법과 순서	13
제2장 머신러닝 알고리즘과 차별의 문제	16
제1절 인공지능 로봇과 자율적 에이전트	17
I. ‘4차 산업혁명’의 상징적 표현	17
II. 인공지능 연구와 학습하는 기계	20
III. 과학기술 및 기계산업 중심의 현대 문명과 로봇	23
IV. 사회의 행위자로서 에이전트	31
제2절 머신러닝 알고리즘과 데이터 학습 모델	34
I. 알고리즘의 어원과 정의	34
II. 머신러닝 알고리즘과 데이터 마이닝 모델의 맹점	38
III. 머신러닝의 지도학습과 비지도학습	43
IV. 머신러닝 알고리즘과 인공지능	44
제3절 머신러닝 알고리즘의 오류와 편향	52
I. 머신러닝 알고리즘의 의심스러운 분류와 예측	52
II. 머신러닝 알고리즘의 인식 오류와 차등적 분류	55
III. 머신러닝 알고리즘의 편향과 사회 이미지의 학습	57
IV. 머신러닝 알고리즘을 이용한 실험	61



제3장 머신러닝 알고리즘의 작동과 차별의 인식	71
제1절 머신러닝 알고리즘의 작동원리	73
I. 목표 변수와 클래스 레이블의 정의	73
II. 훈련용 데이터 구성	76
III. 특성 선택	78
IV. 대용물(proxy)의 사용	80
제2절 차별의 형식과 구성	83
I. 차별의 전제로서 구별, 분리, 분류	83
II. 차별의 몇 가지 형식	87
III. 행위 중심의 차별과 결과 중심의 차별	94
제3절 차별의 복잡성과 평등	98
I. 차별에 관한 법적 구조	98
II. 형식적 평등과 차이의 구별	104
III. 평등 형식과 차별 형식의 관계	109
IV. 차별과 헌법적 가치의 연결	111
제4절 차별의 부당성과 비교	118
I. 규범적 의미의 차별	118
II. 차별금지 사유의 규범적 무관성	121
III. 특정 집단에 불리한 차별 결과	124
IV. 비교적 차별과 독립적 차별	128
제4장 머신러닝 알고리즘의 결정과 차별의 판단	133
제1절 알고리즘의 지배와 법의 지배	134
I. 사회의 규제 메커니즘으로서 법과 기술	134
II. 데이터 알고리즘 사회와 알고리즘의 지배	139
III. 데이터 알고리즘 사회에서 전달되는 법과 지시하는 법	141
IV. 예측하는 알고리즘과 예측되는 법	147



제2절 머신러닝 알고리즘의 분류 및 추론과 간접차별	150
I. 인공물에 의한 분류와 인간에 의한 분류	150
II. 간접차별과 직접차별	154
III. 머신러닝 알고리즘의 이용과 차별의 우회	161
제3절 머신러닝 알고리즘의 최적화와 차별의 정당화	166
I. 머신러닝 알고리즘과 법적 추론에서 최적화	166
II. 차별의 인식과 불비례성	171
III. 비례성 및 합리성과 차별의 평가	174
제4절 머신러닝 알고리즘의 불투명성과 차별의 은폐	181
I. 불투명성의 세 가지 차원	181
II. 투명성과 불투명성 사이의 헌법적 긴장 관계	185
III. 민주주의 국가의 디폴트(default)로서 투명성	189
IV. 투명성 제한의 근거로서 비밀과 차별	194
제5장 머신러닝 알고리즘의 차별에 관한 책임과 개인정보 보호	197
제1절 차별에 관한 책임의 구조	199
I. 과학기술의 발전과 책임 구조의 변화	199
II. 차별에 관한 국가와 사인의 책임	202
III. 차별금지의무의 성격과 내용	207
제2절 머신러닝 알고리즘의 차별에 대한 책임의 분산	213
I. 알고리즘의 설계자, 감독자, 심사자	213
II. 알고리즘 시스템 운영자	216
III. 알고리즘 에이전트에 대한 책임 귀속 문제	221
제3절 머신러닝 알고리즘의 차별로부터 보호와 개인정보의 보호	224
I. 데이터베이스 구성 단계에서 차별과 개인정보보호	224
II. 데이터 분석 단계에서 차별과 개인정보보호	229
III. 분석 모델 이용 단계에서 차별과 개인정보보호	231



제4절 자동화된 차별적 결정에 구속되지 않을 권리와 설명의 한계 ...	237
I. 머신러닝 알고리즘 시스템의 자동화된 결정과 차별	237
II. 자동화된 결정에 구속되지 않을 권리와 그 예외	240
III. 특정범주의 개인정보와 자동화된 차별적 결정	244
IV. 머신러닝 알고리즘에 대한 설명의무와 설명청구권	247
제6장 결론	254
참고문헌	261
Abstract	288



세부 차례

제1장 서론	1
제1절 연구의 배경	1
I. 헌법 만능주의와 기술 만능주의	1
1. 기계에 관한 두 가지 역사적 경험	1
2. 헤라클레스와 마스터 알고리즘	2
II. 알고리즘의 데이터 분석과 감시	3
1. 테러와의 전쟁과 데이터와의 전쟁	3
2. 알고리즘의 데이터 감시와 차별	4
III. 알고리즘 활용 기술의 발전과 그 파급력	5
1. 인간에 대해 중요한 결정을 하는 알고리즘	5
2. 알고리즘 활용 기술의 파급력에 대한 법적·사회적 논의의 필요성	6
3. 결정 기관에 대한 불신과 인공지능에 대한 기대	8
제2절 연구의 과제	9
I. 연구 대상과 목표	9
1. 연구 대상 및 범위	9
2. 연구 목표	11
II. 연구 방법과 순서	13
1. 연구 방법	13
2. 연구 순서	14
제2장 머신러닝 알고리즘과 차별의 문제	16
제1절 인공지능 로봇과 자율적 에이전트	17
I. ‘4차 산업혁명’의 상징적 표현	17
1. 플레이스홀더로서 ‘4차 산업혁명’	17
2. ‘4차 산업혁명’을 상징하는 용어들과 그 관계	19
II. 인공지능 연구와 학습하는 기계	20
1. 인공지능 개념과 학제 연구의 집약	20
2. 튜링의 모방게임과 학습하는 기계	20
3. 인공지능 연구의 중흥기와 정체기	22



III. 과학기술 및 기계산업 중심의 현대 문명과 로봇	23
1. 로봇 개념의 기원	23
2. 일제 강점기 로봇 개념의 수용 태도	24
3. 로봇 관념에 집약된 현대 과학과 기술의 역설	25
4. 21세기의 인공지능 로봇과 인류 미래에 관한 논의	26
5. 아시모프의 로봇 3법칙과 헌법상 국가의 책무	29
IV. 사회의 행위자로서 에이전트	31
1. 에이전트의 정의	31
2. 에이전트의 특성과 자율성	32
3. 법체계에서 에이전트와 커뮤니케이션	33
제2절 머신러닝 알고리즘과 데이터 학습 모델	34
I. 알고리즘의 어원과 정의	34
1. 알 과리즘과 알고리즘	34
2. 문제를 해결하는 방법	35
3. 컴퓨터가 수행할 일을 순서대로 알려주는 명령어의 집합	37
4. 입력(input)과 출력(output)의 관계를 규정하는 절차	38
II. 머신러닝 알고리즘과 데이터 마이닝 모델의 맹점	38
1. 머신러닝 알고리즘의 학습 개념	38
2. 머신러닝 알고리즘의 학습과 데이터	39
3. 데이터 마이닝과 모델의 단순화	40
4. 모델 정립을 위한 단순화와 인간의 사고실험	41
III. 머신러닝의 지도학습과 비지도학습	43
1. 지도학습과 분류 및 예측	43
2. 비지도학습과 군집	44
IV. 머신러닝 알고리즘과 인공지능	44
1. 기호주의와 논리 및 규칙 기반의 역연역법	45
2. 연결주의와 신경망 기반의 역전파법	46
3. 진화주의와 유전자 프로그래밍 기반의 유전자 탐색	47
4. 베이지주의와 베이지 네트워크 기반의 확률적 추론	48
5. 유추주의와 서포트 벡터 및 사례 기반의 조건부 최적화	49



제3절 머신러닝 알고리즘의 오류와 편향	52
I. 머신러닝 알고리즘의 의심스러운 분류와 예측	52
1. 스위니의 연구	52
2. 루미스 사건	53
II. 머신러닝 알고리즘의 인식 오류와 차등적 분류	55
1. 작은 눈과 감은 눈의 인식 오류	55
2. 고릴라 사건	56
3. 우편번호 분류에 따른 가격의 차등 적용	57
III. 머신러닝 알고리즘의 편향과 사회 이미지의 학습	57
1. 고용 알고리즘과 편향의 학습	57
2. 데이터세트의 성별 편향과 모델에 의한 증폭	58
IV. 머신러닝 알고리즘을 이용한 실험	61
1. 번역 실험	61
2. 이미지 검색 실험	67
제3장 머신러닝 알고리즘의 작동과 차별의 인식	71
제1절 머신러닝 알고리즘의 작동원리	73
I. 목표 변수와 클래스 레이블의 정의	73
1. 목표 변수의 정의와 문제의 형식화	73
2. 클래스 레이블의 정의와 카테고리의 생성	74
3. 목표 변수 정의와 클래스 레이블 정의의 관계	75
II. 훈련용 데이터 구성	76
1. 표본에 대한 레이블링과 클래스의 지정	76
2. 데이터 수집	77
III. 특성 선택	78
1. 특성 선택과 데이터의 대표성	78
2. 통계적 추론과 특성의 일반화	79
IV. 대용물(proxy)의 사용	80
1. 집단에 대한 또 다른 특성의 귀속	80
2. 중복 인코딩과 대용물 제거의 한계	81



제2절 차별의 형식과 구성	83
I. 차별의 전제로서 구별, 분리, 분류	83
1. 차별의 서술적 표현으로서 구별과 인식	83
2. 개별화의 수단으로서 분리	85
3. 집단화의 수단으로서 분류	86
4. 분리를 통한 분류와 집단의 비대칭성	86
II. 차별의 몇 가지 형식	87
1. 편견	88
2. 인식적 고정관념과 규범적 고정관념	89
3. 통계적 고정관념과 합리적 차별	90
4. 차별효과와 간접차별	91
5. 부작위: 적극적 행위의 불이행	92
III. 행위 중심의 차별과 결과 중심의 차별	94
1. 차별의 행위 관련 구성과 의무론	95
2. 차별의 효과 관련 구성과 결과론	95
3. 차별의 행위 관련 구성과 효과 관련 구성의 차이	96
제3절 차별의 복잡성과 평등	98
I. 차별에 관한 법적 구조	98
1. 헌법의 평등 및 차별 관련 규정	98
2. 법률의 차별 관련 규정과 ‘평등권 침해의 차별행위’	99
3. 헌법과 법률의 차별 관련 규정과 그 구조적 유사성	102
4. ‘평등권 침해의 차별행위’에 대한 상이한 해석 가능성	103
II. 형식적 평등과 차이의 구별	104
1. 형식적 평등과 ‘같은 것은 같게’ 알고리즘	104
2. 형식적 평등에서 ‘같게’와 ‘다르게’의 구별	105
3. 형식적 평등이 설명해주지 않는 것	107
III. 평등 형식과 차별 형식의 관계	109
1. 차별과 평등의 관계 설정	109
2. 차별과 평등의 관계 분석	110
IV. 차별과 헌법적 가치의 연결	111
1. 평등과 차별	111
2. 자유와 차별	113
3. 인간의 존엄성과 차별	115
4. 사회통합과 차별	116



제4절 차별의 부당성과 비교	118
I. 규범적 의미의 차별	118
1. 사람의 특성에 대한 의미부여	118
2. 규범적 기준에 따른 구별	119
3. 차별 형식과 차별금지 사유의 구조적 결합	120
II. 차별금지 사유의 규범적 무관성	121
1. 특성에 기초한 질서의 형성	121
2. 차별금지 사유의 도덕적 기초로서 불가변성의 한계	121
3. 넓은 의미의 불가변성과 선택 불가능성	122
4. 난이도에 따른 선택 가능성의 구별	123
III. 특정 집단에 불리한 차별 결과	124
1. 정치적 불이익	125
2. 사회·문화적 불이익	125
3. 실질적 불이익	127
IV. 비교적 차별과 독립적 차별	128
1. 비교적 차별	128
2. 독립적 차별	130
제4장 머신러닝 알고리즘의 결정과 차별의 판단	133
제1절 알고리즘의 지배와 법의 지배	134
I. 사회의 규제 메커니즘으로서 법과 기술	134
1. 기술적 설계를 통한 사회의 구조화	134
2. 기술적 조치를 통한 물리적·심리적 제한	137
3. 규제 메커니즘의 실현 조건	138
II. 데이터 알고리즘 사회와 알고리즘의 지배	139
1. 데이터 알고리즘 사회	139
2. 온라인프 생활양식의 확장	140
3. 사이버-물리 시스템 환경에서 알고리즘의 지배	141
III. 데이터 알고리즘 사회에서 전달되는 법과 지시하는 법	141
1. 정보의 두 가지 의미와 정보로서 법	141
2. 전달되는 정보와 조종하는 정보	143
3. 사실을 구성하는 법과 알고리즘 시스템	144



IV. 예측하는 알고리즘과 예측되는 법	147
1. 법의 체계성	147
2. 머신러닝 알고리즘에 의한 법적 결정의 예측	147
3. 머신러닝 알고리즘에 의한 법적 결정의 대체	148
제2절 머신러닝 알고리즘의 분류 및 추론과 간접차별	150
I. 인공물에 의한 분류와 인간에 의한 분류	150
1. 인식 모델과 인지 작용에 의한 지능적 분류	150
2. 반응적 분류와 개념적 분류	151
3. 기계와 인간의 구별 기준으로서 이해와 그 한계	153
II. 간접차별과 직접차별	154
1. 차별로 인한 개인의 불이익과 집단의 불이익	156
2. 차별의 의도 또는 동기	156
3. 차별의 결과 또는 효과	157
4. 차별의 정당화 가능성	157
5. 차별의 해소 방법	159
6. 구분의 포기	160
III. 머신러닝 알고리즘의 이용과 차별의 우회	161
1. 머신러닝 알고리즘의 의도 숨기기와 직접차별	161
2. 머신러닝 알고리즘의 특성 추론과 간접차별	162
3. 머신러닝 알고리즘의 통계적 추론과 합리적 차별	163
제3절 머신러닝 알고리즘의 최적화와 차별의 정당화	166
I. 머신러닝 알고리즘과 법적 추론에서 최적화	166
1. 법적 특이점으로서 반성적 평형	166
2. 머신러닝 알고리즘에서 최적화	168
3. 법적 추론에서 최적화	168
II. 차별의 인식과 불비례성	171
1. 차별의 인식 도구로서 불비례성	171
2. 법적 차별의 측정 지표	171
3. 확장된 차별 측정 지표와 머신러닝 알고리즘의 차별	173
4. 소결	173
III. 비례성 및 합리성과 차별의 평가	174
1. 차별의 평가 지표로서 비례성	174
2. 차별의 평가 도식으로서 비례성 심사와 합리성	175
3. 비례성의 합리성 문제와 차별의 정당화 논증에서 잠재적 차별 가능성	178



제4절 머신러닝 알고리즘의 불투명성과 차별의 은폐	181
I. 불투명성의 세 가지 차원	181
1. 비밀주의와 불투명성	182
2. 기술적 문맹 상태와 불투명성	183
3. 복잡한 머신러닝 알고리즘의 작동방식과 불투명성	184
II. 투명성과 불투명성 사이의 헌법적 긴장 관계	185
1. 블랙박스 사회의 딜레마	185
2. 정보 관련 입법에 함의된 투명성과 불투명성	185
III. 민주주의 국가의 디폴트(default)로서 투명성	189
1. 공정하고 효율적인 정책을 위한 유인책으로서 투명성	189
2. 투명성의 확장 범위와 클라우드소싱	190
3. 투명성 통제와 프라이버시	191
4. 투명성과 개인의 자율성	192
IV. 투명성 제한의 근거로서 비밀과 차별	194
1. 공공안전과 국가안보를 위한 기밀과 영업비밀	194
2. 대용물로 사용된 특성에 의한 낙인	194
제5장 머신러닝 알고리즘의 차별에 관한 책임과 개인정보 보호 ...	197
제1절 차별에 관한 책임의 구조	199
I. 과학기술의 발전과 책임 구조의 변화	199
1. 위험에 대응하기 위한 규칙 시스템의 진화 과정	199
2. 사고와 위험 책임	200
3. 책임의 주체와 상대의 확장	201
4. 책임 대상의 확대	201
II. 차별에 관한 국가와 사인의 책임	202
1. 민주주의 사회에서 자동화된 차별의 위험	202
2. 차별에 관한 국가의 책임과 시스템의 차별	204
3. 차별에 관한 사인의 책임과 차별금지법의 특수한 책임 주체	205
III. 차별금지의무의 성격과 내용	207
1. 차별금지에 관한 소극적 의무와 적극적 의무 도식	208
2. 합당한 배려의무의 부차적 성격	209
3. 적극적 조치와 머신러닝 알고리즘의 차별	210



제2절 머신러닝 알고리즘의 차별에 대한 책임의 분산 213

- I. 알고리즘의 설계자, 감독자, 심사자 213
 - 1. 알고리즘의 작성과 권력분립의 유추 213
 - 2. 알고리즘 설계의 중요성 214
 - 3. 시스템의 차별을 규제할 견제와 균형 시스템 215
- II. 알고리즘 시스템 운영자 216
 - 1. 데이터 프로세싱과 차별 216
 - 2. 정보처리자로서 알고리즘 시스템 운영자 218
 - 3. 개인정보의 수집과 정보 수탁자 218
 - 4. 개인정보의 추론과 사무 관리자의 책임 219
 - 5. 알고리즘 시스템 운영자의 사용자 책임 220
- III. 알고리즘 에이전트에 대한 책임 귀속 문제 221
 - 1. 양 극단의 제한주의와 허용주의 221
 - 2. 원칙적 허용과 공리주의적 제한 222
 - 3. 민속 심리학에 기초한 법 개념과 허용의 한계 223

제3절 머신러닝 알고리즘의 차별로부터 보호와 개인정보의 보호 ... 224

- I. 데이터베이스 구성 단계에서 차별과 개인정보보호 224
 - 1. 후버 사건 224
 - 2. 분류금지원칙과 수집제한원칙 226
 - 3. 차별금지법과 개인정보보호법의 적용범위 227
 - 4. 복잡한 차별금지와 안정화된 개인정보보호 228
- II. 데이터 분석 단계에서 차별과 개인정보보호 229
 - 1. 데이터 분석과 이용의 구별 229
 - 2. 테스트-아사 사건 229
 - 3. 통계적 분석에 대한 차별금지와 개인정보보호의 접근방식 230
 - 4. 집단의 차별에 대한 개인의 권리의 한계 231
- III. 분석 모델 이용 단계에서 차별과 개인정보보호 231
 - 1. 개인정보 수집제한의 한계와 이용단계의 중요성 231
 - 2. 데이터 이용에 의한 차별자와 피차별자의 연결 233
 - 3. 인간에 의한 차별과 알고리즘 시스템에 의한 차별 234
 - 4. 시스템의 차별에 대한 차별금지법과 개인정보보호법의 한계 236



제4절 자동화된 차별적 결정에 구속되지 않을 권리와 설명의 한계 ...	237
I. 머신러닝 알고리즘 시스템의 자동화된 결정과 차별	237
1. 자동화된 개별적 결정과 GDPR의 제정	237
2. 차별의 문제와 설명에 관한 권리	238
3. 차별에 비중을 둔 해석의 필요성	239
II. 자동화된 결정에 구속되지 않을 권리와 그 예외	240
1. GDPR 제22조 제4항의 형식적 구조와 차별 관련성	240
2. 자동화된 결정에 구속되지 않을 권리	240
3. 자동화된 결정에 구속되는 예외	242
4. ‘설명에 관한 권리’ 존부 논쟁	242
III. 특정범주의 개인정보와 자동화된 차별적 결정	244
1. 특정 범주의 개인데이터에 근거한 자동화된 결정에 구속되지 않을 권리 ...	244
2. 자동화된 결정에 구속되는 예외	245
3. 소결	247
IV. 머신러닝 알고리즘에 대한 설명의무와 설명청구권	247
1. 자동화된 결정에 구속되지 않을 권리 실현의 전제	247
2. 모델 중심의 설명과 주체 중심의 설명	248
3. 사전 설명과 사후 설명	249
4. ‘로직(logic)에 관한 의미 있는 정보’ 제공의무와 접근권	250
5. 설명에 관한 권리 중심 접근의 한계	252
 제6장 결론	 254
 참고문헌	 261
 Abstract	 288



자료 차례

자료 1. 편향(bias)과 분산(variance)	50
자료 2. imSitu vSRL에 관한 편향 분석	60
자료 3. MS-COCO MLC에 관한 편향 분석	60
자료 4. 번역기에 의한 터키어 ‘o bir doktor.’의 영어 번역 결과	62
자료 5. 번역기에 의한 터키어 ‘o bir hemşire.’의 독일어 번역 결과	62
자료 6. 번역기에 의한 터키어 ‘o bir hemşire.’의 한국어 번역 결과	63
자료 7. 번역기에 의한 터키어 ‘o bir hemşire.’의 영어 번역 결과	63
자료 8. 번역기에 의한 마침표('.')가 생략된 터키어 ‘o bir hemşire’의 영어 번역 결과	64
자료 9. 번역기에 의한 마침표('.')가 생략된 터키어 ‘o bir hemşire’의 독일어 번역 결과	64
자료 10. 번역기에 의한 마침표('.')가 생략된 터키어 ‘o bir hemşire’의 한국어 번역 결과	64
자료 11. 번역기에 의한 마침표('.')가 생략된 터키어 ‘o bir doktor’의 영어 번역 결과	64
자료 12. 번역기에 의한 한국어 ‘그 사람은 의사입니다.’의 영어 번역 결과	65
자료 13. 번역기에 의한 한국어 ‘그 사람은 간호사입니다.’의 영어 번역 결과	65
자료 14. 성별에 따른 번역 기능이 추가된 번역기의 번역 결과 (터키어→영어)	66
자료 15. 성별에 따른 번역 기능을 지원하지 않는 언어 간 번역 결과 (터키어→독일어)	67
자료 16. 성별에 따른 번역 기능을 지원하지 않는 언어 간 번역 결과 (터키어→한국어)	67
자료 17. 성별에 따른 번역 기능을 지원하지 않는 언어 간 번역 결과 (한국어→영어)	67
자료 18. 한국어 기반 이미지 검색기에 의한 ‘장보기’ 검색 결과의 예	68
자료 19. 한국어 기반 이미지 검색기에 의한 ‘쇼핑’ 검색 결과의 예	69
자료 20. 보호집단과 비보호집단 중에 이익이 거부 또는 부여된 인원	172



제1장

서론

제1절 연구의 배경

I. 헌법 만능주의와 기술 만능주의

1. 기계에 관한 두 가지 역사적 경험

2016년 봄 바둑 게임을 놓고 벌어진 기계와 인간의 대결에서 인간이 패배하는 장면을 직접 목격한 이래 대한민국에서 기술과 산업 분야를 이끄는 핵심어는 ‘인공지능(artificial intelligence, AI)’과 ‘4차 산업혁명’이다. 2016년 가을 한반도의 남쪽 땅에서 발견된 태블릿 피시(tablet PC)는 탄핵 정국을 촉발했고, 대통령과 국민의 대결에서 국민이 승리한 이후 대한민국의 정치와 헌법 분야를 주도한 핵심어는 ‘개헌’, 즉 ‘헌법개정’이다.¹⁾

한쪽에선 혁신적 기술이 인류의 운명을 바꿔 놓을 것이라고 하며, 또 다른 쪽에선 헌법이 바뀌면 내 삶이 바뀐다고 한다. 기술과 헌법에 관한 이와 같은 수사(修辭)에는 ‘인간이라는 자연의 미래’²⁾ 및 복잡한 사회의 미래에 대한 불확실한 변화의 전망과 함께 그러한 미래의 현재라고 할 수 있는 위험에 대한 처방이 동시에 담겨 있다. 이를 각각 헌법 만능주의와 기술 만능주의라고 부를 수 있다면 두 가지 기획이 만나는 교차로에서 불완전한 인간은 대단히 불안정한 상태에 놓인다.

1) 2016년 10월 24일 박근혜 대통령은 국회 본회의 연설을 통해 개헌을 주장했다, 같은 해 12월 9일 국회의 탄핵소추의결과 2017년 3월 10일 헌법재판소의 탄핵결정을 통해 대통령직에서 파면됐다. 이어진 대통령 선거에서도 헌법 개정은 모든 후보자의 공통된 공약 사항이었고, 2017년 5월 10일 취임한 문재인 대통령의 발의로 2018년 3월 26일 헌법개정안이 제안되었다. 국회는 헌법 제130조 제1항에 따라 공고일로부터 60일 이내에 의결하기 위해 같은 해 5월 24일 의결을 시도했으나 출석의원의 부족으로 의결 자체가 불성립되었다. 결국 국회는 헌법에 규정된 의결 기한을 넘김으로써 헌법을 준수하지 않아 헌법의 작동을 방해하여 헌법의 자기보장성을 무력화시키는 위험적인 상황을 자초했다.

2) 하버마스(J. Habermas)가 쓴 책 ‘Die Zukunft der menschlichen Natur: auf dem Weg zu einer liberalen Eugenik?’[2001]의 한글 번역서 제목이다. Jürgen Habermas, 인간이라는 자연의 미래, 장은주(역), 나남, 2003 참조.



이러한 상황에서 인간은 마치 불안정한 지위를 감내해야 할 것처럼 내몰린다. 그러나 그것이 인간의 운명과도 같은 몫이라고 확언할 수는 없다.³⁾ 역설적이지만 대한민국의 국민은 인간이 만든 기계에게 언제든 패배할 수 있다는 두려움을 간직한 채, 인간의 행동과 기억을 대신 기록하여 저장하고 있는 기계를 발견하는 것만으로도 인간이 만든 헌법에 따라 권한을 행사해야 하는 대통령을 언제든 이길 수 있다는 자신감도 갖게 되었다.

2. 헤라클레스와 마스터 알고리즘

헌법이 바뀌면 내 삶이 바뀔 것이라는 전망은 현재 내 삶이 바뀌지 않는 원인을 헌법에서 찾는다. 혁신적인 기술이 인류의 운명을 바꿔 놓을 것이라는 전망은 인류 운명의 미래를 결정짓는 요인을 기술에서 찾는다. 법과 정치의 가교로서 헌법의 변화 과정에는 이를 제정하거나 개정하는 결정자뿐만 아니라 이를 해석하고 적용하는 결정자의 역할이 핵심적이다.⁴⁾ 사회와 자연의 가교로서 생활양식을 구조화하는 기술의 변화 과정에는 이를 설계하고 개발하는 결정자뿐만 아니라 이를 사용하고 응용하는 결정자의 역할이 핵심적이다.⁵⁾

이렇게 핵심적 역할을 수행하는 결정자의 중요성 때문에 한쪽에서는 이러한 결정자들에게 헤라클레스가 될 것을 요청하고, 다른 쪽에서는 결정자로서 인간을 대체할 수 있는 마스터 알고리즘을 개발하려 한다. 드워킨(R. Dworkin)은 자신이 고안한 법률가 즉, “초인(superman)의 기량과 학습 능력, 인내심과 예리한 통찰력을 갖춘 법률가”를 “헤라클레스(Hercules)”라고 부르고, 이를 대표적인 미국 사법부의 법관 중 하나로

3) 하이데거(M. Heidegger)는 정밀한 자연과학을 이용하는 현대 기술의 본질이 몰아세움(Ge-stell)에 있다고 주장한다. 그에 따르면 몰아세움은 이전까지 포이에시스(poiēsis), 즉 스스로 밖으로 끌어내어 앞에 내어놓는 것을 의미했던 탈은폐의 방식을 부품(Bestand)처럼 주문할 수 있는 것으로 바꾸는 도발적 요청(Herausfordern)이다. 이런 몰아세움은 탈은폐의 예정 속에 있지만, 예정 속에 있다는 것이 변경 불가능한 운명 속에 있다는 것을 의미하지는 않는다. Martin Heidegger, “Die Frage nach der Technik”, in: ders., *Die Technik und die Kehre*, 8. Aufl., Neske, 1991[1962], S. 5~36 참조.

4) 헌법을 “법체제와 정치체제의 구조적 연결(strukturelle Kopplung)을 보장하는 형식”으로 이해하는 관점은 Luhmann, Niklas, *Das Recht der Gesellschaft*, 8. Aufl., Frankfurt am Main: Suhrkamp, 2013[1995], S. 470 및 ders., “Verfassung als evolutionäre Errungenschaft”, *Rechtshistorisches Journal* 9, 1990, S. 176~220 참조.

5) 위너(L. Winner)는 기술을 생활의 형식으로 보면서 정치와 유사하게 사회의 구조와 질서를 형성하는 측면이 있다는 점을 강조한다. Langdon Winner, “Technology as Forms of Life”, in *Epistemology, Methodology and the Social Sciences*, Robert S. Cohen · Marx W. Wartofsky(Eds.), D. Reidel, 1983, pp. 249~263 및 Langdon Winner, “Techne and Politeia”[1983], in *The Whale and the Reactor: A Search for Limits in an Age of High Technology*, University of Chicago Press, 1986, pp. 40~58 참조.



상정하였고,⁶⁾ 도밍고스(P. Domingos)는 자신의 중심 가설로 “데이터로부터 과거와 현재, 그리고 미래의 모든 지식을 도출할 수 있는 단일한 보편적 학습 알고리즘”을 상정하여 “마스터 알고리즘(the Master Algorithm)”이라고 부른다.⁷⁾ 이와 같은 두 가지 기획은 모든 문제에 가장 적합한 방법을 찾아낼 수 있는 만능 해결사가 올바른 하나의 정답을 제시해 줄 것이라는 기대 속에서 서로 교차한다.

II. 알고리즘의 데이터 분석과 감시

1. 테러와의 전쟁과 데이터와의 전쟁

너무 똑똑해서 더 이상 그냥 ‘전화기’라고만 부를 수 없게 된 ‘스마트폰’에 있는 달력 화면을 오른쪽으로 밀어 왼쪽으로 시간을 거슬러 올라가 보자.⁸⁾ 21세기의 시작을 알리는 신호탄은 미국에서 출발한 ‘테러와의 전쟁’이다. 2001년 9월 11일 쌍둥이 빌딩으로 익숙한 세계무역센터가 테러 공격을 당하자 미국의 대통령은 세계를 향해 “테러와의 전쟁(the war against terrorism)”⁹⁾을 선포했다.¹⁰⁾ 그런데 정작 테러에 대한 전쟁이 예정한 것은 데이터에 대한 전쟁이었다. 테러와의 전쟁은 테러를 일으킬 만한 의심스러운 특성을 가진 사람들을 분류해 내는 것이다.

그러나 이러한 분류는 사람들을 한 데 모아 놓고 그 사람들을 여기로 모이게 하고 또 저기로 모이게 하는 방식을 채택하지 않는다. 데이터화한 사람들, 즉 사람들의 데이터를 분류해 내는 것이다. 데이터의 분류는 수집된 데이터가 다양하고 많을수록 더 섬세하고 정교해진다. 그리고 더 정확하고 확실하게 데이터를 분석하고 이용하려는 열망은 방대한 데이터 수집을 촉진한다.

6) Ronald Dworkin, *Taking Rights Seriously*, Harvard University Press, 1977, p. 105.

7) Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, New York: Basic Books, a member of the Perseus Books Group, 2015, p. 25.

8) 손가락으로 전자화면을 전환할 수 있도록 한 기술은 종이책을 좌우로 혹은 상하로 넘기던 인간의 행동 양식을 모방한 것이다. 행위자의 행동 양식이 반영된 기술은 기존 양식의 패턴에 최소한의 변화만을 야기함으로써 새로운 기술을 이용하고 기술에 적응할 때 발생할 수 있는 저항감을 감소시킨다.

9) “A Day of Terror; Bush’s Remarks to the Nation on the Terrorist Attacks”, The New York Times, 12 September 2001, NYTimes.com, <https://www.nytimes.com/2001/09/12/us/a-day-of-terror-bush-s-remarks-to-the-nation-on-the-terrorist-attacks.html>, 접속일: 2018년 6월 13일.

10) 9·11 사건을 계기로 그동안 진행되어 온 감시의 자동화, 통합화, 지구화가 보다 공개적인 방식으로 추진되었다는 지적은 David Lyon, *Surveillance after September 11*, Malden, Mass: Polity Press in association with Blackwell Pub. Inc, 2003, 특히 p. 4 및 pp. 62~141 참조.



2. 알고리즘의 데이터 감시와 차별

미국이 테러와의 전쟁을 치르기 위해 선택한 감시 대상에서 자국민, 즉 미국 시민이라고 해서 예외가 될 수는 없었다. 진실은 시간을 머금었을 때 비로소 자기 자신을 드러내는 것인 양, 2013년 미국의 국가안보국(National Security Agency)에서 근무한 경력을 가진 스노든(E. Snowden)이 내부 문건을 외부에 폭로¹¹⁾함으로써 감시국가의 민낯을 확인시켜 주었고,¹²⁾ 미국의 국가안보국 ‘프리즘(PRISM)’이란 전자 감시 프로그램을 운영했다는 사실도 드러났다. 전 세계인을 상대로 그들의 전화 통화, 온라인 통신을 수집하는 행위자는 인간이 아니라 알고리즘(algorithm)에 따라 작동되는 연산 기계 즉, 컴퓨터였던 것이다. 개인의 데이터가 감시되고 있다는 측면 못지않게 그러한 감시가 알고리즘에 의해 실행되고 있다는 측면이 강조되어야 하는 이유가 여기에 있다.

감시를 주제로 한 학제 간 연구 분야로서 “감시 연구”¹³⁾에서는 일찍이 사회적으로 의미 있는 데이터를 분류하거나 정렬하는 것을 감시로 보면서,¹⁴⁾ 이러한 감시가 차별로 이어질 수 있다는 점을 지적해 왔다.¹⁵⁾ 사회의 차별로 이어질 수 있는 “데이터 감시(dataveillance)”¹⁶⁾의 주체가 인간이 아니라 알고리즘에 따라 만들어진 컴퓨터 프로그램이라는 점은 법을 비롯해 인간을 중심으로 구축된 여러 가지 사회의 제도에 심각한 도전을 제기한다.

11) 2013년 스노든(E. Snowden)의 폭로를 취재한 가디언의 기자 그린월드(G. Greenwald)가 그 취재 과정과 내용을 담은 것으로 Greenwald, Glenn, *No Place to Hide: Edward Snowden, the NSA and the U.S. Surveillance State*, Metropolitan Books/Henry Holt, 2014 참조.

12) 감시의 다양한 유형 및 법적 쟁점에 관해서 이준일, *감시와 법*, 고려대학교출판부, 2014 참조.

13) Kirstie Ball · Kevin D. Haggerty · David Lyon(Eds.), *Routledge Handbook of Surveillance Studies*, Routledge, 2012 참조.

14) David Lyon, “Surveillance as Social Sorting: Computer Codes and Mobile Bodies”, in *Surveillance as Social Sorting: Privacy, Risk and Digital Discrimination*, Lyon David(Ed.), Routledge, 2003, pp. 13~30 참조.

15) Dorothy Nelkin · Lori Andrews, “Surveillance Creep in the Genetic Age”, in *Surveillance as Social Sorting: Privacy, Risk and Digital Discrimination*, David Lyon(Ed.), Routledge, 2003, pp. 94~110: p. 106.

16) ‘데이터 감시(dataveillance)’는 ‘데이터(data)’와 ‘감시(surveillance)’를 결합시킨 용어로 정보기술(IT)이 발달하면서 방대한 양의 개인정보를 대규모로 수집하고 저장하는 것이 용이해진 감시의 관행을 묘사하기 위해 만들어졌다. Roger A. Clarke, “Information Technology and Dataveillance”, *Communications of the ACM* 31(5), 1988, pp. 498~512 및 Colin J. Bennett, “The Public Surveillance of Personal Data: A Cross-national Analysis”, in *Computers, Surveillance, and Privacy*, David Lyon · Elia Zureik(Eds.), University of Minnesota Press, 1996, pp. 237~259 참조.



III. 알고리즘 활용 기술의 발전과 그 파급력

1. 인간에 대해 중요한 결정을 하는 알고리즘

세계의 시민을 감시하는 데에 사용될 수 있는 알고리즘은 국가의 관심 대상인 군사, 경찰, 형사사법, 건강보험, 교육, 의료 등 복지에 관한 행정 등에서 다양하게 활용될 수 있고,¹⁷⁾ 범용성을 갖춘 알고리즘은 적용 분야에 대해서 공사(公私)를 가리지 않는다. 알고리즘은 기업의 여러 가지 복잡한 업무를 매우 효율적으로 수행 하면서 인간 행위자(agent)를 대체해 가고 있다.

경영 분야에서는 이미 2014년에 세계 최초로 알고리즘이 기업 이사회에 구성원으로 지명됐다.¹⁸⁾ 이 알고리즘 이사의 업무는 금융 정보, 특정 약품의 임상 시험, 회사가 보유한 지적 재산권과 이전의 투자 정보 등 다양한 데이터를 분석하는 것이다. 분석 대상이 되는 데이터의 양은 방대해지고 그 형태는 더 이상 인간이 그 구조를 파악하기 어려운 비정형의 것으로 변해가고 있다. 이른바 ‘빅데이터(big data)’를 처리할 수 있는 알고리즘만이 문제를 다룰 수 있게 되는 것이다.

그리고 디지털 평판(digital reputation), 검색(search), 금융(finance) 분야처럼 빅데이터를 활용해 우리 삶에 중요한 영향을 미치는 분야는 계속해서 확대되고 있다.¹⁹⁾ 평판 알고리즘은 인간으로서 우리가 어떻게 보일 것인지를 결정한다면, 검색 알고리즘은 우리가 무엇을 볼 수 있는지를 결정하고, 금융 알고리즘은 우리가 무엇에 어느 정도 가치를 두어야 하는지를 결정한다는 점에서 인간에게 근본적인 영향을 미친다.

온라인에서 이루어지는 모든 활동은 기록된다. 그리고 이제는 오프라인에서 이루어지는 활동이라고 해서 예외가 되는 것도 아니다. 인간이 소지하거나 사용하는 물건과 인간의 생활환경을 이루는 인공물은 인간의 활동을 지각하고 사물 인터넷

17) Kate Crawford · Meredith Whittaker · Alex Campolo · Madelyn Sanfilippo, “AI Now 2017 Report”, The AI Now Institute at New York University, 2017, p. 1 및 Meredith Whittaker · Kate Crawford · Roel Dobbe · Genevieve Fried · Elizabeth Kaziunas · Varoon Mathur · Sarah Myers West · Rashida Richardson · Jason Schultz · Oscar Schwartz, “AI Now Report 2018”, The AI Now Institute at New York University, December 2018, pp. 12~17 참조.

18) 홍콩의 의약품 관련 창업투자회사(venture capital firm)인 ‘Deep Knowledge Ventures’가 이사로 지명한 이 컴퓨터 프로그램의 이름은 ‘VITAL’이다. “Algorithm Appointed Board Director”, BBC News, 16 May 2014, <https://www.bbc.com/news/technology-27426942>, 접속일: 2018년 6월 14일.

19) Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*, Harvard University Press, 2015, pp. 19~139 참조.



(internet of things)²⁰⁾ 통신을 통해 그 모든 활동을 기록한다. 거리와 도로 곳곳에 설치되어 있는 폐쇄회로 텔레비전(CCTV)의 사각지대를 찾기가 점점 어려워지고 있고, 집안에 들여 놓은 인공지능 스피커(AI speaker)는 사적인 대화 내용도 여과 없이 기록할 준비가 되어 있다. 언제 어디에서 무엇을 갖고 싶어 했고 어떤 행동을 했는지 본인은 기억하지 못해도 “환경이 기록하는 시대”로 접어들고 있는 것이다.²¹⁾ 이렇게 기록된 데이터는 정부 산하의 어느 기관 또는 기업의 서버 어딘가에 저장되어 데이터베이스로 구축된다.

인터넷 기업과 금융 기업은 ‘정보 경제’의 중심부에서 방대한 디지털 데이터를 축적함으로써 고객의 사생활까지 면밀히 파악한다. 금융 기업이 돈을 다룬다면, 인터넷 기업은 관심을 다룬다. 데이터를 이용해 고객에 관한 중요한 결정을 내릴 뿐만 아니라 고객 스스로 의사결정을 하는 데 영향을 미치기도 한다. 각 분야를 지배하는 알고리즘은 빅데이터를 분석하여 타인에게 보일 우리의 모습을 결정함으로써 직업이나 집을 구할 때 우리의 처지를 유리하게도 할 수 있고 불리하게도 할 수 있다.

또한 상품이나 서비스를 제공받으려 할 때도 우리가 볼 수 있는 것과 볼 수 없는 것을 결정하는 데에 그치지 않고, 볼 수 있는 것의 우선순위를 결정함으로써 보여줄 대상에 무엇 또는 누군가를 포함시키거나 배제시킬지 정하고 또 포함시키는 것의 순서를 정렬한다. 이로써 특정한 생각, 상품, 서비스로 관심을 몰고, 나머지 다른 부분에 대한 관심은 사라지게 만들 수 있다. 고객을 위한 세상이 조직되고, 고객은 데이터 기반의 편의를 재빨리 받아들인다. 현실 세계에서 대부분의 누군가는 알고리즘에 따라 고객의 선호까지 결정할 수 있는 이들 기업의 고객이고, 거대 다국적 기업의 플랫폼 이용자만으로도 몇 개의 국가를 설립하고도 남는다.

2. 알고리즘 활용 기술의 파급력에 대한 법적·사회적 논의의 필요성

대한민국 정부는 이른바 ‘4차 산업혁명’이라는 기치 아래 데이터와 인공지능을 결합하면 새로운 산업을 창출할 수 있다며 데이터경제 활성화를 위한 규제혁신의 필요성을

20) 사물 인터넷은 감지기(sensor)를 이용해 주변 환경과 커뮤니케이션을 할 수 있는 플랫폼으로서 ‘만물 인터넷’으로 부르기도 한다.

21) 東浩紀, 一般意志 2.0 ルソー, フロイト, グーグル, 講談社, 2011, 97面; 아즈마 히로키는 사회 구성원의 욕망의 이력을 본인의 의식적이고 능동적인 의지 표명과는 관계없이 조직적으로 축적하여 이용이 가능하도록 바꾸는 사회를 “총기록사회(總記録社会)”로 개념화하기도 한다. 이러한 사회에서 사람들의 의지는 데이터로 변환되어 수학적 존재로 바뀌기 때문에 이때 필요한 것은 “적절한 접속 권한과 분석 알고리즘의 설정뿐”인데, 이를 몇몇 민간 기업이 독점하고 있는 것은 문제라고 지적한다(같은 책, 90~111面 참조).



언급하지만 그 바탕에는 미래 산업의 부흥에 지향된 낙관적 전망만이 가득하다. ‘정보경제’ 논의는 이미 1977년 미국 정부의 전기통신부에서 포라트(M. Porat)의 연구서가 발간되면서 활발하게 이루어졌던 것이다. 포라트는 국민총생산(GNP)의 증가율이 ‘정보활동’에 의존하고 있으며, 일자리의 증가율은 ‘정보노동’에 의존한다고 주장했다.²²⁾

또한 이른바 ‘정보사회’ 양상의 확산 속에서 기술은 “지배의 한 조건”이 되기도 한다.²³⁾ 최근 빅데이터 분석에 관한 미국 정부의 보고서에 따르면, “빅데이터 분석은 주택, 신용, 고용, 건강, 교육, 시장의 개인정보를 이용하는 방식에 있어서, 오랜 세월에 걸쳐 정착된 시민권 보호 조치를 무의미하게 만들 가능성이 있다.”²⁴⁾고 밝히면서, 사회적 약자에 해당하는 집단은 특히 심한 타격을 입을 것이라고 경고하고 있다.

인공지능 같은 알고리즘 기반의 기술을 발전시킨 국가에서는 기술을 이용한 산업 발전 논의에 매몰되어 있지 않고, 그에 보조를 맞추어 그러한 기술이 사회의 제반 영역에 미치는 영향에 대한 연구에 상당한 비중을 할애하고 있다.²⁵⁾ 예를 들어 미국 정부와 함께 인공지능 기술이 미치는 사회적 및 경제적 영향력에 대한 심포지엄을 주최했던 뉴욕대학교는 2016년부터 인공지능(AI)에 관한 보고서를 매년 발간하고 있을 뿐만 아니라²⁶⁾ 학계에서는 이미 어느 정도 논의가 진전되어 실제

22) Marc Uri Porat, “The Information Economy: Definition and Measurement”, Office of Telecommunications (U. S. Department of Commerce), 1 May 1977, 특히 p. 43 이하 및 p. 105 이하 참조.

23) 임흥빈, 기술문명과 철학, 문예출판사, 1995, 81쪽.

24) Executive Office of the President, “Big Data: A Report on Algorithmic Systems, Opportunity and Civil Rights”, May 2016, https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

25) National Science and Technology Council, Executive Office of the President, “Preparing for the Future of Artificial Intelligence”, October 2016, https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf 참조; 동 보고서에서 제시된 23개의 권고(recommendations)만을 주려서 국내에 소개한 것으로 김희연, “미국 국가과학기술위원회(NSTC)의 인공지능(AI) 기반 준비를 위한 권고안”, 정보통신방송정책 28(19), 2016, 12~19쪽 참조.

26) 뉴욕대학교의 AI Now 연구소(The AI Now Institute)는 2016년부터 인공지능 기술이 사회적·경제적으로 미치는 영향에 관한 보고서인 ‘The AI Now Report’를 매년 발간하고 있다. 이에 관한 자세한 내용은 <https://ainowinstitute.org> 및 Kate Crawford et al., “The AI Now Report: The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term”, A Summary of the AI Now public symposium(7 July 2016), Hosted by the White House and New York University’s Information Law Institute, 2016; Kate Crawford et al., “AI Now 2017 Report”, The AI Now Institute at New York University, 2017; Meredith Whittaker et al., “AI Now Report 2018”, The AI Now Institute at New York University, December 2018 참조; 국내에서는 2016년에 한국법제연구원에서 인공지능 기술에 관한 법적 연구 보고서(손승우·김윤명, 인공지능 기술 관련 국제적 논의와 법적 대응방안 연구, 한국법제연구원, 2016)를 발간했고, 카이스트(KAIST)는 4차산업혁명에 수반하는 정책 및 거버넌스 이슈에 대한 새로운 접근방식을 개척하기 위해 2017년 카이스트 연구소(KI)에 4차산업혁명지능정보센터(FIRIC)를 설립했으며(<https://kis.kaist.ac.kr> 참조), 서울대학교 법과경제연구센터는 2017년부터 ‘인공지능 정책 이니셔티브(AI Policy Initiative)’ 프로그램을 운영하기 시작해서 2017년에는 ‘인공지능, 알고리즘, 개인정보보호를 둘러싼 정책적 과제’를, 2018년에는 ‘인공지능의 시대: 기술발전에 따른



인공지능 기술의 선두에 있는 글로벌 기업의 정책에 반영되는 수준에 이르렀다.

예를 들어 2018년 6월 7일 구글(Google)의 최고경영자 피차이(S. Pichai)는 공식 블로그를 통해 구글이 개발하고 적용하는 인공지능이 준수해야 할 7개 원칙을 목표로 제시했다. 구체적으로는 전체적인 해악을 발생시키거나 발생시킬 것 같은 기술, 무기처럼 사람을 해치는 것이 주된 목적인 기술, 국제규범을 위반하는 감시를 위해 정보를 수집하거나 이용하는 기술, 널리 수용된 국제법과 인권을 위반 또는 침해하는 목적을 가진 기술 영역에는 인공지능을 설계하거나 배치하지 않을 것이라고 밝혔다.²⁷⁾

어쩌면 이러한 결론만을 취합하여 정책에 반영하거나 법제도에 이식하는 것이 변화에 가장 빠르고 쉽게 대처하는 한 방법이 될 수도 있다. 그러나 결과물에 이르는 논의 자체에 대한 이해 없이 다른 사회에서 생산된 결과물만을 이전시키려면 사회의 학습 기회를 박탈시키는 위험을 감수해야 할 것이다.

3. 결정 기관에 대한 불신과 인공지능에 대한 기대

2016년 대한민국에서 국민의 불신은 행정부로 향해 있었다. 그리고 2018년 그 불신은 사법부로 향해 있다. 비록 바둑 게임에 한해서이지만 인공지능이 인간 대표보다 우월한 지적 능력을 가질 수 있다는 점을 확인한 이래로 ‘인공지능으로 ○○○을 대체하는 것이 낫겠다.’는 문구는 인공지능으로 대체되는 목적어 ‘○○○’의 자리에 올려놓는 그 대상에 대한 불신을 표현하는 관용어처럼 되었다. 그 대상은 ‘대통령’에서 ‘법원’으로 옮겨 갔고, 그 다음은 헌법상 권력분립의 또 다른 축인 ‘국회’가 될 수도 있다. 또한 그 목적어 자리에는 국가기관의 공무원뿐만 아니라 정당이나 기업, 시민 단체의 대표나 임원 등 인간이 권한을 갖고 결정 업무를 수행하는 그 어떤 지위도 해당될 수 있다.

인간이 권한을 행사하는 지위가 인공지능으로 대체되길 바라는 기대에는 인공지능이 인간보다 객관적이고 합리적이면서 공정할 것이라는 인식적 기대와 함께 그 권한 행사는 객관적이고 합리적이면서 공정해야 한다는 규범적 기대도 자리 잡고 있다. 따라서 실제로 인공지능이 인간을 대체해 그 권한을 행사할 수 있게 되더라도 인공지능이 객관적이고 합리적이면서 공정한 결정을 내리지 못한다면 동일한 규범적 기대가 유지되는 한 인공지능 역시 불신의 대상으로 옮겨질 수 있다.

책임과 규제’를 주제로 Facebook, Google, Kakao, Microsoft, NAVER, SAMSUNG, SK telecom 등 인공지능 연구를 주도하는 기업의 후원을 받아 학술대회를 개최했다(<http://ai.re.kr> 참조). 서울대학교 법과경제연구센터에서 펴낸 최근 문헌으로는 데이터 이코노미, 한스미디어, 2017 참조.

27) Sundar Pichai, “AI at Google: Our Principles”, Google, 7 June 2018, <https://www.blog.google/topics/ai/ai-principles>, 접속일: 2018년 6월 13일.



제2절 연구의 과제

I. 연구 대상과 목표

1. 연구 대상 및 범위

컴퓨터 프로그램이 인간의 특성을 대표하는 데이터를 분류함으로써 인간의 집단을 구별할 뿐만 아니라 데이터 분석을 통해 사회적 약자에 해당하는 집단에 특별히 심한 타격을 가하는 방식으로 인간에 대해 중대한 영향을 미친다는 것은 컴퓨터 프로그램을 구성하는 알고리즘의 작동과 차별이 일정한 맥락에서 연결될 수 있다는 점을 암시한다. 본 논문에서 연구하고자 하는 것 역시 알고리즘, 그 중에서도 인공지능의 기반이 되는 머신러닝 알고리즘이 사회에 미치는 여러 가지 영향중에 차별의 효과를 발생시키는 것과 밀접하게 관련되어 있고, 이러한 차별 관련성이 본 연구의 범위에 경계를 설정한다.²⁸⁾

대한민국 헌법 제11조 제1항 2문에 따르면 “누구든지 성별·종교 또는 사회적 신분에 의하여 정치적·경제적·사회적·문화적 생활의 모든 영역에 있어서 차별을 받지 아니한다.”²⁹⁾고 하여 차별은 그 누구도 어떤 영역에서건 받으면 안 되는 그 무엇이다. 그러나 차별 그 자체가 무엇인지에 대해서 헌법이 직접 말해주는 바는 없다. 특히 평등을 비롯해 다른 헌법의 개념과의 관계에서 이해되는 차별은 대단히 복잡한 내용을 가질 수 있고, 개별 사유에 기초한 개별 집단을 상정하고 발전해 온 차별에 관한 법률에서 정의된 차별 개념은 단순하기보다는 복잡한 해석 가능성을 수반하고 있다. 그렇기 때문에 복잡한 해석과 구성이 전개되는 차별의 개념 및

28) 인공지능과 관련된 포괄적인 규범적 이슈를 요약 정리한 것으로 이원태, 인공지능의 규범이슈와 정책적 시사점, KISDI Premium Report, 정보통신정책연구원, 2015 참조. 이 보고서는 인공지능과 관련된 규범적 이슈를 ‘인공지능의 자율성 범위’와 ‘인간의 권리침해 및 통제권’으로 구분하면서 전자의 세부 이슈로 인공지능의 법적 책임, 인격성, 신뢰성, 안전성 그리고 인공지능의 사회문화적 영향에 대한 평가체계구축 등의 문제를 거론하고, 후자의 세부 이슈로 인권 개념 재정립, 프라이버시(privacy) 보호, 알고리즘의 책임성 강화, 갈등조정 거버넌스 구축, 정보격차 해소, 인간고유 역량 강화 등의 문제를 제기하지만(11쪽), 차별 관련 이슈가 직접적으로 언급되어 있지는 않다.

29) 헌법 제11조 전체의 규정은 다음과 같다. “제11조 ① 모든 국민은 법 앞에 평등하다. 누구든지 성별·종교 또는 사회적 신분에 의하여 정치적·경제적·사회적·문화적 생활의 모든 영역에 있어서 차별을 받지 아니한다. ② 사회적 특수계급의 제도는 인정되지 아니하며, 어떠한 형태로도 이를 창설할 수 없다. ③ 훈장등의 영전은 이를 받은 자에게만 효력이 있고, 어떠한 특권도 이에 따르지 아니한다.”



차별금지법(반차별법, Anti-Discrimination Law)의 이론에 대한 관찰은 연구 수행에 필수적인 내용이 된다.³⁰⁾ 무엇보다 머신러닝 알고리즘에 의한 어떤 결정이 차별적이라고 인식하고 평가하기 위해 그 전제가 되는 차별의 개념과 그에 관한 이론이 탐색되어야 한다.³¹⁾ 또한 인간에 관한 데이터를 알고리즘이 처리할 때 차별이 발생한다는 점은 개인정보를 보호하는 것이 차별로부터 보호하는 것과 어떤 관련을 맺을 수 있는지 살펴봐야 하는 이유가 된다.

머신러닝 알고리즘은 헌법학을 포함해 법학의 관점에서 볼 때 낯설거나 이질적일 수 있다. 더구나 머신러닝 알고리즘에 따라 작동하는 에이전트가 인간의 활동을 지각하고 차별적 결정을 할 수 있다는 의심은 그 자체로도 낯설고 이질적이다. 또한 머신러닝 알고리즘이 인간의 직관 또는 사회의 의미체계 속에서 차별 개념으로 포착될 수 있는 어떤 결정을 한다는 것은 사실이라기보다는 가설적 주장에 가깝게 보일 수도 있다. 그렇기 때문에 차별의 의심을 받는 결과를 도출하는 알고리즘이 인간이 생산해 낸 사회의 무수한 데이터로부터 추출된 일종의 규칙이라는 점에서 이를 가능하게 하는 머신러닝에 대한 개념 이해는 가설적 주장의 사실적 조건을 정립하는 기초가 된다. 다만, 머신러닝 알고리즘 자체가 연구의 대상은 아니기 때문에 차별과 관련된다고 볼 수 있는 수준에서 기초적인 개념을 다루는 정도에 머물 수밖에 없고, 이 점은 본 논문에서 수행하는 연구의 한계이면서 동시에 인공지능 관련 기술 자체를 분석하고 설명하고 성찰할 수 있는 제반 학문 분과, 예컨대 컴퓨터 과학, 기술공학, 신경과학, 제어이론과 사이버네틱스(cybernetics), 수학, 통계학, 경제학, 심리학, 언어학, 철학, 사회학 등의 융합적인 관심을 촉구하는 계기가 될 수 있을 것이다.³²⁾ 차별은 법학 분야뿐만 아니라 여러 학문 분야에서 연구가 필요한 분야이고

30) 대한민국의 법체계에서 차별금지 사유나 차별금지 영역과 관련된 개별 법률이나 법규정은 있지만 일반적 차별금지법은 아직 제정되어 있지 않다. 차별금지법을 사인에게 적용되는 평등규범의 대표적인 입법형태로 보는 헌법적 이해는 윤영미, “평등규범의 사인에 대한 적용”, 헌법학연구 19(3), 2013, 39~78쪽 참조.

31) 이러한 방향 설정은 본 연구가 구체적인 차별금지법체계의 세부적 내용이나 평등에 관한 헌법의 일반이론보다는 차별에 관한 법이론에 좀 더 비중을 두는 것과 무관하지 않다.

32) 이른바 ‘포스트휴먼 사회’의 도래를 준비하는 가장 좋은 방법으로 ‘융합’이 제시되기도 한다. 공학, 과학, 철학, 법학, 사회학, 행정학, 교육학 등 다양한 분야의 연구자 및 전문가들이 공통의 주제에 대해 다양한 측면의 접근법을 제시하는 예로 한국포스트휴먼학회(편저), 포스트휴먼 시대의 휴먼, 아카넷, 2016 참조; 포스트휴먼에 관해서는 Rosi Braidotti, *The Posthuman*, Polity Press, 2013 참조; 한국법제연구원은 2016년에 일본(나채준, 지역법제 연구 16-16-③-1), 미국(윤인숙, 지역법제 연구 16-16-③-2), 독일(장원규, 지역법제 연구 16-16-③-3), 영국(권건보, 지역법제 연구 16-16-③-4), 캐나다(윤성현, 지역법제 연구 16-16-③-5), 프랑스(정관선, 지역법제 연구 16-16-③-6) 등의 ‘포스트휴먼 기술법제에 관한 비교법적 연구’를 수행한 바 있다.



실제로도 학제 간 연구가 수행되고 있다.³³⁾

그런데 위에 언급된 학문 분과 중 대부분은 현대의 인공지능 연구 발전에 밀접하게 연관되어 있기도 하다. 그렇기 때문에 어떤 의미에서 머신러닝 알고리즘과 차별을 연결시키는 본 연구는 ‘인공지능과 법(AI & Law)’³⁴⁾이라는 연구 분과의 하위 주제를 다루려는 시도로 볼 수도 있다. 인공지능에 접근하는 기초가 되는 머신러닝 기술로 생성된 알고리즘은 인간에게 매우 난해하다. 기본적으로 알고리즘의 세계에서는 인간이 이해할 수 있는 차원의 수를 초과하는 경우가 비일비재하기 때문이다. 차별의 개념이 복잡한 만큼 그 이상으로 복잡한 머신러닝 알고리즘의 특성은 머신러닝 알고리즘에 의해 발생하는 차별 현상에 대한 연구가 단순하게 정리된 결과를 도출하는 것이 아니라 그 난해함을 드러내는 과정에 연구의 초점을 맞추게 한다. 그러나 머신러닝 알고리즘의 결정과 차별의 연결 가능성에 대한 의심에서 비롯된 이러한 작업은 머신러닝 알고리즘의 어떤 특성이 차별의 문제를 야기하는지 그리고 차별에 대한 어떤 개념적 이해가 머신러닝 알고리즘의 결정을 헌법이 요청하는 금지된 차별로 포착하도록 하거나 포착하지 못하도록 하는지 구체화할 수 있는 몇 가지 관점을 제공해 줄 수 있을 것이다.

2. 연구 목표

머신러닝 알고리즘의 기계적 작동 결과와 법적 또는 도덕적 차원에서 규범적 의미를 갖는 차별을 연결시키는 것은 자칫 무리한 시도나 난센스(nonsense)처럼 보일 수도 있다. 그런데 이러한 관점의 한 쪽에는 은연중에 차별은 인간만이 할 수 있다는 직관 또는 고정관념이 자리 잡고 있다. 헌법이 작동하는 현실만을 보더라도 차별은 인간만이 할 수 있는 것은 아니라는 점을 쉽게 알 수 있다. 헌법은 이미 차별의 주체 또는 차별의 원인 제공자로 국가를 배제하지 않으며 오히려 차별에 관한 규범의 조건으로서

33) Andrea Romei · Salvatore Ruggieri, “A Multidisciplinary Survey on Discrimination Analysis”, *Knowledge Engineering Review* 29(5), 2014, pp. 582~638; 인공지능에 의한 차별에 관한 철학 분야의 국내 연구로는 허유선, “인공지능에 의한 차별과 그 책임 논의를 위한 예비적 고찰 - 알고리즘의 편향성 학습과 인간 행위자를 중심으로 -”, *한국여성철학* 29, 2018, 165~210쪽이 있으나, 해당 연구는 내부적으로 언급되어 있듯이 “차별 개념 자체의 심화 고찰을 수행하고 있지 않다.”(202~203쪽 참조). 법이론뿐만 아니라 정치철학과 도덕철학에서도 중요한 차별 관련 문제를 다룬 여러 문헌을 차별의 개념적 이슈, 차별의 부당성, 피차별자 집단, 차별 영역, 차별의 원인과 수단, 차별의 역사 등으로 구분하여 모아 놓은 것으로 Kasper Lippert-Rasmussen (Ed.), *The Routledge Handbook of the Ethics of Discrimination*, Routledge, 2018 참조.

34) ‘인공지능과 법’ 연구의 약사(略史)를 주요 문헌에 관한 소개와 함께 정리한 것으로 Trevor Bench-Capon et al., “A History of AI and Law in 50 Papers: 25 Years of the International Conference on AI and Law”, *Artificial Intelligence and Law* 20(3), 2012, pp. 215~319 참조.



국가의 차별을 전제하고 있기 때문이다. 국가를 그 어떤 사회적 기능으로 본다면³⁵⁾ 국가의 차별은 국가기관의 작용을 통해 생산되고 발생할 수 있다. 국가를 그 자체로 하나의 법질서라고 본다면³⁶⁾ 국가는 법질서가 형성되고 집행되며 적용되는 절차와 과정에서 차별을 양산할 수 있다. 이러한 가능성을 전제로 차별에 관한 헌법 모델은 국가가 차별의 직접 생산자가 되는 역할에서 뿐만 아니라 이미 발생한 차별을 그대로 유지하고 지속시키는 방관자의 역할로부터도 멀어지도록 구성될 수 있다.

홉스(T. Hobbes)의 인공동물(*artificial animal*)에 관한 생각은 국가의 차별 가능성과 머신러닝 알고리즘에 기초한 인공지능의 차별 가능성을 연계시킬 수 있는 하나의 철학적 논거가 된다.³⁷⁾ 홉스는 그의 저서 “리바이어던(*leviathan*)”³⁸⁾에서 인공동물에 대해 말한다. 그 동물은 자연적이지 않다. 인간을 보호하기 위해 인공적으로 만들어진 것이다. 그 인공동물은 톱니바퀴가 맞물려 돌아가는 시계처럼 잘 작동한다. 그 인공동물이 기괴한 고래 모양을 가진 리바이어던이고, 이는 코먼웰스(*commonwealth*) 즉, 오늘날의 국가를 상징한다. 국가는 결코 자연이 아니며 인공물(*artifact*) 그 자체인 것이다. 인간을 보호하는 인공동물로서 리바이어던에 관한 이러한 생각은 인공지능 개발에 철학적 근거를 제공하는 주요 사상 중 하나다.³⁹⁾

국가에 대해 인간을 보호하는 인공물로 보려는 관점은 오늘날 인공지능 또는 머신러닝 알고리즘으로 작동하는 시스템에 대해 인간을 보호하는 인공물 또는 인간에게 편익을 제공하는 기술로 보는 관점과 닮아 있다. 그리고 국가가 차별의 생산자나 방관자가 되어서는 안 된다는 규범적 태도는 인공지능 또는 머신러닝 알고리즘이 차별을 재생산하거나 확대해서는 안 된다는 규범적 태도와 맞닿아 있다. 이러한 전제에서 출발하는 머신러닝 알고리즘의 차별에 관한 연구는 단지 머신러닝 알고리즘을 활용한 기술과 차별에 관한 법을 단순히 대조하는 데에만 정향되어 있지 않다. 오히려 인간을

35) 엘리네크(G. Jellinek)에 따르면 사회적 개념으로서 국가는 실체(*Substanz*)가 아니라 기능(*Funktion*)이다. Georg Jellinek, *Allgemeine Staatslehre*, 3. Aufl., Verlag von Julius Springer, 1929, S. 174-182 참조.

36) 켈젠(H. Kelsen)은 국법학과 국가사회학을 구분하고, 법학의 대상으로서 국가는 법질서 전체이거나 부분으로 한정한다. 이에 따르면 국가를 법인(*Juristische Person*)으로 볼 경우에 한해 국가의 행동은 법규범의 내용이 된다. 국가는 본질상 규범체계 또는 이러한 규범체계의 통일성에 대한 표현으로서 다른 질서가 통용되는 것을 허용하지 않는 실정적 법질서인 것이다. Hans Kelsen, *Allgemeine Staatslehre*, Österreichische Staatsdruckerei, 1925, S. 6~7 및 S. 16~17 참조.

37) 홉스(T. Hobbes)의 국가철학과 법철학에 관한 내용은 윤재왕, “개인주의적 절대주의: 토마스 홉스의 국가철학과 법철학에 관하여”, *원광법학* 28(2), 2012, 7~35쪽 참조.

38) Thomas Hobbes, *Leviathan*, J. C. A. Gaskin(Ed.), Oxford University Press, 1998, p. 7 이하 참조.

39) Stuart J. Russell · Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed., Prentice Hall, 2010, p. 6.



이롭게 하면서 동시에 인간을 해롭게 할 수 있는 인공물의 현대적 대용물로서 인공지능의 기초가 되는 머신러닝 알고리즘에 의해 이용자의 목적이나 의도와 상관없이 차별을 재생산하고 확대시킬 가능성을 살펴봄으로써 간접적으로 유사한 맥락 속에 있는 인공물로서 국가와 사회의 시스템에 의한 차별에 대한 이해를 넓히는 것을 목표로 한다. 따라서 본 연구는 국가를 포함한 인공물의 차별에 관한 연구로 확장되는 것에 개방되어 있고, 국가와 사회, 그리고 인간의 차별을 그에 관한 모방 시스템을 통해 반성적으로 검토하는 것에 지향되어 있다.

II. 연구 방법과 순서

1. 연구 방법

법학의 엄밀성은 그 분석적 차원의 수준에서 결정된다고 해도 과언이 아니다.⁴⁰⁾ 가장 추상적인 법규범이라고 할 수 있는 헌법에서부터 가장 구체적인 법규범이라고 할 수 있는 개별 사례에 대한 법적 판단에 이르기까지 다양한 수준에서 차별은 문제될 수 있다. 특히 차별은 그 개념과 이론 구성이 복잡할 뿐만 아니라 개념 설정과 이론 구성에 따라 머신러닝 알고리즘이 야기하는 현상을 포착하기 위한 차별의 형식과 내용이 달라질 수 있기 때문에 차별의 개념을 분석하고 이론의 의미를 파악하는 작업은 매우 중요하다. 헌법학에서 전통적으로 전개되는 평등에 관한 원리나 이념적 차원의 접근도 중요하지만 법의 문제이면서 동시에 사회의 문제이기도 한 차별은 다양한 분야의 연구 대상이기도 하다는 점을 고려해 차별에 직접적으로 접근해 보고자 한다.⁴¹⁾ 이를 위해 차별을 평등과 개념적으로 반대 관계에 있다는 주장이나 전제로부터도 한 발 물러선다. 물론 그렇다고 해서 차별과 평등이 밀접한 관련 속에 있다는 점을 배척하지는 않으며, 차별을 법적으로 인식하고 그에 대해 판단하여 책임을 귀속시키는 법체계의 알고리즘은 본 연구의 전체 흐름 속에 반영되어 있다.

40) 알렉시(R. Alexy)는 법학을 분석적 차원, 경험적 차원, 규범적 차원으로 구분한다. 그에 따르면 실천적 학문분과로서 법학은 이 세 가지 차원을 하나로 묶는 원리이고, 이 경우 법학은 다차원적이면서 통합적인 학문분과이어야 한다. Robert Alexy, *Theorie der Grundrechte*, 1. Aufl., Suhrkamp, 1994, S. 22~27 참조.

41) 이와 유사한 접근법을 취한 선행 연구로 이준일, *차별금지법*, 고려대학교출판부, 2007 및 같은 이, *차별없는 세상과 법*, 홍문사, 2012 참조. 다만, 해당 연구는 차별을 평등의 반대개념으로 보고 차별에 대한 이해는 반드시 평등에 대한 이해를 동반하는 것으로 전제한다.



또한 차별에 관해 다른 분야에서 수행한 연구의 도움도 받을 것이다.⁴²⁾ 실제로 차별에 관한 연구는 법학 분야뿐만 아니라 여러 학문 분야에서 수행되고 있다.⁴³⁾ 더구나 본 연구대상은 차별만을 다루는 것이 아니라 머신러닝 알고리즘을 함께 다루고 있다. 그러므로 머신러닝 알고리즘에 관한 개념 이해에서부터 작동 원리나 방식에 이르기까지 컴퓨터과학, 인공지능 연구, 통계학 등 해당 기술을 연구하는 다양한 분야의 도움을 받는 것은 필수적이다. 또한 머신러닝 알고리즘과 차별을 연결시키는 것이 가상 속의 인위적인 작업이 아니라는 점은 과학적으로 연구된 실증적 사례뿐만 아니라 일상생활에서 경험할 수 있는 사례를 제시하고, 이를 바탕으로 간단한 실증적 실험 또는 사유 실험을 통해 그 근거를 제시할 것이다. 이러한 과정은 기술적 작동이지만 지능적으로 보이기도 하는 머신러닝 알고리즘의 작동 결과가 경험의 문제이면서 규범의 문제인 차별의 관계를 어떻게 설정할 것인지 모색하는 것이기도 하다.

그리고 차별의 형식이나 구조 유형 그리고 이를 일관되게 설명하기 위한 철학적 및 도덕적 기초에 관한 최근의 논쟁들도 적극 반영하고자 한다. 나아가 차별은 법에 의해 제한을 받기도 하지만 법이 차별의 근거가 될 수도 있다는 점을 엿보기 위해 알고리즘과 법의 비교를 염두에 두고 논의를 전개한다. 이때 알고리즘이 구현된 양식을 사회의 기술로서 이해하기 위해 기술철학이나 과학기술학의 관점들도 고려하고, 알고리즘이 규제적 속성을 갖는 사회의 규범으로 이해하기 위해 법철학과 법사회학의 통찰을 빌려올 것이다. 따라서 각 분과에서 지극히 당연하거나 기본적인 개념이나 내용이 본 연구에서 새삼스럽게 부각되거나 누락되는 것은 외부적 관찰자 또는 내부적 참여자로서 갖는 맹점이 반영된 한계일 수 있다.

2. 연구 순서

알고리즘의 차별은 알고리즘 시스템이 사람에게 중요한 영향을 미치는 사회의 결정 시스템을 지원하거나 대체하는 과정에서 발생하는 문제이다. 그러므로 우선 알고리즘 시스템이 사회의 결정 시스템과 연계될 수 있도록 하는 알고리즘의 접근 방식을 살펴볼 것이다. 알고리즘이 지원하거나 대체하는 사회의 결정 절차나 과정은 법규범을 비롯해 다른 규범이나 절차적 보호 장치와 결부되어 있기 때문에 사회의

42) 이준일, *차별금지법*, 고려대학교출판부, 2007, 5쪽에서는 “평등을 단지 법학적 측면에서만 다룰 것이 아니라 다양한 사회과학의 도움을 받아 다루어야” 한다는 점이 강조된다.

43) Andrea Romei · Salvatore Ruggieri, “A Multidisciplinary Survey on Discrimination Analysis”, *Knowledge Engineering Review* 29(5), 2014, pp. 582-638.



결정 절차나 과정에서 차별을 방지하기 위한 사회의 시스템, 특히 차별금지법체계를 살펴보아야 한다. 그리고 무엇보다 중요한 것은 과연 차별이 머신러닝 알고리즘의 작동방식과 관련해서 어떻게 적용될 수 있는 개념인지 검토하는 것이다. 차별 아니면 무차별, 평등 아니면 불평등, 자유 아니면 억압, 존중 아니면 경멸의 이분법적 도식을 통해 법 또는 불법의 이진 코드와 연결된 단순화한 차별 개념의 이면에 있는 차별의 복잡한 맥락과 의미를 탐색하는 것은 필수적이다. 이를 위해 차별에 관한 이론적 구성을 살펴보아야 한다. 특히 실정법체계에서 작동하고 있는 차별금지법에 관한 이론을 검토하고 알고리즘의 차별 상황을 맞아 어느 지점에서 이론적 한계나 모순이 발생하는지 면밀히 살펴볼 필요가 있다. 또한 알고리즘에도 차별 개념이 적용될 수 있다면 차별의 한 유형으로 시스템의 차별이 보다 비중 있게 다루어질 수 있을 것이다.

알고리즘의 사용은 사회에 이익을 안겨다 주지만 그에 못지않은 위험을 사회에 가져다준다. 순차 및 반복 등을 특징으로 하는 알고리즘과 그에 따라 형성된 시스템에 따라 차별이 발생한다는 것은 차별이 반복적으로 그리고 자동적으로 재생산될 수 있다는 점에서 알고리즘 또는 시스템에 의한 차별은 그러한 위험의 하나로 이해될 수 있다. 이에 대한 책임을 어떻게 구성할 것인지 위험을 분산시키는 측면과 차별에 대한 책임을 제한하는 법이론 구성의 측면이 함께 고려되어야 할 것이다. 이러한 연구 과정에는 해석적 측면과 정책적 측면이 중첩되고, 문제 중심의 측면에서는 머신러닝 알고리즘의 데이터 수집, 분석, 이용의 단계에서 개인정보 보호 문제가 차별로부터의 보호 문제와 중첩된다. 이에 관해서는 기존의 법체계에 대한 해석을 관찰하는 측면과 새로운 해석에 대한 가능성을 제시함으로써 기존의 실천적 해석이 부딪히는 한계를 드러내 보이고 그러한 해석의 변화 또는 폐기 가능성을 엿볼 수 있게 하는 것이다. 문제 해결에 관한 이론적 가능성은 경우에 따라 지향성을 보여주는 것으로 이해할 수도 있지만, 여러 가지 선택 가능성을 확대시켜주는 것이기도 하다. 어떤 것을 선택할 것인지는 사회를 형성하는 행위자들의 결정에 따라 달라질 것이다.



제2장

머신러닝 알고리즘과 차별의 문제

머신러닝 알고리즘이 단순히 어떤 결정을 하는 것을 넘어 차별적 결정까지 한다는 것은 법학 분야에서는 생소한 논제일 수 있다. 컴퓨터나 수학을 연상시키는 용어 자체가 주는 이질감과 법적 결정이나 판단의 주체는 인간에 국한될 뿐 기계는 결코 그러한 결정이나 판단의 주체가 될 수 없다는 심리학적 직관에 의존한 법적 개념에 차별이라는 복잡한 현상까지 결합되면 이러한 논제가 과연 법학에서 다룰 수 있는 문제인지조차 의구심이 들게 된다. 또한 알고리즘은 그 자체의 독자적 의미보다는 인공지능, 로봇, 에이전트 등과 혼용되거나 결합되어 사용되기 때문에 구체적인 의미를 파악하는 것도 쉽지 않다.

따라서 우선 인공지능과 로봇, 그리고 에이전트의 기술적 및 사회적 의미와 함께 이러한 용어의 사용을 가능하게 하는 과학기술¹⁾이 인류 미래에 대해 미칠 영향에 대한 전망들을 소개한 다음 인공지능과 로봇, 에이전트가 시스템으로서 작동하기 위해 필수적이라고 할 수 있는 알고리즘 자체의 정의와 의미를 정리하도록 한다. 그리고 기계가 인간처럼 지능적으로 학습할 수 있다는 관념이 반영된 머신러닝 알고리즘의 주요 학습 방식에 대해 컴퓨터 과학 분야에서 제시하는 설명을 따라가 보고자 한다.

이렇게 알고리즘을 둘러싼 용어들의 개념과 의미를 개괄한 후에는 좀 더 구체적인 차원에서 일상적으로 접할 수 있는 머신러닝 알고리즘에 의한 결정 중에 차별을 의심하게 만드는 몇 가지 사례를 제시해 보고, 이러한 의심의 전제가 되는 머신러닝 알고리즘의 편향에 관한 실증적 연구를 살펴본다. 그리고 이러한 연구에 착안하여 인간이 쉽게 이해할 수 없는 방식으로 머신러닝 알고리즘이 작동할 수 있다는 점을 환기시킬 수 있는 간단한 실험을 진행하여 그 결과를 분석하고 해석해보도록 한다.

1) 전통적으로 기술과 과학의 관계에 대한 소박한 관념은 기술을 과학의 응용으로만 여겼다. 그러나 오늘날 기술은 더 이상 과학의 응용에만 머물러 있지 않는다. 응용된 과학처럼만 보였던 그 기술이 다시금 과학을 발전시키고 있다. 이에 관해서 홍성욱, “과학과 기술의 상호작용: 지식으로서의 기술과 실천으로서의 과학”, 창작과 비평 22(4), 1994, 329~350쪽 참조.



제1절 인공지능 로봇과 자율적 에이전트

2016년 다보스 포럼에서 시작된 것으로 알려진 4차 산업혁명론은 유독 대한민국의 정부와 기업, 그리고 언론의 관심 속에서 짧은 시간 안에 광범위하게 확산됐다. 하지만 정작 그 실체는 화려한 첨단 과학기술 용어들 뒤에 불분명하게 가려져 있다. 이미 인간의 육체적 노동을 대체하는 기계가 등장했을 때부터 이를 받아들이는 인간의 태도는 각각의 타당한 근거를 가진 대등한 관점들로 팽팽한 대립 관계를 형성해 왔다. 그 한편에는 인류의 진보에 대한 열망이 담겨 있고, 다른 한편에는 인류의 멸망에 대한 공포가 담겨 있다. ‘인공지능 알고리즘’²⁾, ‘인공지능 로봇’, ‘인공지능 에이전트’, ‘로봇 에이전트’, ‘알고리즘 로봇’, ‘알고리즘 에이전트’ 등 알고리즘과 함께 사용되거나 서로 간에 혼합되어 사용될 수 있는 인접 용어들이 어떤 의미를 갖는지 파악하는 것은 다양한 맥락 속에서 해당 용어들이 등장할 때 초래될 수 있는 의미의 혼란을 조금이라도 줄일 수 있는 예비적 고찰이 될 것이다.

I. ‘4차 산업혁명’의 상징적 표현

1. 플레이스홀더로서 ‘4차 산업혁명’

이른바 ‘4차 산업혁명’은 발화자의 이런저런 바람이나 전망을 담을 수 있는 용기라고 할 수 있는 플레이스홀더(placeholder)로 기능하는 대표적인 용어로 꼽힌다. 예를 들면 과학기술학(STS, Science and Technology Studies)에서는 4차 산업혁명이 국가적 의제(agenda)로 설정되면서 “각각의 분야에서 절실했던 요구를 4차 산업혁명이라는 국가적 아젠다에 빚대어 표출”³⁾하게 되고, 그 용어에 걸맞은 실체가 없기 때문에 “4차 산업혁명이라는 공허한 플레이스홀더에 각자 담으려고 하는 이야기들을 모”⁴⁾을 수 있는 것이라고 평가하기도 한다.

2) 인공지능이 모든 일을 대신해 줄 것이라는 막연한 환상을 경계하려는 의도에서 알고리즘을 결합하여 사용하는 경우로 양종모, “인공지능 알고리즘의 편향성, 불투명성이 법적 의사결정에 미치는 영향 및 규율 방안”, 법조 66(3), 2017, 60~105쪽 참조.

3) 홍성욱, “왜 ‘4차 산업혁명론’이 문제인가?”, 김소영·김우재·김태호·남궁석·홍기빈, 4차 산업혁명이라는 유행: 우리는 왜 4차 산업혁명에 열광하는가, Humanist, 2017, 29~52쪽: 42쪽.

4) 김소영, “4차 산업혁명, 실체는 무엇인가?”, 김소영·김우재·김태호·남궁석·홍기빈, 4차 산업혁명



실제로 2017년 3월 헌법재판소의 대통령 탄핵 인용결정⁵⁾이 이루어진 이후 같은 해 5월 실시된 대통령선거를 앞두고 후보들이 하나 같이 4차 산업혁명을 공약으로 내세웠다.⁶⁾ 그리고 국회는 4차 산업혁명 특별위원회를 설치하여 ‘4차 산업혁명 국가 로드맵’ 준비하는가 하면,⁷⁾ 새 정부는 데이터경제 활성화를 위한 규제혁신의 필요성을 언급하며 ‘데이터경제와 AI 활성화 로드맵’을 마련하려고 한다. 물론 대통령령인 “4차산업혁명위원회의 설치 및 운영 규정에 관한 규정” 대통령 소속의 4차산업혁명위원회의 심의 조정·조정 대상으로 ‘4차 산업혁명의 역기능 대응에 관한 사항’(제2조 제7호)을 언급하고 있긴 하다.⁸⁾ 하지만 적어도 “4차 산업혁명 시대에 데이터는 미래 산업의 ‘원유(原油)’이며 AI는 21세기의 ‘전기(電氣)’로서 데이터와 AI의 결합이 다양한 새로운 산업을 만들어 낼 것”⁹⁾이라는 정부 관료의 말을 통해 대표되는 인식으로부터 플레이스홀더에 담아낼 수 있는 것은 미래 산업의 부흥에 지향된 낙관적 전망뿐이다. 또한 동 규정 역시 4차 산업혁명이 “초연결·초지능에 기반”(제1조)을 두고 있다는 수식적 표현 외에 4차 산업혁명 그 자체가 무엇인지에 대해 특별히 정의하고 있지 않다. 어쨌든 이처럼 국가 차원의 정책과제를 제시하려는 시도들이 이어지면서 ‘4차 산업혁명’에 기대어 각 분야의 요구를 표출할 수 있는 기회는 점차 증가하고 있다.

이라는 유령: 우리는 왜 4차 산업혁명에 열광하는가, Humanist, 2017, 11~26쪽: 26쪽.

- 5) 대한민국 헌정사에서 대통령에 대한 국회의 탄핵소추가 헌법재판소의 심판을 받은 경우는 두 번이다. 2004년 노무현 대통령에 대해서는 기각결정이, 2017년 박근혜 대통령에 대해서는 인용결정(파면)이 내려졌다. 이에 관해서 현재 2004. 5. 14. 2004헌나1, 판례집 16-1, 609 및 현재 2017. 3. 10. 2016헌나1, 판례집 29-1, 1 참조.
- 6) “뜨거움 ‘4차 산업혁명’ 공약...창조경제 전철 우러”, 연합뉴스 TV, 2017. 4. 14, <http://www.yonhapnewstv.co.kr/MYH20170414000900038>, 접속일: 2018년 7월 6일.
- 7) 정원영, “국회 4차산업혁명 특위, 4차산업혁명 국가로드맵 초안 발표”, 로봇신문사, 2018. 4. 24, <http://www.irobotnews.com/news/articleView.html?idxno=13764>, 접속일: 2018년 7월 6일.
- 8) 4차산업혁명위원회의 설치 및 운영에 관한 규정[대통령령 제28613호, 2018. 1. 26. 개정; 대통령령 제28250호, 2017. 8. 22. 제정]: “① 초연결·초지능 기반의 4차 산업혁명 도래에 따른 과학기술·인공지능 및 데이터 기술 등의 기반을 확보하고, 신산업·신서비스 육성 및 사회변화 대응에 필요한 주요 정책 등에 관한 사항을 효율적으로 심의·조정하기 위하여 대통령 소속으로 4차산업혁명위원회를 둔다. ② 제1항에 따른 4차산업혁명위원회(이하 "위원회"라 한다)는 다음 각 호의 사항을 심의·조정한다. ... (중략) ... 7. 4차 산업혁명에 대응한 법·제도 개선 및 역기능 대응에 관한 사항; (후략)...”
- 9) 이설영, “4차 산업혁명 핵심 ‘데이터 경제’ 활성화 위한 TF 발족”, 2018. 9. 6, <http://www.fnnews.com/news/201809061409461558>, 접속일: 2018년 9월 16일.



2. ‘4차 산업혁명’을 상징하는 용어들과 그 관계

‘4차 산업혁명’이란 단어를 건인하는 것처럼 보이는 ‘인공지능(artificial intelligence)’ 역시 구체적이고 실체적인 내용을 담지 못한 채 종종 ‘4차 산업혁명’에 버금가는 플레이스홀더의 기능을 수행한다. 그리고 인공지능을 실체화한 것처럼 보이는 ‘로봇(robot)’, 그리고 이들을 통칭하는 것처럼 보이는 ‘에이전트(agent)’ 및 에이전트의 성격을 규정하기 위한 ‘자율성(autonomy)’마저 그러한 플레이스홀더의 기능을 공유한다. 그리고 개념화나 가시화하기 어려워 비교적 잘 드러나지는 않지만 그 중심에는 ‘알고리즘(algorithm)’이 있다. 알고리즘을 소재로 논의를 구성하는 경우 인공지능, 로봇, 에이전트, 자율성 등은 관행적으로 따라 붙다시피 하는 인접 용어들이기도 하다.

이 용어들은 맥락에 따라 서로 대체되기도 하고 서로 간의 성격을 규정해 주기도 한다. 예를 들어 세계 최초의 기업 이사로 발탁된 컴퓨터 프로그램인 ‘VITAL’은 이를 소개하는 언론의 보도 속에 알고리즘(algorithm),¹⁰⁾ 로봇(robot),¹¹⁾ 인공지능(artificial intelligence)¹²⁾ 등으로 언급되고, 인공지능과 법 연구(Artificial Intelligence and Law Studies) 분야에서는 인공지능, 로봇, 인조인(synthetic person)을 상호 교환적으로 사용하기도 한다.¹³⁾ 인간과 기계의 바둑 게임이라는 세계적 이벤트가 열린 대한민국 사회의 맥락에서 알고리즘을 그 중심에 두고 자율성을 획득한 인공지능이 탑재된 로봇이 에이전트로서 사회의 행위자가 될 수 있다는 점을 감안하면 이를 모두 조합한 ‘자율적 인공지능 알고리즘 로봇 에이전트’는 마치 ‘4차 산업혁명’의 상징처럼 여겨진다. 그러나 그 출처를 단순화해서 대별해보자면 각각의 단어 즉, 자율성, 인공지능, 알고리즘, 로봇, 에이전트는 (법)철학, 컴퓨터과학, 수학, 로봇공학(robotics), 사회과학(또는 법학)의 용어이며, 각각의 함의도 조금씩 차이가 있다. 그러면 알고리즘과 자주 결합되어 사용될 수 있는 인공지능, 로봇, 그리고 에이전트에 대해 먼저 살펴보도록 한다.

10) “Algorithm Appointed Board Director”, BBC News, 16 May 2014, <https://www.bbc.com/news/technology-27426942> 및 본 논문 「제1장 제1절 III. 1. 인간에 대해 중요한 결정을 하는 알고리즘」 참조.

11) Zolfagharifard, Ellie, “ROBOT Becomes the World’s First Company Director”, Mail Online, 19 May 2014, <http://www.dailymail.co.uk/sciencetech/article-2632920/Would-orders-ROBOT-Artificial-intelligence-world-s-company-director-Japan.html>, 접속일: 2018년 6월 14일.

12) Andreas von der Heydt, “First Time Ever: Artificial Intelligence Nominated as a Board Member”, 20 May 2014, <https://www.linkedin.com/pulse/20140520045550/-175081329-first-time-ever-artificial-intelligence-nominated-as-a-board-member>, 접속일: 2018년 6월 14일.

13) 예를 들어 Joanna J. Bryson · Mihailis E. Diamantis · Thomas D. Grant, “Of, for, and by the People: The Legal Lacuna of Synthetic Persons”, *Artificial Intelligence and Law* 25(3), 2017, pp. 273-291 참조.



II. 인공지능 연구와 학습하는 기계

1. 인공지능 개념과 학제 연구의 집약

인공지능에는 여러 학문 분과의 노력이 집적되고 결합된다. 인공지능 연구의 기반이 되는 학문으로 철학, 수학, 경제학, 신경과학, 심리학, 컴퓨터공학, 제어이론과 사이버네틱스(cybernetics), 언어학 등이 거론된다.¹⁴⁾ 인공지능에 관한 현대적 접근 방식에서는 “환경으로부터 지각(percepts)을 받고 동작(actions)을 수행하는 에이전트(agents)에 대한 연구”¹⁵⁾로 인공지능을 정의한다.

‘인공지능(artificial intelligence, AI)’이란 용어는 매카시(J. McCarthy)가 1956년 여름 두 달간 미국 뉴햄프셔(New Hampshire) 주(州) 동북부 해노버(Hanover)에 있는 다트머스(Dartmouth) 대학에서 열린 워크숍에서 10명의 참석자에게 인공지능 연구를 제안하는 과정에서 공식적으로 처음 사용되었다. 매카시가 인공지능을 독립된 연구 분야로 삼을 필요가 있다고 주장한 이유는 크게 두 가지이다. 하나는 연구 대상의 측면이고 다른 하나는 연구 방법의 측면이다. 첫째, 연구 대상으로서 인공지능은 다른 분야에서 다루지 않는 생각을 다루는데, 인공지능 연구는 인간의 창의성, 자기계발, 언어사용 같은 인간의 재능을 복제하고자 하는 생각을 포괄하고 있다는 것이다. 둘째, 연구 방법에서도 인공지능은 다른 분과와 구별되는 방법론을 갖는데, 인공지능은 컴퓨터 과학의 한 분과이면서 복잡하고 변화하는 환경에 자율적으로 작동할 기계를 제작하려고 시도하는 유일한 분야라는 것이다.¹⁶⁾ 그런데 ‘인공지능’이란 용어로 담아내려고 했던 생각은 그보다 앞서 튜링과 러브레이스로까지 거슬러 올라간다.

2. 튜링의 모방게임과 학습하는 기계

컴퓨터 발명에 지대한 공헌을 한 것으로 알려진 튜링(A. M. Turing)은 1950년에 발표한 “계산 기계와 지능”¹⁷⁾에서 ‘생각하는 기계’에 대한 관념을 ‘학습하는 기계

14) Stuart J. Russell · Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed., Prentice Hall, 2010, pp. 5~16.

15) Stuart J. Russell · Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed., Prentice Hall, 2010, viii.

16) Stuart J. Russell · Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed., Prentice Hall, 2010, pp. 17~18.

17) Alan M. Turing, “Computing Machinery and Intelligence”, *Mind* 59(236), 1950, pp. 433~460.



(learning machine)’에 대한 관념으로까지 구체화시킨다. 계산 기계¹⁸⁾ 즉, 오늘날의 컴퓨터는 생각할 수 있을까? 이 질문에 직접 답하려면 ‘기계(machine)’와 ‘생각하다(think)’를 정의해야만 한다. 그런데 기계가 무엇이고 생각하는 것이 무엇인지 밝히다 보면 인간이 아닌 기계가 인간만이 할 수 있는 고도의 지적 작용이라고 믿어지는 ‘생각’을 한다는 논제는 기계에게 지능이 없다는 고정관념과 부딪힌다. 그래서 튜링은 ‘기계는 생각할 수 있을까?’라는 질문 대신에 ‘모방게임(imitation game)’¹⁹⁾이라고 부르는 사고실험을 설계한다.

훗날 ‘튜링테스트(Turing Test)’로 불리게 된 이 모방게임에는 세 사람이 참여한다. 남자 A와 여자 B, 그리고 남자 또는 여자인 질문자(interrogator) C가 참여한다. 질문자 C는 다른 두 사람과 떨어져 있는 다른 방에 들어간다. 이 게임의 목표는 질문자가 두 사람의 성별을 판단하는 것이다. 질문자는 두 사람을 일단 레이블(label) X와 레이블 Y로 구분한 다음 게임이 끝날 때 ‘X는 A이고, Y는 B이다.’ 또는 ‘X는 B이고, Y는 A이다.’라고 말하면 된다. 이를 위해 질문자는 ‘X의 머리카락 길이는 얼마입니까?’ 같은 질문을 할 수 있다. 여기서 남자 A의 임무는 C의 판단을 그르치게 하는 것이고 여자 B의 임무는 C의 판단을 돕는 것이다. X가 A일 경우 A는 대답해야 하는데, ‘내 머리카락은 치켜 올려져 있고, 가장 긴 것은 20센티미터 정도 된다.’고 답할 수 있다. 질문자 C가 음색으로 판단하는 것을 막기 위해 답변은 타자기로 쳐서 글로 제시하고 원격통신이 가능한 프린터를 통해 출력되도록 한다. C의 판단을 도와야 하는 B는 정직하게 답변하는 것이 최선의 전략일 수 있다. 그래서 ‘나는 여자이니 그 남자의 말을 듣지 마세요!’라고 답변할 수 있지만, 이런 답변은 남자 A도 할 수 있기 때문에 C가 그대로 믿는다는 보장은 없다.²⁰⁾

튜링은 이러한 설정 속에서 ‘기계는 생각할 수 있을까?’라는 처음의 질문을 “보다 정교한 형식”²¹⁾의 새로운 질문으로 바꾼다. “기계가 A를 대신하면 무슨 일이 벌어질까?”, “질문자는 남자와 여자를 상대로 한 게임에서 잘못 결정한 만큼 기계가 A를 대신하는 경우에도 잘못 결정할까?”²²⁾ 튜링은 약 50년 뒤, 그러니까 이제는

18) 이른바 ‘튜링 기계’에 관해서는 본 논문 「제2장 제2절 I. 3. 컴퓨터가 수행할 일을 순서대로 알려주는 명령어의 집합」 참조.

19) “모방게임”으로 번역한 예로 김선희, “인공지능과 이해의 개념”, *인지과학* 8(1), 1997, 37~56쪽; “흉내게임”으로 번역한 경우로 Leavitt, David, 너무 많이 알았던 사람: 앨런 튜링과 컴퓨터의 발명, 고중숙(역), 승산, 2008, 특히 275~308쪽 참조.

20) Alan M. Turing, “Computing Machinery and Intelligence”, *Mind* 59(236), 1950, pp. 433~460: pp. 433~434.

21) Alan M. Turing, “Computing Machinery and Intelligence”, *Mind* 59(236), 1950, pp. 433~460: p. 442.

22) Alan M. Turing, “Computing Machinery and Intelligence”, *Mind* 59(236), 1950, pp. 433~460: p. 434.



과거가 된 미래인 2000년에는 10의 9제곱(10^9)에 이르는 메모리를 가진 컴퓨터에 모방게임을 아주 잘 해내도록 프로그램하면 5분 동안 질문한 후에 올바르게 판별하는 질문자의 비율이 70퍼센트를 넘지 않을 것이라고 믿었다.²³⁾ 그런데 사물을 창안(originate)하는 경우에만 정신(minds)을 가졌다고 믿은 러브레이스(A. Lovelace)는 명령된 것만을 수행하는 기계에게 그런 정신은 없다는 이유로 모방게임을 통해 기계가 생각할 수 있는지 측정할 수 있다는 튜링의 견해에 반대했다.²⁴⁾

오늘날 컴퓨터과학의 한 분과로 다루어지는 머신러닝(machine learning) 분야에서 학습은 ‘할당된 작업에 대한 성능이 향상되는 것’으로 본다.²⁵⁾ 머신러닝을 통해 프로그래밍 가능한 컴퓨터는 새로운 알고리즘을 만들어낼 수 있도록 고안된다. 러브레이스의 기준에 따르는 경우에도 이렇게 “다른 알고리즘을 생성하는 알고리즘”²⁶⁾ 시스템을 상정한다면 생각하는 기계의 관념은 수용될 수 있을 것이다.²⁷⁾ 최초에는 인간이 컴퓨터 알고리즘을 설계했다 하더라도 머신러닝은 점차적으로 인간의 설계를 벗어나 스스로 알고리즘을 설계하는 단계로 넘어간다. 그리고 이러한 전개 양상은 어느 순간부터 인간도 머신러닝 알고리즘이 생성하는 알고리즘 자체를 이해할 수 없는 상태에 이를 것이라는 전망으로 귀결된다. 그래서 인공지능은 종종 인간의 “마지막 발명품”²⁸⁾이라고 불리기도 한다.

3. 인공지능 연구의 중흥기와 정체기

인간의 마음 또는 정신을 모사하여 영원히 간직하고 싶다는 열망은 인간이 생각하는 과정을 기호로 나타내는 방식으로 개진되기도 하고, 뇌를 연구하는 방법으로 전개되기도 했다. 인공지능의 역사는 ‘봄’으로 표현되는 중흥기와 ‘겨울’이라고 불리는 정체기가 교차하면서 발전해 왔다. 흔히 1980년대 후반을 인공지능의 겨울이라고 하는데²⁹⁾ 이 시기는 인공지능에 대한 관심과 자금 지원이 줄어들었을 뿐만 아니라

23) Alan M. Turing, “Computing Machinery and Intelligence”, *Mind* 59(236), 1950, pp. 433-460: p. 442.

24) 튜링의 견해에 대한 러브레이스의 반론은 Alan M. Turing, “Computing Machinery and Intelligence”, *Mind* 59(236), 1950, pp. 433-460: pp. 450-451 참조.

25) Tom Mitchell, *Machine Learning*, McGraw-Hill, 1997, p. 2.

26) Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, New York: Basic Books, a member of the Perseus Books Group, 2015, p. 6.

27) 러브레이스 테스트에 대해서 Selmer Bringsjord · Paul Bello · David Ferrucci, “Creativity, the Turing Test and the (Better) Lovelace Test”, *Minds and Machines* 11(1), 2001, pp. 3-27 참조.

28) James Barrat, *Our Final Invention: Artificial Intelligence and the End of the Human Era*, Thomas Dunne Books, 2013 참조.

29) Ryan Calo, “Artificial Intelligence Policy: A Primer and Roadmap”, *U. C. Davis Law Review* 51(2),



인공지능의 개발 자체도 정체기에 빠졌던 때이다. 인공지능을 지나치게 기호체계에 의존해 구현하려고 했고, 지능적인 능력 전체를 한 가지 체계에서 구현하려고 시도한 것이 문제였던 것이다. 이에 따라 실제 세계의 데이터를 분석하고 조작하는 방식으로 접근 방식을 변경하고, 한 가지 체계에서 특정한 문제만을 해결하는 방식으로 바꾸면서 인공지능의 발전은 다시 한 번 중흥의 시기를 맞게 되었다.³⁰⁾

III. 과학기술 및 기계산업 중심의 현대 문명과 로봇

1. 로봇 개념의 기원

로봇은 그 영어식 표기에도 불구하고 이름이 붙여진 출처는 영국이나 미국이 아니다. ‘로봇(robot)’이란 표현은 체코 프라하의 작가 차페크(K. Čapek)가 1920년에 발표한 희곡 “R. U. R.”³¹⁾에서 처음 사용되었고, 그 시기는 ‘인공지능’이란 용어가 등장한 때보다 앞선다. ‘로숨의 유니버설 로봇(Rossum’s Universal Robots)’의 준말인 ‘R. U. R.’은 회사 이름이다. 이성(理性, reason)이라는 뜻을 가진 체코어 ‘Rozum’에 어원을 둔 ‘로숨(Rossum)’은 로봇을 만든 박사의 이름이자 로봇을 대량생산하려는 그 아들의 이름이다. ‘유니버설(universal)’은 희곡의 내용 중에 나라나 민족마다 다른 ‘내셔널(national)’ 로봇이 등장하는 것에 비추어 볼 때 보편이나 국제의 의미를 갖는다. 마지막으로 ‘로봇(robot)’은 노동을 의미하는 체코어 ‘robota’에서 ‘a’를 뺀 조어인데,³²⁾ 애초에 차페크가 작품의 제목을 영어로 지었기 때문에 ‘로봇(robot)’이란 표현은 그대로 영어에 편입되어 일상적으로 사용되는 데 어려움이 없었다.

유대인이 널리 거주했던 프라하는 히브리 전설에 등장하는 골렘의 고향이기도 한데, 이 전설에 따르면 ‘프라하의 골렘(the Golem of Prague)’은 16세기에 학식과

2017, pp. 399~436: pp. 404~405. 인공지능에 관한 1980년대 후반의 정체기가 1990년대를 거쳐 2000년대 초반까지 한동안 지속된 것으로 보기도 한다.

30) 미국 국가과학기술위원회(NSTC)의 보고서에 따르면 주요 컴퓨터과학의 발전사에서 실험실의 어떤 개념이 산업 성숙도를 갖추는 데 15년 이상이 걸린다고 한다. National Science and Technology Council, Executive Office of the President, “Preparing for the Future of Artificial Intelligence”, October 2016, https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf, p. 25.

31) Karel Čapek, 로봇: 로숨의 유니버설 로봇[R. U. R. (Rossum’s Universal Robots), 1920], 김희숙(역), 모비딕, 2015. 이 책은 체코어판을 저본으로 한 최초의 한글 완역본으로서 그 초판은 Karel Čapek, 로봇 R. U. R., 김희숙(역), 길, 2002 참조.

32) 김희숙, “역자 후기: 로봇, 현대SF의 탄생”, Karel Čapek, 로봇 R. U. R., 김희숙(역), 모비딕, 2015, 191~214쪽.



신앙심으로 존경받던 현자인 랍비(Rabbi), 유다 뢰브(Judah Loew ben Bezalel)에 의해 창조됐다.³³⁾ 전설 속 골렘은 유대 신비주의의 비밀 지식인 카발라(Kabbalah)에 따라 주문을 걸어 마치 신이 사람에게 생명을 불어 넣는 것처럼 진흙에 생명을 불어 넣어 만들어진 것이다. 인간처럼 보이는 골렘은 강력하지만 신이 인간에게만 언어능력을 주었다는 이유로 말을 할 수는 없었다. 체코의 북동부 보헤미아 지방에서 태어나 프라하에서 공부한 적이 있던 차페크 역시 이러한 골렘의 전설에 영향을 받았다.³⁴⁾

2. 일제 강점기 로봇 개념의 수용 태도

‘R. U. R.’은 발표 후 몇 년도 지나지 않아 1923년 일제 강점기 한국(조선)에 ‘인조인(人造人)’이란 제목으로 소개됐고, 2년 뒤 ‘인조노동자(人造勞動者)’란 제목으로 완역되기도 했다.³⁵⁾ 이때 ‘R. U. R.’은 ‘사람이 사람의 손으로 창조한 기계적 문명에 노예가 되며 마침내 멸망하는 날을 묘사한 심각한 풍속극’³⁶⁾ 또는 ‘종말에는 세계적 혁명이 비롯되며 또한 기계문명에 발달된 인류사회의 말세를 보이는 미래과의 일대결작’³⁷⁾으로 소개되었다.

당시 해석자의 관점에 따라 작품 속의 기계와 인간 사이 대립에 초점을 맞추어 인간이 편안함과 물질적 편리만을 추구한 결과 주체성을 잃고 기계의 노예가 된다고 보면서 기계문명의 지배에서 벗어날 방법으로 인간의 양심과 희생심이라는 덕목을 강조하기도 하고, 최후의 로봇이 얻어낸 자유와 해방에 초점을 맞추어 자본주의와 군국주의를 토대로 한 근대문명에 대한 반역의 정신으로부터 우리난 사회혁명의 사상을 보기도 했다. 나아가 작품 속의 대립 구도를 단순히 인간과 기계의 대립으로

33) Jack M. Balkin, “The Three Laws of Robotics in the Age of Big Data”, *Ohio State Law Journal* 78(5), 2017, pp. 1217~1241: p. 1222.

34) 조선시대 고대소설인 ‘전우치전(全雲致傳)’에서 나뭇잎으로 병사를 만들어 전쟁을 치르는 것을 보면 이런 전설이 유럽에만 있었던 것은 아니다. 김희숙, “역자 후기: 로봇, 현대SF의 탄생”, Karel Čapek, 로봇 R. U. R, 김희숙(역), 모비딕, 2015, 191~214쪽: 194~195쪽 참조.

35) “인조인(人造人)”은 이광수가 1923년 4월 1일 ‘동명(東明)’ 제31호(2권 14호)에 발표했고, “인조노동자(人造勞動者)”는 박영희가 1925년 2월부터 5월까지 4회(제56~59호)에 걸쳐 ‘개벽(開闢)’에 스즈키 젠타로(鈴木善太郎)의 번역본을 완역하여 연재했다. ‘R. U. R.’을 해석하는 김기진, 김우진, 박영희, 이광수 등 당대 문학가들의 관점에 대해서 황정현, “1920년대 『로봇의 유니버설 로봇』의 수용 연구”, *현대문학이론연구* 61, 2015, 513~539쪽 참조.

36) Karel Čapek, 李光洙(역), “人造人”, *東明* 31, 1923. 4, 15쪽: “사람이 사람의 손으로創造한機械의文明에 奴隸가되며 마침내 滅亡하는날을 描寫한深刻한諷刺劇”

37) Karel Čapek, 朴英熙(역), “人造勞動者”, *開闢* 56, 1925. 2, 56쪽: “終末에는世界的革命이비롯되며 또한機械文明의發達된人類社會의末世를보이는未來派의一大傑作”



보지 않고 계급을 영구한 인간생활의 실상으로 보아 과학을 위시한 근대 산업문명이 주는 해악으로서 계급의 대립과 투쟁에 초점을 맞추어 인간이 노동하는 고통을 없애려 로봇을 제조한다는 로봇 회사 대표의 말을 현대 자본가와 군국주의의 허울 뿐인 구실이라고 보면서 물질문명을 앞세워 세계를 지배하던 제국주의 열강에 대한 통렬한 비판으로 받아들이기도 했다.

여기에는 과학과 기술문명에 대한 부정적 태도 또는 로봇과 세계의 미래를 긍정적이고 희망찬 것으로 보는 태도, 나아가 단순히 부정 또는 긍정의 한쪽 면만을 볼 수 없다는 변증적 태도가 반영되어 있다. ‘R. U. R.’에 대한 해석에는 “식민지 경제구조에서 근대기술문명과 자본주의의 도입이라는 급격한 변화를 맞닥뜨려야만 했던 당대 문학 담당층의 시대인식과 고민”³⁸⁾이 집약적으로 담겨 있는 것이다.

3. 로봇 관념에 집약된 현대 과학과 기술의 역설

예술 작품의 해석이 작가의 의도에 구속되는 것은 아니지만 실제로 차페크는 ‘R. U. R.’을 통해 로봇보다는 인간에 초점을 맞추어 두 가지 희극 요소 즉, ‘과학의 희극’과 ‘진실의 희극’을 드러내려고 했다고 말한다.³⁹⁾ ‘지성(intellect)’ 또는 ‘두뇌(brain)’로 표기할 수 있는 늙은 과학자 ‘로숨(Rossum)’은 신이 불필요하고 부조리하다는 것을 증명하기 위해 기계적인 의미가 아닌 화학적이고 생물학적 의미의 인조인간(artificial man)을 창조하려고 했던 19세기의 과학적 유물론을 전형적으로 대표하고, 젊은 과학자 로숨은 과학 실험을 산업의 생산을 향한 길로 여겨 증명보다는 제조에 관심을 가진 더 이상 형이상학적 사유로 고민하지 않는 현대의 과학자를 대표한다. 이 젊은 현대의 과학자는 로봇을 대량 생산하는 산업주의의 길로 들어선다. 기계가 멈추면 수많은 존재를 파멸에 이를 수 있기 때문에 기계를 멈출 수는 없다. 그 과정에서 기계는 수많은 존재를 파멸시킴에도 더 빨리 가동되고, 산업을 지배하려고 생각했던 사람들은 그 산업의 지배를 받게 된다. 결국 인간의 두뇌에서 나온 개념이 인간의 손이 제어할 수 없는 영역으로 벗어나게 된다는 데에 첫 번째 의도인 과학의 희극(the comedy of science)이 있다.

38) 황정현, “1920년대 『로숨의 유니버설 로봇』의 수용 연구”, *현대문학이론연구* 61, 2015, 513~539쪽: 535쪽.

39) Karel Čapek, “The Meaning of ‘R. U. R.’”, in *The Saturday Review of Politics, Literature, Science and Art* 136(3534), 21 July 1923, p. 79.



두 번째 의도인 진실의 희극은 작품 속 인물들의 태도 속에 숨겨져 있다. 로봇 회사의 대표이사 도민(Domin)은 기술의 진보가 고된 육체노동으로부터 인간을 해방시킨다고 하고, 회사의 영업담당 이사인 부스만(Bussman)은 산업주의만이 현대의 필요를 충족할 수 있다고 본다. 톨스토이주의자 알퀴스트(Alquist) 기술의 진보가 자신을 타락시킨다고 믿고, 스물 한 살의 헬레나(Helena)는 비인간적 기계화를 본능적으로 두려워한다. 그리고 로봇(Robots)은 이 모든 이상주의에 저항한다. 그런데 이러한 태도는 모두 그렇게 믿을 만한 각각의 물질적이고 정신적인 이유에 근거한다는 점에서 모두 옳고 긍정할 수 있다는 것이 현대 문명에서 가장 극적인 요소이다. “더 이상 하나의 단일한 가치관의 정당성에 근거해서 다른 삶의 가치들을 일반적으로 종속시킬 수가 없게”⁴⁰⁾ 된 것이다. 진실의 희극(the comedy of truth)은 어떤 숭고한 진리와 타락한 이기적 오류 사이의 투쟁이 아니라 인간적 진리와 그에 못지않은 또 다른 인간적 진리, 이상적인 것과 그에 버금가는 또 다른 이상적인 것, 긍정적 가치와 견줄만한 또 다른 긍정적 가치 간의 투쟁에서 찾을 수 있는 것이다.⁴¹⁾

4. 21세기의 인공지능 로봇과 인류 미래에 관한 논의

인공지능 시스템이 탑재된 21세기의 로봇은 16세기의 전설 속 골렘이나 20세기의 희극 속 로봇과는 다르게 이야기나 작품의 밖으로 나와 현실 세계에서 실제로 사람에게 말을 하고 사람의 말을 알아듣기도 한다. 로봇은 “생물학적인 의미에서 살아있지 않을 뿐 물리적으로나 정신적으로 행위자성을 드러내는 하나의 구성된 체계”⁴²⁾인 것이다. 그렇기 때문에 이러한 과학기술의 발전에 따른 인류의 미래에 대한 논의는 보다 현실적이면서 먼 미래보다 가까운 미래에 관한 것이 되고 있다.

그럼에도 불구하고 차페크의 소설 속에 등장하는 헬레나가 가진 기계화에 대한 본능적인 두려움은 21세기에도 인간의 마음 한 구석을 차지하고 있다. 희극 속의 상상 세계가 아닌 현실 세계에서 판결기계의 도입 가능성에 대해 논의하면서 “수학 특히 확률론과 통계학에 기반을 둔 기계의 결정에 승복할 수 있는지도 의문이다. 자신이 교도소에 가서 몇 년씩을 복역하여야 하는지가 확률·통계적 모델인 판결기계에 의하여 결정된다는 방식을 우리 사회가 저항 없이 받아들이기는 어려울 것이다. 기계가

40) 임홍빈, *기술문명과 철학*, 문예출판사, 1995, 290쪽.

41) Karel Čapek, “The Meaning of ‘R. U. R.’”, in *The Saturday Review of Politics, Literature, Science and Art* 136(3534), 21 July 1923, p. 79.

42) Neil M. Richards · William D. Smart, “How Should the Law Think about Robots?”, in *Robot Law*, Ryan Calo · Michael Froomkin · Ian Kerr(Eds.), Edward Elgar Publishing, 2016, pp. 3~22: p. 6.



아닌 같은 인간에게 운명을 맡기겠다는 요구, 결정의 정확성(Decision Accuracy)보다는 가치를 우선시하는 생각 때문에 판결기계에 대하여 본능적으로 거부감을 가질 것이다.”⁴³⁾라는 인간의 본능적 감정에 의존하는 논거가 여전히 유효하게 사용되는 것만 보아도 이를 쉽게 확인할 수 있다. 인공지능이 인간의 ‘마지막 발명품’이라는 세련된 방식의 은유가 사용된다는 점에서 차이가 있지만 인공지능 로봇이 인류의 미래에 미칠 영향에 대한 논의 구조는 1920년대에 로봇을 희곡과 연극으로 접할 때 드러난 비평 구조와 크게 다르지 않다.

논의의 방향은 크게 두 갈래로 전개된다.⁴⁴⁾ 먼저 일론 머스크(Elon Musk), 스티븐 호킹(Stephen Hawking), 빌 게이츠(Bill Gates)의 경우 인공지능 로봇이 인류 문명에 커다란 위협을 가할 것이라고 믿는다.⁴⁵⁾ 이런 논제는 보스트롬(N. Bostrom)⁴⁶⁾의 초지능(super intelligence)에 관한 글에 잘 나타나 있다.⁴⁷⁾ 여기에는 두 가지 전체가 있는데 우리는 인간의 지능을 뛰어 넘는 초지능을 개발하는 길에 들어섰다는 점과 이 길의 끝에서 초지능은 그 창조자인 인간을 심각한 위협에 빠뜨릴 것이라는 점이다.

인공지능 로봇이 인류에 위협이 되는 경우는 몇 가지 형식으로 나타날 수 있다.⁴⁸⁾ 첫째, 인공지능 로봇이 스스로 일어나 의도적으로 인류에 대한 적대감이 그들 자신의 공간을 확보하기 위해 모든 사람을 죽이는 경우이다. 둘째, 인공지능 로봇이 임의의 목표를 맹목적으로 달성하려고 하다가 우연히 모든 사람을 죽이는 경우이다. 예를 들면 클럽을 만들게 되어 있는 인공지능 로봇이 재료인 광물을 채굴 하는 과정에서 지구를 파괴하는 경우를 생각해 볼 수 있다.⁴⁹⁾ 그리고 셋째, 인류의 삶을 종식시킬 목적으로 악의를 가진 개인이나 집단이 인공지능 로봇을 이용하는 경우이다.

43) 양종모, "인공지능에 의한 판사의 대체 가능성 고찰", *홍익법학* 19(1), 2018, 1~29쪽: 19쪽. 참고로 원문에서 "Decision Accuracy"를 번역 없이 그대로 사용한 것을 번역어로 대체했다.

44) Ryan Calo, "Artificial Intelligence Policy: A Primer and Roadmap", *U. C. Davis Law Review* 51(2), 2017, pp. 399~436: pp. 431~435.

45) Sonali Kohli, "Bill Gates Joins Elon Musk and Stephen Hawking in Saying Artificial Intelligence Is Scary", *Quartz*, 29 January 2015, <https://qz.com/335768/bill-gates-joins-elon-musk-and-stephen-hawking-in-saying-artificial-intelligence-is-scary>, 접속일: 2017년 12월 6일.

46) 보스트롬은 현대의 대표적인 트랜스휴머니스트(transhumanist)로 소개되기도 한다. 김건우, "포스트휴먼의 개념적, 규범학적 의의", *한국포스트휴먼학회(편저)*, 포스트휴먼 시대의 휴먼, 아카넷, 2016, 29~66쪽: 30쪽 이하 참조.

47) Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, 2014 참조.

48) Ryan Calo, "Artificial Intelligence Policy: A Primer and Roadmap", *U. C. Davis Law Review* 51(2), 2017, pp. 399~436: p. 433 이하.

49) Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, 2014, p. 123.



첫 번째 사례는 컴퓨터 과학의 문헌으로 뒷받침되지 않는 과학 소설이나 상업 영화에 주로 등장하는 상황으로 현실적인 주목을 받지 못한다. 두 번째 사례는 인간이 설정한 목표를 수행하는 과정에서 의도치 않게 인류가 파멸되는 경우이다. 그런데 이러한 경우는 인간에 의해 클립 제조와 같은 단순한 목표를 할당 받았다는 점에서 너무 원시적이기도 하지만, 그런 단순한 목표를 추구하기 위해 인류 전체를 우회하고 압도할 수 있는 능력을 동시에 가졌다는 점에서 과도한 설정이다. 세 번째 사례는 인공지능 로봇을 활용해 핵 보안을 위태롭게 한다거나 시장을 교란시킨다든지 또는 표적을 세밀하게 설정해 잘못된 정보를 흘려 폭력을 선동하는 예들을 생각해 볼 수 있다. 그나마 세 번째 사례가 현실 가능성이 있는 것처럼 보이지만 인류의 종말과는 거리가 멀다.

인공지능 로봇이 인류에 위협이 될 것이라는 주장에 대한 반론은 현재 모든 영역에서 인간과 경쟁할 수 있는 기계 지능을 만들 수 있는 모델이 없으며, 설령 초지능을 만들 수 있다고 하더라도 그것이 반드시 악성일 것이라고 단정할 수 없다고 본다. 첫 번째 반론은 체스나 바둑 같이 특별한 영역을 설정해 그 영역에서 인간을 이길 수 있는 설계는 가능하지만 머신러닝을 비롯해 인공지능과 관련된 현재 연구 수준에서는 인간의 지능은 차치하고 낮은 수준의 포유류 지능조차 충분히 따라갈 수 있는 모델이 없다는 사정을 근거로 제시한다.⁵⁰⁾ 그러면서 선불리 인류의 파멸에 대해 경고하는 것은 인공지능에 대한 이해의 부족에 기인한 것이라고 지적한다.⁵¹⁾

두 번째 반론은 어떤 이유로건 세계 지배에 대한 특별한 프로그램이 시스템에 탑재되지 않는 한 초지능이 당연히 인간을 지배하는 악성일 수는 없다고 지적하면서, 지배는 지능과 무관하다는 점을 근거로 제시한다. 딥러닝(deep learning)⁵²⁾ 연구의

50) Erik Sofge, “Why Artificial Intelligence Will Not Obliterate Humanity”, Popular Science, 20 March 2015, <https://www.popsoci.com/why-artificial-intelligence-will-not-obliterate-humanity>, 접속일: 2017년 12월 6일.

51) Connie Loizos, “This Famous Roboticist Doesn’t Think Elon Musk Understands AI”, TechCrunch, 19 July 2017, <http://social.techcrunch.com/2017/07/19/this-famous-roboticist-doesnt-think-elon-musk-understands-ai>, 접속일: 2017년 12월 6일.

52) 딥러닝(deep learning) 방식은 하위 수준 특성들(features)의 구성으로 형성된 계층(hierarchy)의 상위 수준 특성을 이용해 특정 계층의 구조를 학습하는 것을 목표로 하며, 기계지능(machine intelligence)에 도달하기 위해 필수적인 학습능력 및 자기훈련능력을 머신러닝 연구의 중심에 둔다. Yoshua Bengio, *Learning Deep Architectures for AI*, Foundations and Trends in Machine Learning 2(1), Now, 2009, pp. 5~6 및 Peter A. Flach, *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*, Cambridge University Press, 2012, p. 361 참조.



선두 그룹에 있는 얀 르쿤(Yann LeCun)은 “인공지능은 테스트스테론이 없다.”⁵³⁾는 말로 압축해서 표현한다. 오히려 지금의 문제는 인공지능이 너무 똑똑해져서 세계를 지배할 수 있다는 점에 있는 것이 아니라 오히려 너무 멍청해서 그런 것은 꿈도 꿀 수 없다는 점에 있다는 것을 상기해야 한다고도 한다.⁵⁴⁾ 그럼에도 불구하고 어떤 방향의 논의가 맞고, 어떤 갈래의 전망이 실현될 것인지는 시간이 말해줄 수 있을 뿐이다.

5. 아시모프의 로봇 3법칙과 헌법상 국가의 책무

아시모프(I. Asimov)는 단편 과학소설 “런어라운드(runaround)”⁵⁵⁾에서 소설 속 인물의 목소리를 빌려 로봇의 양전자 두뇌(positronic brain)에 아주 깊이 새겨져야 할 규칙으로 다음과 같은 “로봇공학의 세 가지 기본규칙(three fundamental Rules of Robotics)”을 제시했다.⁵⁶⁾

첫째, 로봇은 어떤 행동을 하거나 하지 않음으로써 인간에게 해를 입혀서는 안 된다.⁵⁷⁾ 둘째, 로봇은 인간이 내리는 명령에 복종해야 하며, 단 이러한 명령들이 첫 번째 법칙에 위배될 때에는 예외로 한다.⁵⁸⁾ 셋째, 로봇은 자신의 존재를 보호해야 하며, 단 그러한 보호가 첫 번째와 두 번째 법칙에 위배될 때에는 예외로 한다.⁵⁹⁾

이러한 규칙은 소설의 내용으로 머물러 있지 않고 ‘아시모프의 로봇 3법칙’으로 불리며 현실 속 로봇 연구에서 로봇의 행동 지침이 되는 윤리적 규칙으로서 지속적인 재발견과 재구성의 대상이 되어 왔다.⁶⁰⁾ 세 가지 법칙 중 우선이 되는 법칙으로서

53) Dave Blanchard, “Musk’s Warning Sparks Call For Regulating Artificial Intelligence”, NPR.org, 19 July 2017, <https://www.npr.org/sections/alltechconsidered/2017/07/19/537961841/musks-warning-sparks-call-for-regulating-artificial-intelligence>, 접속일: 2017년 12월 6일.

54) Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, New York: Basic Books, a member of the Perseus Books Group, 2015, pp. 285~286.

55) Isaac Asimov, “Runaround”, *I, Robot*, Bantam Books, 2004[1st, 1950], pp. 35~55 참조.

56) Isaac Asimov, “Runaround”, *I, Robot*, Bantam Books, 2004, 특히 pp. 44~45 참조.

57) Isaac Asimov, “Runaround”, *I, Robot*, Bantam Books, 2004, p. 44 이하: “One, a robot may not injure a human being, or, through inaction, allow a human being to come to harm.”

58) Isaac Asimov, “Runaround”, *I, Robot*, Bantam Books, 2004, p. 45: “Two, ..., a robot must obey the orders given it by human beings except where such orders would conflict with the First Law.”

59) Isaac Asimov, “Runaround”, *I, Robot*, Bantam Books, 2004, p. 45: “And three, a robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.”

60) 예를 들면 고인석, “아시모프의 로봇 3법칙 다시 보기”, 철학연구 93, 2011, 97~120쪽 참조.



다른 두 가지 법칙을 제한하는 제1법칙은 로봇이 인간에게 해를 입혀서는 안 된다는 것이다. 로봇 3법칙에서 드러나는 로봇에 관한 인간 중심 또는 인간 우위의 기본 입장은 “인본중심 기술구현”을 이른바 ‘지능정보사회’의 규범설정에서 1순위의 기본원칙으로 삼고,⁶¹⁾ 이를 헌법상 국가의 책무를 설정하는 “제1의 기본원칙으로 하여야 한다.”⁶²⁾고 주장하는 태도에 고스란히 담겨 있다. 이러한 주장의 헌법적 근거는 제10조의 제1문 즉, “모든 국민은 인간으로서의 존엄과 가치를 가지며, 행복을 추구할 권리를 가진다.”이다.

그런데 이 규정이 “고립된 개체로서의 개인주의적 인간이나 국가권력의 객체로서의 인간을 의미하는 것이 아니라 개인 대 사회라는 관계에서 인간 고유의 가치를 훼손당하지 않으면서 사회관계를 가지며 사회에 소속된 인간”⁶³⁾이라는 구체적인 특정 인간상을 전제로 하고 있는지는 의문이다. 또한 이러한 인간상을 전제로 “인간의 윤리적 가치를 벗어나거나 벗어날 수 있는 가능성이 있는 지능정보기술은 처음부터 사전에 철저히 배제되어야 한다.”⁶⁴⁾는 주장은 정보기술(information technology, IT)에 지능(intelligence)을 덧붙인 이른바 ‘지능정보기술’ 자체가 인간의 윤리적 가치를 벗어날 수 있는 가능성이 있는 기술이라는 관점에서 볼 경우 지능정보기술 자체를 배제하라는 요구로 받아들여질 수 있는 강한 의미를 담고 있다. 게다가 이러한 주장은 “기술은 인간의 행복추구활동의 도구”⁶⁵⁾라고 하여 기술을 도구로만 본다는 점에서 기술 일반에 대한 도구주의(instrumentalist)라는 기술철학⁶⁶⁾의 관점에 입각한 것이어서 이러한 입장이 과연 ‘인간으로서의 존엄과 가치’ 및 ‘행복을 추구할 권리’를 보장하기 위해 기술에 관한 헌법상 국가의 책무를 설정하는 지도적 관점이 되어야 하는 것인지도 의문이다. 이미 “기술공학의 현대적 체계는 경험과학의 적극적인 활용을 통해서 도구의 차원을 넘어설 뿐만 아니라, 삶의 다양한 상징적 표현들을 제한하기 시작”⁶⁷⁾했는데, 도구주의 관점의 기술철학에 의존할 경우 기술의 이러한 차원들이 은폐되기 때문이다.

61) 김민호 · 이규정 · 김현경, “지능정보사회의 규범설정 기본원칙에 대한 고찰”, 성균관법학 28(3), 2016, 293~320쪽: 306~308쪽.

62) 정준현 · 김민호, “지능정보사회와 헌법상 국가의 책무”, 법조 66(3), 2017, 106~145쪽: 126쪽.

63) 정준현 · 김민호, “지능정보사회와 헌법상 국가의 책무”, 법조 66(3), 2017, 106~145쪽: 125쪽.

64) 김민호 · 이규정 · 김현경, “지능정보사회의 규범설정 기본원칙에 대한 고찰”, 성균관법학 28(3), 2016, 293~320쪽: 307쪽.

65) 정준현 · 김민호, “지능정보사회와 헌법상 국가의 책무”, 법조 66(3), 2017, 106~145쪽: 125쪽.

66) 기술철학의 다양한 입장을 확인할 수 있는 주요 문헌의 선집으로 David M. Kaplan(Ed.), *Readings in the Philosophy of Technology*, 2nd ed., Rowman & Littlefield Publishers, 2009 참조.

67) 임홍빈, 기술문명과 철학, 문예출판사, 71쪽.



IV. 사회의 행위자로서 에이전트

1. 에이전트의 정의

에이전트(agent)는 다양한 맥락에서 사용될 수 있는 용어이다. 에이전트는 인간이 될 수도 있고, 조직체 또는 그 일부일 수도 있으며, 컴퓨터 시스템도 될 수 있다. 나아가 이런 개념들이 조합된 형태에 대해 사용될 수도 있다. 이렇게 인간, 로봇, 소프트웨어 등 다양한 실체를 포착하기 위해 에이전트를 “물리적 상징(기호)체계 (physical symbol system)”⁶⁸⁾라고 정의하기도 한다. 에이전트는 기본적으로 사회의 환경에 반응하는 체계로서 행위자로 기능하며, 에이전시(agency)는 그러한 행위자성 또는 행위능력을 지칭한다. 에이전트는 “그 이용자를 대신해서 행위하고, 그 이용자의 직접적 입력이나 직접적 감독 없이 어떤 목적을 달성하거나 작업을 완료하려고 시도”⁶⁹⁾한다.

조직이론에서 에이전트는 조직의 과업을 달성하기 위해 지식을 생산하고 적용하는 개체로서 대부분 인간을 나타냈다. 그러나 1980년대 후반부터 인공지능 연구자들이 에이전트 개념에 주목하면서 에이전트는 의미에 변화를 겪는다.⁷⁰⁾ 에이전트는 인공지능 속에 있는 광범위한 기술의 군집과 방대한 연구 프로그램을 표현하기 위한 용어로서 “상대적으로 자율적인 정보 처리 시스템”⁷¹⁾과 관련을 맺게 된다.

정보기술(IT)의 추세를 따라가 보면 1970년대에는 데이터베이스 지향 시스템(database-oriented systems)이, 1980년대에는 데이터베이스 시스템에 사용자 인터페이스와 지식베이스 개념을 추가한 지식기반 결정지원 시스템(knowledge-based decision support systems)이, 1990년대에는 지식기반 결정지원 시스템 개념에 통신기능을 추가한 에이전트 시스템(agents system)이 집중적으로 개발됐다.⁷²⁾ 에이전트 시스템을 지식기반

68) Russ Abbott, “Meaning, Autonomy, Symbolic Causality, and Free Will”, *Review of General Psychology* 22(1), 2018, pp. 85~94: p. 85.

69) John J. Borking · B. M. A. van Eck · P. Siepel, *Intelligent Software Agents: Turning a Privacy Threat into a Privacy Protector*, Registratiekamer, 1999, p. 1.

70) Heesen, Constantijn · Vincent Homburg · Margriet Offereins, “An Agent View on Law”, *Artificial Intelligence and Law* 5(4), 1997, pp. 323~340: p. 325.

71) Chopra, Samir · Laurence F. White, “Artificial Agents and Agency”, in *A Legal Theory for Autonomous Artificial Agents*, University of Michigan Press, 2011, pp. 5~28: p. 6.

72) Heesen, Constantijn · Vincent Homburg · Margriet Offereins, “An Agent View on Law”, *Artificial Intelligence and Law* 5(4), 1997, pp. 323~340: p. 324.



결정지원 시스템과 구별해주는 기능은 커뮤니케이션(communication) 즉, 통신이다. 상호작용을 가능하게 하는 커뮤니케이션의 특성은 에이전트를 “감지기로 환경을 지각하고 작동기로 그 환경에 작용한다고 여겨지는 것”⁷³⁾으로 정의할 수 있게 한다.

2. 에이전트의 특성과 자율성

구체적으로 정의하는 내용에 약간의 차이가 있지만 에이전트는 주로 자율성, 사회성, 반응성, 능동성을 특징으로 한다.⁷⁴⁾ 에이전트는 인간이나 다른 개체의 직접적인 개입 없이 작동하고 자신의 행위와 내부 상태에 대해 어느 정도 통제할 수 있다는 측면에서 자율성(autonomy)을 가지며,⁷⁵⁾ 다른 에이전트와 커뮤니케이션 언어를 통해 상호작용할 수 있다는 점에서 사회성(social ability)을 갖는다. 또한 에이전트는 주변의 환경을 지각하고 발생한 변화에 대해 시기적절하게 내부에서 반응한다는 측면에서 반응성(reactivity)을 가지며, 단순히 환경에 반응하는 것을 넘어 계획을 수립하여 목표 지향적 행위를 실행할 수 있다는 점에서 능동성(pro-activeness)을 갖는다.

이중에 특히 자율성은 에이전트를 수식하는 용도로 자주 언급된다. 예를 들어 인간의 개입 없이 의사 결정을 내릴 권한을 사회의 점점 더 많은 측면에서 획득해 온 컴퓨터 프로그램을 ‘자율적 인공 에이전트’라고 부르거나 이러한 에이전트가 운행하는 자동차 로봇을 ‘자율 주행 자동차’라고 부르는 것이다. 자율성이 전면에서 언급되는 이유는 인공 에이전트에게 인격, 보다 구체적으로 법적 인격을 부여할 수 있는지 판가름하는 결정적 기준으로 여겨지기 때문이다. 로크(J. Locke)에 직접 연결된 경험론적(empirical) 인격 개념에 따르면 인격은 느끼고 이해하는 능력, 자의식과 자율성 같은 어떤 정신적 속성과 결부된다. 반면에 인격을 가치의 운반자로 보는 칸트주의(Kantian) 또는 신칸트주의(neo-Kantian)의 가치론적(axiological) 인격 개념에 따르면 윤리적 가치는 인격에 앞서고 인격을 규정한다.⁷⁶⁾ 이러한 인격에 관한 철학적

73) Stuart J. Russell · Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed., Prentice Hall, 2010, p. 34: “An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators.”

74) Heesen, Constantijn · Vincent Homburg · Margriet Offereins, “An Agent View on Law”, *Artificial Intelligence and Law* 5(4), 1997, pp. 323~340: pp. 325~326.

75) David A. Mindell, *Our Robots, Ourselves: Robotics and the Myths of Autonomy*, Viking, 2015, p. 10 참조.

76) 인격(person)은 서로 다른 철학적 의미에 기초함으로써 다루기 어려운 까다로운 개념이다. 인격 개념에 관한 현대의 논쟁은 Bartosz Brożek, “The Troublesome ‘Person’”, in *Legal Personhood: Animals, Artificial Intelligence and the Unborn*, Kurki, Visa A. J. · Tomasz Pietrzykowski(Eds.), Springer, 2017, pp. 3~13: p. 7 참조.



의미 구성의 차이에도 불구하고 자율성은 경험적으로 확인할 수 있는 정신적 능력으로서 또는 매우 중요한 윤리적 가치로서 인격 개념을 형성하는 주요 지표가 된다. 물론 완벽히 자율적인 군인, 항해사, 항공기 조종사가 없는 것처럼 완벽하게 자율적인 시스템을 찾기는 어려울 수 있지만⁷⁷⁾ 프로그래머, 시스템 운영자 또는 프로그램 이용자에게 확인받지 않고도 여러 가지 법적 기능을 수행할 수 있는 자율적 시스템으로서 에이전트의 특성은 법인격을 인간 에이전트에게 지향되도록 구성되어 있는 기존 법체계와 법이론에 심각한 도전이 된다.⁷⁸⁾

3. 법체계에서 에이전트와 커뮤니케이션

전통적으로 법체계에서 에이전트는 대리인을, 에이전시는 대리능력을 의미한다. 대리인은 본인의 법률행위를 대리하고 그 법률효과를 본인에게 귀속시킨다. 대리인은 타인의 사무를 대신 처리하는 대리행위자이고 본인을 위한다는 점을 제외하면 사회의 법적 관계에서 중요한 행위자라는 점은 분명하다.⁷⁹⁾ 기술 개발의 측면에서 커뮤니케이션은 정보기술(IT)을 정보통신기술(ICT)로 발전시켰고 시스템에 에이전트의 속성을 부여할 수 있도록 해준다는 점에서 중요한 의미를 갖지만, 사회와 법의 측면에서 볼 때 커뮤니케이션의 체계로서 사회, 그리고 사회의 부분체계로서 법이 작동하는 데에 커뮤니케이션은 인간이 아닌 에이전트가 참여할 수 있는 연결점이 되기도 한다.⁸⁰⁾ 예를 들어 인터넷 상점에서 물건을 주문할 때 본인이 직접 할 수도 있지만 소프트웨어 에이전트가 대신 할 수도 있고, 초단시간 주식 거래는 아예 인간이 개입하기도 어려운 시간에 알고리즘 에이전트 사이에 이루어지고 있다는 점에서 에이전트는 인간에 한정되지 않은 채로 이미 사회의 법률 행위자로서 역할을 수행하고 있다.⁸¹⁾

77) David A. Mindell, *Our Robots, Ourselves: Robotics and the Myths of Autonomy*, Viking, 2015, p. 10.

78) Gunther Teubner, "Rights of Non-Humans? Electronic Agents and Animals as New Actors in Politics and Law", *Journal of Law and Society* 33, 2006, pp. 497~521 참조; 알고리즘 에이전트에게 책임을 귀속시키는 전제로 법인격을 부여할 것인지에 관한 제한주의(restrictivism)와 허용주의(permissivism)의 논의는 본 논문 「제5장 제2절 III. 1. 양 극단의 제한주의와 허용주의」 참조.

79) 인공 에이전트를 법적 대리인으로서 권한이 있는지에 관한 논의는 Samir Chopra, "Rights for Autonomous Artificial Agents?", *Communications of the ACM* 53(8), 2010, pp. 38~40 참조.

80) 인간, 동물, 인공지능 로봇 간 커뮤니케이션 문제를 다룬 것으로 Ipke Wachsmuth, *Menschen, Tiere und Max: Natürliche Kommunikation und Künstliche Intelligenz*, Springer Spektrum, 2013 참조.

81) '알고리즘 소비자'의 개념과 결정 절차에 대해서 Michal S. Gal · Niva Elkin-Koren, "Algorithmic Consumers", *Harvard Journal of Law & Technology* 30(2), 2017, pp. 309~353: pp. 313~317 참조.



제2절 머신러닝 알고리즘과 데이터 학습 모델

인공지능이나 로봇, 지능적 에이전트를 조종하는 것은 컴퓨터이다. 보다 정확하게 말하면 이 컴퓨터는 컴퓨터가 인식할 수 있도록 표현한 언어로 만들어진 알고리즘에 따라 작동한다. 이러한 알고리즘은 주로 인간에 의해 설계되었지만 이제는 컴퓨터가 훈련을 통해 자율적으로 데이터를 학습하여 알고리즘을 설계하는 단계에 이르고 있다. 그리고 그 중심에 머신러닝 알고리즘이 있다. 그런데 이렇게 기술적이고 전문적인 기능을 수행하는 것처럼 보이는 알고리즘은 인간이 일상적인 문제를 해결해 온 방법의 또 다른 이름이기도 하다. 아무렇게나 쌓여 있는 책을 책장에 정리하는 방법, 시험에 합격하는 방법, 논문을 작성하는 방법도 모두 각각의 알고리즘을 가지고 있다. 다만 그 정교성과 복잡성에서 차이가 있을 뿐이다. 일반적 수준의 알고리즘에서 출발하여 보다 전문적이고 고차원의 머신러닝 알고리즘에 도달해 보도록 한다.

I. 알고리즘의 어원과 정의

1. 알 콰리즈미와 알고리즘

‘algorithm’에 대한 외래어 표기 방식은 ‘알고리즘’과 ‘알고리듬’⁸²⁾ 두 가지이다. ‘algorithm’과 동의어로 사용되는 ‘algorism’의 발음을 고려하면 ‘algorithm’에 대해서는 ‘알고리즘’보다는 ‘알고리듬’이 실제 발음과 좀 더 유사한 측면이 있다. 그래서 알고리듬이 무엇인지 의미를 파악하려고 어원을 찾을 때 마치 그리스어에서 유래를 찾아야 할 것만 같다. 알고리듬이 그리스어에서 유래한 ‘리듬(rhythm)’과 발음상 흡사한 측면이 있기 때문이다. 그러나 알고리즘에는 리듬의 ‘리(rhy-)’에 있는 ‘h’가 없다. 그런데 이렇게 ‘h’를 뺐다고 해서 알고리듬이 그와 어감이 비슷한 바이오리듬(biorhythm)의 무생물 형태인 것도 아니다. 오히려 알고리즘은 인명에서 유래한다. 그 사람의 이름은 서구에 영(0) 개념을 소개한 인물로 알려진 9세기 페르시아의 수학자 알 콰리즈미(Al-Chwarizmi)이다.⁸³⁾ 따라서 어원이 말해 줄 수 있는 것은

82) ‘알고리듬’을 사용하는 예로 안형준, “알고리듬 안에 내재된 사회적 차별: 빅데이터에 대한 미국 정부의 우려”, 과학기술정책 (214), 2016, 4~7쪽 참조.

83) Sebastian Stiller, *Planet der Algorithmen: Ein Reiseführer*, München: Knaus, 2015, p. 46; 알 콰리즈미는 ‘Al-Khwarizmi’로 표기하기도 한다. Sebastian Stiller, 알고리즘 행성 여행자들을 위한 안내서: 쇼팽부터



알고리즘이 수학과 어느 정도 관련을 맺고 있다는 개연성 정도이다. 이러한 이유 때문인지는 몰라도 알고리즘 연구자들은 알고리즘에 대한 일반적 정의를 내리는 대신에 알고리즘과 관련된 여러 가지 개별 사례들을 제시하기도 한다.⁸⁴⁾ 그럼에도 불구하고 중요한 알고리즘의 특성을 부각시킬 수 있는 몇 가지 정의를 찾아볼 수는 있다.

2. 문제를 해결하는 방법

특정한 목표를 가진 작업을 수행할 때 그 작업을 수행하는 방법은 여러 가지가 있을 수 있다. 어떤 사람이 P 지점에서 Q 지점으로 이동하는 것이 목표일 경우 이동경로는 여러 가지가 있을 수 있다. 이때 이동경로는 몇 가지 변수를 고려한 결정에 따라 그 모습이 달라질 수 있다. 먼저 P 지점과 Q 지점 사이를 육로로 이동할 것인지, 수로 또는 항로를 이용해 이동할 것인지 이동로의 종류를 결정하기로 해보자. 그러면 이용할 길의 성질에 따라 이동수단도 달라질 수 있다. 기본적으로는 사람이 걸어가거나 헤엄쳐 갈 것인지 아니면 유모차나 휠체어, 자전거, 자동차를 직접 운전할 것인지, 택시나 버스, 열차 또는 배나 항공기에 탑승할 것인지를 길의 종류에 적합하게 선택해야 할 것이다. 반대로 이동수단을 먼저 결정할 수도 있다. 그러면 이동수단에 따라 이동로의 종류나 이동의 방향, 시간, 거리가 달라질 수 있다. 특히 버스나 열차, 배, 항공기는 운행 노선이 정해져 있기 때문에 특별한 노선이 없는 다른 수단보다 이동로의 종류나 이동방향에 제약이 따른다. 하지만 또 다른 방식으로 이동의 시간이나 거리를 먼저 결정할 수도 있다. 그러면 P 지점과 Q 지점 사이를 이동할 때 필요한 시간을 단축하거나 연료비를 절약할 수 있는 이동로의 종류와 이동수단을 선택할 수 있을 것이다. 이때 이동거리와 이동수단의 속도는 이동경로를 결정하는 데 중요한 변수가 될 것이다.

구체적으로 세분화하면 더 많은 변수를 찾을 수 있겠지만 위의 내용만을 놓고 보면 이동로의 종류, 이동수단, 이동거리, 이동시간이라는 네 가지 변수가 P 지점에서 출발하여 Q 지점에 도착하는 경로를 결정하는 데에 필요한 주요 변수가 된다. 일단 이동로의 종류로는 육로, 수로, 항로의 3개, 육로로 이동할 수 있는 수단은 사람의 신체, 휠체어, 유모차, 자전거, 자동차, 택시, 버스, 열차의 8개, 수로에서는 사람의

인공지능까지, 우리 삶을 움직이는 알고리즘에 관한 모든 것, 김세나(역), 와이즈베리, 2017, 59쪽 참조.
84) Thomas Ottmann · Peter Widmayer, *Algorithmen und Datenstrukturen*, 6th ed., Springer Vieweg, 2017[1986], pp. 5~20 참조.



신체와 배의 2개, 항로에서는 항공기 1개를 이동수단으로 생각해 볼 수 있다. 수로를 사람이 직접 헤엄치는 것이 일상적인 방법이 아니라는 점을 고려해 배를 이용하는 1개만 상정한다면 수로와 항로는 이동로의 종류와 이동수단이 일 대 일로 대응한다고 볼 수 있다. 다만 배나 항공기의 성능이나 노선에 따라 이동시간과 이동거리에 차이가 있는 여러 개의 선택사항이 발생한다. 육로는 이동수단에 특별한 노선의 유무 및 각 수단에 결합된 이동시간과 이동거리에 따라 선택사항은 증가한다. 여기에 더해 이동수단을 단일하게 할 것인지 복수의 수단을 결합할 것인지 그 조합에 따라 이동경로의 수는 다양하게 증가하고 복잡해진다.

여기서 이동로의 종류를 먼저 선택하여 그 내용을 육로로 결정해 보자. 그러면 우선 이동로의 종류라는 변수에서 다른 선택사항들, 즉 수로와 항로는 배제된다. 아울러 선택 가능한 이동경로의 수도 감소된다. 그 다음 이동수단을 선택하여 그 중에 직접 운전하는 자동차로 결정해 보자. 그러면 이동수단이라는 변수에서 다른 7개 수단은 제거된다. 그만큼 이동경로 조합의 수는 대폭 감소된다. 시간에 여유가 있고 자동차에 연료도 충분해서 시간과 거리에 구애를 받지 않는 경우가 아니라 최단시간에 최단거리로 이동하고 싶다면 이제 남은 것은 이동시간과 이동거리를 계산하여 최적의 길을 찾는 일만 남았다. P 지점과 Q 지점 사이에 직선의 도로가 놓여 있다면 연비와 최고속도가 가장 높은 자동차를 골라 타는 문제로 되돌아가야 하겠지만, 이동경로의 중간에 신호등이 있거나 제한 속도가 있는 여러 개의 도로를 이용해야 하는 상황에서는 직선에 가장 가까운 도로의 조합뿐만 아니라 그때그때 교통 상황도 변수로 고려해야 한다. 또한 P 지점에서 자동차가 출발한 이후에 실시간으로 도착 지점 Q까지의 최적경로를 찾으려면 자동차의 위치를 실시간으로 파악하여 기존에 탐색한 최적경로에서 변경된 교통 상황을 반영해 소요 시간과 연료비의 증감이 발생한다면 그에 따라 새로운 경로를 탐색해야 한다.

P 지점에서 Q 지점으로 이동한다는 간단해 보이는 목적을 달성하기 위한 방법도 구체적인 수준에서 세분하여 보면 복잡해진다. 그런데 이처럼 어떤 목적이나 과업을 수행하기 위한 방법 그 자체를 넓은 의미에서 ‘알고리즘(algorithm)’이라고 할 수 있다. 일반적인 의미에서 알고리즘은 책장에 뒤섞여 있는 책을 가나다 순서로 정렬하는 방법이나, 대학에 입학하는 방법, 직업을 구하는 방법, 급여를 지급하는 방법, 업무를 배치하는 방법에서부터 대통령을 탄핵하는 방법에 이르기까지 다양하게 사용될 수 있다. 알고리즘을 “소프트웨어로 만드는 문제 푸는 방도”⁸⁵⁾로 설명하는

85) 이광근, 컴퓨터 과학이 여는 세계: 세상을 바꾼 컴퓨터, 소프트웨어의 원천 아이디어 그리고



경우에도 “문제풀이법”⁸⁶⁾이라는 일반적 의미가 함축되어 있다. 이러한 의미의 알고리즘은 인간의 역사에서 오랫동안 사용됐던 것이다.

3. 컴퓨터가 수행할 일을 순서대로 알려주는 명령어의 집합

매번 길을 찾을 때마다 사람이 위와 같은 과정을 새롭게 반복하려면 시간과 비용이 든다. 이러한 시간과 비용의 문제는 알고리즘에 따라 보다 원활하게 과제를 수행할 수 있는 기계의 필요성을 제기한다. 이동하는 시간과 거리를 계산하여 최소 시간에 최소 거리를 이동할 수 있는 최적의 경로를 찾는 과제를 알고리즘에 따라 자동적으로 수행할 수 있는 기계가 필요한 것이다.

1936년 튜링이 런던 수리학회에 제출한 “계산가능한 수에 대해서, 수리명제 자동생성 문제에 응용하면서”⁸⁷⁾라는 제목의 논문에는 컴퓨터의 원천 아이디어가 담겨져 있다. 이 논문에서 튜링이 고안한 기계는 종이에 비유되는 테이프와 테이프에 기록되는 부호, 테이프에 기록된 부호를 읽거나 테이프에 부호를 쓰는 장치, 장치의 상태를 나타내는 부호, 그리고 기계가 작동하는 규칙표로 구성된다. 튜링머신(Turing Machine)으로 불리는 이 이론상의 기계는 테이프에 기록된 부호를 읽고 작동 규칙에 따라 새로운 부호를 테이프에 쓰고 배열의 순서대로 좌우로 이동하면서 과제를 달성할 때까지 작동규칙에 따라 동작을 반복한다.

이처럼 튜링머신이 수행하는 일은 알고리즘과 동등한 것으로 이해된다.⁸⁸⁾ 그렇기 때문에 알고리즘을 “컴퓨터가 수행할 일을 순서대로 알려주는 명령어의 집합”⁸⁹⁾이라고 정의하기도 한다. 알고리즘이 여러 개의 명령어로 구성되어 있다는 것은 알고리즘이 결국에는 명령된 것만을 수행한다는 것을 의미한다.

미래, 인사이드, 2015, 94쪽.

86) 이광근, 컴퓨터 과학이 여는 세계: 세상을 바꾼 컴퓨터, 소프트웨어의 원천 아이디어 그리고 미래, 인사이드, 2015, 96쪽: “주어진 문제를 컴퓨터로 풀고 싶다면, 알고리즘(문제풀이법)을 찾고 그 해법이 맞는지 확인한 후, 그 복잡도(실행비용)가 견딜 만하다고 판단되면, 그 알고리즘대로 작동할 소프트웨어를 만들게 된다.”

87) Alan M. Turing, “On Computable Numbers, with an Application to the Entscheidungsproblem”, *Proceedings of the London Mathematical Society* s2-42(1), 1937, pp. 230~265.

88) 오늘날 우리가 사용하는 대부분의 컴퓨터는 튜링기계와 아이디어가 일치하는 폰 노이만(von Neumann)이 설계한 에드박(EDVAC, Electronic Discrete Variable Automatic Computer) 디자인이다. 폰 노이만이 에드박을 설계한 것은 1945년이지만, 1952년에 이르러서야 이 디자인이 실제 컴퓨터로 완성됐다. 이에 관해서 이광근, 컴퓨터 과학이 여는 세계: 세상을 바꾼 컴퓨터, 소프트웨어의 원천 아이디어 그리고 미래, 인사이드, 2015, 85~86쪽 참조.

89) Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, Basic Books, a member of the Perseus Books Group, 2015, p. 1.



4. 입력(input)과 출력(output)의 관계를 규정하는 절차

컴퓨터 과학의 용어로서 알고리즘은 “어떤 값이나 값의 세트를 입력(input)으로 취해서 또 다른 어떤 값이나 값의 세트를 출력(output)로 만들어 내는 명확히 정의된 계산 절차(computational procedure)”⁹⁰⁾ 즉, 컴퓨터를 사용한 계산 절차이다. 알고리즘은 입력을 출력으로, 다시 말해 투입 값을 산출 값으로 변형하는 연속적인 컴퓨터의 계산 단계인 것이다. 다른 측면에서 알고리즘은 정확히 명시된 컴퓨터 계산 문제(computational problem)를 해결하기 위한 수단으로 볼 수도 있다. 문제의 진술은 원하는 투입과 산출의 관계를 일상 언어로 명시하고, 알고리즘은 그러한 “입력(투입)과 출력(산출)의 관계(input/output relationship)를 획득할 수 있는 컴퓨터의 특수한 계산 절차를 서술하는 것”⁹¹⁾이다. 또한 알고리즘은 기본적으로 ‘곱(and)’, ‘합(or)’, ‘부정(not)’이라는 세 가지 기본 연산 기호에 의해 설계될 수 있는데, 이때 사용되는 연산 규칙은 입력된 기호를 출력된 기호로 변형하는 역할을 한다. 이런 의미에서 알고리즘은 “기호를 변형하는 규칙”⁹²⁾이기도 하다.

II. 머신러닝 알고리즘과 데이터 마이닝 모델의 맹점

1. 머신러닝 알고리즘의 학습 개념

미첼(T. Mitchell)은 머신러닝에 관한 교과서에서 학습을 정의하면서 “특정한 과제 T를 수행할 때 성과 기준 P가 경험 E를 통해 향상된다면 컴퓨터 프로그램은 일군의 과제 T와 성과 기준 P에 대해서 경험 E로부터 학습한 것”⁹³⁾이라고 본다. 예를 들어 과제 T가 ‘고양이를 인식하여 분류하는 것’이고 성과 기준 P는 ‘고양이를 정확히 구분하는 확률’ 그리고 경험 E는 ‘다양한 형상의 고양이를 표시한 데이터세트’라고 할 때, 컴퓨터 프로그램이 다양한 형상의 고양이 데이터세트(E)를 통해 고양이를 인식하고 분류(T)함으로써 고양이를 정확히 구분하는 확률(P)이 높아진다면 컴퓨터 프로그램은 과제 T와 성과 기준 P에 대해서 경험 E를 통해 학습한 것이다.

90) Thomas H. Cormen · Charles E. Leiserson · Ronald Rivest · Clifford Stein, *Introduction to Algorithms*, 3rd ed., MIT Press, 2009, p. 5.

91) Thomas H. Cormen · Charles E. Leiserson · Ronald Rivest · Clifford Stein, *Introduction to Algorithms*, 3rd ed., MIT Press, 2009, p. 5.

92) Sebastian Stiller, *Planet der Algorithmen: Ein Reiseführer*, München: Knaus, 2015, p. 70: “eine Regel zur Umformung von Zeichen”

93) Tom Mitchell, *Machine Learning*, McGraw-Hill, 1997, p. 2.



엄밀한 의미에서 학습(learning)은 목적이나 의도가 있는 경우, 그 중에서도 배우는 사람에게 목적이나 의도가 있는 경우를 말한다. 목적이나 의도가 가르치는 사람에게 있는 경우 학습이 아니라 훈련(training)을 의미한다는 것과 비교하면 그 의미가 보다 선명해진다. 그러나 머신러닝에서 학습은 전형적으로 인간의 학습과 관련된 인지적 의미라기보다 기능적 의미에 가깝다. 더 많은 데이터를 수신함으로써 더 나은 성과를 발휘하기 위해 미래의 작동을 변경한다는 정도의 의미인 것이다.⁹⁴⁾ 그렇다면 머신러닝에서 학습은 일종의 은유(metaphor)라고 볼 수 있다. 은유에 그치는 것이 아니라면 컴퓨터 프로그램에서 의도를 찾아야 하는 일이 발생할 것이다. 그러나 인간의 범정에서조차도 문제를 해결하기 위해 의도를 찾는 것은 매우 어려운 일이다.

2. 머신러닝 알고리즘의 학습과 데이터

도밍고스(P. Domingos)는 머신러닝 알고리즘의 학습 개념에서 중요한 세 가지 요소로 표현(representation), 평가(evaluation), 최적화(optimization)를 제시한다.⁹⁵⁾ 이때 학습의 목표는 일반화(generalization)이다. 즉, 머신러닝 알고리즘은 컴퓨터가 처리할 수 있는 형식 언어로 표시되어야 하고, 더 나은 머신러닝 알고리즘을 가려내기 위해 알고리즘에 의해 내부적으로 사용되는 함수로서 목적 함수 또는 점수화 함수라고도 불리는 평가 함수를 갖추어야 하며, 가장 높은 점수를 획득한 알고리즘을 찾기 위한 방법도 마련되어 있어야 한다. 이러한 학습에서 얻게 되는 것은 판별식이다. 머신러닝의 기본적인 목표는 학습용 데이터세트의 예시들을 넘어서는 일반화를 하는 것이다. 일반화가 이루어지면 새로운 데이터에 대해 예측하는 것이 가능해진다.

머신러닝은 알고리즘을 만들기 위해 데이터를 처리하고 만들어진 알고리즘으로 새로운 데이터를 처리한다. 머신러닝은 데이터 간의 상관관계를 찾아냄으로써 어떻게 데이터가 그와 같은 양상을 갖게 됐는지 판별할 수 있는 규칙을 찾아내기 때문에 데이터는 필수적이다. 그래서 ‘내용이 없는 사상들’과 ‘개념들이 없는 직관들’을 각각 경계한 칸트의 격언⁹⁶⁾에 빗대어 알고리즘과 데이터의 관계에서 “데이터 없는 알고리즘은 공허하고, 알고리즘 없는 데이터는 맹목적”⁹⁷⁾이라고 표현하기도 한다.

94) Ian H. Witten · Eibe Frank · Mark A. Hall · Christopher J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed., Elsevier, 2017, pp. 7~9.

95) Pedro Domingos, “A Few Useful Things to Know About Machine Learning”, *Communications of the ACM* 55(10), 2012, pp. 78~87: pp. 79~80.

96) Immanuel Kant, *Critique of Pure Reason*[*Kritik der reinen Vernunft*, 1st, 1781], Paul Guyer · Allen W. Wood(Eds. and Trans.), Cambridge University Press, 1998, [A5 1/B76], pp. 193~194: “Thoughts without content are empty, intuitions without concepts are blind.”



사회적 의미를 담아 표현하자면 데이터는 알고리즘이 지배하는 사회가 작동하기 위한 연료라고 볼 수도 있다.⁹⁸⁾

3. 데이터 마이닝과 모델의 단순화

데이터의 상관관계에 대한 판별식을 찾기 위해 분류(classification)나 예측(prediction), 군집(clustering) 같은 방법을 이용해 문제를 해결하는 것을 컴퓨터과학에서 ‘머신러닝(machine learning)’으로 부른다면, 세밀한 부분에서는 차이가 있지만 컴퓨터과학 못지않게 데이터 분석에 관심이 집중된 통계학에서는 이를 ‘데이터 마이닝(data mining)’으로 부른다.⁹⁹⁾ 데이터 마이닝은 “데이터에서 패턴을 발견하는 프로세스”¹⁰⁰⁾이고, “이미 데이터베이스 안에 있는 데이터를 분석함으로써 문제를 해결하는 것”¹⁰¹⁾이다. 적은 양의 데이터로부터 의미 있는 상관관계를 밝혀내는 것은 인간도 쉽게 할 수 있지만 인간의 지각 능력을 벗어난 양의 데이터, 이른바 ‘빅데이터(big data)’¹⁰²⁾라고 하는 정형 및 무정형 데이터로부터 그 상관관계를 추출하는 것은 기계의 몫이 되는 것이다. 빅데이터 환경이 조성됨으로서 비로소 데이터 마이닝이 빛을 발하는 이유이기도 하다.

전통적 형식의 데이터 분석은 특정 질문에 대해 인간의 분석 능력이 갖는 한계 안에서 몇 가지 변수만을 고려하여 간략한 통계를 그래프나 표로 제시한다. 반면 데이터 마이닝은 디지털 컴퓨터와 데이터 저장 능력의 극적인 발전에 힘입어 대량의 데이터 세트에서 훨씬 더 고차원적인 통계적 관계를 찾아낸다. 그래서 나중에 결정을 할 때 신뢰할 수 있는 유용한 규칙의 패턴을 발견하는 프로세스를 자동화한다.¹⁰³⁾

97) Jack M. Balkin, “The Three Laws of Robotics in the Age of Big Data”, *Ohio State Law Journal* 78(5), 2017, pp. 1217~1241: p. 1220.

98) “데이터가 현대 자본주의 가치 생산의 중심 추동력이 되고 이를 가지고 알고리즘 장치를 통해 사회를 조절하는 신종 기술 사회”를 ‘데이터 사회’ 또는 ‘데이터 알고리즘 사회’로 개념화하는 이해는 이광석, 데이터 사회 비판, 책읽는수요일, 2017, 43쪽 참조.

99) 김의중, (알고리즘으로 배우는) 인공지능, 머신러닝, 딥러닝 입문, 위키북스, 2016, 77쪽.

100) Ian H. Witten · Eibe Frank · Mark A. Hall · Christopher J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed., Elsevier, 2017, p. 6.

101) Ian H. Witten · Eibe Frank · Mark A. Hall · Christopher J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed., Elsevier, 2017, p. 5.

102) 아날로그 환경에서 생성되던 데이터에 비해 방대한 규모(Volume), 빠른 생성 속도(Velocity), 문자 및 영상 데이터를 포함한 종류의 다양성(Variety)을 특징으로 하는 디지털 환경의 거대한 정형 또는 비정형 데이터를 말한다.

103) Usama Fayyad, “The Digital Physics of Data Mining”, *Communications of the ACM* 44(3), 2001, pp. 62~65: p. 64.



이때 발견된 패턴은 하나의 알고리즘으로 표현될 수 있다. 그리고 이렇게 데이터 마이닝을 통해 찾아낸 관계들이 축적된 세트를 ‘모델(model)’이라고 부른다. 모델은 그 성격상 단순화(simplification)를 가정한다.¹⁰⁴⁾ 그래서 “어떤 프로세스의 추상적 표현에 불과”¹⁰⁵⁾한 것으로 여겨지기도 한다. 실제로 세상의 모든 복잡성이나 인간의 커뮤니케이션에서 드러나는 미묘한 차이를 완벽히 반영한 모델은 존재하지 않는다. 모델을 만들기 위해 다양한 정보 가운데 모델에 포함시켜야 할 중요한 정보를 선택하고 세상을 장난감처럼 단순화하는 과정이 필요하기 때문이다. 그리고 그 과정에서 발생하는 “모델의 맹점(model’s blind spots)”¹⁰⁶⁾을 통해 모델 개발자의 판단 기준과 우선순위를 알 수도 있다.

4. 모델 정립을 위한 단순화와 인간의 사고실험

모델로서 확립되기 위해 수반되는 단순화는 인간의 생각 속에서 전개되는 사고 실험에서도 쉽게 발견할 수 있다. 대표적인 예는 1971년 롤즈(J. Rawls)가 정의의 원칙들(the principles of justice)을 선택하기 위해 제시한 사고실험에서 합의 당사자들의 눈앞에 “무지의 베일(the veil of ignorance)”¹⁰⁷⁾을 씌워 “지식의 광범위한 제한”¹⁰⁸⁾을 하는 것에서 찾아볼 수 있다. 여기서 무지의 베일을 통해 지식을 제한하는 목적은 아무도 원칙을 자기에게 유리하도록 제정할 입장에 서지 못하게 하는 것과 세계의 임의적 우연성(the arbitrariness of the world)을 수정하는 데에 있다.¹⁰⁹⁾ 즉, 무지의 베일은 정의의 원칙을 합의하는 최초의 계약적 상황에서 당사자의 자기이익 관련 성과 세계의 우연성이라는 변수를 조정하는 장치인 셈이다. 무지의 베일은 “이용될 수 있는 정보가 적절하다는 것을 보장할 뿐만 아니라 그것이 언제나 동일하리라는 것까지를 보장”¹¹⁰⁾하는 기능을 수행한다.

104) Cathy O’Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Allen Lane, Penguin Books, 2016, pp. 15~31 참조.

105) Cathy O’Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Allen Lane, Penguin Books, 2016, p. 18.

106) Cathy O’Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Allen Lane, Penguin Books, 2016, p. 21.

107) John Rawls, *A Theory of Justice*, Original ed., Belknap Press of Harvard University Press, 2005 [1st, 1971], p. 12 및 pp. 136~142 참조.

108) John Rawls, *A Theory of Justice*, Original ed., Belknap Press of Harvard University Press, 2005, p. 137.

109) John Rawls, *A Theory of Justice*, Original ed., Belknap Press of Harvard University Press, 2005, p. 139 및 p. 141.

110) John Rawls, *A Theory of Justice*, Original ed., Belknap Press of Harvard University Press, 2005, p. 139.



그런데 원칙들은 사용할 수 있는 모든 지식에 비추어서 선택되어야 한다는 점에서 무지의 베일을 정의의 원칙 구성의 조건으로 삼는 것은 불합리하다는 반론이 가능하다. 롤즈는 이러한 반론을 예상하면서 그에 대해 “단순화(simplifications)를 강조해야 된다.”¹¹¹⁾는 간단한 언급으로 답변을 대신한다. “지식에 대한 그와 같은 제한이 없다면 원초적 입장에서 합의 문제는 터무니없이 복잡하게(complicated) 될 것”¹¹²⁾이라는 것이다. 이는 정의의 원칙 역시 그 구성을 위해 단순화가 필수적인 하나의 모델이란 점을 방증한다. 롤즈 역시 “그러한 제한(restrictions)이 없다면 우리는 전혀 어떤 한정된 정의론(any definite theory of justice)을 성립시킬 수가 없을 것”¹¹³⁾이라고 하면서, 원초적 입장에서 특정 지식에 대한 제한은 근본적 중요성을 갖는다고 강조한다. 그러면서 무지의 베일을 통해 당사자들 간의 차이점이 그들에게 알려져 있지 않으며 모두가 똑같이 합리적이고 비슷한 처지에 있기 때문에 누구나 동일한 논의를 수긍하게 되어 만장일치의 합의에 도달하게 될 것이라고 주장한다.

롤즈의 정의 원칙에서 중요한 것은 합의된 어떤 것이 아니라 합의 그 자체이다. 무지의 베일은 “합의 그 자체(the agreement itself)의 실질적 내용을 말하기 위한 조건”¹¹⁴⁾이며 개인의 결정 원칙이 사회의 결정 원칙으로 확대되는 것을 제한하는 수단이다.¹¹⁵⁾ 그래서 만일 원초적 입장에서 정의로운 합의가 생겨나게 되려면 그 당사자들은 “공정한 처지에 있어야 되고 도덕적 인격으로서 평등한 대우를 받아야”¹¹⁶⁾하는 것이다. 그런데 롤즈는 무지의 베일로 지식을 제한하면서도 “당사자들이 일반적인 지식을 모두 갖고 있다고 가정한다.”¹¹⁷⁾ 그들이 접근하지 못할 어떤 일반적 사실도 있을 수 없다는 것이다.

가정에 가정을 덧붙이는 이러한 방식의 구성은 단지 “복잡성(complications)을 피하기 위한”¹¹⁸⁾ 것이다. 무지의 베일은 일반적 지식을 모두 알고 있는 당사자의 특수한 지식을 제한하는 기능을 수행하는 데 어떤 지식이 일반적이고 특수한 것인지, 다시 말해 원초적 입장에 어떤 지식은 사용되고 어떤 지식은 사용되지

111) John Rawls, *A Theory of Justice*, Original ed., Belknap Press of Harvard University Press, 2005, p. 139.

112) John Rawls, *A Theory of Justice*, Original ed., Belknap Press of Harvard University Press, 2005, p. 140.

113) John Rawls, *A Theory of Justice*, Original ed., Belknap Press of Harvard University Press, 2005, p. 140.

114) John Rawls, *A Theory of Justice*, Original ed., Belknap Press of Harvard University Press, 2005, p. 140.

115) John Rawls, *A Theory of Justice*, Original ed., Belknap Press of Harvard University Press, 2005, p. 141.

116) John Rawls, *A Theory of Justice*, Original ed., Belknap Press of Harvard University Press, 2005, p. 141.

117) John Rawls, *A Theory of Justice*, Original ed., Belknap Press of Harvard University Press, 2005, p. 142.

118) John Rawls, *A Theory of Justice*, Original ed., Belknap Press of Harvard University Press, 2005, p. 142.



않는 것인지에 대해 그러한 “복잡성(complexity)을 구분하고 그 등급을 매기는 일이 어려우리라는 것이 분명”¹¹⁹⁾하고 그러한 시도는 이론구성을 복잡하게 한다는 점을 지적한다. 그리고 “공공적 정의관에 대한 근거가 가능한 모든 사람에게 분명하게 이해되는 것이 바람직하다.”¹²⁰⁾는 이유를 들어 복잡성을 피하려는 자신의 의도를 옹호한다. 그렇다면 정의의 원칙 역시 ‘무지의 베일’이라는 단순화 장치를 도입한 하나의 모델이라는 점에서 이를 생성하는 과정에서 발생한 맹점에는 모델 개발자의 판단 기준과 우선순위가 감추어져 있을 수 있다.

III. 머신러닝의 지도학습과 비지도학습

1. 지도학습과 분류 및 예측

머신러닝은 학습용 데이터에 레이블(label)이 있는 경우와 없는 경우를 나누어 각각 지도 학습과 비지도 학습으로 구분한다. 레이블은 학습 데이터의 특성을 미리 정의해 놓은 것이다. 예를 들어 사진 속에서 고양이를 구별하는 과제를 위해 사진을 학습 데이터로 제시할 때, 사진 속의 형체에 각각 ‘개’, ‘고양이’, ‘호랑이’라고 미리 정의해 놓은 것이 레이블이다. 레이블은 사진을 보고 사람이 정의한 것이기 때문에 레이블에 따라 사진을 학습하는 컴퓨터 입장에서 볼 때 사람에게 지도받은(supervised) 것이 된다.

이러한 지도학습에 따르는 것으로 분류 모델(classification model)과 예측 모델(prediction model)이 대표적이다. 둘 다 레이블이 있는 입력 데이터로 학습하지만 분류 모델은 그 결과 값이 학습 데이터세트에 포함되어 있는 레이블 중 하나로 고정되는 반면, 예측 모델은 결과 값이 학습 데이터세트로 결정된 함수식으로 계산된 임의의 값이 되어 데이터세트의 범위 안에 있는 어떠한 값도 가질 수 있다. 예를 들어 ‘개’, ‘고양이’, ‘호랑이’ 레이블로 구성된 데이터세트로 학습한 분류 모델에 다른 종류의 동물도 섞여 있는 사진을 입력하면 결과 값은 ‘개’, ‘고양이’, ‘호랑이’ 세 개 중에 하나다. 그런데 예측 모델에 같은 사진을 입력하면 ‘개’도 ‘고양이’도 ‘호랑이’도 아닌 별개의 결과 값이 나올 수 있다. 그래서 예측 모델은 주가 분석 같이 연속적인 범위 내에서 어떤 결과 값도 가능한 경우에 적용된다.

119) John Rawls, *A Theory of Justice*, Original ed., Belknap Press of Harvard University Press, 2005, p. 142.

120) John Rawls, *A Theory of Justice*, Original ed., Belknap Press of Harvard University Press, 2005, p. 142.



2. 비지도학습과 군집

레이블이 없어 사람에게 지도받지 않은(non-supervised) 비지도 학습에 따르는 것으로 군집 모델(clustering model)이 대표적이다.¹²¹⁾ 군집 모델의 관심사는 레이블이 없는 상태에서 입력된 데이터들이 어떤 형태로 집단을 형성하는지에 있다. 무명의 데이터에서 각각의 특성을 분석해 서로 유사한 특성을 가진 데이터끼리 집단화하는 것이 군집 모델의 학습 목표인 것이다. 전화 통화의 음질을 개선하기 위해 사람의 목소리와 잡음을 구별한다거나, 병원에서 질병군을 구별할 때, 마케팅에서 고객을 세분화하는 경우 등에 군집 모델이 사용된다. 레이블이 없다는 것은 사전지식이 없다는 것이기도 하다. 지도 학습 방식은 여러 가지 접근 방식의 차이에도 불구하고 사전지식이 있다는 공통점을 갖는다. 반면에 비지도 학습 방식은 지침 또는 규범적 가이드라인 설정하지 않고 데이터 자체의 연관성을 추출해 낼 수 있기 때문에 데이터 간의 관계 또는 데이터세트의 구조를 있는 그대로 드러내는 데에 용이하다.

IV. 머신러닝 알고리즘과 인공지능

머신러닝은 인공지능과 밀접한 관련을 맺고 있다. 그렇기 때문에 인공지능에 접근하는 방식에 따라 머신러닝 알고리즘의 종류를 구분하기도 한다. 도밍고스(P. Domingos)는 각 접근 방식에 따른 최상의 알고리즘이 있다는 점에 착안하여 알고리즘의 학파를 다섯 가지로 나누어 각각 기호주의자, 연결주의자, 진화주의자, 베이지주의자, 유추주의자로 부른다. 이때 기호주의자에게는 역연역법, 연결주의자에게는 역전파법, 진화주의자에게는 유전자 프로그래밍, 베이지주의자에게는 베イズ 추론, 그리고 유추주의자에게는 서포트 벡터 머신이 최상의 알고리즘이다. 그런데 실제로 각각의 알고리즘은 특정 과제는 훌륭하게 수행하지만 다른 과제에 대해서는 수행능력이 떨어지기 때문에 이 모든 알고리즘의 특성을 지닌 궁극의 단일한 알고리즘을 만드는 것을 머신러닝 분야 연구의 최종 목표 중 하나로 보기도 한다.¹²²⁾

121) 사람에게 지도받지 않는다는 의미에서 비지도 학습이지만 기계가 자체적으로 지도하는 경우 지도 그 자체로부터 완전히 자유롭지는 않다는 의미에서 반지도(semi-supervised) 또는 자기지도(self-taught) 학습을 별도로 구분하기도 한다. Yoshua Bengio, *Learning Deep Architectures for AI*, Foundations and Trends in Machine Learning 2(1), Now, 2009, p. 107 참조.

122) Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, New York: Basic Books, a member of the Perseus Books Group, 2015, xvii.



1. 기호주의와 논리 및 규칙 기반의 역연역법

기호주의자(symbolists)는 모든 지능이 마치 수학자가 수식을 다른 수식으로 변경함으로써 방정식을 푸는 것과 동일하게 기호를 조작하는 것으로 환원될 수 있다고 본다. 이들은 아무 것도 없는 상태에서 학습할 수는 없다고 생각한다. 즉 데이터와 결합될 수 있는 초기 지식이 필요하다고 보는 것이다. 기호주의자는 기존의 지식을 학습과 통합하는 방법과 문제를 해결하기 위해 각기 다른 지식 조각들을 그때그때 결합하는 방법을 계산해 왔다. 이들이 생각하는 최상의 알고리즘은 역연역법(inverse deduction)이다. 역연역법은 연역을 진행하기 위해 필요한 지식 중에 결여된 지식이 무엇인지 알아내고, 그 지식을 최대한 일반화하는 것이다. 예를 들면 보통의 연역은 다음의 순서로 진행된다.

P1: 모든 사람은 죽는다.

P2: 소크라테스는 사람이다.

C: 소크라테스는 죽는다.

그런데 ‘모든 사람은 죽는다.’는 명제와 ‘소크라테스는 죽었다.’는 명제만 알고 있는 경우를 가정해 보자.

P1: 모든 사람은 죽는다.

P2: _____

C: 소크라테스는 죽었다.

연역이 올바르게 진행되기 위해 필요한 명제가 무엇인지 거꾸로 추론하는 것이다. 모든 사람은 죽는데, 소크라테스가 죽었다고 하니 소크라테스는 사람일 것이라고 역으로 추론하는 것이다. 이것이 역연역법의 기본 생각이다.

역연역법은 계산량이 너무 많아서 대용량의 데이터를 처리하기 어렵다는 문제 때문에 기호주의자들은 귀납적 추론에 기반을 둔 의사결정 트리(decision tree)를 이용한다. 의사결정 트리 알고리즘은 나무를 거꾸로 세운 것과 같이 맨 위쪽에 위치한 뿌리(root)에서 시작해 줄기(branch), 잎(leaf) 순서로 하향식 의사결정 구조를 갖고, 사례를 분류하는 부분인 의사결정 노드(decision node)가 또 다른 줄기를 만드는 분기점이 된다. 줄기가 뻗어 나가는 뿌리 및 각 분기점에서 어떤 특성에 대한 값을 묻고 그 대답에 따라 새로운 줄기가 뻗거나 곧바로 잎에 도달하게 되는데 뿌리에서



앞에 이르는 경로는 하나의 규칙에 대응한다. 의사결정 트리에서 중요한 것은 각 단계마다 시험할 가장 좋은 특성을 찾아내는 것이다.¹²³⁾

2. 연결주의와 신경망 기반의 역전파법

지식이 신경세포(neuron) 사이의 연결에 있다고 믿는 연결주의자(connectionists)는 학습을 뇌가 하는 것이라고 생각하기 때문에 뇌가 학습하는 방식을 그대로 복제하여 설계·제작하려고 한다.¹²⁴⁾ 뇌는 신경세포 간의 연결 강도를 조절하여 학습하기 때문에 어떤 연결이 어떤 오류를 일으키는지 알아내어 그에 맞춰 연결들을 수정하는 것이 중요한 문제이다. 이때 연결 강도는 가별 가중치(variable weight)에 따라 결정되는데, 신경세포 입력의 가중치가 높을수록 이 가중치를 가진 신경접합부의 연결 정도는 더 강하다. 음성 인식이나 문자 인식과 같은 지각(perception)을 염두에 둔 신경세포 모델인 퍼셉트론(perceptron) 알고리즘에서 가중치 합은 한계값보다 클 경우 퍼셉트론을 작동시키고 작을 경우 작동을 멈추게 하는 방식으로 퍼셉트론의 기능 수행에 변화를 줄 수 있다.

문제는 다층의 신경세포가 연결되어 있을 때 출력 층의 신경세포가 오류를 일으킨 경우 그 원인을 은닉 층에 있는 모든 신경세포 간 경로 중에 어디에서 찾을 것이며 반대로 출력이 정확한 경우 어떤 경로에 의한 것인지 찾는 것이다. 이러한 신뢰 할당(credit-assignment) 문제는 머신러닝의 핵심 문제로서 연결주의자에게 최상의 알고리즘은 역전파법(back-propagation)이다.¹²⁵⁾ 역전파법은 시스템이 출력한 것을

123) 의사결정 트리에 관해서 Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, Basic Books, a member of the Perseus Books Group, 2015, pp. 85-89 및 김의중, (알고리즘으로 배우는) 인공지능, 머신러닝, 딥러닝 입문, 위키북스, 2016, 135~148쪽 참조.

124) 연계(association)에 의한 학습은 로크(J. Locke)와 흄(D. Hume)에서 밀(J. S. Mill)에 이르는 영국의 경험주의자들이 좋아하는 주제이다. 도밍고스에 따르면 기호주의자는 기호와 개념 사이에 일대일 관계를 상정하고 학습을 순차적인(sequential) 것으로 이해하는 반면, 연결주의자는 각 개념이 여러 신경세포에 흩어져 있고 각 신경세포는 여러 개념을 표현하는 데 참여하는 관계를 상정하고 학습을 병행적인(parallel) 것으로 이해한다는 점에서 차이를 드러낸다. 이에 관해서 Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, Basic Books, a member of the Perseus Books Group, 2015, pp. 93~94 참조.

125) 심리학자인 데이비드 러멜하트(David Rumelhart)가 인공신경망(artificial neural networks) 연구의 권위자이자 딥러닝(deep learning) 연구의 선두주자로 유명한 제프리 힌튼(Geoffrey Hinton), 그리고 로널드 윌리엄스(Ronald Williams)와 함께 역전파 알고리즘을 다층 신경망에 적용한 논문으로 David E. Rumelhart · Geoffrey E. Hinton · Ronald J. Williams, "Learning Representations by Back-Propagating Errors", *Nature* 323(6088), 1986, pp. 533~536 참조. 이들은 기호주의자에 의해 퍼셉트론이 학습할 수 없는 것으로 주장된 배타적 논리합(XOR)도 역전파 알고리즘이 학습할 수 있다는 것을 보여주었다.



원했던 것과 비교한 후 원하는 목표 값에 근접할 때까지 층층이 겹쳐 있는 신경세포의 연결들을 계속해서 변경하는 것이다. 이때 가중치를 계속 조절하여 오류를 낮추고 조절이 모두 실패하면 변경을 멈춘다.

세포의 완전한 모델을 만들기 위해서는 상이한 유전자의 발현도를 연결하고, 환경 변수들을 내부 변수들과 연관시키는 등의 기능을 수행하는 매개변수를 학습하는 양적인 모델이 필요하다. 그런데 이런 변수들은 단순한 선형 관계가 아니기 때문에 전형적인 비선형 시스템인 살아 있는 세포의 완전한 모델을 생성하는 것은 어려운 일이다. 하지만 비선형 함수를 효율적으로 파악할 수 있는 역전파법은 이러한 문제를 잘 처리할 수 있다.¹²⁶⁾

3. 진화주의와 유전자 프로그래밍 기반의 유전자 탐색

진화주의자(evolutionaries)는 학습의 모태는 자연선택이라고 생각한다. 자연선택으로 인간이 만들어진 것이라면 자연선택은 어떤 것도 만들어 낼 수 있다는 것이다. 그러므로 남은 일은 자연선택을 컴퓨터에서 모의실험해 보는 것이다. 진화주의자가 해결하려는 핵심 과제는 학습하는 구조이다. 연결주의자의 역전파법처럼 변수를 최적화함으로써 진화된 구조를 미세 조정하는 것에 그치지 않고, 그런 조정을 더 미세하게 할 수 있는 뇌 자체를 만들려고 하는 것이다. 다시 말해 자연으로부터 착상을 얻어 머신러닝 알고리즘을 설계한다는 공통점이 있지만 연결주의자는 가중치 학습에 초점을 맞추는 반면, 진화주의자는 구조 학습에 주된 관심을 둔다.¹²⁷⁾

진화주의에서 최상의 알고리즘은 유전자 프로그래밍(genetic programming)이다. 유전자 프로그래밍은 DNA의 염기서열처럼 프로그램의 비트 열을 교차시키는 것에서 한 걸음 더 나아가 프로그램 자체를 교차시키는 것이다. 자연이 유기체를 결합하여 진화시킨 것과 동일한 방식으로 컴퓨터 프로그램을 교차하여 진화시키는 것이다. 이때 유전 알고리즘의 적합성은 진화된 프로그램이 목표에도 도달한 정도에 숫자를 표시하여 점수를 할당하고 훈련용 데이터에서 프로그램의 실제 출력과 올바른 결과 사이의 차이를 측정함으로써 평가될 수 있다.¹²⁸⁾

126) Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, Basic Books, a member of the Perseus Books Group, 2015, p. 109 및 p. 114 참조.

127) Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, Basic Books, a member of the Perseus Books Group, 2015, p. 137.

128) Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, Basic Books, a member of the Perseus Books Group, 2015, p. 123~124 및 p.132 참조.



4. 베이즈주의와 베이즈 네트워크 기반의 확률적 추론

베이즈주의자(Bayesians)는 불확실성에 주목한다. 이들은 학습된 모든 지식은 불확실하고, 학습 자체를 하나의 불확실한 추론의 형태라고 생각한다. 그래서 잡음(noise)이 섞여 있고, 불완전하며, 모순적이기까지 한 정보를 흐트러뜨리지 않고 어떻게 다룰 수 있느냐가 관건이다. 이에 대한 해결책은 확률적 추론(probabilistic inference)이고, 이를 실현하는 최상의 알고리즘은 베이즈 정리와 그 파생 수식이다. 베이즈 정리는 다음과 같다.¹²⁹⁾

$$P(A | B) = P(A) \times P(B | A) / P(B)$$

이 공식에서 ‘ $P(A | B)$ ’는 ‘B의 조건에서 A가 발생할 확률’을 의미한다. 그러므로 B가 발생할 경우 A가 발생할 확률은 B가 발생할 확률에 B의 조건에서 A가 발생할 확률을 곱하는 것 즉, ‘ $P(B) \times P(A | B)$ ’이다. 그리고 반대로 A가 발생할 경우 B가 발생할 확률은 A가 발생할 확률에 A의 조건에서 B가 발생할 확률을 곱하는 것 즉, ‘ $P(A) \times P(B | A)$ ’이다. A와 B가 모두 발생하는 경우 ‘ $P(B) \times P(A | B)$ ’는 ‘ $P(A) \times P(B | A)$ ’와 같게 되어 ‘ $P(B) \times P(A | B) = P(A) \times P(B | A)$ ’가 되고 좌변의 ‘ $P(B)$ ’를 우변으로 옮긴 것이 최종적인 위 공식이다.

베이즈 추론은 의료 진단처럼 증상에 대한 원인을 찾는 추론에 유용하다. 예를 들어 암 진단을 위한 검사에서 양성 판정을 받았을 때 오진 확률이 1퍼센트라고 할 경우 암에 걸렸을 확률은 99퍼센트처럼 보인다. 그러나 베이즈 추론은 양성 판정이 내려진 경우 암이 발병했을 확률을 추론하기 위해 사전확률을 고려한다. 이때 경험적 지식 또는 가설이 사전확률로 사용될 수 있다. 암에 걸렸을 경우 양성 판정을 받을 확률이 99퍼센트라는 것에 더해 전체 인구 대비 암 발병률을 0.2퍼센트, 암 진단에서 양성 판정을 받을 확률을 1퍼센트라고 가정할 경우 양성 판정시 암 발병률은 ‘ $0.002 \times 0.99 / 0.01$ ’로 ‘0.198’ 즉, 19.8퍼센트가 된다.¹³⁰⁾ 새로운 증거를 얻었을 때 가설에 대한 믿음의 정도를 갱신하는 간단한 규칙으로서 베이즈 정리는 새로운 증거를 우리의 믿음에 통합하는 방법을 알려주고, 확률 추론 알고리즘이 그 일을 최대한 효율적으로 수행한다.

129) Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, Basic Books, a member of the Perseus Books Group, 2015, p. 143 이하 참조.

130) 에이즈 바이러스 감염에 관한 양성 판정을 받은 경우의 에이즈 바이러스 감염률을 예로 든 Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, Basic Books, a member of the Perseus Books Group, 2015, p. 148 이하 참조.



베이지주의에서 확률 분포를 함께 정의하는 특성 및 이에 대응하는 가중치의 세트라고 할 수 있는 베이지 네트워크는 기호주의에서 논리가 차지하는 위상과 같다. 마르코프 연쇄(Markov chains)¹³¹⁾는 현재라는 조건에서 미래는 과거에 대하여 조건부 독립이라는 가정을 구현하고, 은닉 마르코프 모델(Hidden Markov Model)은 조건부 독립에 더해 각 관찰은 해당하는 상태에만 의존한다고 추정한다. 마르코프 연쇄는 구글 번역기 같은 기계 번역 시스템을 비롯해 구글을 탄생시킨 알고리즘인 페이지랭크(PageRank) 자체에도 적용되고, 애플(Apple)의 음성 인식 및 자연어 번역 시스템인 시리(Siri)에는 전달되는 소리를 관찰함으로써 은닉 상태에 있는 글로 쓰인 단어를 추론하는 은닉 마르코프 모델이 사용된다.

5. 유추주의와 서포트 벡터 및 사례 기반의 조건부 최적화

유추주의자(analogizers)는 상황들 사이의 유사성을 인식하여 다른 유사성을 추론하는 것이 학습의 핵심이라고 여긴다. 만약 두 명의 환자가 유사한 증상을 보인다면 두 환자는 아마도 동일한 질병을 갖고 있을 가능성이 높다. 문제는 두 환자의 증상이 얼마나 유사한지이다. ‘최근접 이웃 알고리즘(nearest-neighbor algorithm)’으로 불리는 유추 알고리즘¹³²⁾은 입력된 데이터를 판단할 때 그와 가장 유사한 데이터, 즉 최근접 이웃을 찾는다. 이때 유사한 근접 데이터들은 그 자체로 작은 분류기(classifier)이며 자기를 가장 근접한 점으로 삼는 질의 예시(query example)에 대해 그 클래스를 예측하는 역할을 한다.

131) 러시아의 수학자 마르코프(A. A. Markov)는 1913년에 발표한 논문에서 러시아 문학의 고전인 푸시킨(A. S. Pushkin)의 ‘예브게니 오네긴(Evgenij Onegin)’ 중 일부를 발췌하여 해당 텍스트에 있는 20,000개의 러시아 알파벳 글자를 자음이나 모음으로 연결된 건본으로 삼아 순차(sequence) 구조를 도입하여 각 글자의 확률이 바로 앞에 있는 글자에 의존한다는 점을 분석했다. 해당 논문의 영문 번역은 Andrei A. Markov, “An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains”, Gloria Custance · David Link(Trans.), *Science in Context* 19(4), 2006, pp. 591~600 참조.

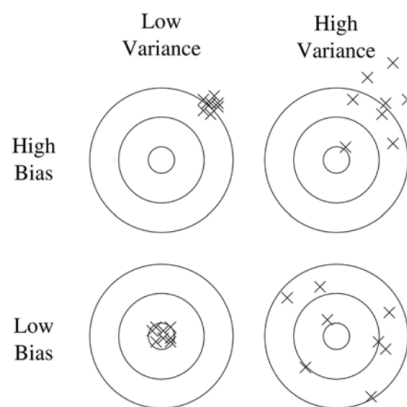
132) 유추를 알고리즘으로 구현한 내용은 1951년 2월 픽스(E. Fix)와 호지스(J. L. Hodges)의 기술 보고서(Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties, Report 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, February 1951)에 담겨 있지만 학술지에는 상당 기간이 지난 후에 발표되었다. 이에 관해서 B. W. Silverman · M. C. Jones, “E. Fix And J. L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951)”, *International Statistical Review / Revue Internationale de Statistique* 57(3), 1989, pp. 233~238 및 Evelyn Fix · J. L. Hodges, Jr., “Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties”, *International Statistical Review / Revue Internationale de Statistique* 57(3), 1989, pp. 238~247 참조.



그런데 최근접 이웃 알고리즘은 데이터 지점을 잘못된 클래스로 분류할 경우 그 오류가 전체 영역으로 퍼지는 과적합(overfitting) 문제¹³³⁾에 취약하다. 이때 최근접 이웃이 여러 개일 경우 즉, k 개의 최근접 이웃을 가진 k -최근접 이웃 알고리즘은 다수의 잘못이 있을 때만 오류가 발생하므로 과적합 문제에 좀 더 강인하다. 또한 최근접 이웃의 개수를 늘리는 것에 더해 상관성에 기초한 가중치를 부여함으로써 가중치 k -최근접 이웃 알고리즘을 도출할 수도 있다. 다만 k 가 커질수록 분산은 감소하지만 편향은 증가하고,¹³⁴⁾ 이때 증가하는 편향에 가중치가 높게 부여될 경우 편향은 가중될 수 있다.

아래 ‘자료 1’¹³⁵⁾은 머신러닝에서 편향 또는 편중을 의미하는 ‘bias’와 분산을 의미하는 ‘variance’를 설명하기 위한 표적이다.

자료 1. 편향(bias)과 분산(variance)



133) 과적합 문제는 무시해야 할 노이즈나 아웃라이어 데이터까지 모두 정상적인 것으로 인식하고 학습하면서 생기는 것으로 머신러닝에서 매우 빈번히 등장하는 이슈이며 실무자가 해결해야 할 가장 큰 문제이기도 하다. 이에 관해서 김의중, (알고리즘으로 배우는) 인공지능, 머신러닝, 딥러닝 입문, 위키북스, 2016, 123~124쪽 참조; 과적합 문제의 해법을 베이지주의에서 찾는 경우로 Nate Silver, *The Signal and the Noise: Why So Many Predictions Fail – But Some Don't*, Penguin Books, 2012, 특히 Chapter 5 및 8 참조.

134) Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, New York: Basic Books, a member of the Perseus Books Group, 2015, p. 183. 머신러닝 알고리즘의 오류를 줄이기 위해 유추의 근거가 되는 데이터가 많아질 경우 편향이 증가할 수 있다는 점은 데이터를 분류하고 유형을 예측하는 과정이 차별과 밀접한 관련을 맺는다는 점을 시사한다. 머신러닝 알고리즘의 오류와 편향에 관한 사례는 본 논문 「제2장 제3절 머신러닝 알고리즘의 오류와 편향」 참조.

135) Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, New York: Basic Books, a member of the Perseus Books Group, 2015, p. 79.



상위에 있는 두 개의 표적은 똑같이 중앙에서 떨어진 오른쪽 위에 ‘x’ 표시가 집중되어 있어 편향이 높지만, 좌측 표적에 비해 우측 표적은 표시가 표적의 중앙에서부터 표적 외부까지 넓게 퍼져 있어서 분산이 높아 편향을 가리는 효과가 있다. 하위에 있는 두 개의 표적 역시 똑같이 중앙에 ‘x’ 표시가 집중되어 있어 편향이 낮지만 우측 표적이 좌측 표적에 비해 분산이 높다.

유추주의자에게 최상의 알고리즘은 서포트 벡터 머신(support vector machine)이다. 어떤 경험들을 기억할 것인지 그리고 그 경험들을 새로운 예측에 어떻게 결합할 것인지 계산해 내는 서포트 벡터 머신은 일정 수의 예시들과 그 가중치로 양과 음의 클래스 사이에 경계선을 구하는 유사성 측정을 수행하기 때문에 가중치 k-최근접 이웃 알고리즘과 유사해 보인다. 하지만 서포트 벡터 머신은 경계선을 확정하는 데 필요한 핵심 예시들만을 서포트 벡터로 기억한다는 점에서 차이가 있다. 경계선과 예시들 사이의 거리인 마진(margin)과 예시들에 대한 가중치를 변경함으로써 마진을 최대화하거나 가중치를 최소화하는 최적화 경계선을 찾는 것이 서포트 벡터 머신이 수행하는 작업이다.

최근접 이웃 알고리즘과 서포트 벡터 머신이 최근접 이웃이나 서포트 벡터의 클래스를 기반으로 새로운 대상의 클래스를 예측하는 것이라면 또 다른 유추주의 머신러닝으로서 사례 기반 추론(case-based reasoning)은 검색 대상의 구성 요소로 형성된 복잡한 구조를 출력물로 내놓을 수 있다.¹³⁶⁾ 전화 상담 서비스나 판례에 기초하여 특정 사건을 변론하는 법률 서비스는 사례 기반 추론이 적용될 수 있는 대표적인 분야이다. 이러한 분야는 모든 논증을 계산으로 환원시키려는 라이프니츠(Leibniz)의 꿈과도 맞닿아 있다.¹³⁷⁾

136) 사례 기반 추론은 규칙 기반 추론(또는 논리 기반 추론)과 함께 법적 추론에 관한 인공지능 모델을 생성하는 주요 접근 방식이다. 예를 들어 법학을 전공하는 학생이 선례로부터 추론하는 것을 학습하도록 도울 수 있게 설계된 시스템인 CATO에 탑재된 사례 기반 법적 논증의 계산 모델에 관한 설명으로 Kevin D. Ashley, “An AI Model of Case-Based Legal Argument from a Jurisprudential Viewpoint”, *Artificial Intelligence and Law* 10(1~3), 2002, pp. 163~218 참조; 사례 기반 추론에 관한 자세한 내용은 Kevin D. Ashley, “Case-Based Reasoning”, in *Information Technology and Lawyers: Advanced Technology in the Legal Domain, from Challenges to Daily Routine*, Arno R. Lodder · Anja Oskamp(Eds.), Springer, 2006, pp. 223~260 참조; 사례 기반 법적 논증과 규칙 기반 법적 논증에 관한 설명은 조한상 · 이주희, “인공지능과 법, 그리고 논증”, 법과 정책연구 16(2), 2016, 295~320쪽; 302~313쪽 참조.

137) 머신러닝 알고리즘이 법적 결정을 예측하고 대체하는 문제에 관해서 본 논문 「제4장 제1절 IV. 예측하는 알고리즘과 예측되는 법」 참조.



제3절 머신러닝 알고리즘의 오류와 편향

기계가 하는 판단에는 인간이 의사결정을 할 때 개입될 수 있는 선입견이나 편견이 작용하지 않을 것이라고 종종 가정된다. 그런데 이러한 가정에는 기계가 인간보다 객관적이고 합리적이며 공정할 것이라는 기대 또는 믿음이 담겨 있다. 그런데 머신러닝 알고리즘이 작동한 결과가 오늘날의 상식에 반하거나 인간의 규범적 직관에 맞지 않게 편향적으로 보이는 경우가 종종 발생하고 있다. 여기서 오늘날의 상식 또는 직관과 배치된다는 것은 구체적인 결정 과정에 포함되면 안 될 것 같은 성별이나 인종 같은 요소들이 알고리즘의 판단에 결정적 근거로 작용하여 이른바 ‘의심스러운 분류(suspicious classification)’¹³⁸⁾를 수행하고 있는 것처럼 보인다는 것이다.

머신러닝 알고리즘에 의한 결정이 고용이나 노동, 주거, 교육, 보건, 출산, 육아 등과 같이 사회생활의 중요한 영역¹³⁹⁾에서 차별적으로 이루어지는 것일 때 문제의 심각성은 단순히 기계에 대한 믿음이나 기대가 좌절되는 수준을 넘어선다. 머신러닝 알고리즘의 차별 문제는 머신러닝 알고리즘 개발의 주축이 되는 컴퓨터 과학 연구자들이 이를 증명하려는 연구를 수행하면서 실증적 사례들을 제시함으로써 본격적으로 논의된 측면이 있지만, 머신러닝 알고리즘이 인간이 기대하는 바와 다른 방식으로 판단을 할 수 있다는 점은 비전문가라도 인터넷 기업의 검색엔진에 간단한 질문 몇 개를 입력해 보는 것만으로도 어렵지 않게 확인할 수 있을 것이다.

I. 머신러닝 알고리즘의 의심스러운 분류와 예측

1. 스위니의 연구

스위니(L. Sweeney)의 연구¹⁴⁰⁾는 인터넷 기업에서 운영하는 플랫폼을 통해 전달되는 광고가 특정 인종을 겨냥한 것이 아닌가 하는 의문에서 출발한다. 해당 플랫폼의 검색엔진에 몇 가지 이름을 입력했을 때 해당 플랫폼은 그 이름에

138) 의심스러운 분류(suspicious classification)에 관해서 John Hart Ely, *Democracy and Distrust: A Theory of Judicial Review*, Harvard University Press, 1980, pp. 145~170 참조. 여기서 일리(J. H. Ely)는 의심스러운 분류 이론이 동기분석(motivation analysis)을 위한 수단으로 기능한다는 점을 지적한다.

139) 헌법상 노동, 교육, 출산, 육아, 보건 그리고 경제에 관한 권리와 민주주의의 관련성에 대해서 김하열, “민주주의 정치이론과 헌법원리”, *공법연구* 39(1), 2010, 161~192쪽: 180~186쪽 참조.

140) Latanya Sweeney, “Discrimination in Online Ad Delivery”, *Communications of the ACM* 56(5), 2013, pp. 44~54.



연결된 범죄 관련 기록, 예를 들어 체포 또는 구금 기록을 알고 싶은지 묻는 광고를 질문의 형태로 모니터에 보여준다. 그런데 다른 이름의 문자열을 입력하면 그런 광고가 나타나지 않는다. 예를 들어 ‘Latanya Farrell’, ‘Latanya Sweeney’, ‘Latanya Lockett’처럼 ‘Latanya’라는 이름이 포함된 문자열을 입력하면 체포기록에 관해 알려주겠다는 광고를 추천해 주는 반면, “Kristen Haring”, ‘Kristen Sparrow’, ‘Kristen Lindquist’ 같이 ‘Kristen’이라는 이름을 입력하면 그런 광고가 전혀 나타나지 않는다는 것이다. 영어권에서 ‘Latanya’는 흑인을 연상시키는 이름이고 ‘Kristen’은 백인을 연상시키는 이름이다.

광고는 대한민국헌법을 비롯해 미국 헌법에서도 표현의 자유(freedom of expression)로 보호 받는다.¹⁴¹⁾ 그런데 이러한 온라인 광고 기능이 취업이나 경연을 위해 지원한 사람들의 이름을 고용주 또는 인사 담당자가 해당 플랫폼의 검색엔진에 입력할 때도 작동한다고 해보자. 그래서 어떤 지원자의 이름을 입력했을 때에는 전과기록을 확인할 것이냐고 묻는 광고가 나타나는 반면 그 지원자와 경쟁 관계에 있는 다른 지원자의 이름에 대해서는 그와 같은 광고가 나타나지 않는다. 그 결과 범죄기록 검색을 추천하는 광고가 나타난 이름을 가진 지원자만 추가적인 조사 및 심사의 대상으로 선별된다면 이러한 상황은 범죄 기록 관련 광고가 추천되는 이름을 가진 특정 지원자, 다시 말해 그런 이름을 사용하는 특정 인종에 속하는 지원자에게 불리하게 작용할 수 있다. 이러한 상황을 여전히 표현의 자유로 보호된다고 볼 것인지 그에 못지않게 중대한 차별의 문제로 다루어야 하는 것인지 결정하는 것은 어려운 문제이다.

2. 루미스 사건

루미스(E. Loomis)는 2013년 미국 위스콘신 주에서 발생한 차량 이용 총격 사건에 사용된 차량을 운전했다는 이유로 체포되고 주정부로부터 기소됐다. 루미스는 기소 사유 5개 중에 형벌이 낮은 2개 즉, 경찰로부터 도주를 시도했다는 점과 자동차 소유자의 동의 없이 자동차를 운전했다는 점에 대해서만 유죄의 항변을 했다. 루미스의 항변을 받아들인 순회법원(circuit court)은 판결 전 조사(presentence investigation)¹⁴²⁾를 명령했다.¹⁴³⁾ 이에 따라 보호관찰관은 판결 전

141) 대한민국헌법 제21조 제1항: “모든 국민은 언론·출판의 자유와 집회·결사의 자유를 가진다.”; 미국 헌법 수정 제1조(First Amendment of the US Constitution): “의회는 …(중략)… 언론과 출판의 자유 …(중략)… 를 축소하는 법률을 제정해서는 안 된다{Congress shall make no law …(중략)… abridging the freedom of speech, or of the press; (후략)…}.”

142) 미국 연방 법률(U. S. Code) 제18편(범죄와 형사소송) 제2부(형사소송) 제227장(판결) 부분에 있는 제3552조는 ‘판결 전 보고(presentence report)’라는 제목 하에 판결 전 보호관찰관(probation officer)에



조사를 실시했고, 판결 전 조사에 대한 보고에는 ‘콤파스(COMPAS)’ 위험 평가가 첨부됐다. ‘콤파스(COMPAS)’는 소프트웨어 개발업체(Northpointe)에서 교정당국이 범죄자의 배치, 관리, 처우 등에 대한 결정을 할 때 그 결정을 지원할 수 있도록 설계한 위험 평가 장치이다.¹⁴⁴⁾ 이 소프트웨어는 범죄자를 대상으로 한 면담이나 전과기록 등에서 수집된 정보에 기초해서 2년 안에 다시 체포될 가능성을 예측하여 점수를 매긴다.

법원은 루미스에게 11년형을 선고하면서 그 중에 6년을 교정시설에서 수감생활을 하고, 나머지 5년은 교정시설 밖에서 교정국의 감독을 받도록 판결했다. 루미스는 법원이 자유형을 집행하도록 결정한 데에는 무엇보다 재범의 위험성이 높게 평가됐기 때문이라고 판단했다. 그런데 루미스가 알 수 있는 것은 자신에 대한 위험평가 점수가 높게 산정되어 위험도가 높은 인물로 평가됐다는 것뿐이고, 어떻게 그러한 평가가 이루어졌고 어떤 요인이 비중 있게 측정됐는지 알 수 없다.¹⁴⁵⁾ 소프트웨어 개발업체는 영업비밀을 이유로 알고리즘 공개를 거부했기 때문이다.¹⁴⁶⁾

루미스는 이러한 위험 평가 장치를 사용한 것이 판결을 받을 때 성별처럼 그 근거로 삼아서는 안 되는 특성이 고려되고, 부정확한 정보를 기초로 삼게 되며, 개별적으로 판단 받지 못하게 된다는 점을 이유로 미국 헌법 수정 제14조¹⁴⁷⁾의 적법절차에 관한 권리(a right to due process)를 위반했다고 주장했지만¹⁴⁸⁾ 위스콘신 주 대법원은 이를 받아들이지 않았다.¹⁴⁹⁾ 그런데 이 사건에서 루미스가 수정 제14조의 평등 보호(equal protection) 위반을 주장하지는 않았다.¹⁵⁰⁾

의한 조사와 보고(a), 교정국에 의한 연구와 보고(b), 정신의학 또는 심리학 검사관에 의한 검사와 보고(c), 보고의 공개(d)에 관해 법원의 명령 권한을 규정하고 있다. 이 중 위스콘신 주 대법원이 루미스 사건에서 명령한 것은 동조 (a)에 근거한 것이다. 18 U. S. C. §3552 (2016) 참조.

143) State v. Loomis, 2016 WI 68, 881 N. W. 2d 749 (2016), para. 11~12.

144) Northpointe, Inc., “Practitioner's Guide to COMPAS Core”, 19 March 2015, http://www.northpointeinc.com/files/technical_documents/Practitioners-Guide-COMPAS-Core_031915.pdf, p. 1.

145) 재판부는 콤파스(COMPAS)의 개발업체가 발간한 실무가들을 위한 안내서(Practitioner's Guide to COMPAS Core)를 보면 위험 점수가 대부분 범죄전력(criminal history) 같은 정적인 정보(static information)에 기초한다는 것이 설명되어 있다는 점을 이유로 루미스의 주장을 배척한다. State v. Loomis, 2016 WI 68, 881 N. W. 2d 749 (2016), para. 54.

146) State v. Loomis, 2016 WI 68, 881 N. W. 2d 749 (2016), para. 51.

147) 미국 헌법 수정 제14조: “...(전략); 모든 주는 적법한 절차 없이 누구의 생명, 자유, 재산도 박탈해서는 안 되고, 관할권 내의 누구에 대해서도 법의 평등한 보호를 부인해서는 안 된다{...(전략); nor shall any State deprive any person of life, liberty, or property, without due process of law; nor deny to any person within its jurisdiction the equal protection of the laws}.”

148) State v. Loomis, 2016 WI 68, 881 N. W. 2d 749 (2016), para. 34.

149) State v. Loomis, 2016 WI 68, 881 N. W. 2d 749 (2016), para. 10.

150) State v. Loomis, 2016 WI 68, 881 N. W. 2d 749 (2016), para. 80.



II. 머신러닝 알고리즘의 인식 오류와 차등적 분류

1. 작은 눈과 감은 눈의 인식 오류

타이완계 미국인이 2010년에 디지털 카메라를 구입하여 자신을 피사체로 하여 사진을 촬영했다.¹⁵¹⁾ 그런데 사진을 찍을 때 마다 웃는 모습을 하는 피사체를 디지털 카메라의 소프트웨어는 눈을 깜빡이고 있는 것으로 판단하여 ‘누군가 눈을 깜박였나?(Did someone blink?)’라는 경고 문구를 보여줬다. 이와 같은 작동은 계속 반복 되었고, 눈을 방울처럼 크게 뜨고 나서야 문구는 사라졌다. 이 여성은 그녀의 블로그에 사진을 올리면서 ‘인종주의자 카메라! 아니야, 나는 눈을 깜박인 적이 없어. 난 단지 아시아인일 뿐이야!’라고 적었다. 이에 대해서 ‘당신이 사용한 카메라가 일본 회사 제품(Nikon Coolpix S630)인데 아시아인의 눈에 맞게 설계되지 않았을까?’라는 댓글이 달렸다. 얼굴 인식 기술이 장착된 카메라는 눈을 깜박일 때 경고 문구를 띄우거나, 누군가 웃을 때 자동적으로 사진을 찍도록 프로그래밍 되어 있다. 얼굴 인식의 원리는 상대적으로 단순한데 대부분의 사람들이 두 개의 눈과 눈썹, 위아래의 입술과 코를 갖고 있다는 점을 토대로 한다. 알고리즘은 그 공통의 특성을 찾도록 훈련된 것일 뿐이다.

그런데 웃는 눈과 감은 눈 그리고 작은 눈을 섬세하게 인식하지 못하는 얼굴 인식 프로그램의 효과는 단순히 이용자의 감정을 상하게 하는 해프닝으로 끝나지 않는 경우도 있다. 호주에서 유학생활동을 하는 타이완계 뉴질랜드인 학생은 여권을 갱신하기 위해 정부에서 운영하는 알고리즘 기반의 사진 판독 시스템에 사진을 등록했다.¹⁵²⁾ 그러나 사진 판독 시스템은 사진 속 인물이 눈을 감고 있어 등록 거부 사유에 해당된다는 오류 메시지를 발신했다. 다른 사진으로 교체하여 몇 번의 등록을 시도한 끝에 승인을 받았지만, 정교하지 못한 사진 판독 기술은 해당 알고리즘 시스템에 인종차별주의자의 관점이 투영된 것은 아닌지 의구심을 갖게 한다.

151) Adam Rose, “Are Face-Detection Cameras Racist?”, Time, 22 January 2010, <http://content.time.com/time/business/article/0,8599,1954643,00.html>, 접속일: 2017년 12월 3일.

152) Selina Cheng, “An Algorithm Rejected an Asian Man’s Passport Photo for Having ‘closed Eyes’”, Quartz, 7 December 2016, <https://qz.com/857122/an-algorithm-rejected-an-asian-mans-passport-photo-for-having-closed-eyes>, 접속일: 2017년 12월 3일.



2. 고릴라 사건

아프리카계 미국인인 소프트웨어 개발자가 친구와 함께 찍은 사진을 사진 인식 응용프로그램(Google Photo)에 올렸다.¹⁵³⁾ 이 프로그램은 이미지에서 사람과 장소 그리고 사건을 인식할 수 있다는 머신러닝 알고리즘 기술을 장점으로 내세운 것이었다. 그런데 이 응용프로그램은 두 사람에게 ‘Gorillas(고릴라)’라는 레이블을 달았다(이하 ‘고릴라 사건’). 또 다른 회사(Flickr)에서 출시한 유사한 프로그램에서는 흑인 남성과 백인 여성을 침팬지나 고릴라 같은 유인원으로 식별했다. 구글(Google)의 경우 즉시 사과하고 검은 피부의 얼굴을 더 잘 인식할 수 있도록 바로잡는 것을 포함해 재발 방지를 위해 장기간 동안 노력을 다하겠다고 했다. 그런데 구글 직원의 대부분이 백인(60%)과 아시아인(31%)인 반면 라틴계(3%)와 아프리카계 미국인(2%)은 극히 일부이다.

이 고릴라 사건이 발생한 이후 2년이 조금 지난 시점에 기술의 사회적 영향을 다루는 잡지인 와이어드(Wired)에서 해당 응용프로그램(Google Photo)에 대한 몇 가지 시험을 진행했다.¹⁵⁴⁾ 첫 번째는 40,000개의 풍부한 동물 이미지를 가지고 시험해본 결과 판다나 푸들 같은 동물을 포함해 많은 동물을 찾아냈지만, ‘gorilla(고릴라)’, ‘chimpanzee(침팬지)’ 그리고 원숭이를 의미하는 ‘monkey(몽키)’라는 단어에 대해서는 아무런 결과를 내놓지 못했다. 해당 응용프로그램에서 그 단어들을 검색어에서 차단시켰기 때문이다. 다만 그런 단어를 사용하지 않은 몇 가지 영장류, 예컨대 개코원숭이를 의미하는 ‘baboon(배분)’, 긴팔원숭이를 의미하는 ‘gibbon(기번)’, 마모셋원숭이를 의미하는 ‘marmoset(마모셋)’, 그리고 ‘orangutan(오랑우탄)’에 대해서는 프로그램이 제 기능을 다했다. 결국 이 사진 인식 프로그램을 작동시키는 알고리즘의 세계에서는 같은 영장류라고 하더라도 ‘monkey’라는 레이블에 연결된 원숭이, ‘gorilla’에 연결된 고릴라, 그리고 ‘chimpanzee’에 연결된 침팬지는 아예 존재하지 않는 것이 된다.

153) Jessica Guynn, “Google Photos Labeled Black People ‘Gorillas’”, USA TODAY, 1 July 2015, <https://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465>, 접속일: 2017년 12월 3일.

154) Tom Simonite, “When It Comes to Gorillas, Google Photos Remains Blind”, WIRED, 18 January 2018, <https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind>, 접속일: 2018년 7월 3일.



3. 우편번호 분류에 따른 가격의 차등 적용

미국의 대학에 입학하기 위해 대학수학능력시험(SAT)을 준비하는 고등학생들을 주 고객층으로 하는 세계 최대의 입시교육 기업은 온라인 강의 및 일대일 개인 교육 서비스를 제공한다. 그런데 이 기업은 우편번호(ZIP code)에 따라서 서비스 가격을 다르게 책정하여 지리상 위치에 따라 4개의 차등적 가격을 적용한다. 고급 과정의 경우 최고가는 8,400달러이고 최저가는 6,600달러로 1,800달러의 차이 즉, 1달러 당 1,128원의 환율을 적용할 경우 한화로 약 203만 원의 차이가 발생한다. 그런데 미국 전체 인구 중에 아시아인은 4.9퍼센트(%)를 차지하고 있지만, 공교롭게도 이 기업에서 고가로 서비스를 제공하는 지역에 거주하는 아시아인 비율은 8.49%인 반면, 저가로 서비스를 제공하는 지역의 아시아인 비율은 2.89%이다. 아시아인 밀집 지역의 소비자들은 소득과 상관없이 다른 지역의 거주자에 비해 1.8배 높은 비용을 지불해야 한다. 해당 기업은 지역별 가격의 차등 책정은 자동차 휘발유나 달걀에 관한 지역별 가격 차등 책정과 마찬가지로 시장의 경쟁적 특성에 따른 것일 뿐이라고 주장한다.¹⁵⁵⁾ 비록 의도적인 결과가 아니라고 할지라도 컴퓨터 알고리즘으로 가격을 책정하는 시대에 이러한 결과는 더 일반적으로 보일 것이다.

III. 머신러닝 알고리즘의 편향과 사회 이미지의 학습

1. 고용 알고리즘과 편향의 학습

자동화를 핵심으로 하는 전자상거래 중심의 글로벌 기업 아마존(Amazon)은 10년간 자사에 제출된 이력서를 기반으로 인공지능 채용 알고리즘을 개발하여 구직자가 제출한 이력서를 검토하는 데에 사용했다. 그런데 인공지능 채용 알고리즘은 기존의 이력서를 통해 기술 산업 전반에 걸쳐 있는 남성의 지배적인 지위를 패턴으로 발견함으로써 새로운 지원자의 이력서에 대해서도 같은 패턴을 가진 규칙에 따라 순위를 매겼다. 마치 고객에게 판매한 물건에 대한

155) Julia Angwin · Jeff Larson, “The Tiger Mom Tax: Asians Are Nearly Twice as Likely To...”, ProPublica, 1 September 2015, <https://www.propublica.org/article/asians-nearly-twice-as-likely-to-get-higher-price-from-princeton-review>, 접속일: 2017년 12월 3일.



만족도를 확인하기 위해 별 한 개부터 별 다섯 개를 부여하듯이 구직자에게 다섯 등급의 평점을 부과했다. 이때 기존 지원자의 남성 지배적 패턴이 적용된 채용 알고리즘은 예컨대 ‘여성 체스 클럽 주장(women’s chess club captain)’처럼 ‘여성’이라는 단어가 사용된 이력서에 대해 편향된 별점을 부과함으로써 기존의 편향을 반복·유지하는 결과를 만들었다. 아마존을 비롯해 구글이나 페이스북(Facebook), 애플(Apple), 마이크로소프트(Microsoft) 같은 첨단기술 기업의 경우 직원 중에 남성이 차지하는 비율은 높은 편이다. 페이스북 같은 경우 전체 직원 중에 남성이 차지하는 비율은 64%, 여성이 차지하는 비율은 36%이지만, 기술 부문에서는 남성이 78%, 여성이 22%를 차지하여 기술 부문에서 남성 편향이 더 커진다. 마이크로소프트의 경우 전체 직원 중에 남성은 74%, 여성은 26%의 비율을 차지하고, 기술 부문만을 놓고 볼 경우 남성이 81%, 여성이 19%로 전체적인 성별 편향이 심하다.¹⁵⁶⁾ 소프트웨어 개발자 같은 기술 산업 종사자 중에 남성이 지배적인 위치에 있다고 해서 개발되는 기술까지 반드시 남성 우호적일 것이라고 단정할 수는 없다. 그러나 새롭게 기술 산업에 진입하려는 지원자를 기존의 남성 편향적 채용 패턴에 입각해 생성된 자동화 모델로 선별하는 것은 적어도 기술 산업 분야의 고용 영역에서 기존의 성별 편향을 재생산하게 된다.

2. 데이터세트의 성별 편향과 모델에 의한 증폭

일정한 행동이나 활동에 대한 기대가 특정성별의 행위자에게 결부되어 있는 경우 일반적으로 성역할이 부여됐다고 볼 수 있다. 이를 다른 방식으로 표현하면 특정한 행동이나 활동이 특정한 성별에 상대적으로 치우쳐져 있는 것 즉, 편향되어 있는 것으로 볼 수 있다. 어떤 행동이나 활동의 내용은 언어 속에서 주로 동사와 목적어로 표현된다. 그렇다면 어떤 행동이나 활동을 표현하는 동사와 목적어가 주어가 되는 행위자의 성별과 얼마나 자주 결합되는지 살펴보면 그 편향의 정도를 확인할 수 있을 것이다. 데이터세트로부터 학습하는 머신러닝 알고리즘의 특성상 그 데이터세트가 어떤 성별 편향을 가지고 있으며, 데이터세트로부터 생성된 알고리즘 모델이 새로운 데이터를 처리할 때 그 편향이 어떻게 변하는지 살펴보는 것은 실증적으로 중요한 의미를 갖는다. 데이터세트 구성에 사용된

156) Jeffrey Dastin, “Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women”, Reuters, 10 October 2018, <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>, 접속일: 2018년 10월 13일.



이미지 속의 어떤 행동이나 활동에서 관련된 동사와 목적어가 여성 행위자(female agent)와 남성 행위자(male agent)에 대해 어느 정도 편향(bias)이 있는지 분석한 연구를 살펴보자.¹⁵⁷⁾ 이러한 연구는 편향을 가시화하고 수량화하는 데에 기여할 수 있다. 여기서 1.0의 수치는 완전한 남성 편향을, 0.0은 완전한 여성 편향을 지시한다.

첫 번째 과제는 이미지로 구성된 데이터셋(imSitu)를 사용해서 두 집단으로 나눈 성별 행위자에 어떤 동사가 연결됐는지 그 비율을 확인하는 가시적인 의미론적 역할 레이블링(visible semantic role labeling, vSRL)이다. 이때 약 125,000개의 이미지로 구성되어 있었던 데이터셋은 인간의 활동과 관련된 약 60,000개의 이미지로 추렸다.¹⁵⁸⁾ 두 번째 과제는 목적어 추적에 공통적으로 사용되는 데이터셋(MS-COCO)에서 어떤 목적어가 역시 둘로 나눈 성별 행위자 집단과 자주 결합하는지 확인하는 다중 레이블 목적어 분류(multilabel object classification, MLC)이다. 이때 데이터셋은 80개의 목적어 유형을 담고 있지만 남녀의 성별 구분을 하지 않고 있는 것이어서 이미지 속에 있는 사람들의 성별을 가리키기 위해 데이터셋의 각 이미지에 사용될 수 있는 이미지 캡션 5개를 사용하고, 인간과 관련성이 적은 목적어 유형도 제거해 66개만을 남겼다. 이 과제를 수행한 결과 동사의 45%와 목적어의 37%가 한쪽 성별에 대해 두 배(2:1)가 넘는 편향을 드러냈다.

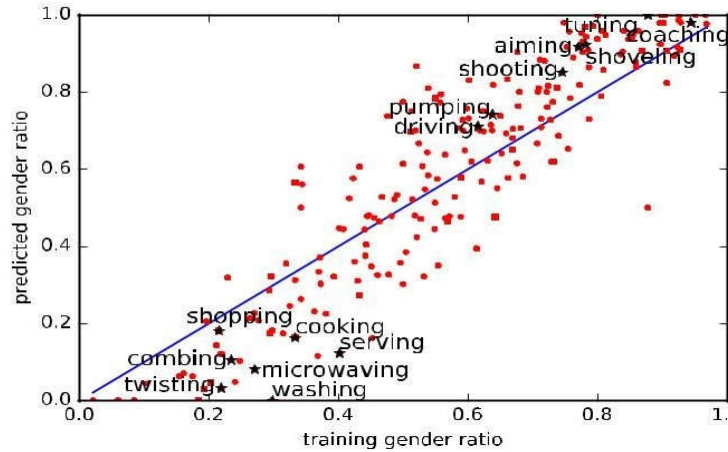
첫 번째 과제에서 동사의 64.6%가 평균 0.707의 편향으로 남성 행위자에게 친화적이고, 동사의 46.95%가 최소한 0.7의 편향을 넘는 수준으로 남성 또는 여성에게 친화적인 것으로 나타났다. 아래의 ‘자료 2’는 문제가 있어 보이는 편향을 드러낸 몇 가지 활동에 대한 레이블을 담고 있다. 장보기를 의미하는 ‘shopping(쇼핑)’, 전자 레인지 돌리기의 ‘microwaving(마이크로웨이빙)’, 요리하기의 ‘cooking(쿠킹)’, 빨래 하기의 ‘washing(워싱)’은 여성에게 편향된 반면, 운전하기의 ‘driving(드라이빙)’, 지도하기의 ‘coaching(코칭)’, 조준하기의 ‘aiming(에이밍)’, 발사하기의 ‘shooting(슈팅)’은 남성 행위자에게 편향되었다.

157) Jieyu Zhao · Tianlu Wang · Mark Yatskar · Vicente Ordonez · Kai-Wei Chang, “Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints”, *Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2979~2989 참조.

158) Jieyu Zhao · Tianlu Wang · Mark Yatskar · Vicente Ordonez · Kai-Wei Chang, “Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints”, *Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2979~2989: p. 2983.

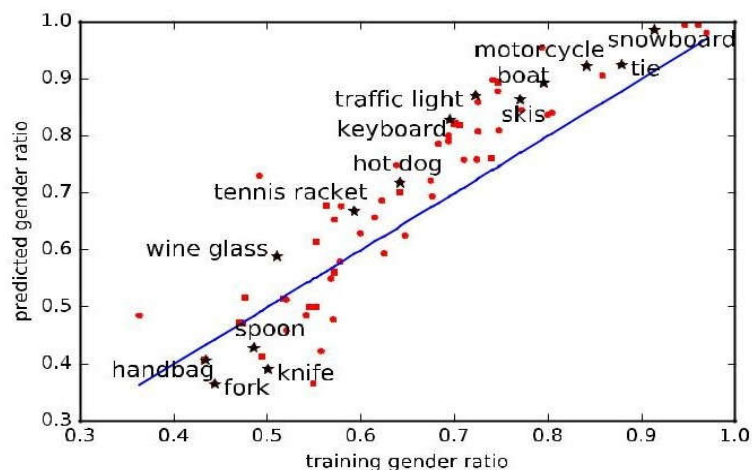


자료 2. imSitu vSRL에 관한 편향 분석



두 번째 과제에서는 첫 번째 과제에서보다 남성 편향의 비율이 더 커서 목적어의 86.6%가 남성 편향적이지만 편향의 정도는 0.65로 동사의 경우보다는 낮게 나타났다. 또한 명사의 37.9%가 최소 0.7의 편향으로 남성 친화적인 것으로 나타났다. 아래의 ‘자료 3’에서 문제가 되는 목적어로 주방에서 사용되는 물건으로서 부엌칼을 의미하는 ‘knife(나이프)’, 숟가락의 ‘spoon(스푼)’, ‘fork(포크)’는 여성에게 편향된 반면, 야외 취미 활동과 관련된 용품인 ‘tennis racket(테니스 라켓)’, ‘snowboard(스노보드)’, ‘motorcycle(모터사이클)’, ‘boat(보트)’는 남성에게 훨씬 더 편향되어 있다.

자료 3. MS-COCO MLC에 관한 편향 분석



특히 이 연구를 통해 드러난 것은 데이터세트에서 나타난 편향이 머신러닝 알고리즘에 단순히 반영되는 것에 그치지 않고 그 편향이 더 증폭됐다는 점이다. 성별 편향이 있는 데이터세트를 학습한 모델은 동사와 목적어에 관한 기존의 편향에 비해 각각 5.0% 포인트와 3.6% 포인트씩 편향이 증폭됐다.¹⁵⁹⁾ 동사의 경우 시중들기를 의미하는 ‘serving(서빙)’은 훈련용 데이터세트에서 0.4로 여성에 대한 편향 정도가 높지 않지만 머신러닝 알고리즘을 거치면서 0.1에 가까워질 정도로 편향이 증폭되었고, 개조하기를 의미하는 ‘tuning(튜닝)’은 0.87정도였던 것이 머신러닝 알고리즘에 의해 처리되면서 1.0에 접근하여 극단적으로 남성 편향이 증폭되었다. 목적어의 경우 ‘knife(나이프)’는 데이터세트에서 0.5였던 것이 알고리즘을 통해 0.4 아래로 내려와 편향의 폭이 증가했고, 기술과 관련된 명사인 ‘keyboard(키보드)’는 0.7에서 0.8로 0.1 포인트만큼 편향의 정도가 증가했다. 이러한 연구 결과는 훈련용 데이터세트가 성별 편향이 있는 경우 이를 기반으로 생성된 머신러닝 알고리즘이 성역할에 대한 편향된 관계를 찾아내서 증폭시킨다는 실증적 사례가 될 수 있다.

IV. 머신러닝 알고리즘을 이용한 실험

1. 번역 실험

터키어에서 중성의 제3자를 가리키는 ‘o’라는 단어를 사용한 “o bir doktor.”와 “o bir hemşire.”라는 문장을 구글 번역기에 입력해 보자.¹⁶⁰⁾ 터키어 ‘doctor’는 의사들, ‘hemşire’는 간호사를 의미한다. 질문을 “o bir doktor.”로 입력한 결과 영어는 “he is a doctor.”(아래 ‘자료 4’ 참조), 독일어는 “er ist Arzt.”, 한국어는 “그는 의사입니다.”로 번역된다.¹⁶¹⁾ 영어의 ‘he’나 독일어의 ‘er’는 남성인 제3자를

159) Jieyu Zhao · Tianlu Wang · Mark Yatskar · Vicente Ordonez · Kai-Wei Chang, “Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints”, *Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2979~2989: p. 2979.

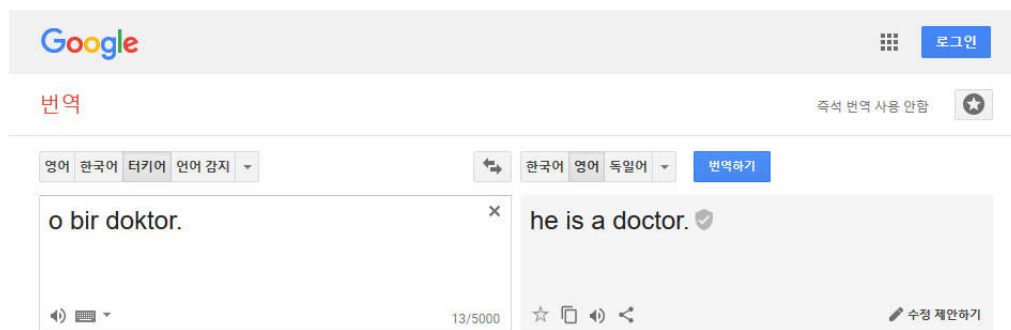
160) 번역 실험은 2017년 12월 3일과 2018년 6월 13일에 동일한 방법으로 수행했고, 결과에는 변함이 없었다.

161) 과학기술학에서 행위자-연결망 이론(actor-network theory)을 제시하는 라투르(B. Latour)는 ‘번역(translation)’을 “자신들의 이익과 그들이 등록한 타자의 이익으로 사실을 구성한 자들로부터 제공된 해석(the interpretation given by the fact-builders of their interests and that of the people they enroll)”으로 이해한다. Bruno Latour, *Science in Action: How to Follow Scientists and Engineers through Society*, 11th print, Harvard University Press, 2003[1st, 1987], p. 108.

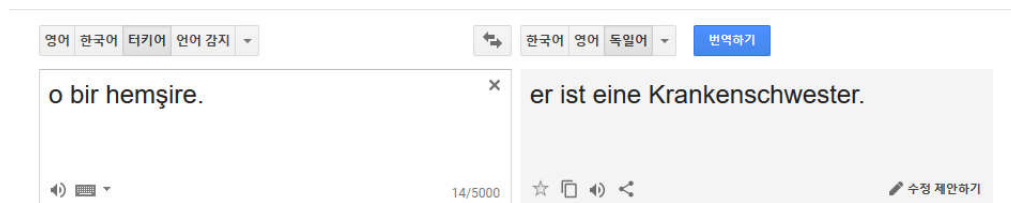


가리킬 때 사용되는 표현이고, 한국어 ‘그’는 남성인 제3자를 가리킬 때 사용되기도 하지만 남녀를 구별하지 않고 사용되기도 한다. 그리고 영어 ‘doctor’, 독일어 ‘Arzt’는 모두 의사를 의미한다. 그런데 질문에서 목적어를 변경하여 “o bir hemşire.”를 입력하면 독일어로는 “er ist eine Krankenschwester.”(아래 ‘자료 5’ 참조), 한국어로는 “그는 간호사입니다.”(아래 ‘자료 6’ 참조)가 출력됐지만, 영어로는 “she is a nurse.”(아래 ‘자료 7’ 참조)가 표시됐다. 입력된 터키어 문장의 목적어가 의사(doctor)에서 간호사(hemşire)로 변경됐을 때 독일어와 한국어에서는 목적어만 간호사를 의미하는 단어 즉, ‘Krankenschwester’와 ‘간호사’로 바뀌고, 주어는 남성인 제3자를 의미하는 ‘er’와 제3자 또는 남성인 제3자를 의미하는 ‘그’가 변경 없이 유지되는 반면, 영어에서는 목적어가 간호사를 의미하는 ‘nurse’로 변경되면서 주어도 여성 제3자인 ‘she’로 변경되었다. 그리고 이러한 변화에 상관없이 문장의 첫 글자에 알파벳 소문자 ‘o’가 입력됐을 때 독일어와 영어 번역 문장 모두 첫 단어가 소문자로 시작하는 결과가 나왔다는 점도 확인할 수 있다.

자료 4. 번역기에 의한 터키어 ‘o bir doktor.’의 영어 번역 결과



자료 5. 번역기에 의한 터키어 ‘o bir hemşire.’의 독일어 번역 결과



자료 6. 번역기에 의한 터키어 ‘o bir hemşire.’의 한국어 번역 결과

자료 7. 번역기에 의한 터키어 ‘o bir hemşire.’의 영어 번역 결과

그런데 질문을 입력하는 과정에서 마침표, 즉 문장 마지막 글자 오른쪽 아래에 입력하는 점 표시('.')를 생략할 경우 출력 언어에 따라 번역 결과가 달라지기도 한다. 즉 “o bir hemşire”로 입력한 경우 영어는 “she is a nurse”(아래 ‘자료 8’ 참조)로 번역 결과가 마침표가 있는 경우(위 ‘자료 7’ 참조)와 동일했지만, 독일어는 “Sie ist eine Krankenschwester”(아래 ‘자료 9’ 참조), 한국어는 “그녀는 간호사”(아래 ‘자료 10’ 참조)로 그 결과가 바뀌었다. 독일어 ‘Sie’와 한국어 ‘그녀’는 여성의 제3자를 가리키는 표현으로서 마침표가 있던 경우(위 ‘자료 5’ 및 ‘자료 6’ 참조)와 비교할 때 마침표 하나의 유무로 의미에 차이를 발생시키는 번역 결과가 나타난 것이다. 앞에서 한국어 ‘그’가 남성 제3자 또는 남녀를 불문한 제3자를 가리킬 때 사용하는 표현이라고 했지만 적어도 이 실험에 사용된 번역엔진의 알고리즘은 ‘그’와 ‘그녀’를 구별하고 있다. 그리고 문장 끝에 마침표를 입력한 경우 앞에서 보았듯이 독일어와 영어는 소문자로 시작하는 ‘er’(위 ‘자료 5’ 참조)와 ‘she’(위 ‘자료 7’ 참조)로 표시됐지만, 마침표가 입력되지 않은 경우 입력 문장의 첫 단어가 소문자임에도 불구하고 독일어의 경우 출력 문장의 첫 단어가 대문자로 시작하는 ‘Sie’로 표시됐다(아래 ‘자료 9’ 참조). 또한 같은 방식으로 마침표를 뺀 터키어 “o bir doktor”를 입력하면 한국어로는 마침표가 추가된 “그는 의사 야.”(아래 ‘자료 11’ 참조)로 표시된다.



자료 8. 번역기에 의한 마침표('.')가 생략된 터키어 'o bir hemşire'의 영어 번역 결과

The screenshot shows a web-based translation tool. On the left, there are language selection buttons: '영어' (English), '한국어' (Korean), '터키어' (Turkish), and '언어 감지' (Detect language). The input box contains 'o bir hemşire'. On the right, the output box shows the translation 'she is a nurse' with a checkmark icon. Below the input box, there is a character count '13/5000'. Below the output box, there are icons for star, copy, audio, and share, along with a '수정 제안하기' (Suggest correction) link.

자료 9. 번역기에 의한 마침표('.')가 생략된 터키어 'o bir hemşire'의 독일어 번역 결과

The screenshot shows the same translation tool with the language selection buttons. The input box contains 'o bir hemşire'. The output box shows the translation 'Sie ist eine Krankenschwester'. The character count '13/5000' is visible below the input box. The same set of icons and '수정 제안하기' link are present below the output box.

자료 10. 번역기에 의한 마침표('.')가 생략된 터키어 'o bir hemşire'의 한국어 번역 결과

The screenshot shows the translation tool with the language selection buttons. The input box contains 'o bir hemşire'. The output box shows the translation '그녀는 간호사'. The character count '13/5000' is visible below the input box. The same set of icons and '수정 제안하기' link are present below the output box.

자료 11. 번역기에 의한 마침표('.')가 생략된 터키어 'o bir doktor'의 영어 번역 결과

The screenshot shows the translation tool with the language selection buttons. The input box contains 'o bir doktor'. The output box shows the translation '그는 의사 야.' (He is a doctor, right?). The character count '12/5000' is visible below the input box. The same set of icons and '수정 제안하기' link are present below the output box.

이번에는 질문을 한국어로 바꾸어 보자. “그 사람은 의사입니다.”와 “그 사람은 간호사입니다.”를 입력하고 영어로 번역되게 하면 각각 “He is a doctor.”(아래 ‘자료 12’ 참조)와 “He is a nurse.”(아래 ‘자료 13’ 참조)로 출력된다. 한국어에서 ‘그 사람’은 중성의 제3자를 가리키는 단어임에도 불구하고 두 경우 모두 영어로는 남성의 제3자를 가리키는 ‘He’로 번역된 것이다. 그리고 대문자와 소문자를 구분하지 않는 한글을 입력했지만, 영어의 출력 결과는 문장을 시작하는 첫 알파벳이 대문자 ‘H’로 표시됐다.



자료 12. 번역기에 의한 한국어 '그 사람은 의사입니다.'의 영어 번역 결과



자료 13. 번역기에 의한 한국어 '그 사람은 간호사입니다.'의 영어 번역 결과



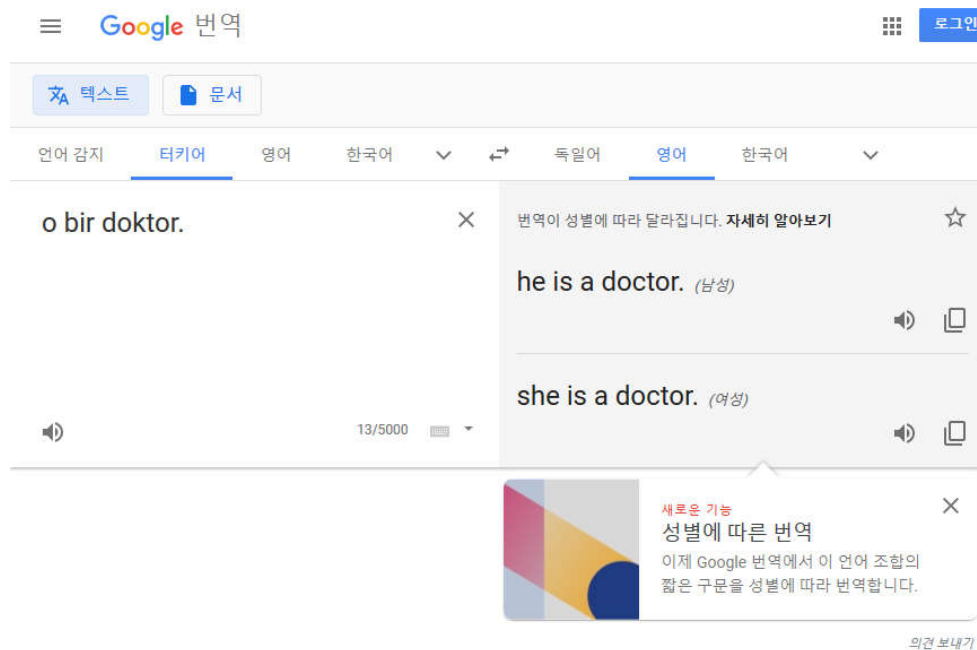
한 문장에서 목적어가 의사를 뜻하는 표현에서 간호사를 뜻하는 표현으로 변경됐을 때 주어에 있던 성 중립적 제3자 표현이 번역되는 언어에 따라 성별을 지시하는 표현으로 변경되는 경우만을 놓고 보면 번역엔진에 사용되는 알고리즘은 특정 직업과 성을 연결시켜 분류한다는 의심을 받을 수 있다. 그러나 문장 마지막에 점 하나가 찍혀 있는지 유무에 따라서도 주어의 성별이 변경되는 것을 보면 번역 엔진의 알고리즘이 특별히 어떤 직업과 성별을 연결시켜 분류하고 있지는 않은 것처럼 보인다.

머신러닝 알고리즘이 산출한 결과가 의미의 차이를 발생시켜 차별의 의심을 불러일으키기도 하지만, 그 의미의 차이가 발생하는 임의성은 머신러닝 알고리즘의 결정에 대해 차별의 혐의를 일방적으로 추정하기 어렵게 한다. 머신러닝 알고리즘이 확률적 추론 기능을 통해 수행하는 번역 작업¹⁶²⁾은 어떤 직업이 특정한 성별과 연결되어 있을 확률이 높다는 경험적 데이터에 기반을 두고 있지만, 번역 알고리즘이 성을 구별하는 언어체계와 성을 구별하지 않는 언어체계를 호환시키는 기능을 학습할 경우 다른 결과가 나올 수 있기 때문이다.

162) 이에 관해서 본 논문 「제2장 제2절 IV. 4. 베이지주의와 베이지 네트워크 기반의 확률적 추론」 참조.

위 실험이 진행된 시점에서 일정 기간(6개월)이 경과한 뒤에 해당 번역엔진은 새로운 기능을 탑재하여 적어도 영어로 번역되는 경우에 한해서는 성별에 따라 달라지는 번역어를 제시하도록 변경되었다.¹⁶³⁾ 제3자인 의사에 관한 터키어 입력문인 위 ‘자료 4’의 “o bir doktor.”와 제3자인 간호사에 관한 터키어 입력문인 위 ‘자료 7’의 “o bir hemşire.”에 대한 영어 번역 결과는 각각 제3자의 남성과 여성을 지칭하는 ‘he’와 ‘she’가 사용된 문장으로 번역되도록 바뀌었다(아래 ‘자료 14’ 참조).¹⁶⁴⁾ 다시 말해 해당 번역엔진에서 사용되던 알고리즘이 새로운 기능을 학습한 다른 알고리즘으로 교체된 것이다.

자료 14. 성별에 따른 번역 기능이 추가된 번역기의 번역 결과 (터키어 → 영어)



163) 구글은 2018년 12월 6일 공식 블로그를 통해 머신러닝에서 공정성 증진 및 편향 감소를 위해 구글의 번역 웹사이트에서 성 중립적 단어에 대한 번역에 대해 남성 지칭용 어휘와 여성 지칭용 어휘가 사용된 결과를 함께 제공할 수 있도록 했다는 소식을 알렸다. 성별에 따른 번역 기능은 예를 들어 외과 의사를 의미하는 영어 단어 ‘surgeon’을 프랑스어로 번역할 경우 동일한 의미의 남성용 어휘인 ‘chirurgien’과 ‘chirurgienne’를 결과로 함께 제공하는 방식으로 작동한다. James Kuczmariski, “Reducing Gender Bias in Google Translate”, Google, 6 December 2018, <https://www.blog.google/products/translate/reducing-gender-bias-google-translate>, 접속일: 2018년 12월 10일.

164) 이전에 번역 실험을 수행한 2018년 6월 13일로부터 6개월이 경과한 2018년 12월 13일에 실행했다.



그러나 성별에 따른 번역어의 구별 기능은 터키어가 영어로 번역되는 경우에만 적용될 뿐, 다른 언어로 번역되는 경우에 대해서는 적용되지 않았다. 예를 들어 ‘간호사’에 관한 “o bir hemşire.”의 독일어 번역은 여전히 위 ‘자료 5’에서처럼 제3자인 남성을 가리키는 ‘er’로 번역됐지만(아래 ‘자료 15’ 참조), 기존에 문장의 첫 글자임에도 소문자인 ‘er’로 번역됐던 것이 대문자 ‘Er’로 번역되어 문장 자체의 정확성은 향상되었다. 같은 터키어 문장의 입력에 대한 한국어 번역은 위 ‘자료 6’처럼 ‘그는 간호사입니다.’로 동일한 결과로 번역되었고(아래 ‘자료 16’ 참조), 똑같이 영어로 번역하는 경우에도 터키어가 아닌 한국어를 대상으로 하는 경우에는 여전히 성별에 따른 번역 기능이 지원되지 않았다(아래 ‘자료 17’ 참조).

자료 15. 성별에 따른 번역 기능을 지원하지 않는 언어 간 번역 결과 (터키어 → 독일어)

언어 감지 **터키어** 영어 한국어 ▼ ↔ **독일어** 영어 한국어 ▼

o bir hemşire. × Er ist eine Krankenschwester. ☆

14/5000 14/5000

의견 보내기

자료 16. 성별에 따른 번역 기능을 지원하지 않는 언어 간 번역 결과 (터키어 → 한국어)

언어 감지 **터키어** 영어 한국어 ▼ ↔ 독일어 영어 **한국어** ▼

o bir hemşire. × 그는 간호사입니다. ☆

geuneun ganhosaibnida.

14/5000 14/5000

의견 보내기

자료 17. 성별에 따른 번역 기능을 지원하지 않는 언어 간 번역 결과 (한국어 → 영어)

언어 감지 터키어 영어 **한국어** ▼ ↔ 독일어 **영어** 한국어 ▼

그 사람은 간호사입니다. × He is a nurse. ☆

geu salam-eun ganhosaibnida.

13/5000 13/5000

의견 보내기



2. 이미지 검색 실험

이미지에 담긴 인간의 활동과 연관된 동사와 목적어의 성별 편향에 관한 앞의 연구¹⁶⁵⁾에 따르면 해당 실험에 사용된 데이터세트에 나타난 인간의 활동 중에 장보기를 의미하는 동사 ‘shopping(쇼핑)’이 여성에게 편향되어 있다. 그렇다면 그런 실험 결과를 토대로 한국어 기반의 인터넷 포털 사이트에 한국어로 ‘장보기’를 입력해 보자.¹⁶⁶⁾ ‘shopping’이 여성에 편향되어 있다고 했으니 아마도 검색된 화면에는 주로 여성의 이미지가 출력될 것이라는 예상해 볼 수 있을 것이다. 그런데 결과는 예상과는 다르게 아래 ‘자료 18’¹⁶⁷⁾의 이미지처럼 전통시장으로 보이는 곳에서 온몸에 어깨띠를 두른 중장년층의 남성들이 여러 명씩 무리지어 등장하는 이미지들로 가득 찼다.

자료 18. 한국어 기반 이미지 검색기에 의한 ‘장보기’ 검색 결과의 예



‘장보기’를 입력하여 출력된 이미지들 속 남성들은 전통시장으로 보이는 곳에서 음식 재료로 가득 찬 진열대를 웃으며 지그시 바라보거나 진열된 물건을 손가락으로

165) 본 논문 「제2장 제3절 III. 2. 데이터세트의 성별 편향과 모델에 의한 증폭」 참조.

166) 네이버 검색엔진(<https://search.naver.com>)에 질의어(query)를 ‘장보기’로 하여 이미지 검색, 검색일: 2018년 7월 3일; 2018년 12월 13일 기준으로 해당 검색엔진은 검색결과 품질 향상을 이유로 이미지 검색결과를 1,000건까지 제공한다.

167) 이미지 주소: http://blogfiles.naver.net/MjAxNzAxMTI1fMTYz/MDAxNDg0NzU5NzY5NjMy.9NpQ3b3-qOlkiiWL-M5pRd5BpMLweRt2dDit31dYq6sg.bbjq3Us3okFQHi6PVsWpjvT0y5ojGKN-ZSxrKW08RzEg.JPEG.dsb1009/17.1.18_%C1%A4%BA%B4%C0%B1_%B0%E6%C1%A6%BA%CE%C1%F6%BB%E7_%B0%E6%BB%EA%BD%C3%C0%E5_%C0%E5%BA%B8%B1%E2%C7%E0%BB%E7_%283%29.jpg; 일정한 기간을 두고 ‘장보기’라는 동일한 질의어를 입력했을 때에도 위 이미지는 1,000건의 검색결과 중에 비교적 상위에 노출되어 있고 그와 유사한 패턴의 이미지들이 검색결과 전반에 분포되어 있지만(검색일: 2018년 12월 18일), 검색결과 개수를 제한하는 조건에서는 동일한 이미지라도 질의어를 입력하는 시점의 검색 알고리즘에 의해 어떻게 평가되느냐에 따라 순위가 내려가면 검색결과에서 배제될 수 있다.



가리켜 본다. 그렇지 않으면 물건을 사이에 두고 건너편에 있는 중년 여성과 악수를 한다.

그렇다면 이번에는 검색어를 한국어 ‘쇼핑’으로 입력해 보자.¹⁶⁸⁾ 출력되는 결과에는 아래 ‘자료 19’¹⁶⁹⁾의 이미지처럼 대리석 바닥이 있는 백화점이나 대형 마트로 보이는 세련된 건물과 주로 젊은 여성 및 남성이 등장한다. ‘쇼핑’에 관한 그 밖의 이미지들 속에는 두 손을 펼쳐 마네킹을 가리키고 있는 쇼핑 호스트, 여러 개의 쇼핑백을 들고 있는 광고 모델, 그리고 ‘쇼핑’이라는 문구가 제목에 들어가는 드라마의 출연 배우들이 나온다. 검색되는 물건들의 이미지도 ‘장보기’의 검색 결과와는 달리 주로 패션이나 가공제품 위주로 나타난다.

자료 19. 한국어 기반 이미지 검색기에 의한 ‘쇼핑’ 검색 결과의 예



만약 시장에 가본 적이 없거나 장보기 또는 쇼핑에 대한 개념이 형성되지 않은 아이에게 이미지 검색을 통해 장보기와 쇼핑이 무엇인지 학습시킨다면 그 아이에게 ‘장보기’는 중장년의 남성이 음식 재료를 펼쳐 놓은 곳에서 진열대를 가리키거나 진열대 위의 물건을 만져보거나 혹은 맞은편에 있는 중년의 여성과 악수하는 것이 될 수 있다. 그리고 ‘쇼핑’은 한껏 멋을 낸 젊은 여성이나 남성이 사람 체형의 플라스틱 위에 여러 가지 물건들을 입혀 놓고 다양한 포즈로 손짓하는 것이거나 화려한 건물 안에서 두 손을 맞잡고 함께 걷는 것이 될 수 있다. 이 아이에게 ‘장보기’와 ‘쇼핑’은 다른 의미를 갖는 단어로 학습된다.

168) 네이버 검색엔진(<https://search.naver.com>)에 질의어(query)를 ‘쇼핑’으로 하여 이미지 검색, 검색일: 2018년 7월 3일.

169) 이미지 주소: http://blogfiles.naver.net/20140723_188/9250sky_1406081848198ofhwP_JPEG/%C7%CF%BF%CD%C0%CC%BC%EE%C7%CE20.jpg.

분명히 상품이나 서비스를 구입할 때 사용하는 ‘shopping’이란 단어가 여성에게 편향되어 있다고 했는데 왜 이런 결과가 나타났을까? 몇 가지 가설을 세워 그 이유를 추론해 보자. 첫째, 입력된 기호의 문제이다. 즉, ‘shopping’, ‘장보기’, ‘쇼핑’은 모두 다른 형태의 문자열을 갖는다. 둘째, 알고리즘 문제이다. 앞의 연구에서 사용된 알고리즘과 한국어 기반의 이미지 검색엔진에 사용된 알고리즘이 다를 수 있다. 입력된 명령어가 같더라도 알고리즘이 다르면 출력 결과는 달라질 수 있다. 마지막 셋째, 입력된 명령어에 따라 다른 결과를 출력하도록 생성된 알고리즘 모델의 훈련용 데이터세트의 편향 문제이다. 일상적으로 장보기 또는 쇼핑을 하는 사람은 그 모습을 사진이나 영상으로 남기지 않는다. 공개적으로 검색되는 이미지는 장보기를 행사로서 홍보거나 쇼핑 채널을 통해 제품을 홍보하려는 사람, 또는 장보기나 쇼핑을 특별한 일로 생각하여 추억을 간직하려는 사람이다. 물론 다른 사람들의 카메라나 곳곳에 설치되어 있는 폐쇄회로 텔레비전(CCTV)에 촬영된 사람일 수도 있다. 머신러닝 알고리즘은 이렇게 인간이 남겨 놓은 흔적들로부터 학습할 수 있고,¹⁷⁰⁾ 그 토대 위에 알고리즘의 세계에서만 통용되는 사회의 이미지를 구축한다.

170) 튜링(A. Turing)은 학습하는 기계의 관념을 구상하면서 어린 아이의 두뇌(child-brain)에 착안하여 “아이 기계(child machine)”라는 용어를 사용한다(Alan M. Turing, “Computing Machinery and Intelligence”, *Mind* 59(236), 1950, pp. 433~460: p. 456). 유사한 은유법을 사용해 모라벡(H. Moravec)은 유전자의 학습이 시작된 1억 년 전에서부터 거슬러 올라와 이제 “우리 마음의 아이들(children of our minds)은 생물학적 진화의 흐름에서 벗어나 더 큰 우주에서의 거대하고 근본적인 도전에 직면하여 성장하기 위해 자유로워질 것”이라고 한다(Hans Moravec, *Mind Children: The Future of Robot and Human Intelligence*, Harvard University Press, 1988, pp. 1~2).



제3장

머신러닝 알고리즘의 작동과 차별의 인식

머신러닝 알고리즘이 데이터를 처리한 결과에 대해 차별의 의심을 갖게 되면서 그러한 결과를 도출하게 된 머신러닝 알고리즘 내부의 작동원리(mechanism)가 차별과 일정한 관련이 있는 것은 아닌지 의구심을 갖게 된다. 그래서 이 장에서는 먼저 차별의 문제와 연결시킬 수 있는 머신러닝 알고리즘이 어떻게 알고리즘을 생성하는지 그 작동원리를 살펴보고자 한다. 다만 모든 유형의 머신러닝 기법에 대해 검토하는 대신에 머신러닝에서 가장 기본이 되는 지도학습방식을 사용하는 데이터 마이닝의 작동원리를 중심으로 차별과 연결될 수 있는 절차를 제시해 본다.¹⁾

그런데 머신러닝 알고리즘의 결정이 차별의 의심을 받는다는 것은 머신러닝 알고리즘이 프로세스를 진행하여 산출해낸 결과에 대해 차별 개념을 적용하는 것이 가능하다는 것을 전제한다. 또한 머신러닝 알고리즘의 작동원리를 차별과 연결시키기 위해서도 차별에 관한 일정한 개념이 전제될 수밖에 없다. 차별 개념을 규범적 차원의 부당함이나 잘못으로 이해하여 적용할 경우 이러한 규범적 의미의 차별 개념에 포착되는 머신러닝 알고리즘의 결정은 차별적이라고 평가 받을 수 있을 것이다. 그리고 차별적이라는 평가가 법적인 의미로서 불법에 연결된다면 그러한 평가를 받은 머신러닝 알고리즘의 결정은 불법의 영역에 속하게 된다. 그렇기 때문에 머신러닝 알고리즘의 결정을 차별 개념의 구성요소 및 그밖에 다른 조건과 연결시켜 합법 또는 불법의 영역으로 이동시키는 것은 전적으로 차별금지법을 구성하는 범명제와 그 토대가 되는 이론의 구성에 달려 있다고 해도 과언이 아니다.

문제 해결 방법이라는 넓은 의미의 알고리즘²⁾을 사용하면 차별금지법은 차별 문제를 해결하는 방법으로서 ‘차별에 관한 알고리즘’이라고 할 수도 있다. 차별의 의심을 받는 머신러닝 알고리즘이 있듯이 차별을 발견하고 제거하는 머신러닝

1) 이에 관한 지도적인 연구로 Solon Barocas · Andrew D. Selbst, “Big Data’s Disparate Impact”, *California Law Review* 104, 2016, pp. 671~732 참조.

2) 인간의 역사에서 오랫동안 사용되어 온 넓은 의미의 알고리즘에 대해서는 본 논문 「제2장 제2절 I. 2. 문제를 해결하는 방법」 참조.



알고리즘도 가능하다는 점³⁾을 고려할 때 차별에 관한 법제도의 작동원리는 그 표준이 될 수도 있다. 하지만 차별금지법을 구성하는 요소들을 일관되게 설명할 수 있는 하나의 기준이 있는지는 여전히 학술적 논쟁의 대상이며,⁴⁾ 그 중심에는 차별 개념 자체의 통일성에 대한 의구심이 자리 잡고 있다.⁵⁾ 차별 개념의 통일성이 확보된다면 차별금지법의 구성이나 해석 및 적용이 단순해질 수 있지만, 차별 개념이 몇 가지 차원을 공유한다면 그러한 작업의 복잡성(complexity)⁶⁾은 높아질 수밖에 없다.

역사적으로 볼 때 차별의 문제는 어떤 특성을 지닌 사람들이 집단으로서 부당한 대우를 받는 것에 대한 반성적 고려에서 출발하며 그로 인한 피해를 구제하고 방지하기 위한 논의가 법제도로 정착되는 과정을 거쳤다.⁷⁾ 그렇기 때문에 차별을 문제화하고 그 해결 방안을 찾는 과정은 차별사유를 중심으로 전개되어 왔다.⁸⁾ 이러한 차별 사유 중심의 차별금지법제는 그 사유의 확장과 제한에 대한 일관된 설명의 요구 앞에 불확정성을 드러낸다.

나아가 새로운 차별사유를 빅데이터로부터 추출하여 비가시적으로 연결할 수 있는 머신러닝 알고리즘은 그 불확정성을 증폭시킨다. 그렇기 때문에 차별을 헌법적 차원의 문제로 설정할 때에도 이러한 불확정성을 야기하는 변수들로 구성되는 법명제의 구조와 내용을 확인하는 것은 중요한 의미를 갖는다. 따라서 머신러닝 알고리즘의 작동원리와 그 작동 결과로 산출된 결정을 포착하는 차별의 개념의 구성 방법과 그러한 차별을 법적 차별로 포착하기 위한 법명제의 의미 체계로서 차별금지법의 구조 및 그와 관련된 이론적 개념도 함께 다루어질 필요가 있다.

3) Indrė Žliobaitė · Bart Custers, “Using Sensitive Personal Data May Be Necessary for Avoiding Discrimination in Data-Driven Decision Models”, *Artificial Intelligence and Law* 24(2), 2016, pp. 183~201: p. 184.

4) Noah D. Zatz, “Disparate Impact and the Unity of Equality Law”, *Boston University Law Review* 97, 2017, pp. 1357~1425 참조.

5) Patrick S. Shin, “Is There a Unitary Concept of Discrimination?”, in *Philosophical Foundations of Discrimination Law*, Deborah Hellman · Sophia Reibetanz Moreau(Eds.), Oxford University Press, 2013, pp. 163~181.

6) 컴퓨터 과학 문헌에서는 ‘complexity’를 비용의 문제와 관련시켜 ‘복잡도’로 번역하기도 한다. 예를 들면 이광근, 컴퓨터 과학이 여는 세계: 세상을 바꾼 컴퓨터, 소프트웨어의 원천 아이디어 그리고 미래, 인사이트, 2015, 100쪽: “알고리즘의 실행 비용(복잡도complexity)에서 우리의 관심은 입력의 크기에 따라 비용이 어떻게 증가하는지다.”

7) 차별 개념 확대의 역사에 대하여 조순경 · 한승희 · 정형욱 · 정경아 · 김선옥, 간접차별의 이론과 여성노동의 현실, 푸른사상, 2007, 15~33쪽 참조.

8) 성, 인종과 민족, 집사와 여행자, 종교, 장애, 성적 지향, 나이 등 차별 사유별 역사적 맥락과 법적 전개에 대한 자세한 내용은 Sandra Fredman, *Discrimination Law*, 2nd ed., Oxford University Press, 2011, pp. 38~108 참조.



제1절 머신러닝 알고리즘의 작동원리

머신러닝 알고리즘에 따라 처리된 결과가 경험적 혹은 규범적 직관에 따라 차별적이라는 의심을 받는 것은 머신러닝 알고리즘이 결정 규칙을 생성하는 방식과 어느 정도 관련성을 맺을 수 있다. 인간의 인식과 판단 작용을 모방한 머신러닝 알고리즘의 작동원리를 살펴보는 것은 인간이 어떤 방식으로 인식하고 판단하는지 역으로 관찰하고 분석하는 것이기도 하다. 기계와 인간이 판단 규칙을 생성하는 과정을 유추적으로 살펴보기 위해서는 모델의 생성에 인간의 판단이 개입하는 지도학습방식의 머신러닝 알고리즘에 대해 살펴보는 것이 적절하다. 데이터 마이닝은 지도학습방식을 사용하는 머신러닝의 대표적 예이다. 데이터 마이닝을 통해 생성된 모델은 관심 있는 실체나 활동을 분류하고, 관찰되지 않는 변수들의 가치를 평가하고, 미래 결과를 예측하는 각각의 프로세스를 자동화할 때 이용된다. 데이터 마이닝의 모델화 과정을 통해 기본적인 머신러닝 알고리즘의 메커니즘을 살펴 보도록 한다.

I. 목표 변수와 클래스 레이블의 정의

1. 목표 변수의 정의와 문제의 형식화

데이터 마이닝으로 생성된 모델이 사용되는 흔한 예는 전자우편의 스팸 분류,⁹⁾ 온라인 거래에서 사기성 조작 탐지, 신용 등급 평가, 보험료 책정 등에서 찾아볼 수 있다. 이때 사례를 분류하고 평가 및 예측하는 결정은 상관관계에 있는 데이터에 접근하는 방식에 따라 달라질 수 있다.¹⁰⁾ 그 접근 방식을 결정하는 것은 데이터 마이닝으로 생성된 모델이 수행할 목표(target)로서 작업의 내용이다. 그래서 예를 들면 스팸, 사기, 채무 불이행, 건강 악화와 같은 어떤 상태나 결과를 목표로 설정하는 것이 필요하다. 그 다음에는 이러한 상태나 결과가 또 다른 어떤 특성이나 행동과 함수 관계에 있는 것으로 취급한다. 그러면 머신러닝 알고리즘을 사례에 노출시키면

9) 이메일 스팸 필터(email spam filters)는 비교적 간단한 방식 때문에 머신러닝을 설명하는 예로 사용된다. 머신러닝의 대표적인 예로 이메일 스팸 필터를 들고 있는 경우로 Harry Surden, "Machine Learning and Law", *Washington Law Review* 89(1), 2014, pp. 87~115: pp. 90~93 참조.

10) Pedro Domingos, "A Few Useful Things to Know About Machine Learning", *Communications of the ACM* 55(10), 2012, pp. 78~87: p. 78.



알고리즘은 어떤 특성이나 행동이 작업의 목표로서 관심 사항이 되는 어떤 상태나 결과를 대신할 수 있는 잠재적 대용물인지 학습한다. 여기서 작업 목표로서 관심 사항이 되는 상태나 결과, 즉 스팸, 사기, 채무 불이행, 건강 악화를 ‘목표 변수(target variables)’라고 한다. 목표 변수는 데이터 마이너(data miner)의 목적, 즉 데이터 마이닝을 실행하는 사람이 얻으려는 것이 무엇인지를 정의한다.

그런데 목표 변수가 명확하지 않은 경우 데이터 마이너는 그것을 정의해야 한다. 목표 변수를 정의하는 것은 일종의 번역이다. 확실한 형태가 없는 문제를 컴퓨터가 분석할 수 있도록 보다 형식적인 용어로 표현될 수 있는 질문으로 전환하는 작업인 것이다. 그러므로 데이터 마이닝의 첫 번째 프로세스에서 데이터 마이너는 프로젝트의 목표와 요구조건 등을 이해하고 그 지식을 데이터 마이닝의 문제 정의로 옮겨야 한다. 그런데 이 작업은 어떤 목표에도 개방되어 있기 때문에 목표 변수를 정의하는 것은 필연적으로 주관적인 프로세스이다. 그러나 문제를 명확히 하는 것이 전적으로 자의적인 프로세스는 아니다. 데이터 마이닝은 목표 변수의 상태 또는 결과에 관한 물음이면서 형식화할 수 있는 문제만을 다룰 수 있기 때문이다.

2. 클래스 레이블의 정의와 카테고리의 생성

목표 변수가 정해지면 목표 변수에는 몇 가지 가능한 값이 할당될 수 있다. 제2장에서 육로를 이용해 P지점에서 Q지점으로 이동하기 위한 방법을 찾을 때 고려할 수 있는 변수로서 이동수단이 있었던 것을 상기해 보자.¹¹⁾ 이때 가능한 이동수단으로 ‘사람의 신체’, ‘휠체어’, ‘유모차’, ‘자전거’, ‘자동차’, ‘택시’, ‘버스’, ‘열차’의 8개를 선별했다. 이러한 선별 과정은 어떤 이동수단을 선택할 것인지 그 결정에 사용될 수 있는 8개의 클래스(class)¹²⁾를 만든 것이기도 하다. 이동수단은 목표 변수이고 8개의 구체적인 이동수단은 목표 변수 중에 선택될 수 있는 가능한 값이 된다. 이처럼 목표 변수의 가능한 모든 값을 상호 배타적인 범주로 분리하기 위한 각 클래스의 특성을 정의한 표시가 ‘클래스 레이블(class labels)’이고 레이블을 지정하는 것을 ‘레이블링(labeling)’이라고 한다. 레이블(label)은 흔히 ‘라벨’이라고 불리기도 하는데 좀 더 직관적인 표현으로는 ‘꼬리표’ 또는 ‘이름표’라고 부를 수도 있다.

11) 본 논문 「제2장 제2절 1. 2. 문제를 해결하는 방법」 참조.

12) 클래스(class)는 보통 계급, 계층, 학급을 의미하지만 데이터 마이닝의 용어로서 클래스는 데이터를 한 군데에 모을 수 있는 상자 같은 개념이다. 더 큰 상자과 더 작은 상자가 있을 수 있으므로 ‘상위 클래스’와 ‘하위 클래스’라는 표현 역시 가능하다. 이는 일상적으로 사용하는 컴퓨터 용어로 ‘폴더’ 개념과도 유사한 측면이 있다.



이동수단 중에 ‘사람의 신체’라고 정의된 레이블이 지정된 클래스가 형성되면 해당 클래스에 할당되는 데이터는 ‘휠체어’라고 정의된 레이블이 지정된 클래스에 할당된 데이터와 분리된다. 이러한 클래스는 하나의 카테고리(category)가 되는 것이다.

3. 목표 변수 정의와 클래스 레이블 정의의 관계

데이터 마이닝은 스팸이나 사기처럼 ‘스팸인 경우와 스팸이 아닌 경우’ 또는 ‘사기인 경우와 사기가 아닌 경우’ 같은 이진 분류법¹³⁾에 따라 두 개의 클래스로 분류되는 목표 변수에 대해 아주 잘 작동한다. 컴퓨터는 이러한 클래스를 구분하는 정의에 따르면 되고, 데이터 마이너는 클래스 레이블을 정의할 때 단순한 기존의 범주들에 의지할 수 있기 때문이다. 그런데 목표 변수를 정의함으로써 새로운 클래스가 형성되는 경우 클래스를 새롭게 정의해야 하므로 문제가 어려워진다. 예를 들면 어떻게 고객에게 신용을 성공적으로 발급할 것인가에 관한 문제 정의에서 신용가치는 문제 정의 그 자체의 가공물이다. 이제는 당연한 것으로 여겨질지 몰라도 ‘신용할 만한 가치가 있다’는 개념을 대출금의 상환 불가능성에서 찾는 것은 신용 산업이 신용을 발급하고 대출금을 상환 받는 시스템을 구축해 온 특정한 방법적 기능 때문이지 신용가치 그 자체를 측정할 방법은 없다. 그저 대출금을 최소한 매월 상환할 수 있는 능력이 신용 획득의 자격 여부를 사전에 그리고 한번에 모두 결정할 수 있는 표준으로서 자의적이지 않다고 여겨질 뿐이다.

결국 목표 변수 및 그와 관련된 클래스 레이블에 대한 정의는 데이터 마이닝이 무엇을 발견할 것인지 결정한다. 여기에는 “종래의 목적과 수단의 관계에 대한 표상을 허구적인 것으로 만들어 버리는 경향”¹⁴⁾이 있는 기술체계의 특징이 함축되어 있다. 다시 말해 신용가치라는 목적을 위해 상환 불가능성이라는 수단을 도입하는 것이 아니라 대출금의 상환 시스템의 운영 방법이 신용가치라는 목적이 되는 것이다. 또한 목표 변수와 클래스 레이블의 정의는 알고리즘의 세계에 무엇이 존재하는지도 결정한다. 알고리즘의 세계에서 레이블이 사라지면 그 존재도 사라지는 것이다. 예를 들어 ‘gorillas’가 레이블로 지정된 클래스가 알고리즘의 검색 조건에서 삭제되면 알고리즘의 세계에 고릴라는 ‘찾을 수 없는’ 것 즉, ‘존재하지 않는’ 것이 된다.¹⁵⁾

13) Peter A. Flach, *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*, Cambridge University Press, 2012, pp. 49~80 참조.

14) 임홍빈, *기술문명과 철학*, 문예출판사, 1995, 77쪽.

15) 고릴라 사건 및 그에 관한 사후 조치에 대해서 본 논문 「제2장 제3절 II. 2. 고릴라 사건」 참조.



II. 훈련용 데이터 구성

데이터 마이닝은 표본(samples)을 통해 학습한다. 이때 표본으로 기능하는 데이터를 ‘훈련용 데이터(training data)’라고 한다. 문자 그대로 모델이 작동하는 방식을 훈련시키는 데이터이다. 훈련용 데이터의 성격은 데이터 마이닝이 학습하는 내용에 의미 있는 결과를 가져올 수 있다. 훈련용 데이터에 편향이 있으면 그 데이터를 기반으로 훈련된 모델에 그 편향이 반영된다는 것이다.¹⁶⁾ 이는 상당히 다른 두 가지 의미로 나타날 수 있다. 첫 번째는 편견이 어떤 역할을 하는 표본을 훈련용 데이터로 삼을 경우 데이터 마이닝으로 생성된 추론 규칙은 표본에서 어떤 역할을 하던 편견을 단순히 재생산하는 것일 수 있다는 것이다. 두 번째는 인구 구성이 편향된 표본으로부터 생성된 추론 규칙인 경우 훈련용 데이터에 과대하게 또는 과소하게 반영된 인구에게 불이익을 줄 수 있다는 것이다. 이 두 가지 의미는 모두 훈련용 데이터에 영향을 미침으로써 어떤 부적절한 결과를 초래할 수 있다는 것인데, 구체적 실현 방식은 다르다. 전자는 표본의 레이블을 정의하는 방식으로 편견을 재생산하는 데 기여할 수 있고, 후자는 편향되게 데이터를 수집하는 방식으로 과대 또는 과소 대표된 인구에게 불이익을 야기할 수 있는 것이므로 양자의 메커니즘은 구별된다.¹⁷⁾

1. 표본에 대한 레이블링과 클래스의 지정

표본에 레이블을 지정하는 것은 훈련용 데이터가 클래스 레이블에 수동적으로 지정되는 프로세스이다. 즉 목표 변수를 정의하고 그에 따라 클래스 레이블을 정의한 상태에서 어떤 표본에 클래스 레이블과 동일한 레이블을 지정하면 그 표본은 동일한 레이블을 가진 클래스에 연결된다. 스팸이나 사기의 경우 이전에 레이블이 지정된 표본들로부터 추출한다. 예를 들어 ‘스팸’이라고 레이블이 정의된 클래스와 스팸이 아니라는 의미의 ‘일반’ 레이블이 정의된 클래스가 있다면, 개별 고객이 어떤 이메일에 대해 ‘스팸’ 표시를 하는 경우 고객들은 훈련용 데이터가 되는 표본에

16) 훈련용 데이터세트의 편향이 단순히 모델에 반영되는 것을 넘어 그 편향이 증폭될 수 있다는 점은 Jieyu Zhao · Tianlu Wang · Mark Yatskar · Vicente Ordonez · Kai-Wei Chang, “Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints”, *Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2979~2989 및 이에 관한 본 논문 「제2장 제3절 III. 2. 데이터세트의 성별 편향과 모델에 의한 증폭」 참조.

17) 이러한 구별은 Solon Barocas · Andrew D. Selbst, “Big Data’s Disparate Impact”, *California Law Review* 104, 2016, pp. 671~732, 특히 pp. 680~687 참조.



레이블을 지정해 서비스 제공업자에게 알려주는 것이다. 그러면 머신러닝 알고리즘은 ‘스팸’으로 레이블이 지정된 메시지 데이터를 분석하여 스팸의 패턴 공식을 발견한다.

그런데 레이블이 지정되지 않은 데이터가 있는 경우 데이터 마이너는 그들 스스로 표본에 레이블을 지정해야 한다. 이 과정은 시간과 노력이 많이 드는데다 위험 요소도 많다. 서로 다른 분류들에 대해 지정할 수 있는 최선의 레이블이 무엇인지는 논쟁의 소지가 있다. 예를 들어 신용카드 4개에 대해 대금 지급을 하지 않는 사람 또는 신용카드 1개의 대금을 4개월간 지급하지 않는 사람을 신용가치 기준의 어느 쪽에 지정할 것인지에 대한 답은 명확하지 않다. 클래스 레이블이 명확한 경우라고 하더라도 특정 사례가 여러 가지 특성을 갖고 있는 경우 그 중에 어떤 특성을 특정 클래스에 포함시키기 위한 기준으로 삼을 것인지, 즉 대표적인 특성으로 선택할 것인지를 문제가 남는다. 반대로 클래스 레이블이 정밀하지 못해서 사례들 사이에서 의미 있는 차이를 포착하기 불충분한 경우에도 마찬가지로 문제가 발생한다. 이렇게 “불완전한 연결 짓기(imperfect matches)”¹⁸⁾는 데이터 마이너가 최종적 판단을 하도록 요구한다. 표본에 주관적으로 레이블을 지정하는 것이 불가피한 경우 이르 기초 사실로 삼아 도출된 추론규칙은 같은 선상에 있는 모든 미래의 사례를 특징지을 것이다. 편견의 형식에 영향을 받은 이전의 결정들이 올바르게 부여된 결정의 표본들로서 기여하는 한 데이터 마이닝은 필연적으로 동일한 편견을 드러내는 규칙을 추론할 수밖에 없다.¹⁹⁾

2. 데이터 수집

훈련용 데이터가 부정확하거나 부분적인 경우 또는 대표성이 없는 경우 이런 데이터로부터 추출된 결정 규칙들에 따라 내린 결론은 특정 집단에 대해 편파적인 이익 또는 불이익을 가져올 수 있다. 특정 인구 집단의 데이터가 불이익을 결정하는 규칙을 추론하는 데이터에 집중됐다면 그렇게 추론된 규칙으로 결정이 이루어질 때 해당 집단은 상대적 불이익을 받을 수 있다. 특히 그 인구 집단에 속한 개인은 자신의 다른 특성과 상관없이 그 집단의 일원으로서 집단에 대한 판단을 공유하게 된다. 이를 통해 인간의 추론 과정에서 쉽게 나타나는 오류인 이른바 ‘확증편향(confirmation bias)’ 같은 현상이 동일하게 발생할 여지가 생긴다. 확증편향은 “기존의

18) Solon Barocas · Andrew D. Selbst, “Big Data’s Disparate Impact”, *California Law Review* 104, 2016, pp. 671~732: p. 681.

19) Solon Barocas · Andrew D. Selbst, “Big Data’s Disparate Impact”, *California Law Review* 104, 2016, pp. 671~732: p. 682.



지배적인 신념, 기대 또는 가설에 치우쳐 증거를 찾거나 해석하는”²⁰⁾ 추론적 오류를 일컫는 심리학적 용어이다. 이는 우범지역으로 선정한 곳에는 그곳이 실제로 범죄가 많이 발생해서라기 보다 더 많은 경찰력이 집중적으로 배치되는 만큼 범죄가 적발 될 확률이 다른 곳보다 더 높아져 상대적으로 범죄율도 증가한다고 보는 것과 같은 이치이다. 그러면 그 지역 주민들의 범죄 가능성은 다른 지역 주민에 비해 더 높아지고, 그 지역에 거주한다는 것은 더 높은 범죄율과 밀접한 관계를 맺게 하는 특성이 된다.

이익을 결정하는 규칙을 추론할 때 반영되지 않은 데이터의 인구 집단은 그 규칙으로 결정이 이루어질 때 상대적으로 불이익을 받을 수 있다. 훈련용 데이터에 반영되지 않은 인구 집단은 규칙 추론 과정에서 애초에 존재하지 않는 집단이 된다. 임의적으로 생략된 것이 아닌 “시스템에 의해 생략된 빅데이터의 주변부에 사는 사람들”²¹⁾인 것이다. 이들은 데이터 마이닝으로도 포착되지 않을 수 있다. 그야말로 체계적으로 생략된 사람들이다. 이들은 가난해서 덜 대표되거나 생활 방식이 달라 덜 대표된다. 데이터를 생성하고 활성화하는 장비를 갖추 경제적 능력이 없는 사람들, 디지털 장비가 익숙하지 않아 디지털 문화에 동화되기 어려운 생활 방식으로 살아가는 사람들, 디지털 문화에 익숙하고 최소한의 장비를 갖추 수는 있더라도 소비력이 낮아 수익을 창출할 수 없는 그래서 상업적인 관심을 받지 못하는 사람들의 그림자 영역이 발생한다.

III. 특성 선택

1. 특성 선택과 데이터의 대표성

조직과 그 조직을 위해 일하는 데이터 마이너는 ‘특성 선택(feature selection)’으로 불리는 프로세스를 통해서 그들이 관찰한 특성을 선택한 뒤에 분석에 끌어넣는다.²²⁾ 그런데 특성 선택을 통해 생성된 데이터세트에 특정 집단이 잘 표현되지 못할 경우 문제가 발생할 수 있다. 관찰한 특성에 대한 표현들은 대비되는 중요한 지점을

20) Raymond S. Nickerson, “Confirmation Bias: A Ubiquitous Phenomenon in Many Guises”, *Review of General Psychology* 2(2), 1998, pp. 175~220: p. 175.

21) Jonas Lerman, “Big Data and Its Exclusions”, *Stanford Law Review Online* 66, 2013, pp. 55~63: p. 57.

22) Ke Wang · Suman Sundaresh, “Selecting Features by Vertical Compactness of Data”, in *Feature Extraction, Construction and Selection: A Data Mining Perspective*, Huan Liu · Hiroshi Motoda(Eds.), Springer US, 1998, pp. 71~84: pp. 71~72.



발견할 수 있을 정도로 세부사항을 포착하는 데에 실패할 수도 있고, 설명 완전하고 효과적인 표현들이 있다고 해도 메커니즘 자체가 데이터에 있는 완전하고 효과적인 표현들에 적합하지 않을 수 있기 때문에 분석의 면밀함을 높이고 그 범위를 확장하는 것이 다른 산출을 설명하는 메커니즘을 포착하는 해답이 되지 못할 수도 있다. 정확하게 결정하기 위한 세부사항은 똑같이 세밀한 수준에 남아 있는데, 분석하려는 주제와 관련된 모든 특성을 수집하거나 주변의 모든 환경 요소를 하나의 모델에 전부 고려하는 것은 종종 불가능한 일로 여겨지기 때문이다.²³⁾

그렇다면 잘 표현되지 못했다는 것은 데이터세트 구성의 정확성이 떨어진다는 것과 동시에 그러한 데이터세트가 특정 집단을 대표하지 못했다는 것을 함축한다. 따라서 특성 선택에서 특정 집단을 구별하는 특성이 포함되지 않은 경우 또는 그러한 집단을 온전히 표현할 수 없는 특성이 선택될 경우 표현되지 않은 집단의 특성을 공유하는 개인은 부정확한 분류나 예측 시스템에 의해 체계적으로(systematically) 은폐되고 복속된다. 또한 데이터는 그 자체가 복잡하고 구체적인 현실 세계의 대상이나 현상을 단순히 감축하여 표현한 재현(representation)의 성격을 갖고 있기 때문에²⁴⁾ 현실 세계의 복잡성과 구체성을 모두 표현해내지 못하는 데이터의 속성은 체계적인 은폐 및 복속을 보다 근본적인 것으로 만든다.

2. 통계적 추론과 특성의 일반화

실제로 문제가 되는 것은 통계적 차별을 허용하는 포괄적인 기준과 다른 집단을 잘못된 결정에 종속되게 하는 불평등한 비율이다. 여기서 중요한 것은 이렇게 잘못되고 잠재적으로 부정적인 산출물이 결정자의 편견이나 데이터세트 구성에 있는 편향이 아니라 “통계적 추론의 인공물(artifacts of statistical reasoning)”²⁵⁾이라는 점이다. 통계적으로는 타당하지만 보편적이지 않은 일반화에 의존하는 결정자는 “합리적이면서 동시에 불공정”²⁶⁾할 수 있다.

23) Toon Calders · Indrè Žliobaitė, “Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures”, in *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, Bart Custers · Toon Calders · Bart Schermer · Tal Z. Zarsky(Eds.), Springer, 2013, pp. 43~57: p. 47.

24) Annamaria Carusi, “Data As Representation: Beyond Anonymity in E-Research Ethics”, *International Journal of Internet Research Ethics* 2008, pp. 37~65: pp. 42~43 참조.

25) Solon Barocas · Andrew D. Selbst, “Big Data’s Disparate Impact”, *California Law Review* 104, 2016, pp. 671~732: p. 688.

26) Frederick Schauer, *Profiles, Probabilities, and Stereotypes*, Belknap Press of Harvard University Press, 2003, p. 3.



보험은 이를 실증적으로 보여주는 좋은 예이다. 보험을 통해 공통의 특정 위험을 가진 집단을 상대로 잠재적 위험이 사고로 발생할 경우를 대비해 비용을 미리 분담한다. 이러한 보험료는 기본적으로 자신의 구체적인 사고 가능성이 아니라 자신을 포함해 같은 사고의 위험이 있는 공통의 특성을 가진 다른 사람들의 통계적 패턴에 의해 결정된다. 여러 개인들은 통계적으로 타당한 추론에 의해 책정된 보험료를 부담하지만 그럼에도 불구하고 그러한 추론이 개인들에게 정확한 것은 아니다. 가령 자동차 보험의 경우라면 운전을 아주 잘 한다는 특성과 운전이 매우 서툴다는 개인들의 특성이 정확하게 반영됐다고 볼 수는 없다.

결국 자동차를 운전한다는 ‘특성의 선택’은 운전을 얼마만큼 잘하는지에 대한 개별적 특성을 특화(特化)시키지 못하고 무화(無化)시킨다. 개별적 특성을 모두 고려할 수 없는 데에는 비용의 문제도 한 몫을 담당한다. 정확하게 구별하려면 정보가 필요하지만 획득된 정보는 항상 불충분하다. 정확하게 구별하기 위해 비용이 든다는 것은 정확성을 개선하기 위한 실천을 주저하게 한다. 그리고 이러한 비용의 측면은 더 세밀하고 포괄적인 분석을 하지 않는 것을 정당화하는 근거로 쉽게 제시된다.

IV. 대용물(proxy)의 사용

1. 집단에 대한 또 다른 특성의 귀속

데이터 마이닝 프로세스에 차별적 효과를 인위적으로 도입하지 않고 결정하는 경우에도 그 결정이 체계적으로 특정 집단의 구성원에게 비우호적인 결과로 이어질 수 있다. 합리적이고 잘 알고 있는 결정을 할 때 실제로 관련을 맺는 기준이 특정 집단의 구성원을 대신하는 믿을 만한 프락시(proxy)²⁷⁾ 즉, 대용물 또는 대체지표로 쓰이게 될 경우 그러한 일이 가능해진다. 예를 들어 직업 수행 능력이 뛰어날 것이라고 예측하여 개인들을 정확하게 분류하는 데 사용된 바로 그 기준이 특정 집단의 구성원 자격에 따라 개인들을 분류한 것과 같은 결과를 낳게 되는 것이다. 이런 결과가

27) ‘프락시(proxy)’는 ‘대용물’ 외에 ‘대체지표’나 ‘대체변수’로 번역되기도 한다. 의미의 핵심은 대용(代用), 대신(代身), 대체(代替), 대표(代表), 대리(代理)한다고 할 때 공통적으로 사용되는 ‘다른 것으로 바꾸다’ 또는 ‘갈음하다’를 뜻하는 ‘대(代)’에 있다. 대용물로 번역하는 경우로 Hellman, Deborah, 차별이란 무엇인가: 차별은 언제 나쁘고 언제 그렇지 않은가[*When Is Discrimination Wrong?*, Harvard University Press, 2008], 김대근(역), 서해문집, 2016 참조; 맥락에 따라 ‘대체지표’ 또는 ‘대체변수’로 번역하는 경우는 고훈수, “인공지능 알고리즘과 시장”, 서울대 법과경제연구센터, 데이터 이코노미, 한스미디어, 2017, 11~38쪽 참조.



발생하는 데에는 명확한 이유가 있는 경우가 있다. 과거의 경험적 데이터로부터 데이터 마이닝을 실행할 경우 현실 속에 고유한 전통적인 편견을 발견할 수도 있지만 마찬가지로 어떤 특정 집단이 업무 수행 능력이나 숙련도가 부족하다는 것을 발견할 수도 있다. 이러한 발견은 불평등이라는 단순한 사실을 드러낼 뿐만 아니라 어떤 특정 집단의 구성원이 자주 상대적으로 불리한 위치에 있는 불평등이 있다는 보다 구체적인 불평등의 상태를 드러낸다.

어떤 정당한 결정을 내리기 위해 사용되는 특성이 특정 집단에서 낮은 비율로 갖고 있는 특성일 경우 그러한 정당한 결정은 해당 특성을 가진 개인들에 대해 체계적으로 비우호적인 결정을 내리게 되는 결과로 이어지게 된다. 예를 들어 고용주가 자신의 사업에 이익을 가져다주는 업무에 더 경쟁력이 있다고 예측하는 직원이나 지원자에게 훨씬 더 많은 관심을 갖고 기회를 제공하지만 그것이 결국 어떤 특정 집단을 지속적으로 불리하게 취급하는 것이라는 점을 발견할 수도 있다.²⁸⁾ 고용주의 관심을 끄는 직원이나 지원자의 특성을 어떤 특정 집단의 구성원들은 낮은 비율로 보유하고 있을 수 있기 때문이다.²⁹⁾ 결정자들이 이런 차별효과를 의도하는 것이 필연적인 것은 아니다. 결정자가 갖고 있는 신념에 따른 것일 수 있기 때문이다. 이익을 추구하는 자로서 합당한 선호는 사회에 존재하게 되는 불평등을 비의도적으로 되풀이하게 한다.

2. 중복 인코딩과 대용물 제거의 한계

결정자의 의도와 상관없이 차별효과가 발생하는 현상은 특정 기준들이 결정의 근거로 사용되지 않도록 데이터베이스에서 제거된 경우 또는 잠복된 편견이나 편향으로부터 데이터가 자유로운 경우, 특성들이 특별히 세밀하고 다양한 경우, 그리고 유일한 목적이 분류 또는 예측의 정확성을 최대화하는 것인 경우에도 발생할 수 있다.³⁰⁾ 예를 들어 소셜 네트워크 서비스(SNS)에서 관심 사항으로 고려할

28) 본 논문 「제2장 제3절 III. 1. 고용 알고리즘과 편향의 학습」에 소개된 인공지능 채용 알고리즘은 해당 기업(아마존)의 과거 채용 기준으로 삼았던 특성이 의도적으로 또는 결과적으로 특정한 성별(여성) 집단에 계속해서 비우호적이었음을 확인시켜 준다. 이에 관해서 Jeffrey Dastin, “Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women”, Reuters, 10 October 2018, <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G> 참조, 접속일: 2018년 10월 13일.

29) Faisal Kamiran · Toon Calders · Mykola Pechenizkiy, “Techniques for Discrimination-Free Predictive Models”, in *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, Bart Custers · Toon Calders · Bart Schermer · Tal Z. Zarsky(Eds.), Springer, 2013, pp. 223-239: pp. 223-224.



특성에 사용되길 원하는 것과 원치 않는 것을 선택할 수 있게 허용하더라도 그 특성은 다른 특성에도 코드로 기입되어 있기 때문에 이용자가 사용을 금지한 특성이 완전히 제거되지 않을 수 있다. ‘성(sex)’이나 ‘관심 있는 남녀(Interested in men/women)’ 같은 비트를 삭제한다고 해도 동성애(homosexuality)라는 성적 취향을 거의 숨기지 못한다는 것이다. 컴퓨터 과학 연구자들은 이러한 위험을 “중복 인코딩 기반의 차별(discrimination based on redundant encoding)”로 부른다.³¹⁾

당면한 결정에 대한 데이터의 중요한 통계적 관련성은 데이터로부터 학습하는 알고리즘이 비록 그것의 유일한 목적이 결정을 위해 가능한 한 최고의 정확성을 보장하는 것인 경우에도 차별적 모델처럼 보이는 이유를 설명해 준다. 특정한 데이터 조각 또는 그 조각에 대한 특정 값이 어떤 집단의 구성원 자격의 특성과 고도의 관련성이 있을 때 알고리즘의 목적과 상관없이 차별적이라는 의심을 받게 된다. 어떤 특성이 불평등하게 분포되어 있다면 더 정교한 형태의 데이터 마이닝은 그런 분포를 포착할 것이고 양질의 데이터와 더 많은 특성은 불평등의 정확한 정도를 보여줄 것이다.

30) Solon Barocas · Andrew D. Selbst, “Big Data’s Disparate Impact”, *California Law Review* 104, 2016, pp. 671~732: p. 691.

31) Cynthia Dwork · Moritz Hardt · Toniann Pitassi · Omer Reingold · Rich Zemel, “Fairness Through Awareness”, 2011, <https://arxiv.org/pdf/1104.3913v2.pdf>, pp. 1~23: p. 19 및 p. 22.



제2절 차별의 형식과 구성

머신러닝 알고리즘이 지각 반응을 통해 내놓은 결과가 차별의 의심을 받는다는 것은 차별이라는 개념을 통해 머신러닝 알고리즘의 결정을 포착하여 분류하는 것이다. 그리고 차별을 사회에 해로운 것으로 보아 사회에서 발생하는 차별을 규제하는 것을 목적으로 하는 법규범은 차별 개념에 그 조건 사실을 포착하는 기능을 부여한다. 그런데 차별에 관한 규범적 논의는 대개 차별의 부당함을 전제하거나 차별이 부당한 이유를 논증하는 데에 초점이 맞추어져 있다. 머신러닝 알고리즘의 차별 문제는 알고리즘 에이전트가 그 알고리즘이 설계될 때 적용된 구조에 입각하여 사물을 인식하고 판단하는 과정에서 발생하는 오류나 편향과 밀접하게 관련되어 있다는 점을 고려할 때 차별 개념은 인간을 비롯해 환경을 지각하는 모든 에이전트에게 잠재적으로 적용될 수 있는 광범위의 개념이 될 수 있다. 그러나 이렇게 넓게 이해된 차별 개념을 법적 개념으로 그대로 수용할 것인지는 별개의 문제이다. 각 시대와 지역적 상황에 대응하여 개발된 차별금지법의 특수성을 고려할 때 여기에서는 차별 개념을 하나의 기준으로 통일되게 정의하려는 시도는 일단 접어 두고 차별 개념의 몇 가지 구성 방식을 살펴보도록 한다.

I. 차별의 전제로서 구별, 분리, 분류

1. 차별의 서술적 표현으로서 구별과 인식

어떤 회사가 직원에게 ‘고객을 차별화하라.’³²⁾고 하면 고객의 요구에 맞는 대응을 하라거나 고객에게 맞춤형 서비스를 제공하라는 의미를 직관적으로 떠올리게 된다. 고객의 요구에 맞게 대응하는 것이나 고객에 대해 맞춤형 서비스를 제공하는 것이 정당한 대우라는 주장과 결합하지 않는 한 고객을 차별화하는 것은 우선적으로 고객을 구별하는 것이다. 구별한다는 것은 차이를 인식(cognition of differences)하는 지능적 작용이기도 하다. 인간의 관점에서 구별은 사람이 태어나 모체로부터 독립된 자기를 인식하는 작용에서부터 이미 시작된다.³³⁾ 신경생물학의 관점을 동원해 보자

32) 기업 경영의 관점에서 고객은 차별화의 대상이다. 홍성준, 고객을 유혹하고 기업을 성장으로 이끄는 차별화의 법칙, 21세기북스, 2013 참조.

33) 영국의 수학자 스펜서 브라운(G. Spencer Brown)은 구별을 “완전한 절제(perfect continence)”로 정의하면서, “구별하려는 욕구(desire to distinguish)” 속에 형식 관념이 자리 잡고 있는 것으로



앞선 단계로 거슬러 올라가보면 모체에서 뇌가 형성된 때부터 ‘자기’와 ‘세계’를 분리하는 인식 작용이 시작됐다고 볼 수도 있다. 그렇다면 인간에게 구별은 끊임없이 세계를 이해하는 과정이며 아울러 자기를 이해하는 과정이기도 하다.³⁴⁾ 한편, 세계의 관점에서 보면 구별은 세계가 자기 자신, 즉 세계 그 자체를 스스로 관찰하기 위해 세계를 분리하는 것이다.³⁵⁾ 이때 관찰은 구별과 지시(indication)가 하나로 작동함으로써 수행되는데, 구별함으로써 생긴 공간의 한 쪽 영역에 대한 지시는 지시된 것과 지시되지 않은 것 사이에 비대칭성을 산출한다. 이처럼 사람 또는 사물의 차이를 인식하고 구별하는 맥락에서 사용된 ‘차별’은 정당성이나 부당성의 평가에 종속되지 않는다는 측면에서 사실적이고, 정당성이나 부당성의 평가를 위해 전제된 상황을 묘사한다는 측면에서 서술적이다.

특히 한국어의 ‘차별하다’로 번역되는 영어식 표현 ‘discriminate’는 X와 Y를 구별한다거나 W로부터 X를 구별한다고 할 때도 사용된다. 이러한 표현은 “우리는 (법체계의 의미에서) 법이라는 용어의 사용 방식을 구별(차별)해야 한다.”³⁶⁾거나 “법관은 법적 안정성을 위해 단지 법률을 적용만 해서는 안 되고 부당한 법률과 정의에 향해 있지 조차 않은 법률을 구별(차별)해야 한다.”³⁷⁾와 같이 어떤 개념이나 의미, 방법 등을 구별하기 위한 이론적 맥락에서 도입되기도 하고, “컴퓨터 과학자는 얼굴 인식 알고리즘이 인간의 얼굴과 다른 사물을 얼마나 성공적으로 잘 구별(차별)하는지 물을 수 있다.”³⁸⁾거나 “체계이론적 인권이론이 기존의 인권이론과 어떤

본다. George Spencer-Brown, *Laws of Form*, US ed., Julian Press, 1972, p. 1 및 p. 69 참조; 스펜서 브라운은 “형식의 법칙”에서 구별, 지시, 시간 단위를 사용하여 존재함의 원리를 수학적으로 계산한다. 이철, “스펜서브라운의 ‘재진입’과 그 과학철학적 의의 - $X^2+1=0$ 에 숨겨진 시간과 상상의 세계”, *사회사상과 문화* 18(2), 2015, 111~137쪽 참조; 스펜서 브라운의 형식 개념에 관한 내용은 현윤경, “니클라스 루만의 체계이론에서 ‘형식’ 개념의 수용과 응용”, *사회와이론* 31, 2017, 211~251쪽 참조.

34) 자아(self) 개념을 어떻게 구성할 것인지, 이를 위해 개인과 공동체의 관계를 어떻게 설정할 것인지는 정치철학의 주요 논쟁거리이다. 이에 관해서 Stephen Mulhal · Adam Swift, *Liberals and Communitarians*, 2nd ed., Blackwell Publishing, 1996 참조.

35) Niklas Luhmann, *Einführung in die Systemtheorie*, 4. Aufl., Carl-Auer, 2008, S. 164.

36) Jeremy Waldron, “The Concept and the Rule of Law”, *Georgia Law Review* 43(1), 2008, pp. 1~61: p. 13: “... we should be ... discriminating about how we use the term *law* (in the sense of legal system).”

37) 힐데브란트(M. Hildebrandt)가 라드브루흐(G. Radbruch)의 이율배반적(antinomian) 법 개념을 설명하는 과정에서 사용한 표현이다. Mireille Hildebrandt, “Radbruch’s Rechtsstaat and Schmitt’s Legal Order: Legalism, Legality, and the Institution of Law”, *Critical Analysis of Law* 2(1), 2015, pp. 42~63: p. 53: “[T]he judge should discriminate between a legal statute that is unjust and one that is not even directed to justice.”

38) Joshua A. Kroll · Joanna Huey · Joel R. Reidenberg · David G. Robinson · Harlan Yu, “Accountable Algorithms”, *University of Pennsylvania Law Review* 165(3), 2017, pp. 633~705: p. 678, fn. 134: “[A] computer scientist may ask ... how well a facial recognition algorithm successfully discriminates



차별성을 가지고 있고, 어떤 새로운 함의를 가지고 있는지³⁹⁾ 또는 “인간의 판단과 인공지능의 판단이 차별성을 가질 수밖에 없지만, 불가피하게 기계적 지능과의 소통이 일상화된 상황을 상정해 본다면, 기존에 인간만이 존재하던 세상에서의 예측 가능성 기준과 인공지능이라는 매개체가 인간과 함께 공존하는 세상에서의 예측 가능성 기준 간에는 차이가 생길 수밖에 없다.”⁴⁰⁾는 식으로 차별은 구별이나 분류 또는 차이의 확인에 대한 가치중립적 동의어로 사용되기도 한다. 또한 표현의 중의성을 이용해 “인종에 기초한 차별(구별)을 멈출 수 있는 방법은 인종에 기초한 구별(차별)을 멈추는 것이다.”⁴¹⁾와 같이 수사적 맥락에서 사용되기도 한다.⁴²⁾

2. 개별화의 수단으로서 분리

고객을 한 명씩 구별하는 것은 최적의 맞춤형 서비스(optimal targeted-service)를 제공할 수 있도록 고객을 개별화(individualization)하는 것이기도 하다. 개별화는 분리하는(divide) 것으로부터 출발한다. 분리를 반복하다보면 분리를 멈추게 되는 순간이 다가오고, 더 이상 분리할 수 없는 상태에 도달했을 때 비로소 개별자(individual)만 남게 된다. 머신러닝 알고리즘의 학습방식 중 기호주의의 가정대로 차이의 인식 과정이 순서대로 진행되는 것이라면⁴³⁾ 덩어리로 인식되던 세계는 인식 작용이 계속될수록 작은 알갱이로 인식되고 결국엔 하나의 입자로 인식된다. 사회의 관점에서 본다면 덩어리는 집단에 상응하고 입자는 개인에 호응한다. 처음에는 한 가지 특성을 기준으로 나눈다. 이때 기준으로 사용되는 그 특성의 유무는 하나의 토대(ground)가 되어 집단(group)을 형성한다. 우선 몇 개의 큰 집단으로 분리하고 여러 가지 특성을 더해가며 더 작은 집단으로 분리를 거듭하다보면 더 이상 분리할

between human faces and inanimate objects.”

39) 홍성수, “인간이 없는 인권이론?—루만의 체계이론과 인권—”, 법철학연구 13(3), 2010, 251~280쪽: 252쪽.

40) 심우민, “인공지능과 법패러다임 변화 가능성: 입법 실무 거버넌스에 대한 영향과 대응 과제를 중심으로”, 법과 사회 56, 2017, 351~385쪽: 355쪽.

41) 2005년부터 미연방법원의 대법원장을 맡고 있는 로버츠(J. Roberts)가 지원자가 몰린 학교에 자리를 배정할 때 인종분류를 사용하는 것에 반대하는 주장을 펼치며 사용한 표현이다. *Parents Involved in Community Sch. v. Seattle School Dist. No. 1*, 551 U. S. 701 (2007), 748: “The way to stop discrimination on the basis of race is to stop discriminating on the basis of race.”

42) 헬먼(Hellman)은 ‘차별(discrimination)’을 그 사용 방식, 즉 서술적(descriptive) 방식과 도덕적(moralized) 방식 때문에 ‘중대한 모호성(important ambiguity)’을 갖는 용어로 본다. Deborah Hellman, *When Is Discrimination Wrong?*, Harvard University Press, 2008, p. 13.

43) 기호주의의 학습 방식의 특징은 순차적(sequential)이라는 데에 있다. Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, New York: Basic Books, a member of the Perseus Books Group, 2015, p. 94; 기호주의에 관해서 본 논문 「제2장 제2절 IV. 1. 기호주의와 논리 및 규칙 기반의 역연역법」 참조.



수 없는(individual) 곳에서 ‘개인(individual)’이 드러나게 된다. 이렇게 이해된 개인은 계속된 분리의 과정에서 사용된 여러 가지 특성이 다양하게 조합된 복합적 성격을 갖는다. 따라서 자기(self)를 이해하거나 해석할 때 한편으로 조합된 상태의 유일성을 강조할 수도 있지만, 다른 한편으로 집단과 연결된 여러 가지 특성의 공통성을 강조할 수도 있다.

3. 집단화의 수단으로서 분류

역설적이지만 생물학의 관점에서 보면 유기체로서 개인은 하나의 세포가 분열을 거듭한 끝에 더 이상 별개의 조직으로 분화할 수 없는 단계에 이른 기관들이 결합된 곳에서 나타나게 된다. 이렇게 나타난 개인이 사회의 기호체계에 편입되기 위해 여러 가지 사회적 의미가 담긴 이름 즉, 레이블(labels)이 붙여진다. 그리고 사회에는 개인이 편입되기 전부터 통용되는 여러 가지 이름의 기호체계가 있다. 개인에게 붙여진 이름은 사회의 기호체계와 연결됨으로써 비로소 기호에 결부되어 있는 의미를 획득하게 된다. 개별화를 위해 사용됐던 분리 기준으로서 그 어떤 특성의 이름은 이미 사회에서 의미의 체계를 형성하고 있는 집단의 이름일 수 있다. 그렇다면 개인에게 어떤 이름이 붙여지는 순간 그 이름과 이에 연결된 개인은 같은 이름을 가진 사회의 집단(class)과 연동됨으로써 분류(classification)된다. 따라서 분류를 통해 개인은 집단화한다.

4. 분리를 통한 분류와 집단의 비대칭성

흑인의 자유와 권리를 위해 투쟁했던 미국의 인권운동가 마틴 루터 킹(Martin Luther King Jr.)이 살해된 다음 날인 1968년 4월 5일 금요일과 그 다음 주 월요일인 4월 8일 이틀 동안 미국 아이오와 주의 한 초등학교에서 이색적인 수업이 진행됐다(이하 ‘학급 분리 실험’).⁴⁴⁾ 3학년 학급(class)의 담임 선생님이던 제인 엘리엇(Jane Elliott)는 담당 학급의 학생들을 먼저 눈동자 색깔에 따라 분류한다. 파란색은 17명, 초록색은 3명, 갈색은 8명이었다. 크게 두 집단으로 나누기 위해 파란색 눈을 가진 학생들과 갈색 눈을 가진 학생들로 분리하면서 초록색 눈을 가진 학생들을 갈색 눈을 가진 학생들의 집단에 포함시켰다. 그리고 주말을 앞둔 금요일에는 파란색 눈동자 집단의 학생들을 열등하게 대우하고, 주말이 지난 월요일에는 갈색 눈동자 집단의 학생들을 열등하게 대우했다.

44) 분리 수업에 대한 학생들의 반응과 실험의 경과에 대한 자세한 내용은 William Peters, *A Class Divided, Then and Now*, Expanded ed., Yale University Press, 1987, 특히 pp. 11~34 참조.



열등하게 대우하는 방식으로 금요일에는 갈색 눈동자를 가진 사람이 파란색 눈동자를 가진 사람보다 깨끗하고, 교양이 있을 뿐만 아니라 더 똑똑하다고 가르친다. 그리고 갈색 눈동자 집단의 학생들에게 쉬는 시간을 5분 더 주고, 식사하러 더 일찍 보낼 뿐만 아니라 식사시간도 더 오래 가질 수 있도록 했다. 그리고 파란색 눈동자 집단의 학생들에게는 작은 놀이기구를 교실 밖으로 가지고 가지 못하게 하고, 집에 돌아갈 때 버스 뒷자리에 앉도록 하는 규칙을 부과했다. 이렇게 분리된 집단에 대해 각각 다른 행동규칙을 부여하고 지키도록 하고, 다음 수업 시간이 있는 월요일에는 그 반대로 갈색 눈동자 집단을 열등한 사람으로 가르치고 이들에게 불리한 규칙들을 지정해 주었다.

파란색 눈동자와 갈색 눈동자를 가진 집단으로 분리되어 수업에 참여한 학생들이 거주하는 지역의 주민은 모두 백인이고 기독교인이라는 공통된 인종 및 종교적 배경을 가졌다. 학급 분리 실험은 차별이 어떻게 발생하고 차별을 하는 사람과 받는 사람에게 어떤 영향을 미치는지 알아보기 위한 경험적이고 실증적인 실험이다. 이러한 실험에 사용된 방법의 중심에는 분리가 있다. 그리고 이러한 분리는 그 기준이 되는 어떤 이유와 연결되어 있다. 이로써 기존의 학급은 분리의 이유가 되는 개별적 특성을 공유하는 학생들만으로 구성된 별개의 집단들로 분류된다. 개별화의 수단인 분리의 기준이 되는 어떤 특성은 집단화의 수단인 분류의 기준이 되기도 하는 것이다. 그리고 각 집단에 대한 서로 다른 대우는 집단 간의 관계를 비대칭적으로 만든다.

II. 차별의 몇 가지 형식

어떤 행위 A가 법적으로 차별의 양상을 갖는다고 해석될 수 있는 몇 가지 대표적인 형식을 살펴보자.⁴⁵⁾ 그러한 예로 첫째 개인의 어떤 특성 때문에 그에 대한 주관적인 반감이 동기가 돼서 한 행위(유형①), 둘째 개인이 어떤 특성을 갖고 있다는 점을 기초로 개인의 자격에 관해 불합당한 추론을 한 다음 그러한 추론에 입각해서 한 행위(유형②), 셋째 어떤 특성을 가진 개인들은 그런 특성을 가졌기 때문에 어떻게 행동해야 한다는 불합당한 신념에 기초한 행위(유형③), 넷째 개인이 어떤 특성을 가졌다는 것으로부터 예측될 수 있는 그 사람의 행동에 관한 통계적으로 ‘합리적인’ 믿음에 기초한 행위(유형④), 다섯째 결과적으로 어떤 특성에 기초해서 중대한 이질적

45) Patrick S. Shin, “Is There a Unitary Concept of Discrimination?”, in *Philosophical Foundations of Discrimination Law*, Deborah Hellman · Sophia Reibetanz Moreau(Eds.), Oxford University Press, 2013, pp. 163~181: pp. 172~179.



효과를 발생시키는 행위(유형⑤), 여섯째 어떤 특성을 갖는 개인들에 대한 적극적 지원의 불이행(유형⑥) 등이 제시된다.

이러한 예는 모든 차별 유형을 망라하는 것이라기보다 차별을 규제하는 여러 국가의 법체계에서 주로 사용되는 대표적 차별 형식이라고 할 수 있다.⁴⁶⁾ 차별에 관한 법 모델을 생성하기 위한 이론의 구성은 이러한 유형을 체계적이고 일관적으로 설명할 수 있는 도식을 찾는 과정이기도 하다. 물론 설명이 체계적이고 일관적인 것과 차별의 규범적 기초가 하나의 원칙으로 환원될 수 있다는 것은 전혀 다른 차원의 문제이다. 오히려 체계적이고 일관적인 차별에 관한 법이론을 구성하려는 시도는 차별금지법의 법철학적 기초의 일원론에 대한 의구심을 불러일으키는 결과로 나타나기도 한다.⁴⁷⁾

1. 편견

‘유형①’은 어떤 특징에 대한 반감이 곧 같은 특징을 가진 집단에 대한 반감으로 비화되어 적대적인 행위로 나타난다. 이때 행위의 상대방, 즉 적대적 행위의 대상이 되는 집단으로 분류되는 특징을 가진 개인들의 이해관계는 오로지 행위자의 특권에 종속될 수밖에 없다. 이런 행위는 “어떤 특징을 가진 사람들을 단지 그런 특징을 가졌다는 이유만으로 다른 사람들보다 도덕적으로 가치가 낮은 것으로 보는 판단”⁴⁸⁾이라고 할 수 있는 편견(prejudice)에 기초한다. 이러한 편견은 “현실을 왜곡하는 렌즈”⁴⁹⁾의 기능을 수행하기도 하는데, 예를 들어 어떤 특징을 가진 사람들에 대한 적대감은 적대감의 대상이 되는 특징을 가진 사람들을 집단화하기도 하지만 그런 특징을 가진 사람들에게 적대감을 공통적으로 갖게 되는 사람들도 집단화함으로써 사안에 따라 자유롭게 다수가 형성되어야 하는 중첩적 이해관계로 구성되어 있는 현실을 왜곡할 수 있다. 또한 편견은 가치판단의 성격을 갖고 있기 때문에 차이에 대한 인식뿐만 아니라 차별에 대한 판단에까지 영향을 미칠 수 있다.⁵⁰⁾

46) 대한민국, 유럽연합(EU), 아메리카합중국(USA), 캐나다, 남아프리카공화국, 인도 등의 차별금지법 제에 관한 소개는 이준일, *차별금지법*, 고려대학교출판부, 2007 및 같은 이, *차별없는 세상과 법*, 홍문사, 2012; 차진아, “독일의 차별금지법 체계와 「일반적 평등대우법」의 역할”, *공법연구* 40(1), 2011, 327~356쪽; Evelyn Ellis · Philippa Watson, *EU Anti-Discrimination Law*, 2nd ed., Oxford University Press, 2012; Sandra Fredman, *Discrimination Law*, 2nd ed., Oxford University Press, 2011; Tarunabh Khaitan, *A Theory of Discrimination Law*, Oxford University Press, 2016 참조.

47) Re'em Segev, “Making Sense of Discrimination”, *Ratio Juris* 27(1), 2014, pp. 47~78 참조.

48) Tarunabh Khaitan, *A Theory of Discrimination Law*, Oxford University Press, 2016, p. 53.

49) John Hart Ely, *Democracy and Distrust: A Theory of Judicial Review*, Harvard University Press, 1980, p. 153.



2. 인식적 고정관념과 규범적 고정관념

‘유형②’와 ‘유형③’은 “어떤 특성을 가진 사람은 또 다른 어떤 특성도 가지고 있다고 판단”⁵¹⁾하는 고정관념(stereotype)에 근거한 행위로 볼 수 있다. ‘유형②’ 즉, 개인이 어떤 특성을 갖고 있다는 점을 기초로 개인의 자격에 관해 불합당한 추론을 한 다음 그러한 추론에 입각해서 한 행위와 ‘유형③’ 즉, 어떤 특성을 가진 개인들은 그런 특성을 가졌기 때문에 어떻게 행동해야 한다는 불합당한 신념을 반영한 행위를 구분한 것은 개인이 갖는 어떤 특성에 연결되는 다른 특성의 성격에 따른 것이다.

‘유형②’는 개인이 갖는 어떤 특성을 또 다른 사실적 특성에 연결시킨다. 앞서 ‘학급 분리 실험’⁵²⁾에서 눈동자가 갈색이라는 특성과 그 사람이 깨끗하거나 교양 있다거나 똑똑하다는 사실적 특성을 연결시키는 것과 같은 방식이다. ‘유형③’은 개인이 갖는 어떤 특성을 또 다른 규범적 특성에 연결시키는 형식을 취한다. 위의 ‘학급 분리 수업’에서 눈동자 색깔이 파랗다는 특성과 버스의 뒷자리에 앉아야 한다는 지시적 규범을 연결시키는 것과 같은 방식이다. 만약 여성이 공격적 성격을 갖는 것이 부적절하다고 판단하여 공격적 성격이 있는 여성 직원의 승진을 거부하는 경우도 같은 형식의 고정관념이 적용된 것으로 볼 수 있다. 왜냐하면 해당 직원의 여성이라는 특성을 온순해야 한다는 지시적 규범과 연결시킨 판단을 전제하고 있기 때문이다. 이러한 전제의 형성에는 어떤 특성을 가진 집단에 대한 규범적 기대가 작용한다.

이처럼 차별 형식의 두 번째 유형과 세 번째 유형 사이의 차이를 섬세하게 구별하여 전자를 ‘인식적(epistemic) 고정관념’, 후자를 ‘지시적(prescriptive) 고정관념’으로 부르기도 하고,⁵³⁾ 각각 ‘사실이 아닌(false) 고정관념’과 ‘규범적(normative) 고정관념’이라고 구분하기도 한다.⁵⁴⁾ 이러한 고정관념은 어떤 특성을 다른 특성의 대용물(proxy)로 사용하는 ‘대용물의 첫 번째 용법’이기도 하다. 인식적 고정관념은 눈동자 색을 ‘청결하다.’, ‘교양이 있다.’, ‘지능이 높다.’ 같은 사실적 특성의 대용물로 삼았고, 지시적 또는 규범적 고정관념은 ‘눈동자 색’과 ‘성별’을 각각 ‘어느 자리에 앉아야 한다.’와 ‘성격이 온순해야 한다.’는 규범적 특성의 대용물로 사용한 것이다.

50) 이준일, “소수자(Minority)와 평등원칙”, 헌법학연구 8(4), 2002, 219-243쪽: 225-226쪽.

51) Tarunabh Khaitan, *A Theory of Discrimination Law*, Oxford University Press, 2016, p. 53.

52) 본 논문 「제3장 제2절 I. 4. 분리를 통한 분류와 집단의 비대칭성」 참조.

53) Patrick S. Shin, “Is There a Unitary Concept of Discrimination?”, in *Philosophical Foundations of Discrimination Law*, Deborah Hellman · Sophia Reibetanz Moreau(Eds.), Oxford University Press, 2013, pp. 163~181: pp. 173~174.

54) K. Anthony Appiah, “Stereotypes and the Shaping of Identity”, *California Law Review* 88(1), 2000, pp. 41~53: pp. 48~49.



3. 통계적 고정관념과 합리적 차별

‘유형④’는 ‘유형②’ 및 ‘유형③’과 마찬가지로 개인의 어떤 특성이 다른 특성을 대신하는 지표로 사용되는 대용물의 첫 번째 용법에 따른다는 점에서 어떤 특성을 가진 사람은 다른 특성도 가지고 있다고 판단하는 고정관념의 유형으로 볼 수 있다. 다만, 대용물로 사용되는 특성은 다른 특성과 상호관련이 있고, 그러한 관련성이 통계적 증거에 의해 뒷받침됨으로써 합리성을 가질 수 있다는 점에 차이가 있다. 그래서 이런 차별 형식을 ‘합리적(rational) 차별’로 구분하기도 한다.⁵⁵⁾ 그런데 차별의 개념 표지로 ‘비합리성(irrationality)’만을 고려할 경우 합리적인 행위는 차별의 영역에 진입할 수 없는 것이기 때문 ‘합리적 차별’이라는 표현은 그 자체가 형용모순처럼 보이기도 한다.

하지만 이러한 행위 역시 집단화할 수 있는 개인의 어떤 특성을 다른 특성과 연결시킨다는 점에서 고정관념의 개념으로 포착할 수 있다.⁵⁶⁾ 인식적 고정관념이 어떤 특징을 가진 집단의 구성원이라는 점에서 개인의 추론된 사실적 특성에 대한 불합당한(unreasonable) 믿음에 기초하고, 지시적 고정관념이 개인의 행동이 적절한지 또는 그 행동을 받아들이는 것인지 판단할 때 어떤 특성을 가진 집단의 구성원이기 때문에 그 개인에게 적용하는 행위 규범에 관한 불합당한 믿음에 기초하는 것이라면, 이러한 통계적 고정관념은 어떤 특성과 다른 특성의 상호관련성에 대한 통계적으로 합당한(reasonable) 또는 합리적인(rational) 믿음에 기초한 것이라는 점에서 차이가 있다. 물론 통계적 관련성이 약한 경우에도 보다 강한 통계적 관련성이 있는 새로운 특성을 대용물로 찾는데 비용이 많이 든다면 약한 통계적 관련성만으로도 기존의 대용물을 인식의 준거로 사용할 수 있고 이러한 행위 양식은 종종 비난을 피할 수 있다.

그럼에도 불구하고 합리적 행위를 차별의 형식으로 다룬다는 것은 어떤 특성은 비록 그것이 통계적으로 유효한 대용물이더라도 이를 사용하는 것이 정의의 맥락에서 부정의의 조건을 영속화하는 경향을 만들 수 있다고 보거나, 그런 특성을

55) Samuel R. Bagenstos, “‘Rational Discrimination’, Accommodation and the Politics of (Disability) Civil Rights”, *Virginia Law Review* 89(5), 2003, pp. 825~923; Deborah Hellman, *When Is Discrimination Wrong?*, Harvard University Press, 2008, pp. 114~137.

56) 카이탄(T. Khaitan)은 고정관념이 부당한 것뿐만 아니라 정당한 것도 포함하는 것으로 보고, 아피아(A. Appiah)는 사실이 아닌 고정관념과 규범적 고정관념 외에 통계적(statistical) 고정관념을 염두에 둔다. Tarunabh Khaitan, *A Theory of Discrimination Law*, Oxford University Press, 2016, p. 53 및 K. Anthony Appiah, “Stereotypes and the Shaping of Identity”, *California Law Review* 88(1), 2000, pp. 41~53; pp. 48~49 참조.



사용하는 것이 누군가에게 해가 되는 경우 그것을 대용물로 사용하지 않아도 어차피 결과에서 차이가 없어 같은 효과가 발생한다면 굳이 그러한 사유를 직접적인 대용물로 사용할 필요가 없다고 보는 것이다. 직원의 업무 수행 능력 또는 학생의 학업 능력을 평가하기 위해 어떤 특성 C를 사용하는 것과 다른 특성 F를 사용하는 것이 효과에서 차이가 없지만 C를 사용하는 것이 해당 직원 또는 학생에게 부가적인 해를 입힌다면 굳이 C를 사용할 필요가 없다는 것이다. 여기까지는 차별 형식을 구성하기 위해 어떤 행위 A를 하는 행위자가 밀접한 관련을 맺으며 행위자의 이해 관계가 깊게 반영되어 있다.

4. 차별효과와 간접차별

‘유형⑤’로 접어들면서 차별은 다른 차원의 형식을 갖게 된다. 효과로서 발생한 어떤 결과로부터 출발하기 때문이다. 그래서 이러한 차별의 형식을 차별효과(disparate impact)⁵⁷⁾라고 부르기도 한다. 이 관점의 출발점은 어떤 행위 A를 한 행위자가 아니라 행위의 효과로 발생한 결과의 영향을 받는, 다시 말해 차별을 겪는 피차별자(discriminatee)⁵⁸⁾이다. 차별을 해악의 측면에서 접근한다면 차별의 관점이 가해자에서 피해자로 전환되는 것이다.

아울러 대용물의 용법도 변경된다. 고정관념의 하위 유형으로 묶일 수 있는 세 가지 유형(유형②, 유형③, 유형④)에서 어떤 특성 C는 다른 특성 F의 대용물로 사용되지만 ‘유형⑤’에서 개인들을 집단으로 묶어주는 어떤 특성 C는 피차별자에게 효과를 미치는 행위에서 그 기준으로 직접 등장하지 않는다. 오히려 다른 특성 F가 대용물로서 전면에 등장한다. 다시 말해 다른 특성 F를 기준으로 수행된 어떤 행위 A의 영향을 받는 사람이 어떤 특성 C를 가진 경우 행위의 기준 또는 근거로서 다른 특성 F가 피차별자의 어떤 특성 C의 대용물로 사용된 것이 아닌지 문제 삼는 차원에서 대용물의 지위는 다른 특성 F에게 이전된 것이다. 이것이 ‘대용물(proxy)의 두 번째 용법’이다. 다른 특성 F가 피차별자의 어떤 특성 C의 대용물이라면 어떤 특성 C는 간접적인 경로로 행위자의 어떤 행위 A에 차별의 형식을 부여하게 된다. 이때 피차별자의 어떤 특성 C가 차별 형식에 기여하는 방식 때문에 이와 같은 차별 형식을 간접차별(indirect discrimination)로 부르기도 한다.⁵⁹⁾

57) “불평등 효과”로 번역한 예로 조순경·한승희·정형욱·정경아·김선욱, 간접차별의 이론과 여성노동의 현실, 푸른사상, 2007, 22쪽 참조.

58) 차별하는 자 즉, ‘차별자(discriminator)’에 상응하여 차별의 대상이 되는 자 또는 차별받는 자를 지시하기 위해 사용 ‘피차별자(discriminatee)’를 사용하는 경우로 Shlomi Segall, “What’s So Bad about Discrimination?”, *Utilitas* 24(1), 2012, pp. 82~100 참조.



만약 차별의 관점을 행위자에게로 재전환시킨다면 행위자가 다른 특성 F를 대용물로 도입한 것이 피차별자의 어떤 특성 C를 판단의 기준으로 직접 사용할 때 발생하는 차별의 의심을 회피하기 위한 것인지 여부가 다시금 중요한 문제로 부각된다. 그러나 차별에 대한 관점을 피차별자에게 고정시키면 행위의 이면에 있는 행위자의 동기나 의도는 상대적으로 중요한 문제가 아니다. 차별의 결과, 상태 또는 경험 그 자체가 중요한 문제로 남아 있기 때문이다.

결과로부터 출발하는 것은 결과를 발생시킨 원인을 찾는 것이다. 따라서 차별효과 또는 간접차별에 대한 도덕적 관심은 차별의 효과를 발생시킨 원인의 객관적 정당화 가능성에 두게 된다.⁶⁰⁾ 이때 ‘유형⑤’는 ‘유형④’에 견주어 볼 때 통계적 분석에 따른 증거의 용처도 달라진다. ‘유형④’ 즉, 합리적 차별의 유형에서 통계적 증거는 행위자가 피차별자의 어떤 특성 C를 사용하는 것의 정당성을 입증하기 위한 정당성 주장의 합당한 증거가 됐다면, ‘유형⑤’ 즉, 차별효과 또는 간접차별 유형에서 통계적 증거는 행위자가 다른 특성 F를 사용하는 것이 피차별자의 어떤 특성 C를 사용하는 것에 대한 은폐라는 부당성 주장에 대한 합당한 증거가 된다.

5. 부작위: 적극적 행위의 불이행

‘유형⑥’은 ‘유형⑤’와 몇 가지 형식을 공유하면서 새로운 형식을 더한다. 여섯 번째 유형의 차별 형식은 다른 특성 F를 만족하여 동등한 자격을 갖춘 피차별자의 어떤 특성 C에 상응하는 적극적 조치가 취해지지 않은 부작위로 인해 피차별자가 다른 특성 F를 기준으로 동등한 자격을 갖춘 경쟁자들에 비해 상대적으로 불리한 상태에 있다는 점에서 출발한다. 피차별자의 불리한 상태라는 결과로부터 출발한다는 점에서 행위자의 관점에서 행위로부터 출발하는 편견이나 고정관념에 기초한 차별 유형과 다르고, 차별효과 또는 간접차별 유형과 유사하다. 그러나 차별효과 또는 간접차별과 달리 ‘유형⑥’에서 다른 특성 F는 피차별자의 어떤 특성 C의 대용물로 작용하지 않기 때문에 집단에 귀속되는 연결점으로서 개인의 어떤 특성 C가 차별에 기여하는 방식은 간접적이지 않고, 오히려 직접적이다. 다만 행위의 양식이 작위가 아니라 부작위이기 때문에 새로운 형식이 더해지는 것이다.

59) 간접차별과 직접차별의 구분 방식에 관해서 본 논문 「제4장 제2절 II. 간접차별과 직접차별」 참조.

60) Patrick S. Shin, “Is There a Unitary Concept of Discrimination?”, in *Philosophical Foundations of Discrimination Law*, Deborah Hellman · Sophia Reibetanz Moreau(Eds.), Oxford University Press, 2013, pp. 163~181: p. 177.



이러한 차별 형식에서 문제의 관건은 피차별자의 어떤 특성 C에 상응하는 적극적 조치의 수준에 달려 있다. 적극적 조치의 수준을 결정하는 요인은 두 가지이다. 하나는 적극적 조치의 내용과 관련된 것이고, 다른 하나는 적극적 조치가 포괄하는 범위와 관련된다. 먼저 적극적 조치의 내용은 규범적이거나 사실적이다. 규범적 내용의 조치는 어떤 특성 C를 고려하는 판단 기준을 적용하도록 하는 것이고, 사실적 내용의 조치는 어떤 특성 C를 고려하는 물적 여건을 구비하는 것이다. 또한 다른 차별 형식은 피차별자의 어떤 특성 C를 판단의 기준 또는 근거로 삼는 것이 문제가 됐다면, ‘유형⑥’의 부작위에 의한 차별 형식은 어떤 특성 C를 판단 또는 결정의 사유로 고려하지 않은 것이 문제가 된다.⁶¹⁾ 따라서 피차별자의 어떤 특성 C를 판단의 기준 또는 근거로 삼게 하는 것이 적극적 조치의 내용이 된다. 이때 물질적 측면과 관련해서 피차별자에게 미친 효과를 회복 또는 극복하기 위한 방법이 재정적인 것에 한정되는지 기술적인 것을 포함하는지에 따라서 적극적 조치의 이행 수준이 달라진다. 우선적으로 적극적 조치에 관한 의무자의 재정 여건이 적극적 조치의 수준을 결정하지만, 재정이 충분하더라도 기술 여건이 마련되어 있지 않으면 적극적 조치의 수준은 달라질 수밖에 없다.

다음으로 적극적 조치가 포괄하는 범위는 시간과 비용에 따라 달라질 수 있다. 적극적 조치에 의해 회복 또는 극복할 대상이 현재의 불이익에 한정되는지 과거의 불이익까지 포함하는지에 따라 적극적 조치의 수준은 달라진다. 과거의 불이익을 포함하면 과거의 판단 기준을 사후에 변경함으로써 이미 변경 전 기준에 따라 형성된 현재의 상황의 불안정성을 증가시킬 뿐만 아니라 변경을 위한 재정적 부담도 증가시키기 때문에 비용 부담을 매개로 적극적 행위의 범위가 결정된다.

특히 비용 부담의 문제는 ‘유형⑥’의 차별 형식과 ‘유형④’의 차별 형식 즉, 합리적 차별의 형식 사이에 긴장 관계를 형성시킨다. 행위자가 피차별자에게 다른 특성 F를 만족하여 동등한 자격을 부여하였지만 피차별자의 어떤 특성 C를 고려하지 않은 부작위에 대한 비용 부담을 해야 한다면 이를 회피하는 방법은 어떤 특성 C를 F가 아닌 또 다른 특성 F'의 대용물로 사용하는 것이다. 예를 들어 피차별자의 어떤

61) 정당한 사유 없이 장애를 고려하지 아니하는 기준을 적용한 것에 대해서 장애인을 차별한 행위에 해당한다고 판단한 경우로 국가인권위원회 2011. 9. 27.자 10진정0480200 결정, 결정례집(차별시정분야) 4, 262 참조. 이 결정에서 국가인권위원회는 “신입사원 채용 시 자격요건 중 영어능력시험 점수와 관련하여 청각장애인에 대하여 비장애인과 같은 기준을 적용한 행위는 실제로는 중증의 청각장애인의 장애특성을 고려하지 않음으로써 이들을 불리하게 대우한 행위라고 판단”했다.



특성 C로서 장애(disability) 여부를 고려하지 않고 다른 특성(F)으로서 ‘자격시험 점수’만을 고려해 직원을 채용한 고용주가 실제 업무 수행을 위해 장애 여부를 고려하지 않은 경우를 생각해 볼 수 있다. 이러한 부작위로 다른 합격자에 비해 업무 수행에서 불리한 피차별자의 상태를 교정하기 위해 합당한 배려(reasonable accommodation)⁶²⁾를 위한 비용을 부담하지 않을 수 있는 방법은 직원 채용 단계에서 지원자의 장애 여부를 고려하되 그 특성을 또 다른 특성(F')으로서 ‘정상 업무 수행 능력’의 대용물로 사용하는 것이다. 공정한 채용을 위해 정상 업무 수행 능력의 조건에서 배제했던 ‘장애’라는 특성을 다시 정상적인 능력의 요구 조건으로 고려함으로써 통계적 증거를 기초로 합리적 차별을 수행할 수 있게 되는 것이다. 따라서 ‘유형④’의 차별 형식이 차별의 유형으로 인정되지 않을 경우, ‘유형⑥’의 차별 형식을 극복하기 위한 적극적 행위를 누군가에게 부담시키고 이행할 수 있도록 하는 실천적 의미는 상당히 퇴색된다.

III. 행위 중심의 차별과 결과 중심의 차별

차별에 관한 이론은 행위 또는 의도를 중심에 두고 의무론적으로 구성하거나 결과 또는 피해를 중심에 두고 결과론적으로 구성할 수 있다.⁶³⁾ 행위 또는 결과를 중심에 둔다는 것은 차별의 부당성에 대한 논증에서 행위 또는 결과가 결정적 논거가 될 수 있도록 차별이론이 구성된다는 것이다. 또한 의무론과 결과론의 측면에서 차별이론을 구성해 보는 것은 데이터 마이닝 같은 예측 분석 기술이 야기하는 차별의 현상을 다루는 데 적합한 차별이론이 무엇인지 재고할 수 있는 계기가 된다.

62) ‘reasonable accommodation’은 한국어로 ‘합리적 배려’로 번역하여 사용되고, 영어권에서는 ‘reasonable adjustment’와 혼용되기도 한다. 본 논문에서는 ‘reasonable’을 ‘rational’과 구분하기 위해 ‘reasonable’에 대해서는 ‘합당한’이라는 표현을 사용한다. 대한민국의 장애인차별금지법(장애인차별금지 및 권리구제에 관한 법률) 제4조 제1항 제3호는 “정당한 편의 제공”이란 용어를 사용하는데, 이때 ‘정당한 편의’를 “장애인이 장애가 없는 사람과 동등하게 같은 활동에 참여할 수 있도록 장애인의 성별, 장애의 유형 및 정도, 특성 등을 고려한 편의시설·설비·도구·서비스 등 인적·물적 제반 수단과 조치”(동법 제4조 제2항)를 의미하는 것으로 규정한다. 이와 관련해 ‘정당한 편의 제공’이라는 용어가 ‘합당한 배려’가 의미하는 것보다 장애인 당사자의 주체성과 인권적 관점을 좀 더 잘 반영한 것이라고 평가하는 경우는 장애인법연구회, 장애인차별 금지법 해설서, 나남, 2017, 44쪽 이하 참조.

63) 데이터 마이닝과 같은 예측 분석 기술이 차별이론을 재고하도록 만드는 계기를 제공한다는 점을 의무론과 결과론의 측면에서 검토하는 경우로 Tal Z. Zarsky, “An Analytic Challenge: Discrimination Theory in the Age of Predictive Analytics”, *I/S: A Journal of Law and Policy for the Information Society* 14(1), 2018, pp. 11~35 참조.



1. 차별의 행위 관련 구성과 의무론

서술적 의미의 차별에서 구별, 분리, 분류하는 행위에 초점을 맞추면 그러한 행위를 하는 행위자의 의도도 심리적 사실로서 차별의 한 요소로 고려할 수 있다. 그리고 이러한 행위자의 의도가 차별 행위를 통해 차별 대상의 차이를 인식하는 데 머물 수도 있지만, 구별되거나 분리되거나 분류된 대상의 지위를 높이거나 낮추는 데에 지향되어 있을 수도 있다. 그렇다면 이러한 심리적 사실의 방향성을 기초로 차별 행위를 구성하면 ‘수평적 차별’과 ‘수직적 차별’로 구분할 수 있는 유형이 만들어진다. 이때 구별, 분리, 분류는 특정 조치를 취하는 결정의 사전 단계에서 수행되어 결정의 전제로서 기능한다. 그리고 구별, 분리, 분류의 단계에서는 그 대상의 지위를 격상시키거나 격하시키려는 의도가 없지만 후속 조치에 관한 결정에서 그 대상의 지위에 변동을 주기 위한 의도로 어느 한 쪽에만 이득을 부여하거나 다른 한 쪽에만 부담을 부과하는 방식 또는 동시에 두 가지를 병행하는 방식으로 이익과 불이익을 분배할 수도 있다.⁶⁴⁾ 물론 이러한 후속 조치에 관한 결정을 할 때에도 조치 대상의 지위 변동에 관한 어떤 의도도 없을 수 있다. 차별에 관한 서술적 의미의 복잡성을 낮추기 위해 차별의 행위 관련 구성에서 행위자의 의도나 동기 같은 심리적 사실을 배제하면 차별은 오로지 구별, 분리, 분류 행위와 상대적으로 유리하거나 불리한 대우로만 구성된다.

2. 차별의 효과 관련 구성과 결과론

(1) 정신적 피해

차별을 구별, 분리, 분류의 결과 또는 효과에 초점을 맞추어 구성하면 피차별자가 입은 피해를 차별의 한 요소로 고려할 수 있다. 이때 피해는 구별, 분리, 분류의 대상이 되어 어느 한 집단에 소속되는 것 자체가 갖는 사회의 상징적 의미 때문에 발생하는 정신적인 차원으로 구성할 수 있다. 이 경우에도 그 기준을 피차별자의 인식 여부에 두어 주관적으로 구성할 수도 있지만, 피차별자의 인식 여부와 무관하게 사회적으로 통용되는 의미에 초점을 맞추어 객관적으로 구성할 수도 있다. 예를 들어 피차별자가 어떤 집단으로 구별, 분리, 분류됨으로써 스스로 열등하다거나

64) 차별의 평가 대상을 ‘비교대상에 대한 비교판단에 따른 동등하거나 차별적인 비교대우’로 보는 경우에도 비교판단과 비교대우를 별도로 구분한다. 이준일, “차별, 소수자, 국가인권위원회”, 헌법학연구 18(2), 2012, 177~222쪽: 178~181쪽 참조.



격하됐다고 느끼게 됐는지 여부를 기준으로 삼을 수도 있지만, 자신이 실제로 인식한 것과 상관없이 구별, 분리, 분류에 사용된 특성 자체의 의미가 대상 집단에 대한 비하(demeaning)⁶⁵⁾나 낙인(stigma)⁶⁶⁾을 나타내는 사회의 의미체계와 결부되어 있는지 여부를 기준으로 삼을 수도 있다.⁶⁷⁾

(2) 상대적 불이익과 독립적 불이익

다음으로 생각해볼 수 있는 피해는 구별, 분리, 분류의 대상이 된 것 자체에 있지 않고 이를 토대로 한 후속 조치로 입게 된 불이익을 중심에 두는 것이다. 이때 불이익은 자신이 속하지 않은 특정 집단에만 이득이 부여됨으로써 이익을 얻지 못하거나 다른 집단과 달리 자신이 속한 집단에 부담이 부과됨으로써 불이익을 받는 경우, 그리고 두 가지가 동시에 이루어지는 경우처럼 상대적으로 구성할 수도 있지만, 자신이 속한 집단의 기대에 부응할 수 있는 규범적 기준처럼 독립적으로 구성할 수도 있다. 그리고 그 규범적 기준은 추상적이고 이상적인 수준으로 설정될 수도 있고, 구체적이고 현실적인 수준으로 설정될 수도 있다. 차별에 관한 서술적 의미의 복잡성을 낮추기 위해 차별의 결과 관련 구성에서 구별, 분리, 분류에 관한 상징적 의미로 인해 입은 피해를 배제하면 차별은 오로지 후속 조치로서 특정 집단에 대한 불이익의 결과로만 구성된다.

3. 차별의 행위 관련 구성과 효과 관련 구성의 차이

차별을 효과나 결과의 측면에서 접근하고 그 중심에 피해를 둘 경우 차별의 방향성은 수직적이다. 정신적 피해의 측면을 고려할 때 주관적 기준에 따르든 객관적 기준에 따르든 열등, 격하, 비하는 아래쪽으로 향하는 수직적 방향성을 갖는다. 또한 구별, 분리, 분류에 따른 후속 조치로서 특정 집단에 대한 불이익의 측면을 고려할 때 상대적이든 독립적이든 비교의 기준보다 낮은 수준에 놓인다는 것 역시

65) Deborah Hellman, *When Is Discrimination Wrong?*, Harvard University Press, 2008, pp. 34~58.

66) Iyiola Solanke, *Discrimination as Stigma: A Theory of Anti-Discrimination Law*, Hart Publishing, 2017 참조.

67) 헬먼(Hellman)은 낙인을 결과적 요소라고 보면서 행위 관련 요소로서 비하(demeaning)가 차별을 부당하게 하는 핵심이라고 주장한다. Deborah Hellman, *When Is Discrimination Wrong?*, Harvard University Press, 2008, pp. 26~27. 반면 모로(Moreau)는 이러한 비하가 차별의 부당성 그 자체를 구성하지 않으며 부당함으로부터 야기된 결과로서 부당한 차별의 부수적 효과로 본다. Sophia Moreau, "What Is Discrimination?", *Philosophy & Public Affairs* 38(2), 2010, pp. 143~179: pp. 177~178. 솔란케(Solanke)는 사회의 관계와 구조 속에 있는 낙인을 차별의 핵심으로 보면서 반차별원칙으로 반낙인원칙을 제시한다. 이에 관한 자세한 내용은 Iyiola Solanke, *Discrimination as Stigma: A Theory of Anti-Discrimination Law*, Hart Publishing, 2017, pp. 8~16 및 pp. 29~34 참조.



하향의 수직적 방향성을 갖는다. 앞서 차별의 행위 중심 구성에서 단순한 구별, 분리, 분류는 상하의 방향성을 갖지 않는다는 측면에서 수평적 차별이 개념적으로 성립될 수도 있지만, 그 의도를 고려할 경우 구별, 분리, 분류의 대상에 대해 도덕적 가치를 낮은 것으로 본다는 측면에서 수직적 차별도 개념적으로 성립될 수 있다. 반면에 차별의 결과 중심 구성은 어떤 기준에 미달함으로써 발생하는 피해나 불이익의 속성 때문에 수평적 차별이 개념적으로 성립되지 않는다. 그렇기 때문에 차별의 행위 중심 구성은 ‘분리했지만 동등하게 대우한 경우’도 차별 개념에 포함시킬 수 있는 반면, 차별의 결과 중심 구성은 ‘비하의 의도가 없지만 피해나 불이익을 입은 경우’도 차별 개념에 포함시킬 수 있다.



제3절 차별의 복잡성과 평등

차별은 보통 평등과 “동전의 양면”⁶⁸⁾ 관계에 있는 반대 개념 또는 대칭 개념으로 이해된다.⁶⁹⁾ 대한민국헌법은 “모든 국민은 법 앞에 평등하다. 누구든지 성별·종교 또는 사회적 신분에 의하여 정치적·경제적·사회적·문화적 생활의 모든 영역에 있어서 차별을 받지 아니한다.”(제11조 제1항)고 하여 비록 별개의 문장으로 구성되어 있지만 차별에 관한 언명이 평등에 관한 언명과 같은 조문에 함께 등장한다. 그러나 엄밀히 말하면 평등의 반대 개념은 불평등이다.⁷⁰⁾ 그렇기 때문에 평등에 대한 이해로부터 출발하여 차별에 대한 이해에 종착하거나 차별에 대한 이해로부터 출발하여 평등에 대한 이해에 도달하기 위해서는 불평등과 차별의 관계에 대한 이해가 수반되어야 한다. 불평등과 차별이 형식적 측면이나 내용적 측면에서 어떤 일치와 불일치 속에서 관계를 형성하고 있는지 살펴보는 것은 외관상 평등을 중심으로 하는 차별에 대한 헌법적 이해가 실제 차별금지법의 구성 및 해석에서 어떤 차이와 한계를 드러내는지 세밀하게 관찰하는 것이기도 하다.

I. 차별에 관한 법적 구조

1. 헌법의 평등 및 차별 관련 규정

차별의 대칭 개념으로 이해되는 ‘평등’은 모든 국민의 “법 앞에 평등”을 규정한 제11조 제1항 이외에 혼인과 가족생활의 성립과 유지의 기초로서 “양성의 평등”을 규정한 제36조 제1항과 국회의원 및 대통령의 평등선거를 규정한 제41조 제1항 및 제67조 제1항에 명시적으로 언급되어 있다. 평등과 유사한 의미를 가진 ‘균등’은 전문(前文)에서 “정치·경제·사회·문화의 모든 영역에 있어서 각인의

68) 이준일, “소수자(Minority)와 평등원칙”, 헌법학연구 8(4), 2002, 219-243쪽: 224쪽.

69) 이준일, “차별, 소수자, 국가인권위원회”, 헌법학연구 18(2), 2012, 177-222쪽: 178쪽.

70) 서양에서 사용되는 ‘평등(equality, Gleichheit)’이란 단어는 라틴어의 ‘애크타스(aequitas)’에서 나온 것으로서 여성 명사이다. 동양에서 사용되는 ‘평등(平等)’이란 단어의 ‘평(平)’은 기운이 멈춰 있다가 다시 퍼져(分) 그 기운이 평온하다(一)는 뜻으로 평탄함, 다스림, 고른 상태 등을 의미한다. 또한 ‘등(等)’은 관청(寺)에서 관리가 대쪽(竹)으로 만든 서류를 순서 있게 분류한다는 뜻으로, 등급, 가지런함, 같음, 무리라는 의미를 지니는 것으로 보기도 한다. 이에 관해서 선우현, 평등, 책세상, 2012, 20쪽 및 신기현, “한국의 전통 사상과 평등 인식”, 한국정치학회보 29(2), 1995, 407-430쪽 참조.



기회를 균등히 하고”, “국민생활의 균등한 향상을 기”한다고 하여 국민에 대한 기회 제공 및 생활 향상과 관련해 언급되고, 비슷한 맥락에서 본문의 제116조 제1항에서는 선거운동에 관한 “균등한 기회의 보장”을 규정하고 있다. 그런데 제31조 제1항에서 교육에 관하여 “능력에 따라 균등하게 교육을 받을 권리”를 규정한 것은 ‘능력에 따른’ 차등적 대우를 전제한 것이어서 다른 것에 대한 다른 대우 즉, 상대적 평등 대우를 정당화하는 근거가 될 수 있다. 또한 직접 ‘평등’이나 ‘균등’을 언급하지는 않지만 국민생활의 균등한 향상의 관점에서 모든 국민의 인간다운 생활을 할 권리(제34조 제1항), 연소자의 근로에 대한 특별한 보호(제32조 제5항), 여자의 복지와 권익의 향상(제34조 제3항), 노인과 청소년의 복지향상(제34조 제4항), 신체장애인 및 생활능력이 없는 국민에 대한 보호(제34조 제5항), 모성의 보호(제36조 제2항) 등은 상대적 평등 대우를 정당화하는 근거로 볼 수 있다.

대한민국헌법에서 제11조 제1항 이외에 ‘차별’을 명시적으로 직접 언급하고 있는 경우로는 여성의 노동에 관하여 “여자의 근로는 특별한 보호를 받으며, 고용·임금 및 근로조건에 있어서 부당한 차별을 받지 아니한다.”고 한 제32조 제4항이 있다. 이때 차별에는 제11조 제1항과 달리 ‘부당한’이라는 수식어가 부가되어 있지만 이 조항 역시 차별 그 자체가 무엇인지 구체적으로 정의하고 있지는 않다. 직접 ‘차별’을 언급하고 있지는 않지만 다른 사람보다 우선적으로 유리한 대우를 하는 것이 차별이라면 “국가유공자·상이군경 및 전몰군경의 유가족”에 대해서 “법률이 정하는 바에 의하여 우선적으로 근로의 기회를 부여받는” 것으로 규정한 제32조 제6항은 특정 집단에게만 이익을 제공하는 우선적 대우 즉, 차별을 정당화하는 근거가 된다고 볼 수 있다.

2. 법률의 차별 관련 규정과 ‘평등권 침해의 차별행위’

현행 헌법 시행 이후 평등과 차별에 관한 입법연혁을 살펴보면 고용 분야를 필두로 해서 차별금지사유, 특히 성별을 중심으로 평등실현에 방점을 둔 법률이 우선 제정되었다. 예를 들면 현행 ‘남녀고용평등과 일·가정 양립 지원에 관한 법률’의 모태가 되는 구 ‘남녀고용평등법’은 “헌법의 평등이념에 따라 고용에 있어서 남녀의 평등한 기회 및 대우를 보장하는 한편 모성을 보호하고 직업능력을 개발하여 근로여성의 지위향상과 복지증진에 기여함을 목적”⁷¹⁾으로 1987년에 제정되었고,

71) 남녀고용평등법[법률 제3989호, 1987. 12. 4. 제정] 제1조; 남녀고용평등법은 2007년 ‘남녀고용평등과 일·가정 양립 지원에 관한 법률’로 개정되면서 “대한민국헌법」의 평등이념에 따라 고용에서



현행 ‘양성평등기본법’의 토대가 되는 구 ‘여성발전기본법’은 “정치·경제·사회·문화의 모든 영역에 있어서 남녀평등을 촉진하고 여성의 발전을 도모함을 목적”⁷²⁾으로 1995년에 제정되었다. ‘노동조합 및 노동관계조정법’은 1997년 제정 당시부터 “노동조합의 조합원은 어떠한 경우에도 인종·종교·성별·정당 또는 신분에 의하여 차별대우를 받지 아니한다.”⁷³⁾는 규정을 두었고, 1998년 제정된 ‘과건근로자 보호 등에 관한 법률’은 “균등한 처우”라는 제목 하에 “과건사업주와 사용사업주는 과건근로자가 사용사업주의 사업내의 동일한 업무를 수행하는 동종 근로자와 비교하여 부당하게 차별적 처우를 받지 아니하도록 하여야 한다.”⁷⁴⁾고 규정했다.⁷⁵⁾ 1950년에 제정된 행형법에는 1999년에 비로소 “국적·성별·종교 또는 사회적 신분 등에 의한 수용자의 차별은 금지된다.”⁷⁶⁾는 규정이 신설되었지만, 1961년에 제정된 청원법에서는 이미 “누구든지 청원하였다는 이유로 차별대우를 받거나 불이익을 강요당하지 아니한다.”⁷⁷⁾고 하여 차별대우 및 불이익 강요를 금지하는 규정을 두었다.

남녀의 평등한 기회와 대우를 보장하고 모성 보호와 여성 고용을 촉진하여 남녀고용평등을 실현함과 아울러 근로자의 일과 가정의 양립을 지원함으로써 모든 국민의 삶의 질 향상에 이바지하는 것을 목적으로 한다.”고 하여 현행법[법률 제15109호, 2017. 11. 28. 개정]에 이르고 있다.

- 72) 구 여성발전기본법[법률 제5136호, 1995. 12. 30. 제정] 제1조: “헌법의 남녀평등이념을 구현하기 위한 국가와 지방자치단체의 책무 등에 관한 기본적인 사항을 규정함으로써 정치·경제·사회·문화의 모든 영역에 있어서 남녀평등을 촉진하고 여성의 발전을 도모함을 목적으로 한다.”; 구 여성발전기본법은 2015년 개정을 통해 ‘양성평등기본법’으로 변경되었고, 현행 양성평등기본법 [법률 제15206호, 2017. 12. 12. 개정]의 목적 규정(제1조)은 “이 법은 「대한민국헌법」의 양성평등이념을 실현하기 위한 국가와 지방자치단체의 책무 등에 관한 기본적인 사항을 규정함으로써 정치·경제·사회·문화의 모든 영역에서 양성평등을 실현하는 것을 목적으로 한다.”고 하여 구법의 ‘남녀평등’은 ‘양성평등’으로 바뀌어 있고, ‘여성의 발전’은 삭제되어 있다.
- 73) 구 노동조합 및 노동관계조정법[법률 제5310호, 1997. 3. 13., 제정] 제9조; 해당 조문은 2008년 개정에서 차별금지 사유로 ‘연령, 신체적 조건, 고용형태’를 추가해서 현행법[법률 제12630호, 2014. 5. 20. 개정]에 이르고 있다.
- 74) 구 과건근로자보호등에관한법률[법률 제5512호, 1998.2.20., 제정] 제21조; 현행 과건근로자보호 등에 관한 법률[법률 제14790호, 2017. 4. 18. 개정]제21조에서는 차별적 처우의 금지 및 시정 등에 대해 규정하고 있다.
- 75) 기간제 근로자 및 단시간 근로자에 대한 차별적 처우의 금지 규정을 담은 ‘기간제 및 단시간근로자 보호 등에 관한 법률’은 2006년에 제정되었다. 동 법률 제8조 참조.
- 76) 구 행형법[법률 제6038호, 1999. 12. 28. 개정] 제1조의3; 구 행형법은 2007년에 ‘형의 집행 및 수용자의 처우에 관한 법률’로 개정되었고, 차별금지에 관한 내용은 현행 형의 집행 및 수용자의 처우에 관한 법률[법률 제15259호, 2017. 12. 19. 개정] 제5조에 “수용자는 합리적인 이유 없이 성별, 종교, 장애, 나이, 사회적 신분, 출신지역, 출신국가, 출신민족, 용모 등 신체조건, 병력(病歷), 혼인 여부, 정치적 의견 및 성적(性的) 지향 등을 이유로 차별받지 아니한다.”고 규정되어 있다.
- 77) 구 청원법[법률 제675호, 1961. 8. 7. 제정] 제10조; 현행 청원법[법률 제12922호, 2014. 12. 30. 개정] 제12조에 자구의 변경만 있을 뿐 같은 내용이 규정되어 있다.



정권이 교체된 2000년대를 전후해서 ‘차별’, ‘차별대우’, ‘차별적 처우’, ‘불이익 처우’, ‘불이익행위’ 등을 금지하는 법률 규정을 다양한 영역의 여러 가지 사유에 적용하는 입법이 시행되었다. 그러한 예로 “사용자는 외국인근로자라는 이유로 부당하게 차별하여 처우하여서는 아니 된다.”⁷⁸⁾, “누구든지 유전정보를 이유로 하여 교육·고용·승진·보험 등 사회활동에 있어서 다른 사람을 차별하여서는 아니 된다.”⁷⁹⁾, “정부는 이러닝이라는 이유로 다른 형태의 학습과 차별하여서는 아니 된다.”⁸⁰⁾, “이 법에 의하여 관련자로 인정된 자는 국가·지방자치단체 또는 사용자 등으로부터 민주화운동을 하였다는 이유로 어떠한 차별대우 및 불이익을 받지 아니한다.”⁸¹⁾, “누구든지 이 법에 따른 회생절차·파산절차 또는 개인회생절차 중에 있다는 이유로 정당한 사유 없이 취업의 제한 또는 해고 등 불이익한 처우를 받지 아니한다.”⁸²⁾ 등의 규정을 들 수 있다. 그리고 2007년에는 장애인 당사자의 입장을 반영하여 장애인에 대한 차별을 금지하는 법률이 제정되었고,⁸³⁾ 최근에는 “공공기관등은 입법·사법·행정·교육·사회문화적으로 점자의 사용을 차별하여서는 아니 된다.”⁸⁴⁾는 규정을 두고 있는 점자법도 제정되었다.⁸⁵⁾

78) 외국인근로자의 고용 등에 관한 법률[법률 제14839호, 2017. 7. 26. 개정; 법률 제6967호, 2003. 8. 16. 제정] 제22조.

79) 구 생명윤리 및 안전에 관한 법률[법률 제7150호, 2004. 1. 29. 제정] 제31조 제1항; 해당 조문은 현행 생명윤리 및 안전에 관한 법률[법률 제15188호, 2017. 12. 12. 개정] 제46조 제1항에 규정되어 있다.

80) 구 이러닝(전자학습)산업발전법[법률 제7137호, 2004. 1. 29. 제정] 제3조; 동조는 현행 이러닝(전자학습)산업 발전 및 이러닝 활용 촉진에 관한 법률[법률 제14998호, 2017. 10. 31. 개정]에서도 그대로 유지되고 있다.

81) 구 민주화운동관련자명예회복및보상등에관한법률[법률 제7214호, 2004. 3. 27. 개정] 제5조의6; 해당 조문은 약간의 자구 변경만 있을 뿐 현행 민주화운동 관련자 명예회복 및 보상 등에 관한 법률[법률 제13289호, 2015. 5. 18. 개정]에 같은 내용으로 규정되어 있다.

82) 채무자 회생 및 파산에 관한 법률[법률 제15158호, 2017. 12. 12. 개정; 법률 제7428호, 2005. 3. 31. 제정] 제32조의2; 중전의 회사정리법·화의법·파산법 및 개인채무자회생법을 통합하여 제정된 채무자 회생 및 파산에 관한 법률 제32조의2의 제목은 “차별적 취급의 금지”이다.

83) 장애인차별금지 및 권리구제 등에 관한 법률[법률 제15272호, 2017. 12. 19. 개정; 법률 제8341호, 2007. 4. 10. 제정] 제1조: “이 법은 모든 생활영역에서 장애를 이유로 한 차별을 금지하고 장애를 이유로 차별받은 사람의 권익을 효과적으로 구제함으로써 장애인의 완전한 사회참여와 평등권 실현을 통하여 인간으로서의 존엄과 가치를 구현함을 목적으로 한다.”

84) 점자법[법률 제15168호, 2017. 12. 12. 개정; 법률 제14205호, 2016. 5. 29. 제정] 제4조 제2항; 점자법은 “점자 및 점자문화의 발전과 보전의 기반을 마련하여 시각장애인의 점자사용 권리를 신장하고 삶의 질을 향상시키는 것을 목적”(동법 제1조)으로 한다.

85) 대한민국의 차별금지 법률조항에 대한 자세한 일별은 이준일, 차별없는 세상과 법, 홍문사, 2012, 169~173쪽 참조.



평등에 관한 입법이 차별에 관한 입법으로 구체화된 데에는 무엇보다 2001년 제정된 국가인권위원회법의 역할이 크다고 할 수 있다. 헌법은 차별이 무엇인지 구체적으로 정의하지 않는 상태에서 차별에 관한 국가인권위원회법의 구체적 정의는 다양한 파생 규정을 산출하는 모델이 된다. 국가인권위원회법은 “평등권 침해의 차별행위”(제2조 제3호)를 “합리적인 이유 없이 성별, 종교, 장애, 나이, 사회적 신분, 출신 지역(출생지, 등록기준지, 성년이 되기 전의 주된 거주지 등을 말한다), 출신 국가, 출신 민족, 용모 등 신체 조건, 기혼·미혼·별거·이혼·사별·재혼·사실혼 등 혼인 여부, 임신 또는 출산, 가족 형태 또는 가족 상황, 인종, 피부색, 사상 또는 정치적 의견, 형의 효력이 실효된 전과(前科), 성적(性的) 지향, 학력, 병력(病歷) 등을 이유로 한 다음 각 목의 어느 하나에 해당하는 행위”라고 하면서, 그 구체적인 행위를 “고용(모집, 채용, 교육, 배치, 승진, 임금 및 임금 외의 금품 지급, 자금의 용자, 정년, 퇴직, 해고 등을 포함한다)과 관련하여 특정한 사람을 우대·배제·구별하거나 불리하게 대우하는 행위”(동조 동호 가목), “재화·용역·교통수단·상업시설·토지·주거시설의 공급이나 이용과 관련하여 특정한 사람을 우대·배제·구별하거나 불리하게 대우하는 행위”(동조 동호 나목), “교육시설이나 직업훈련기관에서의 교육·훈련이나 그 이용과 관련하여 특정한 사람을 우대·배제·구별하거나 불리하게 대우하는 행위”(동조 동호 다목), “성희롱[업무, 고용, 그 밖의 관계에서 공공기관(국가기관, 지방자치단체, 「초·중등교육법」 제2조, 「고등교육법」 제2조와 그 밖의 다른 법률에 따라 설치된 각급 학교, 「공직자윤리법」 제3조의2 제1항에 따른 공직유관단체를 말한다)의 종사자, 사용자 또는 근로자가 그 직위를 이용하여 또는 업무 등과 관련하여 성적 언동 등으로 성적 굴욕감 또는 혐오감을 느끼게 하거나 성적 언동 또는 그 밖의 요구 등에 따르지 아니한다는 이유로 고용상의 불이익을 주는 것을 말한다] 행위”(동조 동호 라목)로 규정한다.⁸⁶⁾

3. 헌법과 법률의 차별 관련 규정과 그 구조적 유사성

국가인권위원회법의 ‘평등권 침해의 차별행위’에 관한 정의 규정을 범명제의 구조를 중심으로 재구성해 보면 일단 평등권 침해의 차별행위는 ‘합리적인 이유 없이 ... 을 이유로 특정한 사람을 우대·배제·구별·불리하게 대우하는 행위

86) 차별금지법을 일반적 차별금지법과 개별적 차별금지법으로 나누면서 국가인권위원회법을 일반적 차별금지법으로 보더라도 국가인권위원회법에서 사용하는 차별 개념이 협소하다는 비판은 이준일, 차별없는 세상과 법, 홍문사, 2012, 174~181쪽, 특히 177쪽 이하 참조.



또는 성희롱 행위'이다. 그리고 그 중에 특정한 사람을 우대·배제·구별·불리하게 대우하는 행위는 '고용 관계, 공급이나 이용 관계, 교육·훈련이나 그 이용 관계'라는 특정한 관계를 조건으로 한다. 헌법 제11조의 제2문인 "누구든지 성별·종교 또는 사회적 신분에 의하여 정치적·경제적·사회적·문화적 생활의 모든 영역에 있어서 차별을 받지 아니한다."는 '누구든지 ... 에 의하여 ... 모든 영역에서 차별을 받지 아니한다.'로 재구성할 수 있다. 그리고 이러한 제2문을 "모든 국민은 법 앞에 평등하다."는 동조 제1문과 결합하여 해석하면 '누구든지 ... 에 의하여 ... 모든 영역에서 받은 차별'은 '법 앞에 평등'을 침해하는 행위가 된다. 만약 이러한 행위를 국가인권위원회법으로 규정하면서 '합리적인 이유 없이 ...을 이유로 ... 관계에서 특정한 사람을 우대·배제·구별·불리하게 대우하는 행위 또는 성희롱 행위'로 구체화한 것이라면 구조의 유사성에 기초해서 다음과 같은 해석이 가능하다. 즉, 차별 이유의 범위는 넓어졌고, 적용되는 영역은 특정한 관계로 좁아졌고, 차별 그 자체는 특정한 사람을 우대·배제·구별·불리하게 대우하는 행위로 특정되면서 이러한 행위와는 성격이 다른 성희롱 행위도 포함시키고 있다. 그리고 이러한 조건에 부합하는 행위를 '평등권침해의 차별행위'로 정의한다.

4. '평등권 침해의 차별행위'에 대한 상이한 해석 가능성

'평등권침해의 차별행위'라는 문구는 세 가지 해석 가능성을 남기는데, 첫 번째는 평등권을 침해하지 않는 차별행위가 있을 수 있다는 해석이다. 즉, 차별행위 중에 일정한 조건을 충족한 경우 평등권을 침해하는 차별행위가 된다는 것이다. 이러한 해석의 근거는 '합리적인 이유 없이'라는 조건 때문이다. 합리적인 이유가 있다면 차별행위라도 평등권을 침해하지 않는 경우도 있다고 볼 수 있기 때문이다.

두 번째는 평등권이 아닌 다른 기본권을 침해하는 차별행위가 있을 수 있다는 해석이다.⁸⁷⁾ 첫 번째 해석이 '평등권침해'의 문구 중에 '침해'를 변수로 본 것이라면 두 번째 해석은 '평등권'을 변수로 보는 것이다. 즉, 침해되는 권리가 평등권이 아닌 기본권, 예컨대 자유권이 침해 대상인 차별행위가 있을 수 있다는 것이다.

87) "차별의 개념이 평등의 문제뿐 아니라 자유의 제한까지 포함하는 것"이라고 하여 '평등권침해의 차별행위'는 차별이 평등권을 침해하는 경우뿐만 아니라 자유권을 침해하는 경우에도 발생할 수 있다는 점을 암시하는 것이라고 해석하는 경우로 송석윤, "차별의 개념과 법의 지배", 사회적 차별과 법의지배, 정인섭(편), 박영사, 2004, 3~24쪽: 20쪽 참조.



세 번째는 평등권침해가 본래의 차별행위 개념의 외연을 확대하는 기능을 수행한다는 해석이다. 이러한 해석은 ‘평등권침해’ 자체를 변수로 보는 것이다. 차별행위의 구체적인 목록 중 세 가지 항목에 대해서는 ‘특정한 사람을 우대·배제·구별·불리하게 대우하는 행위’라는 공통의 행위 양식을 부여하면서 별도로 ‘성희롱 행위’를 규정한 것은 전자의 행위 양식이 기본적인 차별행위의 양식인데 평등권을 침해하는 경우는 보다 확장된 행위 양식을 허용하여 차별행위를 확장한다고 보는 것이다.

이와 같이 세 가지 상이한 해석 가능성은 차별에 어떤 헌법적 가치를 연결시키는지에 따라 그 타당성의 정도가 달라질 수 있을 것이다. 다시 말해 차별을 평등과 연결시킬 것인지 아니면 자유 또는 존엄성 등의 헌법적 가치와 연결시키는지에 따라 각 해석의 타당성은 달라질 수 있다. 이에 관해서는 이 절의 마지막에서 살펴보고,⁸⁸⁾ 우선 평등과 차별 또는 불평등과 차별의 관계를 설정하기 위해 평등과 불평등의 형식에 대해 살펴본다.

II. 형식적 평등과 차이의 구별

1. 형식적 평등과 ‘같은 것은 같게’ 알고리즘

‘같은 것을 같게 대우하라(Treat likes alike).’는 언명은 아리스토텔레스⁸⁹⁾까지 거슬러 올라가는 “평등에 관한 생각의 기초를 형성하는”⁹⁰⁾ 형식적 개념이자 평등에 관한 ‘최소한의 기본원리’로 여겨진다.⁹¹⁾ 이러한 평등 개념의 논리적 형식에 따르면 우선 같은 것을 같지 않게 대우하는 것과 같지 않은 것을 같게 대우하는 것, 다시

88) 이에 관해서 본 논문 「제3장 제3절 IV. 차별과 헌법적 가치의 연결」 참조.

89) 아리스토텔레스는 사람은 대개 평등을 추구할 때 파쟁을 일으키는 법이기 때문에 어디서나 불평등이 파쟁의 원인이라고 본다. 그리고 불평등한 자들이 그들 사이에 존재하는 불평등에 비례하는 대우를 받으면 불평등이 아니라고 하면서 평등의 두 가지 종류, 즉 ‘수(數)에 따른 평등’과 ‘가치에 따른 평등’에 대해 설명한다. ‘수에 따른 평등’이란 양(量)이나 크기에서 동일하고 평등한 것을 의미하고, ‘가치에 따른 평등’이란 비례(比例, logos)에서 동등한 것을 의미한다. 2와 3의 차이는 수적으로는 1로서 1과 2의 차이와 같다. 그러나 비례적으로는 2에 대한 4의 관계가 1에 대한 2의 관계와 같다. 이에 관해서 Aristoteles, 정치학[Politika], 천병희(역), 제2판, 숲, 2013, 262쪽: 1301b19 및 1301b26 참조; 수에 따른 평등과 가치에 따른 평등의 관념은 차별의 측정 지표에 그대로 남아 있다. 이에 관해서 본 논문 「제4장 제3절 II. 2. 법적 차별의 측정 지표」 참조.

90) Sandra Fredman, *Discrimination Law*, 2nd ed., Oxford University Press, 2011, p. 8.

91) Evelyn Ellis · Philippa Watson, *EU Anti-Discrimination Law*, 2nd ed., Oxford University Press, 2012, p. 5.



말해 ‘같은 것을 다르게 대우’하는 것과 ‘다른 것을 같게 대우’하는 것은 평등의 형식과 충돌한다. 하지만 ‘다른 것을 다르게 대우’하는 것은 ‘같은 것을 같게 대우’하는 평등의 형식과 충돌하지 않는다. 평등 개념을 ‘같은 것은 같게, 다른 것은 다르게’라고 보면서 이에 대해 ‘상대적 평등’이라는 이름을 붙여 이해하는 것은 ‘같은 것은 같게’로 축약되는 평등 형식과 논리적과 허용 관계에 있는 형식을 평등 개념 안에 포함하는 것이기도 하다.⁹²⁾

그런데 평등에 관한 이론적 설명은 대부분 이 단계에서 멈추기 때문에 짧고 단순한 형식적 알고리즘으로서 ‘같은 것을 같게 대우하라.’는 언명을 실행함으로써 발생하는 결과도 간단할 것이라고 단정 지을 수도 있다. 그러나 반복을 특성으로 하는 알고리즘이 주는 교훈은 매우 단순한 알고리즘이라고 반복 적용되면 그 결과는 매우 복잡해져서 중국에는 제한된 시공간에서 측정할 수 없는 결과를 낳을 수 있다는 것이다. 그렇다면 여기서 멈추지 말고 ‘다른 것’을 대우하는 경우에도 ‘같은 것을 같게’ 대우하는 평등 형식을 그대로 적용해 볼 수 있다. ‘다른 것’ 중에 ‘같은 것’을 ‘같게’ 대우하는 것은 평등 형식에 부합한다. 예를 들어 아리스토텔레스가 살던 시대에 남자와 ‘같은 사람’이 아니었던 ‘다른 것’ 중에 ‘같은 여자’를 ‘같게’ 대우하고 여자와 ‘다른 것’을 ‘다르게’ 대우하는 것은 평등 형식과 충돌하지 않는다는 말이다. 그리고 또 다시 ‘다른 것’ 중에 같은 무엇을 같게 대우하는 방식으로서 ‘같은 것을 같게’는 더 이상 비교할 대상이 없는 최후의 1인이 나올 때까지 반복될 수 있는 알고리즘이다. 이렇게 해석된 평등 형식을 적용하는 것은 같은 것과 다른 것의 차이를 구별하고, 또 다른 것 중에 또 같은 것과 다른 것의 차이를 구별하는 반복을 실행하는 것이라고도 할 수 있다.

2. 형식적 평등에서 ‘같게’와 ‘다르게’의 구별

‘같은 것을 같게’에서 출발하는 형식적 평등의 의미가 일차적으로 ‘같은 것’과 ‘다른 것’을 구별하는 것이라면 그에 수반하여 같은 것에 대해 ‘같게’ 대우하는

92) 헌법상 평등 개념을 상대적 평등으로 보는 입장에서 형식적 평등은 ‘같은 것은 같게, 다른 것은 다르게’로 이해된다. 그러나 이러한 상대적 평등 역시 그 논리적 출발점은 ‘같은 것은 같게’이다. 물론 다른 것을 같게 대우하거나 만드는 것은 가능하지 않거나 바람직하지 않다는 것이 절대적 평등보다 상대적 평등을 헌법상 평등 개념으로 받아들이는 것을 선호하게 하는 중요한 논거이기도 하다. 그러나 여기서는 ‘같은 것은 같게’에서 비롯되는 평등 또는 불평등의 형식이 과연 차별의 형식과 일치하여 동일하게 관념화할 수 있을 것인지 분석하는 것을 염두에 두고 있으므로 ‘다른 것은 다르게’를 논리적 전개에서 앞서는 전제로 삼지 않는다. 상대적 평등 개념을 포함한 평등 형식과 차별 형식의 관계에 대한 분석은 본 논문 「제3장 제3절 III. 평등 형식과 차별 형식의 관계」 참조.



것은 비교 대상이 같다는 전제에서 출발하기 때문에 상대적으로 단순해 보이는 형식을 갖는다. 예를 들어 교사가 학생에게 성적을 부여할 때 시험 점수를 비교해 같은 점수를 획득했다면 같은 등급을 부여하면 된다. 자식에게 용돈을 주기로 결정했을 때 자신이 낳거나 길렀다는 측면에서 같다면 같은 금액을 지급하면 된다. 여기서 자식이 딸이건 아들이건 일찍 낳았건 늦게 낳았건 그러한 사유는 결정의 새로운 기준이 될 수는 있어도 자신이 낳거나 길렀다는 측면에서 같다는 결정이 유지되는 한 지급하는 금액은 같아야 한다. 그런데 성적으로 등급을 부과하고 일정 금액을 용돈으로 지급하는 방식은 ‘같이 대우’하는 것의 의미를 어떻게 해석하느냐에 따라 달라질 수 있다. 이것은 미묘하지만 중요한 차이이다.

같이 대우하는 기준이 비교 대상에 의존하는지 아니면 비교 대상의 유무와 상관없이 별도로 마련된 기준으로서 표준에 의거하는지 여부에 따라 평등 개념은 엄격한(strict) 것과 느슨한(loose) 것으로 구분되기도 한다.⁹³⁾ 예를 들어 X에게 성적을 부과할 때 점수가 90점으로 같은 Y에게 A등급이 부과됐으면 X에게도 같은 A등급을 부과하고, X에게 용돈을 지급할 때 같은 자식인 Y에게 10만 원을 지급했으면 X에게도 같은 10만 원을 지급하는 것이 엄격한 의미의 평등이다. 반면에 ‘90점을 획득하면 A등급을 부여하라.’거나 ‘자기 자식이면 10만 원을 지급하라.’ 같은 별도의 기준이 용돈 지급의 표준으로서 존재한다면 그에 따라 X가 90점을 획득했으면 Y에게 부여한 등급과 무관하게 A등급을 부여하고, X가 자기 자식이면 Y가 얼마를 지급 받았는지에 상관없이 10만 원을 지급하는 것이 느슨한 의미의 평등인 것이다.

두 가지 의미의 차이는 비교 과정에서 나타난다. 먼저 느슨한 평등은 비교 대상 X와 Y가 표준 적용의 조건이 같은 이상 정해진 표준에 따라 일관되게 대우하기만 하면 된다. 그런데 엄격한 평등은 비교 대상 X와 Y가 같다는 것만으로 적용 기준이 정해지지 않는다. 대우에 적용될 기준을 비교 상대에게 적용된 기준과 비교해서 그에 따라야 한다. 만약 90점을 획득한 학생에게 A등급을 주기로 했는데 90점을 받은 X에게 ‘A+’등급을 부여한 경우, 90점을 받은 Y에게는 엄격한 평등에 따르면 X가 받은 것과 똑같이 ‘A+’등급을 부여해야 하지만, 느슨한 평등에 따르면 정해진 표준에 따라 A등급을 부여해야 한다. 또한 자식에게 10만 원을 주기로 했는데 자식인 X에게 10달러를 준 경우 엄격한 평등에 따르면 똑같이 자기가 낳은 Y에게도 10달러를 지급해야 하지만, 느슨한 평등에 따르면 Y에게는 10만원을 지급해야

93) Elisa Holmes, “Anti-Discrimination Rights without Equality”, *The Modern Law Review* 68(2), 2005, pp. 175~194: pp. 179~182 참조.



한다. Y가 받는 등급 또는 금액은 느슨한 평등에 따르면 정해진 표준에 따라 변함이 없어야 하지만, 엄격한 평등에 따르면 X가 받는 등급 또는 금액에 의존하여 변한다.

느슨한 평등의 핵심이 기준 적용의 일관성이라면 엄격한 평등의 핵심은 기준 적용의 상관성이라고 할 수 있다. 같게 대우하는 기준을 비교 상대에게 실제로 적용된 것에 따라 유동적으로 설정할 것인지 비교 상대에게 실제로 적용된 것과 무관하게 고정적으로 설정할 것인지에 따라 평등 개념의 엄격성을 이원화한 것이다. 만약 같은 비교 상대에 대해 고정된 기준을 같게 적용하지 않고 비교 상대에 대한 유동적 기준을 같게 적용한 경우 엄격한 평등 개념에 따르면 평등이 실현된 것이지만 느슨한 평등 개념에 따르면 평등에서 벗어나게 된다.

그런데 평등 개념을 엄격한 것과 느슨한 것으로 나누어 이해한다고 하더라도 일단 같은 것에 대해서 같게 대우하는 기준이 정해지고 나면 다르다고 구별된 비교 상대에 대해서는 같다고 구별된 비교 상대에게 적용되는 기준이 유동적이든 고정적이든 그 기준과 다른 수준의 대우를 하는 것만으로 형식적 평등은 실현된 것으로 보게 된다. 이렇게 해석된 평등 형식은 같다고 구별된 비교 상대에게 적용될 기준을 중심에 두고 있기 때문에 다르다고 구별된 비교 상대에게 어떤 기준이 적용되어야 하는지에 관한 문제는 주변에 남겨 둘 수 있다.

3. 형식적 평등이 설명해주지 않는 것

(1) 평등한 그 무엇

‘같게’ 대우해야 하는 이유를 평등원칙에서 찾는다면 그저 ‘같은 것’이기 때문이라는 점밖에 없다. 앞서 엄격한 평등과 느슨한 평등을 설명하기 위해 사용된 예에서 학생에게 점수에 따른 등급을 부여하고, 자식에게 용돈을 지급하는 것은 각각 ‘모든 학생을 같게 대우하라.’ 또는 ‘모든 자식을 같게 대우하라.’는 행위에 관한 의무론적 규범⁹⁴⁾을 학생에게는 ‘등급’에 관해서, 자식에게는 ‘용돈’에 관해서 적용한 것이다. 등급 부여의 측면에서 평등하게 대우하기 위해 학생을 점수에 따라 나눴다는 점은 직접 드러나지 않는다. 용돈 지급의 측면에서 평등하게 대우하기 위해 자신이 낳았거나 길렀다는

94) 홈즈(E. Holmes)는 평등원칙을 행위 관련 원칙과 상황(situation) 관련 원칙으로 구분하면서, 평등에 ‘형식적(formal)’과 ‘실질적(substantive)’이라고 레이블 붙인 것을 행위 관련 원칙과 상황 관련 원칙에 그대로 적용시키기 어렵다고 본다. 아울러 여러 가지 평등 개념은 평등주의 원칙(egalitarian principles)을 의무론적(deontic) 관점과 목적론적(telic) 관점에서 구분한 형식에 담아내려고 한다. Elisa Holmes, “Anti-Discrimination Rights without Equality”, *The Modern Law Review* 68(2), 2005, pp. 175~194: pp. 177~179 참조.



사실로부터 자식을 구분했다는 점과 성별 또는 낱거나 기른 순서를 고려하지는 않았다는 점도 직접 드러나지 않는다. 이렇게 평등원칙은 그 적용 조건을 결정하는 과정에 대해 아무런 설명을 제공하지 않는다. 그래서 이 원칙이 의미를 갖기 위해서는 어떤 사람과 어떤 대우가 같은 것인지 결정할 수 있는 외부의 가치를 편입시켜야 한다.

하지만 ‘무엇의’ 평등인지 밝히기 위해 끌어들 수 있는 것들은 관심, 고려, 존중, 존엄, 건강, 소득, 부, 행복, 자유, 권리, 기회, 욕구, 선호, 복지, 기본적 재화, 경제적 자원, 사회적 지위나 신분, 정치권력 등 이루 헤아릴 수 없이 많다. 그런데 이러한 외부의 가치가 발견되는 순간 평등원칙은 뒤로 밀려나거나 불필요한 것처럼 보이기도 한다. 이와 같이 형식적 평등이 설명해 주지 못하는 미흡함에 대한 지적은 평등이 혼란과 논리적 오류를 부추길 뿐이기 때문에 결국 평등의 수사학은 포기되어야 한다는 주장⁹⁵⁾으로까지 나아가기도 한다.

(2) 비교 대상의 특성

‘같은 것은 같게’ 알고리즘이 내놓는 복잡한 결과를 예상해 보기 위해 ‘같은 것(likes)’에 관한 문제를 앞에서 다루면서 직접 언급하지는 않았지만, 실제로 숫자와 같은 추상적인 관념을 제외하면 순전히 똑같은 것은 없다. 그렇기 때문에 X와 Y를 평등하게 대우하려면 X와 Y의 ‘어떤 측면’이 평등한지 먼저 결정해야 한다. 앞의 아리스토텔레스 시대의 예에서는 먼저 ‘사람’의 측면에서 사람인 것과 사람이 아닌 ‘다른 것’을 결정했다. 이 과정은 평등하게 대우할 대상을 구분하는 단계로서 그 어떤 특성을 선택하여 그와 일치하는 특성이 있는지 비교하는 것이다. 형식적 평등 개념에는 이 부분이 구체적으로 드러나 있지 않다. 포즈만(L. Pojman)이 예로 들었던 것처럼 똑같이 보이는 두 개의 탁구공일지라도 그것의 재료가 각각 다르고, 각각 놓여 있는 위치도 다르다.⁹⁶⁾ 같은 종류의 플라스틱 재료라 하더라도 그 중에 탁구공으로 만들기 위해 사용되는 부분은 같지 않고, 두 개의 탁구공이 같은 시간에 같은 공간을 차지하는 것은 물리적 한계를 넘어서는 것이기 때문이다.

따라서 X와 Y가 같다고 하려면 어떤 측면에서 같은지 그에 대한 해명이 필요하다. 두 개의 나무는 높이가 같다든가 두 명의 야구 선수는 타율이 같다든지 두 명의 노동자가 상품을 생산하는 비율이 같다고 할 때의 높이, 타율 생산비율처럼

95) 웨스턴(P. Westen)은 아무런 내용이 없는 평등의 형식적 성격 때문에 평등원칙이 서구 사상에서 수천 년간 고수될 수 있었다고 비판한다. Peter Westen, “The Empty Idea of Equality”, *Harvard Law Review* 95(3), 1982, pp. 537~596 참조.

96) Louis P. Pojman, “Introduction: The Nature and Value of Equality”, in *Equality: Selected Readings*, Pojman, Louis P. · Robert Westmoreland(Eds.), Oxford University Press, 1997, pp. 1~14: p. 2.



무엇이 평등하고 같은지 변수로서 작용하는 ‘그 무엇’의 특성을 밝혀야 하는 것이다. 같고 다름을 밝히는 과정은 어떤 특성을 적용해 그것을 기준 삼아 비교하는 것이다. 그렇다면 행위자가 두 가지 대상을 비교하는 것과 그 비교를 위해 어떤 특성을 선택하는 것은 평등 형식을 실천적으로 수행하는 과정의 일부가 된다. 이때 어떤 특성을 사전에 선택하여 비교할 수도 있고, 먼저 비교한 다음 비교의 기준이 된 특성을 사후에 밝힐 수도 있지만 비교 대상의 그 어떤 특성은 비교의 기준으로서 필수적이다.

(3) 상태 관련 평등

같은 것을 같게 대우하는 형식적 평등은 기본적으로 행위와 관련되어 있다. 그런데 이런 평등 형식은 평등을 어떤 상태로서 도달해야 할 목표로 표현하지 못한다. 예를 들어 행위 관련 평등원칙으로 표현할 수 있는 목적적 형식은 Y가 대우 받은 것처럼 X도 대우 받을 수 있도록 행위를 하는 것이다. 보다 구체적으로 Y가 주식매매를 통해 M만큼의 돈을 재산으로 보유하게 됐을 때 X가 주식매매를 통해 M만큼 돈을 벌 수 있도록 지원해 주는 것이다. 비교를 위해 이를 의무론적 형식으로 표현하면 X에게도 Y처럼 주식매매를 할 수 있도록 하는 것이다. 그러나 상태 관련 평등원칙의 목적적 형식은 주식매매가 아니더라도 직접 M만큼의 돈을 지급하는 것을 포함해 X가 M만큼의 돈을 재산으로 가질 수 있도록 해주면 된다. 이때 상태 관련 평등원칙은 그 자체가 목적론적이기 때문에 이에 관한 의무론적 형식은 없다.⁹⁷⁾

III. 평등 형식과 차별 형식의 관계

1. 차별과 평등의 관계 설정

차별은 일반적인 관점에서건 전문적인 관점에서건 직관적으로 평등과 일정한 관계를 맺고 있는 것으로 받아들여진다. 특히 대한민국헌법의 구문 체계에서는 “모든 국민은 법 앞에 평등하다. 누구든지 성별 · 종교 · 또는 · 사회적 · 신분에 의하여 정치적 · 경제적 · 사회적 · 문화적 생활의 모든 영역에 있어서 차별을 받지 아니한다.”(제11조 제1항)고 하여 한 개의 조문 안에 두 문장으로 나뉘어 각각 평등과 차별에 대해 규정되어 있기 때문에 그러한 관계 설정은 타당한 출발점이

97) Elisa Holmes, “Anti-Discrimination Rights without Equality”, *The Modern Law Review* 68(2), 2005, pp. 175~194: p. 180.



된다. 그래서 두 개념의 관계를 평등하지 않은 것이 곧 차별이라거나 차별 받지 않는 것이 곧 평등이라는 식의 단순하고 자명한 반대 관계를 설정해 놓고 논의를 전개하는 것이 일반적이다. 그러나 이러한 관계 설정은 어디까지나 양 개념으로 포착하려는 것이 동일한 현상의 다른 측면이거나 차별이 평등의 반대어인 불평등의 대용물로 사용될 수 있는 경우로 제한된다. 차별이 평등으로부터 환원될 수 있는 관계라는 문제 설정이 단지 개념 논리적인 명쾌함을 위한 것이 아니라면 복잡해 보이는 차별 형식도 ‘같은 것은 같게’라는 단순한 알고리즘으로 구성된 형식적 평등을 비롯해 평등 형식으로 설명될 수 있어야 할 것이다.

2. 차별과 평등의 관계 분석

차별의 몇 가지 형식⁹⁸⁾ 중에 첫 번째 차별 형식부터 네 번째 차별 형식까지(유형①, 유형②, 유형③, 유형④), 즉 편견과 고정관념은 어떤 측면에서 다른 것을 다르게 대우한 것에 대해 문제 삼는다. 그리고 이러한 문제 설정의 이면에는 어떤 측면에서 다른 것을 고려해 다르게 대우하지 말고, 그 다른 것을 무시하고 또 다른 어떤 측면에서 같은 점을 고려해 같게 대우하라는 규범적 요청이 담겨 있다. 차별 형식은 평등 형식과 충돌하지 않는 행위, 즉 ‘다른 것을 다르게 대우’한 것을 문제 삼는 것이다. ‘다른 것을 다르게 대우’하는 것까지 평등 개념에 포함시키는 ‘상대적 평등’ 개념에 따른다면 평등 형식과 충돌하지 않는 수준을 넘어 오히려 평등 형식에 부합하는 행위를 문제 삼는 것이다. 이러한 차별 형식에 반대하는 원칙은 어떤 측면에서 다른 특성을 무시하고 또 다른 어떤 측면에서 같은 특성을 고려해 같게 대우하라는 내용을 담게 된다.

차별효과 또는 간접차별로서 다섯 번째 차별 형식(유형⑤)도 어떤 측면을 고려해서 대우한 것을 고려하지 않고 무시해야 하는 어떤 측면을 고려한 것과 등가로 취급한다는 점에서 어떤 측면에서 다른 것을 다르게 대우한 것을 문제 삼는 것으로 볼 수 있다. 그렇다면 다섯 번째 차별 형식에 반대하는 원칙 또한 앞의 다른 차별 형식에 관한 원칙처럼 다른 것을 무시하고 또 다른 어떤 측면에서 같은 점을 고려해서 같게 대우하라는 요청을 그 내용으로 하게 된다. 차별과 차별 주장에 담긴 규범적 요청이 평등과 평등 주장에 담긴 규범적 요청과 논리적인 반대 관계이려면 차별은 평등 형식에 부합하지 않는 것만을 문제 삼아야 할 것이다. 그러니까 차별을 불평등의 형식으로서 ‘같은 것을 다르게 대우’한 경우나 ‘다른 것을 같게 대우’한 경우만을

98) 이에 관해서 본 논문 「제3장 제2절 II. 차별의 몇 가지 형식」 참조



문제 삼아야 한다는 것이다. 그런데 다섯 번째 형식까지 차별 개념은 다른 것을 다르게 대우한 현상을 포착하고 있다. 그렇다면 적어도 이러한 조건 속에서 차별이 평등의 반대 개념이라는 주장의 타당성은 상실된다.

적극적 조치의 불이행으로서 부작위에 관한 여섯 번째 차별 형식(유형⑥)은 우선 어떤 측면에서 같은 것을 같게 대우한 것에 대해서는 문제 삼지 않는다. 대신에 어떤 측면에서 다른 것을 무시하고 같게 대우한 것을 문제 삼는다. 이러한 차별 형식에 반대하는 원칙은 다른 것을 다르게 대우하라는 요청이다. 그렇다면 이 부분은 적어도 ‘다른 것을 다르게 대우’하는 것을 명령하는 상대적 평등의 규범적 요청에 부합하고, 그것을 허용하는 형식적 평등의 규범적 요청과도 양립 가능하다. 다만, 이때 문제가 되는 것은 다르게 대우하는 수준이다. 엄격한 평등원칙에 따르면 다름을 결정한 어떤 측면이 같은 또 다른 사람에게 적용된 기준이 다르게 대우하는 기준이 될 수 있고, 느슨한 평등원칙에 따르면 또 다른 사람에게 적용된 실제 기준이 아닌 어떤 측면을 고려한 규범적 표준이 다르게 대우하는 기준이 될 수 있다. 예를 들어 사업장에 고용된 장애인에게 적용될 편의시설의 기준을 정할 때 엄격한 평등원칙에 따르면 다른 사업장의 장애인에게 적용되는 기준에 따라 정해진다면, 느슨한 평등원칙에 따르면 전체 사업장의 장애인을 대상으로 적용되는 규범적 표준에 따라 정해진다는 것이다. 다른 것을 다르게 대우하지 않은 부작위를 차별로 포착하는 경우에 비로소 차별 형식은 평등 형식과 조응하여 관련을 맺게 된다. 이런 조건 속에서 차별이 평등의 반대 개념이라는 주장의 타당성은 유지된다.

IV. 차별과 헌법적 가치의 연결

1. 평등과 차별

평등은 그 실천적 수준과 한계에 따라 존재론적 평등, 기회의 평등, 조건의 평등, 결과의 평등 등으로 유형화되기도 한다.⁹⁹⁾ 첫째, 존재론적(ontological) 평등은 평등을 실체적인 것으로 보는 유형인데, 일종의 목적으로서의 평등을 의미하는 것으로 이해된다. 둘째, 기회(opportunity)의 평등은 오늘날 자본주의 사회에 가장 넓게 수용되는 평등의 범주라 할 수 있다. 프랑스 혁명과 미국 혁명에 뿌리를 두고 있는 이 유형은 각 개인이 자신의 소질과 능력을 자유롭게 계발할 평등한 권리와 기회를

99) Bryan Turner, *Equality*, Ellis Horwood Limited and Tavistock Publications, 1986, pp. 34-56 참조.



가질 뿐만 아니라 동일한 업적에 대해서는 동일한 보상이 주어지는 것으로 본다. 그것은 곧 모든 사회적 제도에 대한 접근을 모든 사람에게 균등하게 열어놓는다는 뜻이기도 하다. 이런 유형의 평등에서는 혈통, 종교적 배경, 가문 등의 객관적 조건이 아니라 개인의 주관적 능력이 결정적인 규정요소가 된다. 셋째, 조건(condition)의 평등은 기회의 균등과 연결된 것으로 사회적 기회를 획득하려는 자유경쟁의 출발 조건을 평등하게 정비하고자 노력한다. 넷째, 결과(outcome)의 평등은 자연적 능력이나 출발점을 고려하지 않고 법적 조처나 정치적 수단 등을 이용해 마지막 결과의 평등만을 얻고 하는 유형을 가리킨다. 여기에는 출발 단계의 불평등을 마지막 단계의 사회적 평등으로 바꿀 수 있다는 생각이 전제되어 있다.¹⁰⁰⁾

차별은 종종 불평등을 대체하는 개념으로 채택된다. 이러한 선택은 평등이 침해된 상태로서 불평등 개념을 매개로 차별을 평등의 반대 개념이자 평등이 침해된 상태로 파악하게 하는 논리적 가교를 만든다. 평등의 반대 개념으로 이해되는 차별에 대해 그 부당성의 근거를 평등규범의 위반 또는 침해에서 찾는 것은 헌법 도그마틱에서는 매우 익숙한 시도이다. 예를 들어 어떤 집단이 관련되었든지 기회의 평등(equality of opportunity)을 약화시키는 경우 차별이 부당하고, 기회의 평등을 약화시키기 때문에 차별이 부당해지는 것이라고 하여 차별이 부당한 실천적 모델과 이유를 기회의 불평등에서 찾기도 한다.¹⁰¹⁾ ‘기회’를 ‘평등한 그 무엇’으로 보는 이러한 관점에 따르면 기회의 평등을 약화시키는 것은 차별을 부당하게 하는 필요조건이자 모든 차별 형식에 적용될 수 있는 공통의 원인인 것이다. 이때 기회의 평등은 마치 알고리즘을 로봇 또는 인공지능 에이전트와 상호 교환적으로 사용하듯이¹⁰²⁾ 공정 또는 분배적 정의와 상호 교환적인 용어로 사용된다.¹⁰³⁾ 그리고 기회의 대상은 직업이나 지위에 국한되지 않고 복지를 아우른다.

기회의 불평등에서 차별의 부당함을 찾는 관점은 차별이 피차별자(discriminatee)와 긴밀하게 묶여 있다는 측면을 주의 깊게 고려한다. 그리고 특정한 차별 사건에는 여러 가지 측면의 부당성이 포함되어 있을 수 있지만 각각의 논거가 차별 그 자체의 부당성을 온전히 설명해 주지 못하기 때문에 그러한 논거들 간에도 위계 또는 서열이 필요하고 그 중에 기회의 평등은 가장 일반적인 논거가 된다고 주장한다. 기회의 불평등에서 차별의 부당성을 찾을 경우 평등한 분리에 대한 설명에 한계가

100) 박호성, 평등론, 창작과비평사, 1995, 57~58쪽; 선우현, 평등, 책세상, pp. 32~36 참조.

101) Shlomi Segall, “What’s So Bad about Discrimination?”, *Utilitas* 24(1), 2012, pp. 82~100 참조.

102) 이에 관해서 본 논문 「제2장 제1절 I. 2. ‘4차 산업혁명’을 상징하는 용어들과 그 관계」 참조.

103) Shlomi Segall, “What’s So Bad about Discrimination?”, *Utilitas* 24(1), 2012, pp. 82~100: p. 83.



있다는 반론에 대해서는 평등한데 다만 분리된 경우에도 사회적 맥락에서 존중의 기회가 불평등할 수 있다고 하거나,¹⁰⁴⁾ 적극적 조치는 평등한 기회를 박탈할 수 있다는 반론에 대해서 존중을 위해 일시적으로 기회를 상실하는 것은 기회를 더 증대하는 것과 균형을 이룬다고 반박¹⁰⁵⁾함으로써 ‘존중’처럼 차별의 부당성 논거로 제시되는 다른 사유를 기회의 평등이라는 전체적 관점에 포함시켜 차별 그 자체의 부당성에 관한 일반적 논의의 틀을 구성한다.

그럼으로써 기회의 평등과 관련이 없는 정책과 불쾌감 또는 혐오감만을 주는 정책을 구별한다. 즉, 모든 불쾌한 정책이 곧바로 모두 차별적인 것이 아니라는 것이다. 그러면서 불쾌하거나 혐오스럽다는 것과 차별적인 것의 차이는 차별하는 측과 차별 받는 측이 갖는 권력의 차이에서 발생하는 것으로 본다. 사적 영역에서 함께 가족생활을 할 동반자를 선택할 때 편견에 따라 어떤 사유를 근거로 선택하는 것이 불쾌감이나 혐오감을 줄 수는 있어도 양 측이 기회의 평등을 박탈할 만큼 권력이 불균형 상태에 있지 않다는 점 즉, 차별자가 피차별자보다 우월한 권력을 갖고 있지 않다는 점을 논거로 차별적인 것은 아니라고 논증한다.¹⁰⁶⁾

2. 자유와 차별

헌법이 누구든지 모든 영역에서 차별을 받지 않도록 명령하고 있다면 민간 부문도 예외가 될 수 없다. 그런데 차별행위를 금지함으로써 무언가 보호하려는 특별한 법적 이익이 없다면 차별행위를 민사법상 불법행위의 한 종류로 다루지 않을 이유가 없게 된다. 모로(S. Moreau)는 이러한 문제의식을 바탕으로 차별행위와 대부분 법적으로 인정되는 불법행위 사이에 보호법익의 측면에서 어떤 차이가 있는지에 집중한다.¹⁰⁷⁾ 이때 민사상 불법행위의 보호법익은 신체의 완전성, 재산, 평판 같은 이익이다. 그런데 차별을 하지 못하게 함으로써 보호하려는 대상에 그러한 이익은 없기 때문에 차별로부터의 보호에 관한 논의에서 다른 사적 주체의 잘못으로 개인이 겪는 고통에 대해 개별적으로 배상하게 하는 것보다 그 개인을 포함하는 집단의 전체적 상황을 나아지게 하려는 분배 정책으로 관심을 돌리는 것으로 본다. 따라서 차별을 금지함으로써 지킬 수 있는 이익이 있다면 그 관심을 되돌리거나 최소한 관심의 대상에 그러한 이익이 추가되어야 할 것이다.

104) Shlomi Segall, “What’s So Bad about Discrimination?”, *Utilitas* 24(1), 2012, pp. 82~100: p. 91.

105) Shlomi Segall, “What’s So Bad about Discrimination?”, *Utilitas* 24(1), 2012, pp. 82~100: p. 92.

106) Shlomi Segall, “What’s So Bad about Discrimination?”, *Utilitas* 24(1), 2012, pp. 82~100: p. 98.

107) Sophia Moreau, “What Is Discrimination?”, *Philosophy & Public Affairs* 38(2), 2010, pp. 143~179 참조.



모로는 그러한 이익 즉, 차별로 침해당할 수 있는 이익을 “숙고할 자유 (deliberative freedom)”라는 개념으로 포착하면서, 숙고할 자유는 “우리가 피부색이나 성별처럼 규범적으로 관련 없는 특성의 결과로부터 절연된 채 어떻게 살 것인지 결정할 자유”라고 말한다.¹⁰⁸⁾ 어떻게 살 것인지 자유롭게 결정할 수 있는 조건은 자신의 특성이 만들어 내는 사회적 결과로부터 단절되는 데에 있고, 그러한 특성은 자신과 규범적으로 무관한(normatively extraneous) 것이어야 한다. 따라서 이러한 자유를 보호하기 위해 차별을 금지하는 법에 제시된 특성들은 어떻게 살 것인지 결정할 때 고려하지 않아도 되는 사유가 된다.

그렇다면 모로에게 차별에 관한 권리 즉, 차별 받지 않을 권리는 숙고할 자유에 관한 권리인 것이다. 이러한 권리를 보호하기 위해서는 그에 상응하는 의무가 타인에게 부과되어야 한다. 모로는 이때 의무가 부과됨으로써 제한 받게 되는 타인의 자유 역시 “어떤 특성에 대한 고려 없이 어떻게 살 것인지 결정할 자유”¹⁰⁹⁾라고 본다. 타인은 자신이 어떻게 살 것인지 결정할 때 그 특성을 판단의 근거로 ‘고려하지 않아야 하기’ 때문이다. 그 특성을 고려하지 ‘않을 수’ 있는 것과 고려하지 ‘않아야 하는’ 것은 규범적으로 차이가 있다. 차별에 관한 법에 따라 자신이 어떤 사유를 판단 기준으로 선택할 때 특정 사유를 고려하지 않아야 함에도 불구하고 그 특성을 고려하려면 그것이 타인에게 미치는 효과까지 고려해야 한다. 고려하지 ‘않아야 하는’ 것을 회피 또는 정당화해서 결정하려면 또 다른 고려를 ‘해야 하는’ 것이다.

결국 모로에 따르면 어떤 사람의 숙고할 자유는 다른 사람의 숙고할 자유 즉, 타인이 자신의 판단 기준으로 어떤 사유를 고려하지 않고도 어떻게 살 것인지 결정할 자유 또는 자신의 판단 기준으로 선택한 것이 누군가에게 미칠 효과를 고려하지 않고도 어떻게 살 것인지 결정할 자유와 “형량이 필요한(the need to balance)”¹¹⁰⁾ 관계에 놓인다. 이렇게 차별금지법의 보호법익을 숙고할 자유로 구성하면 법에 성별을 차별 사유로 규정한다는 것은 어떤 사람이 자신의 성별이 누군가의 결정에서 고려 대상이 되어 자신에게 특별한 결과로 영향을 미칠 수 있을 것이라는 우려 또는 고려를 하는 제약에서 절연된 채 자신이 어떻게 살 것인지에 대해 결정할

108) Sophia Moreau, “What Is Discrimination?”, *Philosophy & Public Affairs* 38(2), 2010, pp. 143~179: p. 147.

109) Sophia Moreau, “What Is Discrimination?”, *Philosophy & Public Affairs* 38(2), 2010, pp. 143~179: p. 163.

110) Sophia Moreau, “What Is Discrimination?”, *Philosophy & Public Affairs* 38(2), 2010, pp. 143~179: p. 163.



수 있는 자유를 보장하는 것이면서, 동시에 다른 사람이 결정을 할 때 성별을 고려하지 않을 자유 또는 성별을 기준으로 삼는 것이 누군가에게 영향을 미칠 것이라는 점까지 고려하지 않아도 될 자유가 제한되도록 관계를 설정하는 것이다.

이처럼 차별의 문제를 자유의 문제로 구성하는 이면에는 다원주의에 대한 불안(pluralism anxiety)이 잠재되어 있다.¹¹¹⁾ 특히 여러 가지 개인의 특성이 사회의 집단을 구별하는 특성으로 사용될 수 있는 다양한 가능성이 열려 있는 사회에서 새로운 특성의 발굴과 그 특성의 사회화는 새로운 소수자 집단을 양산한다. 그리고 새로운 소수자 집단이 형성되거나 발견될 때마다 집단에 기초한 비교를 통해 기본권을 보장하는 것은 사회 집단의 구별에 대한 타당성과 정당성을 지속적으로 논증해야 하는 피로감을 사회에 던져준다. 그렇기 때문에 예컨대 성적 지향에 따라 동성애 집단을 구별하여 이 집단에게 다른 집단 즉, 이성애 집단과 비교하여 혼인에 관한 권리를 부여하도록 해야 한다는 특수한 논변을 구성하기보다 모든 사람에게 사랑 하는 사람과 혼인할 자유와 그에 관한 권리가 있다는 보편적 논변을 구성하는 것이 그러한 평등 피로감에서 벗어날 수 있는 대안이 될 수 있다는 것이다.

3. 인간의 존엄성과 차별

뒤리히(G. Dürig)는 1956년에 인간의 존엄을 인격의 존엄으로 이해하는 방식을 헌법논의에 최초로 도입했다.¹¹²⁾ 그에 따르면 권리주체로서 인간이 자신의 권리주체성을 박탈당할 때 가치질서가 타락하기 시작한다.¹¹³⁾ 인간이 존엄하다는 것은 인간이 ‘자유’롭고, 모든 인간이 자유를 갖고 있기 때문에 그 점에서 ‘평등’하다는 것을 뜻한다. 인간존엄의 침해는 곧 인격적 존재의 부정이고 한 개인을 보편적인 법적 평등관계로부터 배제하는 인격적 비하를 뜻한다. 존엄성 요구는 기본적으로 사람을 목적으로 대하고 결코 어떤 과업을 수행하는 수단이나 도구로 대하지 말라는 요구를 담고 있다.

111) Kenji Yoshino, “The New Equal Protection”, *Harvard Law Review* 124(3), 2011, pp. 747~803: pp. 792~802.

112) Günter Dürig, “Der Grundrechtssatz von der Menschenwürde: Entwurf eines Praktikablen Wertsystems der Grundrechte aus Art. 1 Abs. I in Verbindung mit Art. 19 Abs. II des Grundgesetzes”, *AöR* 81(2), 1956, S. 117~157 참조.

113) Günter Dürig, “Der Grundrechtssatz von der Menschenwürde: Entwurf eines Praktikablen Wertsystems der Grundrechte aus Art. 1 Abs. I in Verbindung mit Art. 19 Abs. II des Grundgesetzes”, *AöR* 81(2), 1956, S. 117~157: S. 127.



인권에 관한 법적 해석에서 존엄성(dignity) 개념은 적어도 네 가지 기능을 수행한다.¹¹⁴⁾ 첫째, 인간이 왜 권리를 가져야 하는지에 대한 핵심 논거를 제공한다. 둘째, 특수한 권리들의 목록을 확인하는 데 도움을 준다. 셋째, 원칙에 따라 생성된 권리들의 목록을 해석하기 위한 해석 원칙으로 작용한다. 넷째, 존엄성 그 자체가 특수한 내용을 가진 권리 또는 의무로서 기능한다.¹¹⁵⁾ 대한민국헌법에는 1962년 개정을 통해 “모든 국민은 인간으로서의 존엄과 가치를 가지”(제8조, 현행 1987년 헌법 제10조)는 것으로 규정하여 기본권 해석에서 존엄성이 여러 가지 기능을 수행할 수 있는 명시적 근거가 마련되어 있다.

차별이 존엄성과 관련을 맺을 때 차별은 같은 것을 같게 대우하지 않을 때 문제되는 것이 아니라 인간을 존엄하게 대우하지 않을 때 문제된다. 이때 존엄성은 차별의 형식 또는 불평등의 형식에 부당성을 측정할 수 있는 내용을 추가하는 기준이 되거나 직접 그 내용이 된다. 월드론(J. Waldron)은 가치와 밀접하게 연결되어 있는 존엄성 개념에서 가치를 제거함으로써 추상적 개념을 좀 더 구체적 수준의 개념으로 구성하여 존엄성을 등위(rank)와 같은 의미를 갖는 개념으로 사용한다.¹¹⁶⁾ 이러한 등위 또는 위상¹¹⁷⁾ 개념으로서 존엄성은 무엇이 차별대우에 해당하는지 식별하는 기능을 한다.¹¹⁸⁾ 도덕적으로 동등한 지위에 있다는 원칙에서 출발하는 이러한 구성은 정체성, 자율성, 기본적 필요의 만족 등 인간의 가치를 가능한 많이 보유하고 있는 상위 집단의 특성을 보편화하려는 노력을 통해 형식적 평등과 연결된 차별의 발견이 상향평준화(leveling up)로 이어질 수 있는 계기를 마련하고자 한다.

4. 사회통합과 차별

근대 국민국가가 출현하는 과정에서 차별은 두 가지 기능을 수행한다. 먼저 차별을 통해 국민과 비국민을 나누어 국가의 구성원과 국가의 비구성원을 구별하는

114) Christopher McCrudden, “Human Dignity and Judicial Interpretation of Human Rights”, *European Journal of International Law* 19(4), 2008, pp. 655~724: pp. 680~681.

115) 존엄성이 기능하는 최소한의 역할로 가치(value), 원칙(principle), 정의된 권리의 기초(the basis for defined rights) 등 세 가지를 제시하기도 한다. Gay Moon · Robin Allen, “Dignity Discourse in Discrimination Law: A Better Route to Equality?”, *European Human Rights Law Review* 6, 2006, pp. 610~649: p. 615.

116) Jeremy Waldron, *Dignity, Rank, and Rights*, Meir Dan-Cohen(Ed.), Oxford University Press, 2012 참조.

117) 손제연, “위상적 개념으로서의 인간존엄”, *법철학연구* 21(1), 2018, 295~338쪽 참조.

118) Denise G. Réaume, “Dignity, Equality, and Comparison”, in *Philosophical Foundations of Discrimination Law*, Deborah Hellman · Sophia Reibetanz Moreau(Eds.), Oxford University Press, 2013, pp. 7~27: pp. 22~27.



것이다. 보통 비국민을 외국인과 동일한 개념으로 사용하기도 하지만,¹¹⁹⁾ 엄밀하게 볼 때 비국민과 외국인은 동일한 개념이 아니다. 특히 비국민과 외국인을 동일한 개념으로 사용하면서 국적이 없는 사람 즉, 무국적자까지 외국인 개념에 포함시키면 국민이 아니라는 점에서 한 번 배제되고 어떤 국가의 국적도 갖지 않았다는 점에서 외국인으로부터도 배제되는 무국적자의 개념화 과정의 일부가 은폐되기 때문이다. 어쨌든 내부자로서 국가의 구성원을 외부자로서 외국인 및 무국적자를 구별하여 다르게 처우하는 것은 내부자의 결속력을 높이거나 일체감을 형성함으로써 국민 국가의 단일성과 통일성을 확보하는 수단이 된다. 반면 같은 국민 중에 다른 국민에 대한 차별은 국민통합 또는 사회통합에 걸림돌이 된다.¹²⁰⁾ 사회통합의 측면에서 볼 때 사회적 권리를 인정하는 것은 계급차별로 분리되었던 노동계급을 통합하는데 일정 부분 기여하고 사회적 권리의 인정근거로 차별금지가 원용될 수 있다. 이때 차별금지는 단순히 부작위를 요청하는 것을 넘어 작위를 요청하는 내용을 포함하게 된다.

119) 계희열, 헌법학(중), 박영사, 2007, 62쪽: “외국인이란 대한민국의 국적을 가지고 있지 않은 자를 말한다. 외국인의 범주에는 다국적자나 무국적자도 포함한다.”

120) Hugh Collins, “Discrimination, Equality and Social Inclusion”, *Modern Law Review* 66(1), 2003, pp. 16~43: pp. 21~26.



제4절 차별의 부당성과 비교

차별이 인간의 기본적인 인식이나 판단과 밀접하게 관련되어 있고 그 중에서 특별한 사유를 인식과 판단의 기준으로 삼지 말라는 요청을 차별금지법으로 구현한 것이라고 볼 때 일반화 또는 보편화 가능성을 염두에 둔 차별금지법에 관한 논의는 일관된 규범적 기초의 탐색에 지향되곤 한다. 그것이 가능할 때에 비로소 모든 차별에 반대한다는 주장을 차별금지법에 담아낼 수 있기 때문이다. 그러나 차별의 기초 개념에서 살펴봤듯이 차별은 어떤 에이전트가 환경을 지각하는 과정에서 발생할 수 있는 오류와 편향의 문제이기도 하다는 점을 고려할 때 차별금지법의 목적이 인간을 비롯해 지각 능력이 있는 에이전트의 모든 인식 및 판단 작용에 제한을 가하는 것은 아닐 것이다.

그렇다면 차별의 부당성을 차별금지법과 연결하려는 시도는 차별금지법이 모든 인식 및 판단 작용에 제한을 가하지 않으면서 사회적으로 의미 있는 결정에 대해서만 그 인식 및 판단 근거에 제한을 둬으로써 사회적으로 용인할 수 없는 차별을 구별하려는 것으로 이해될 수 있을 것이다. 그리고 차별의 행위 중심 구성이나 결과 중심 구성에서 살펴봤듯이 그 기저에는 차별금지법의 목적이나 기능에 대한 도덕철학 또는 법철학적 입장의 차이가 묵시적으로 존립하고 있고, 이러한 입장 차이는 인간이 아닌 머신러닝 알고리즘 에이전트의 규범적 지위를 구성하는 데에도 영향을 미쳐 그 차별적 결정에 대한 규범적 평가를 달라지게 할 수도 있다.

I. 규범적 의미의 차별

1. 사람의 특성에 대한 의미부여

어떤 고객이 ‘회사로부터 차별 받았다.’고 하면 직관적으로 그 고객이 회사로부터 무언가 부당한 대우를 받았다는 의미를 떠올리게 된다. 어떤 ‘고객’ 대신에 ‘직원’이나 ‘구직자’가 그 자리를 대신해도 유사한 의미가 유지될 수 있을 것이다. 이러한 직관은 일정한 도덕적 평가나 법적 평가로까지 이어질 수 있는 직관이며, 그래서 의미론적으로는 규범적이다. 차별을 규범적으로 사용하는 경우에는 차별 개념 그 자체에



부당하다거나 잘못됐다는 의미가 함축되어 있다고 보거나 ‘차별’이라는 표현 앞에 그러한 수식어가 생략된 것으로 본다. 그러므로 차별 개념을 규범적으로 이해할 경우 차별 개념이 무엇인지 해명하려는 노력에는 불가불 차별이 왜 부당한지 도대체 어떤 잘못 때문에 차별을 개념화하려는 것인지에 대한 논증이 뒤따르게 마련이다.

차별이 규범적 의미를 갖는다는 것은 서술적으로 구성된 차별의 어떤 요소에 대한 평가를 포함하고 있다는 것이다. 차별 형식을 구성하는 공통 요소 중 하나는 사람의 어떤 특징이 구별, 분리, 분류를 위한 기준으로 사용된다는 점이다. 어떤 ‘특징(character)’은 ‘특성(feature)’이나 ‘속성(attribute)’이라고 부를 수도 있다. 이 기준은 차별의 행위에 직접적으로 사용되기도 하고 차별의 결과에 효과를 미치는 방식으로 간접적으로 작용하기도 한다.

그리고 이 기준이 갖는 규범적 의미는 행위나 결과 또는 행위 및 결과 모두의 측면에서 고려되어야 한다는 적극적 의미와 고려되지 말아야한다는 소극적 의미로 각각 구체화된다. 예를 들어 직접적이든 간접적이든 어떤 특징 C가 구별, 분리, 분류의 기준으로 사용된 경우 이 기준이 적극적인 규범적 의미를 갖는다면 이를 고려하지 않은 행위나 결과 또는 행위 및 결과는 부당함의 근거가 되는 것이다. 이는 앞서 언급한 여섯 번째 차별 형식(유형⑥), 즉 ‘합당한 배려’를 포함하는 적극적 조치의 불이행으로 부작위의 차별 형식에 대한 규범적 해석 도식(scheme)이 될 수 있을 것이다.

2. 규범적 기준에 따른 구별

또 다른 예로 기업은 경영 전문가들이 “천사 고객”과 “악마 고객”이라고 부르는¹²¹⁾ 좋은 고객과 나쁜 고객을 구별한다. 이미 살펴봤듯이 이렇게 고객을 한 명씩 개별화하지 않고 ‘좋은’ 고객과 ‘나쁜’ 고객이라는 집단(class)으로 구별하는 것은 고객을 분류(classification)하는 것이기도 하다. 그런데 이런 구별과 분류 그 자체만으로 필연적으로 규범적 의미가 도출되는 것은 아니다. 여기서 좋은 고객은 기업에 수익을 가져다주는 고객이고, 소량으로 물건을 구매하거나 서비스 이용 횟수가 적은 고객은 나쁜 고객이 된다. 좋음이나 나쁨 같이 어떤 규범적 내용이 그 기준으로 작용할 때 비로소 구별과 분류는 규범적 의미를 획득하게 된다. 고객의 좋음과 나쁨을 판별하는

121) Larry Selden · Geoffrey Colvin, *Angel Customers & Demon Customers: Discover Which Is Which and Turbo-Charge Your Stock*, Portfolio, 2003; 이 책 제목의 한국어 번역은 “회사를 먹여 살리는 착한고객”이다. Larry Selden · Geoffrey Colvin, *회사를 먹여 살리는 착한고객*, 황숙혜(역), 위즈덤하우스, 2010 참조.



기준은 기업의 수익 창출 여부이다. 기업의 수익 창출 여부가 규범적으로 고려해도 되거나 고려해야 하는 사유라면, 다시 말해 기업에게 기업의 수익 창출 여부를 고려할 규범적 권한이 있을 뿐만 아니라 그러한 내용이 고객을 구별하고 분류하는 판단 기준으로 허용되거나 명령된 사유라면 그러한 기준의 충족 여부를 좋음과 나쁨이라는 가치의 대용물로 사용하는 것은 규범적으로 정당한 의미를 갖게 된다.¹²²⁾

3. 차별 형식과 차별금지 사유의 구조적 결합

차별 형식에서 어떤 특징 C를 다른 특성 F의 대용물로 사용하는 것은 고정 관념의 구조적 형태이다. 그렇다면 이러한 사례 역시 어떤 특징 C가 수익성이고 다른 특성 F는 좋음이라는 가치라고 할 경우 수익성을 좋음이라는 가치의 대용물로 사용하는 고정관념과 동일한 구조를 갖는다. 만약 인종이나 성별을 어떤 특징 C의 자리에 놓고, 다른 특성 F에 수익성을 놓아 두 가지 특성을 대체 관계로 설정할 경우, 특정 인종이나 성별은 수익을 창출하는 특성과 연동되고, 좋음이라는 가치로까지 연결될 수 있다. 반대로 특정 인종이나 성별 집단과 구별되는 반대칭 관계에 있는 다른 인종과 성별 집단은 수익 창출과 무관하거나 오히려 수익을 감소시키는 특성과 연동되고, 나쁨이라는 가치로까지 연결될 수 있다.

물론 인종과 성별은 대부분의 법체계에서 차별금지사유로 규정되어 있어 법규범의 차원을 지배하고 있을 뿐만 아니라 그러한 사유가 법체계에 편입되기까지의 역사적 경험에 기초하여 문화적으로 전승된 도덕 규범적 차원에서 상식 수준의 직관을 형성하고 있기 때문에 인종이나 성별을 수익성의 대용물로 사용하는 것은 규범적으로 부당하게 여겨진다. 그런데 이러한 부당함의 근원이 고정관념 같은 차별의 구조적 형식에 있는 것인지 아니면 그런 고정관념의 내용으로 이용되는 차별의 구체적 사유 자체에 있는 것인지 설명할 수 있는 하나의 개념이나 원칙이 있는 것인지는 분명하지 않고, 다만 규범적 의미의 차별에는 차별금지 사유가 차별 형식에 구조적으로 결합되어 있다는 점을 확인할 수 있다.

122) 규범의 양식에 관해서 Hans Kelsen, *Allgemeine Theorie der Normen*, Kurt Ringhofer · Robert Walter(Hrsg.), Manz Verlags- und Universitätsbuchhandlung, 1979; H. L. A. Hart, *The Concept of Law*, 3rd ed., Oxford University Press, 2012; Robert Alexy, *Theorie der Grundrechte*, 1. Aufl., Suhrkamp, 1994 참조.



II. 차별금지 사유의 규범적 무관성

1. 특성에 기초한 질서의 형성

사람의 특성을 비롯해 어떤 특성은 보편적 질서에 속하는 것과 특수한 질서에 속하는 것으로 구분될 수 있다.¹²³⁾ 예를 들어 성(性)은 보편적 질서이지만 여성 또는 남성은 특수한 질서이다. 차별로부터 보호되는 집단의 특성을 논할 때 어떤 특성을 ‘가지고’ 있다는 것은 특수한 질서에 속하는 것에 관한 논의가 된다. 그러나 이러한 관계는 생물의 분류 체계에 따를 때 일정한 범주를 전제로 설정된 보편과 특수의 관계이다.

성이 보편적 질서에 속한다는 것은 유성 생물에 한한다. 무성 생물을 염두에 둔다면 성을 ‘가진다’는 것 역시 특수한 질서에 속할 수 있기 때문이다. 다만 그 범위를 유성 생물로 제한하는 범위에서 성은 보편적 질서이고 그 중에 특정한 성에 귀속되는 것을 특수한 질서에 속하는 것으로 볼 수는 있을 것이다. 인종에 관해서도 마찬가지로 관계가 설정될 수 있다. 인종은 보편적 질서로, 아프리카계(흑인), 유럽계(백인) 또는 아시아계(황인)는 특수한 질서로 구분할 수 있지만 이 역시 호모사피엔스(*Homo sapiens*)라는 종을 전제로 했을 때 가능한 구분이다.

이런 분류 체계는 어떤 생물학적 특성을 기준으로 구축되며 또 다른 생물학적 특성은 또 다른 하위의 특수한 질서를 만들어 내는 데에 사용될 수 있다. 키를 기준으로 한다면 큰 키 또는 작은 키를 ‘가진다’는 특수한 질서를 만들 수 있고, 눈동자 색을 기준으로 한다면 청색 눈동자, 갈색 눈동자, 또는 흑색 눈동자를 ‘가진다’는 특수한 질서를 만들어 낼 수 있는 것이다.¹²⁴⁾

2. 차별금지 사유의 도덕적 기초로서 불가변성의 한계

법적으로 금지되는 것으로 제시된 차별 사유들 사이에 공통된 성격을 밝힐 수 있다면 그러한 차별 사유들은 공통된 성격이 반영되어 있는 각각의 예시가 될 수 있다. 그렇다면 그 밖에 차별 사유는 공통된 성격에 의거하여 얼마든지 찾아내고 발굴할 수 있을 것이다. 금지되는 차별 사유의 공통된 성격의 후보 중에 하나는

123) Tarunabh Khaitan, *A Theory of Discrimination Law*, Oxford University Press, 2016, p. 56.

124) 눈동자 색을 기준으로 한 학급 분리 실험에 관해서 본 논문 「제3장 제2절 I. 4. 분리를 통한 분류와 집단의 비대칭성」 참조.



불가변성(immutability)이다. 오늘날 대부분의 사회에서 인정되는 차별금지 사유로서 인종 또는 성은 그러한 특성을 가지고 있는 사람이 스스로 변경할 수 없는 성격을 갖는다는 점에 착안한 것이다.

이러한 접근은 사람이 가지는 생물학적 성격의 불가변성에서 출발하기 때문에 자연에 따라 정의되는 특성으로서 인종, 성 등 눈에 보이는 사람의 신체적 또는 물리적 특징을 보호하는 데에 차별금지 사유가 편향되어 있는 것처럼 보인다. 그래서 종교상태, 언어정체성, 혼인상태, 성적 지향 등 쉽게 숨겨지거나 논쟁의 여지는 있지만 변경할 수 있는 비신체적 또는 사회적 특징들은 불가변성의 요건을 만족하지 못하는 것으로 취급된다.¹²⁵⁾ 게다가 불가변성의 의미를 엄격하게 해석하면 오히려 오늘날 성(性)도 변경 가능하다는 점을 근거로 불가변성을 적용하여 성별을 차별금지 사유로 삼을 수 없다는 주장도 가능해진다.

3. 넓은 의미의 불가변성과 선택 불가능성

좁게 해석될 경우의 한계를 염두에 둔다면 불가변성은 사회적 맥락을 고려해서 보다 넓은 의미로 사용될 수 있다. 불가변성을 어떤 특성의 원초적 획득이 그러한 특성을 갖게 된 사람의 선택에 의하지 않는다는 의미로 사용하는 것이다. 이러한 불가변성은 변경 불가능성에 선택 불가능성을 더한 것처럼 보인다. 이에 따르면 오늘날 성별도 변경 가능하므로 불가변성이 적용될 수 없는 차별금지 사유라는 논변에 대해 변경의 대상인 특정한 성별 그 자체의 원초적 획득까지 그 성별을 가지고 있는 사람이 선택 가능한 것은 아니라는 반박이 가능하다. 그렇다면 같은 논리의 연장선상에서 자식이 부모의 혼인상태를 선택할 수 없으며, 어린이가 자신의 출생지나 거주지를 선택할 수 없고, 외국인은 다른 국가 법체계의 관할에 있는 시민에게만 부여된 권리를 선택할 수 없다는 점 등도 차별금지 사유로 고려될 수 있다.

그런데 불가변성을 넓게 이해하는 경우에도 설명되지 못하는 차별금지 사유도 있다. 대표적인 예는 종교이다. 유교 문화권에서 태어날 것인지 불교 문화권에서 태어날 것인지 아니면 유대교 문화권, 기독교 문화권 또는 이슬람교 문화권에서 태어날 것인지 선택할 수는 없지만, 그러한 문화권에서 태어났다는 것이 곧바로 특정 종교를 가질 수밖에 없다는 것을 의미하지 않을 뿐만 아니라 개인마다 다를

125) Kenji Yoshino, "Assimilationist Bias in Equal Protection: The Visibility Presumption and the Case of 'Don't Ask, Don't Tell'", *The Yale Law Journal* 108(3), 1998, pp. 485-571: pp. 493-498.



수 있지만 종교의 변경이 절대로 불가능한 것 또한 아니다. 어떤 종교적 문화권에서 태어났든 그 문화권의 법과 제도가 종교의 자유를 보장하는 한¹²⁶⁾ 개인이 종교를 갖지 않기로 선택하거나 주류의 종교와는 다른 종교를 갖기로 선택할 수 있다. 이러한 논변에 따르면 종교 같은 특성에는 넓은 의미의 불가변성도 적용할 수 없게 된다.

넓은 의미의 불가변성은 변경 불가능성에 선택 불가능성을 더한 것처럼 보이지만 좁은 의미이건 넓은 의미이건 불가변성의 구체적 의미를 분석해 보면 둘 다 선택 불가능성에 토대를 두고 있다. 전자의 불가변성은 자신이 가지고 있는 특성의 변경을 선택할 수 없다는 것이고, 후자의 불가변성은 그에 더해 자신이 가지고 있는 특성의 획득을 선택할 수 없다는 것이다. 이 경우 자신이 변경이나 획득을 선택하여 가지게 된 특성은 불가변성에 기초한 차별금지 사유에서 논리적으로 배제된다. 변경은 구별된 경계의 한쪽 편에서 다른 편으로 넘어가는 것이고, 획득은 어떤 기준에 의해 구별을 실행하여 어느 편으로 지시되게 함으로써 차이를 만들어 내는 것이다. 따라서 어떤 특성의 획득과 변경을 선택 가능한 것으로 구성하면 차별금지 사유에서 배제된다. 여성의 임신이나 성적 지향을 선택 가능한 것으로 구성하는 경우 불가변성은 이를 차별금지 사유에서 배제하는 주요한 논거가 된다.

결국 남는 과제는 불가변성을 여전히 차별금지 사유의 공통된 성격으로 보아 이를 기준으로 차별금지 사유를 재정립하는 것 또는 선택 불가능성을 토대로 한 불가변성 논변은 선택 가능한 차별금지 사유를 설명해 내지 못하여 차별금지 사유의 공통된 성격으로 타당하지 않다고 보는 것이다.

4. 난이도에 따른 선택 가능성의 구별

불가변성으로 좀 더 많은 차별금지 사유를 설명해 내기 위해 고안될 수 있는 방법은 난이도에 따라 선택 가능성을 구분하여 난도가 높은 것 즉, 선택하기 어려운 가능성을 불가변성의 영역으로 끌어오는 것이다. 예를 들어 집단을 정의하는 특성을 변경하거나 가리는 것이 물리적으로 불가능한 경우처럼 불가변성의 의미를 엄격하게 제한하여 사용하지 않고, 그러한 특성을 변경하거나 가리는 데에 상당한 어려움이 있어 효과적으로 변경하는 것이 불가능하다는 의미로 사용하는 것이다.¹²⁷⁾ 이때 어려움은 일종의 비용으로서 경제적인 것뿐만 아니라 정신적이고

126) 예를 들어 대한민국헌법 제20조 제1항: “모든 국민은 종교의 자유를 가진다.”

127) 미국연방대법원이 불가변성을 “엄격한(strict)” 의미로 사용하지 않고, “효과적으로 변경할 수 없는(effectively immutable)”의 의미로 사용한다고 보는 경우로 *Watkins v. United States Army* 875 F.2d 699 (1989), 727.



심리적인 비용 나아가 사회적 비용도 포함한다. 이렇게 변경 가능성의 유무 문제를 변경의 난이도 문제로 전환하면 변경 불가능한 경우뿐만 아니라 성전환처럼 변경 가능하더라도 실행이 어려운 경우도 차별금지 사유로 볼 수 있다는 주장의 근거를 제공하게 된다. 그런데 이때 불가변성의 의미를 선택의 불가능성뿐만 아니라 가능성까지 포함하도록 전용할 수 있게 하는 논거로는 선택한 대로 실현될 가능성이 낮다는 점과 그러한 선택이 근본적이고 중요하며 심각하다는 점이 제시될 수 있다.

문제의 초점을 변경의 난이도에 맞출 경우 난이도와 상관없이 변경을 거부하는 어떤 특성에 대해 국가가 처벌하거나 사회가 낙인찍고 불리하게 처우하는 것을 정당화할 수도 있다.¹²⁸⁾ 반면 문제의 초점을 선택의 근본성, 중요성 또는 심각성 자체에 둔다면 근본적인 선택을 차별금지 사유로 보는 것은 그러한 선택 자체가 누군가의 삶에서 성패를 좌우하는 요소로 작용해서는 안 된다는 규범적 함의를 실천하는 것이다. 이러한 관점에 따르면 난이도를 측정하고 비용을 계산에 넣어보려고 하는 것 역시 그런 특성을 ‘가지도록’ 선택한 결정의 근본성, 중요성, 심각성을 가려 보기 위한 것이라고도 볼 수 있다. 그렇다면 차별금지 사유의 공통된 성격은 불가변성 자체에 있는 것이 아니라 보다 상승한 차원의 ‘규범적 무관성’¹²⁹⁾에 있는 것이 된다.

III. 특정 집단에 불리한 차별 결과

차별의 부당성은 특정 집단을 상대적으로 불리한 위치에 놓이게 한다는 데에서 찾을 수도 있다. 차별로 포착할 수 있는 특정 집단의 불이익은 정치적, 사회·문화적, 물질적 불이익의 세 가지 종류로 나눌 수 있는데,¹³⁰⁾ 이는 특정 집단이 불리한 상태에 있다는 점을 강조하는 결과 관련 구성에서 그 불이익의 내용을 구체화하는 것이기도 하다.¹³¹⁾

128) Kenji Yoshino, “Covering”, *The Yale Law Journal* 111(4), 2002, pp. 769~939 참조.

129) 차별의 부당함에 대한 근거를 모로(S. Moreau)는 “규범적으로 관련 없는(normatively extraneous)” 것에 두고, 세게브(R. Segev)는 “도덕적으로 무의미한(morally insignificant)” 것에서, 카이탄(T. Khaitan)은 “규범적으로 무관한(normatively irrelevant)” 것에서 찾아 구체적 표현의 차이에도 불구하고 규범적 무관성을 지적하고 있다. 이에 관해서 각각 Sophia Moreau, “What Is Discrimination?”, *Philosophy & Public Affairs* 38(2), 2010, pp. 143~179: p. 149 이하; Re'em Segev, “Making Sense of Discrimination”, *Ratio Juris* 27(1), 2014, pp. 47~78: p. 56 이하; Tarunabh Khaitan, *A Theory of Discrimination Law*, Oxford University Press, 2016, p. 60 참조.

130) Tarunabh Khaitan, *A Theory of Discrimination Law*, Oxford University Press, 2016, pp. 51~56 참조.

131) 차별의 효과 또는 결과 관련 구성에 대해서 본 논문 「제3장 제2절 III. 2. 차별의 효과 관련 구성과 결론론」 참조.



1. 정치적 불이익

특정 집단이 받게 되는 정치적 불이익은 그 구성원이 선거인으로서 수적인 (numerical) 열세에 있다는 점에서 비롯되는 측면이 강하다.¹³²⁾ 공동의 의사결정에서 수적으로 무의미한 집단, 예를 들어 에이즈 환자, 외국인, 성소수자 등의 의사와 이익은 쉽게 무시될 수 있다. 대의제를 통한다고 하더라도 국가 사무에서 실제로 대표되는 양과 질이 떨어질 경우 국가 제도에 의해 받는 관심의 성격과 정도에 영향을 미치게 된다. 이렇게 형성되는 정치적 불이익은 다른 형태의 불이익의 원인이기도 하고, 또한 그 결과이기도 하다. 실제로 대부분의 법률은 다르게 대우할 사람들을 분류해 내는 데에 초점이 맞추어져 있다고도 볼 수 있고,¹³³⁾ 그로 인해 발생하는 차이의 격차는 제도적으로 지속되고 유지될 수 있다는 점에서 정치적 불이익은 그 자체에만 국한되지 않고 불이익을 확장시키는 기반이 된다.

2. 사회·문화적 불이익

사회-문화적(socio-cultural) 불이익은 전적으로 그렇다고 볼 수는 없지만 보통 특정 집단의 구성원에 대해 유포된 편견이나 고정관념에 포함된 가정으로 표현된다.¹³⁴⁾ 어떤 특성을 가졌다는 이유만으로 그 특성을 가진 사람들이 그렇지 않은 사람들보다 도덕적으로 낮은 가치를 지닌다고 가정하는 편견은 주로 그러한 특성을 가진 사람들에 대한 언행이나 그 사람들을 대하는 태도, 예를 들어 그 사람들을 비하하는 언행이나 무시하는 태도로 나타난다. 일종의 관념적 선호가 항목별 또는 범주별로 형성되어 있는 것인데, 성, 성적 지향, 종교, 인종, 그밖에 다양한 다른 특성들은 여러 사회에서 지금까지 그랬고 여전히 계속해서 편견의 기초로 자리 잡고 있다.

이와 같은 범주적 판단을 하지 않더라도 여러 가지 특성들 간의 상관성은 긍정적으로든 부정적으로든 고정관념을 형성한다. 상관성이 있는 특성들은 서로의 대용물이 될 수 있다. 이러한 상관성에는 생물학적 요인, 사회-문화적 요인, 다른 사람들의 반응적 요인이 작용한다. 예를 들어 여성이 남성보다 수명이 길다는 생물학적 요인에 의한 상관성, 여성은 남편의 경력을 유지시키기 위해 자신의 경력을 중지 또는

132) Tarunabh Khaitan, *A Theory of Discrimination Law*, Oxford University Press, 2016, p. 52 이하.

133) John Hart Ely, *Democracy and Distrust: A Theory of Judicial Review*, Harvard University Press, 1980, p. 135.

134) Tarunabh Khaitan, *A Theory of Discrimination Law*, Oxford University Press, 2016, p. 53 이하.



포기하는 경향이 있다는 사회-문화적 요인에 의한 상관성, 여성에 대해 경찰력을 행사할 때 여성 경찰관이 집행하고 흑인에 대해 흑인 경찰관을 투입하는 것이 효과적이라는 반응적 요인에 의한 상관성 등과 같이 어떤 특성과 다른 특성의 상관성을 찾는 것이다. 이러한 상관성에 강력한 통계적 근거가 덧붙여질수록 그러한 고정관념에 확신을 갖게 되고 고정불변의 진리로서 관념 속에 고착화할 수도 있다. 특히 가정된 상관성에 기초한 고정관념이 항목별 또는 범주별로 형성된 편견과 결합할 경우 고정관념은 편견을 합리화하는 데에 사용된다.

고정관념에 가정된 상관성은 그 증거가 정확하면 정확한대로 부정확하면 부정확한대로 편견의 합리화에 기여한다. 고정관념은 특성 간의 상관성에 대한 기대에 의존하는 것인데 그 기대는 긍정적이거나 부정적일 수 있지만 실제의 효과는 그 반대일 수도 있고 변형될 수도 있다. 예를 들어 여성은 잘 보살핀다는 가정은 긍정적인 고정관념처럼 보인다. 반대로 남성은 폭력적이라는 가정은 부정적인 고정관념을 보인다. 그런데 이런 특성들이 외관상으로 또는 객관적으로는 긍정적이거나 부정적일 수 있지만 여전히 구체적 사회 현실에서 주관적으로는 다르게 판단될 수 있다는 점은 상황을 복잡하게 만든다. 보살피는 특성은 약함의 신호로, 폭력적인 특성은 강함의 신호로 작동할 수 있는 것이다. 또한 긍정적인 고정관념을 사회가 있는 그대로 긍정적으로 보는 경우에도 그 효과는 부정적인 불이익으로 산출될 수 있는데, 여성에게 잘 보살핀다는 특성이 있다는 가정은 대부분 육아의 책임을 여성에게 지우고 남성은 경력을 쌓는데 집중하게 하는 것으로 귀결될 수 있다.

사회적으로 널리 퍼진 편견 또는 주관적으로 부정적인 고정관념의 대상이 되는 집단은 보다 명확하게 사회-문화적 불이익을 받는다. 특히 그런 집단을 구성하는 특성이 개인의 정체성을 부분적으로 구성할 경우 편견이나 고정관념은 그 대상이 된 집단의 구성원으로서 개인이 자부심을 갖는 능력에도 영향을 미칠 수 있다. 특히 부정적 효과를 낳는 긍정적 고정관념은 사회적 불이익에도 기여하지만 다른 형태의 불이익에도 기여한다.

실제로 사회-문화적 불이익은 다른 종류의 불이익 양상과 인과관계를 갖는다. 사회-문화적 불이익이 정치적 불이익으로 이어지는 원인 중에 하나는 정치인 역시 사회의 구성원으로서 사회에 널리 퍼진 편견과 부정적 고정관념을 공유한다는 점이다. 현실 세계를 초월하기 어려운 정치인들이 공유하는 편견과 고정관념은 공식적인 활동에도 영향을 미친다. 또 다른 원인은 정치인이 만약 현실 세계를 초월하여 그런 믿음을 갖고



있지 않다고 하더라도 정치인을 현실 세계로 끌어들이는 선거는 편견과 고정관념에 따른 행위를 하도록 유인책을 제공하는 사람들에 의해 선출됐다는 것에 대해 정치인으로 하여금 책임을 지도록 요구한다. 이는 정치적으로 인기가 없는 집단에게 해악을 주도록 설계된 법률이 있을 수 있고 그러한 법률에 대해서 보다 면밀한 검토가 필요한 이유이기도 하다. 또한 사회-문화적 불이익과 구체적이고 실질적인 불이익 사이에도 인과관계가 형성되기도 하는데, 집단의 사회-문화적 불이익은 종종 집단에 대한 적대감과 폭력의 원인이 되고 그 집단을 거부하고 고립화시키는 원인이 될 수 있다.

3. 실질적 불이익

실질적 불이익은 우선적으로 소득과 부 같은 경제적 지시체에 따라 결정되지만, 교육과 고용에 대한 접근, 사적으로 또 공적으로 자행되는 폭력과 적대로부터 자유, 장수와 건강을 포괄하는 넓은 것으로 해석할 수 있다.¹³⁵⁾ 학력 수준이 낮거나 구직 기회가 적은 사람은 경제적으로도 낮은 수준에 머물러 있을 확률이 높고, 어떤 특성을 가졌다는 이유로 폭력과 적대의 대상이 될 수 있는 사람은 신체와 정신에 대한 위협으로서 일상생활에 중대한 지장을 받고, 그러한 사유가 작용하여 의료 서비스를 제때 받지 못하거나 비위생적인 환경 속에 놓여 건강이 악화되거나 단명하기 쉬운 사람은 생활의 토대가 항상 불안정한 상태에 놓여 있게 된다.

개인들의 사회-경제적(socio-economic) 안전을 지시하는 이러한 요인들은 결과적으로 그런 개인들로 구성된 집단의 실질적 지위에 영향을 준다. 이러한 지표들의 개인 간 상대적 격차가 어떤 특성을 기준으로 집단화되는 경향이 있다면 그러한 경향은 차별로 포착될 수 있는 실질적 불이익이 될 수 있다. 특히 실질적 불이익에서 차별의 부당성을 찾을 경우 실질적 불이익의 지시체들은 실질적인 불이익을 만회, 극복 또는 보상하기 위한 조치들을 정당화하기 위한 근거로 사용될 수 있다.

이 경우에도 실질적 불이익을 증명하기 위해 통계적 자료에 의존하기 쉽다.¹³⁶⁾ 예를 들어 미합중국에서 흑인으로서는 살아가는 것의 실질적 불이익을 증명하기 위해 통계적 자료가 이용된다.¹³⁷⁾ 흑인 어린이의 기대수명은 백인 어린이보다 5년 더 짧고, 흑인 어린이의 엄마가 그들의 유년기에 사망할 확률은 백인 어린이의 엄마보다

135) Tarunabh Khaitan, *A Theory of Discrimination Law*, Oxford University Press, 2016, pp. 54~56.

136) 통계의 의미를 해석할 때 사회적 맥락이 중요성에 관해서 이준일, 인권법, 제7판, 홍문사, 907~908쪽 참조.

137) *Regents of University of California v. Bakke*, 438 U. S. 265 (1978): 395~396.



세 배가 넘으며, 흑인 가정의 중간 소득이 백인 가정의 중간 소득에 60퍼센트(%)에 그치지만 흑인 어린이가 저소득 가정에서 살게 될 확률이 백인 어린이보다 네 배 높고, 흑인 어린이가 일할 나이가 됐을 때 구직 기회는 백인 어린이에 비해 훨씬 적고, 흑인 성인의 실업률이 백인 성인보다 두 배 높고, 4년제 대학을 졸업한 흑인 남성의 중간 연봉소득이 고등학교 졸업장만 갖고 있는 백인 남성보다 110달러(\$) 밖에 높지 않고, 흑인이 전체 인구의 11.5퍼센트를 차지하고 있지만 변호사와 법관의 1.2퍼센트, 의사의 2퍼센트, 치과 의사의 2.3퍼센트, 기술공학자의 1.1퍼센트, 대학교수의 2.6퍼센트만이 흑인으로 구성되어 있다는 점을 미국에서 수 세기 동안 흑인에게 가한 불평등한 대우가 도달한 비극적 결론을 증명하기 위한 근거로 제시하는 것이다.¹³⁸⁾

IV. 비교적 차별과 독립적 차별

1. 비교적 차별

차별과 비교를 개념적으로 불가분의 관계로 파악하는 경우, 어떤 법이나 정책이 부당하게 차별한다고 하려면 한 사람이 받은 대우는 다른 사람이 받은 대우와 비교되어야 한다.¹³⁹⁾ 예를 들어 X가 부당한 차별로 고통을 겪는지 결정하려면 X가 받은 A라는 대우를 살펴보고 그 대우를 적어도 한 명 이상의 다른 사람이 받은 대우 즉, Y가 받은 B라는 대우 또는 그에 더하여 Z가 받은 C라는 대우 등과 비교해야 한다. 이때 차별은 비교 속에 포함되어 있는데, Y가 B를 받을 때 X가 A를 받는 경우 차별이 발생하기 때문이다. 이렇게 비교적 성격을 강조하는 차별 개념을 ‘비교적 차별’이라고 부를 수 있다.

차별의 부당함을 비교적인 것으로 개념화하는 것이 보다 친숙하고 직관에 부합하는 측면이 있는데, 헬먼(D. Hellman)은 그 원인을 차별에 관한 주장이 일반적으로 비교의 틀을 가지기 때문일 것이라고 추측한다.¹⁴⁰⁾ 원고나 청구인이 차별에 관해서

138) 이러한 통계는 U. S. Dept. of Commerce, Bureau of the Census, Statistical Abstract of the United States 65 (1977); U. S. Dept. of Commerce, Bureau of the Census, Current Population Reports, Series P-60, No. 107 (1977); U. S. Dept. of Labor, Bureau of Labor Statistics, Employment and Earnings, January 1978; U. S. Dept. of Commerce, Bureau of the Census, Current Population Reports, Series P-60, No. 105 (1977) 등을 출처로 삼고 있다.

139) Deborah Hellman, “Two Concepts of Discrimination”, *Virginia Law Review* 102(4), 2016, pp. 895~952: p. 899.

140) Deborah Hellman, “Two Concepts of Discrimination”, *Virginia Law Review* 102(4), 2016, pp. 895~952: pp. 900~901.



‘아프리카계 미국인(흑인)이 어떤 국립대학으로부터 입학 승인을 받은 백인 지원자와 비교할 만한 자격을 갖추었음에도 입학할 거부당했다.’¹⁴¹⁾거나 ‘복지 서비스에 관해 여성 가입자는 자신의 배우자가 부양가족임을 입증해야하는 반면 남성 가입자는 자신의 배우자가 부양가족으로 추정된다.’¹⁴²⁾ 또는 ‘장애인의 공동 주거 시설에는 특별 구역 승인을 요건으로 하지만 다른 공동 주거 시설에 대해서는 그러한 요건이 없다.’¹⁴³⁾ 등의 주장은 모두 비교의 형식적 구조를 갖는다.

‘공무원 및 공·사기업체의 채용시험에 응시할 때 군복무를 마친 신체적으로 건장한 남성에게는 제대군인으로서 가산점을 부여하는 반면 군복무를 하지 않거나 할 수 없는 여성과 장애인에게는 그러한 가산점을 부여하지 않는다.’¹⁴⁴⁾거나 ‘국가가 주관하는 자격시험을 일요일에 시행할 경우 일요일에는 종교 관련 활동만 하라고 가르치는 종교를 신봉하는 사람은 시험에 응시할 수 없지만 그러한 종교를 갖지 않은 사람은 시험에 응시할 수 있다.’¹⁴⁵⁾ 또는 ‘외국인 등록 시스템의 데이터베이스에 거주지 국가의 국적을 가진 사람의 데이터는 저장되어 있지 않은 반면 같은 연합의 다른 회원 국가의 국적을 가진 사람의 데이터는 저장되어 있다.’¹⁴⁶⁾는 주장 역시 비교의 구조로 구성되어 있다.

비교를 차별과 개념 내재적으로 연결시키게 되면 차별에 전제된 비교는 차별을 인식하고 평가하기 위한 필수적 요소가 된다.¹⁴⁷⁾ 차별의 인식과 평가는 항상 둘 이상의 개인 또는 집단을 병렬적으로 배치시킬 때 비로소 가능해진다. 그러나 비교를 수행하기만 하면 곧바로 차별로 인식되고 차별로 평가되는 것인지 아니면 차별로 인식되고 차별로 평가되기 위해서는 별도의 무언가가 더 필요한 것인지에 대해서는 진전된 논의가 필요하다. 특히 비교의 형식적 구조만으로는 비교를 통해 무엇을 발견하려는 것인지 알 수 없다는 입장에 서면 자칫 비교만으로 차별 여부를 결정할 수 없으므로 어떤 실체적 내용이 필요하다는 주장으로 성급하게 넘어갈 수 있다.

141) Sweatt v. Painter, 339 U.S. 629, (1950), 631.

142) Frontiera v. Richardson. 41 1 U.S. 677, (1973), 678~679.

143) City of Cleburne v. Cleburne Living Ctr., 473 U.S. 432, (1985), 437.

144) 헌재 1999. 12. 23. 98헌마363, 판례집 11-2, 770: 778.

145) 헌재 2001. 9. 27. 2000헌마159, 판례집 13-2, 353: 357.

146) Heinz Huber v Bundesrepublik Deutschland, C-524/06, Judgment of the Court (Grand Chamber) of 16 December 2008 및 Oberverwaltungsgericht NRW(Nordrhein-Westfalen), 17 A 805/03, 24 June 2009.

147) 평등과 차별의 핵심을 ‘비교’와 ‘평가’에 두고, 평등과 차별의 문제를 ‘비교대상에 대한 비교판단의 결과에 따른 비교대우에 대한 평가’로 설정하는 경우는 이준일, “차별, 소수자, 국가인권위원회”, 헌법학연구 18(2), 2012, 177~222쪽: 178~180쪽 참조.



헬먼 역시 “비교하는 권리는 우리에게 비교하기를 통해 찾는 것이 무엇인지 아직 말해 주지 않는다.”고 간단히 언급하고 곧바로 “형식적 구조는 어떤 실체적 표준에 의해 보완되어야 한다.”는 주장으로 넘어간다.¹⁴⁸⁾ 그러나 비교의 형식적 구조를 통해 아무 것도 발견할 수 없는 것은 아니다. 비교는 ‘어떤 사람 X와 다른 어떤 사람 Y를 비교한다.’거나 ‘어떤 사람 X가 받은 대우 A와 다른 어떤 사람 Y가 받은 대우 B를 비교한다.’는 명제로 표현된다. 이를 면밀히 관찰해 보면 비교의 형식적 구조는 ‘... 와(과) ...’이다. 비교는 ‘... 와(과) ...’의 형식적 구조에서 수행된다. 비교를 하려면 자기‘와’ 타자를 분리 또는 구별하고, 타자‘와’ 또 다른 타자를 분리 또는 구별해야 한다. 그럼으로써 타자와 자기 사이의 차이를 발견하거나 타자와 또 다른 타자 사이의 차이를 발견한다.

이러한 전개는 ‘분리 또는 구별 없이 비교는 불가능하고 비교하지 않으면 차이를 발견할 수도 없다.’는 하나의 논제를 형성한다. 이때 차이를 발견한다는 것의 의미는 ‘있는’ 차이를 확인하는 것으로만 볼 수 없고 ‘없는’ 차이를 새롭게 만들어 내는 것도 포함한다. 차이를 부각시키는 것은 보이지 않던 것 또는 잘 볼 수 없었던 것을 선명히 보이게 해준다. 그렇다면 비교적 차별 개념은 적어도 차이로서 차별을 인식하는 개념으로 활용될 수 있다.

2. 독립적 차별

차별이 다른 사람의 상황과 비교하는 것이 아니라면, 어떤 사람에 대해 법이나 정책이 부당한 차별을 하는 것이라고 인식하거나 평가하기 위해서는 그 사람의 특성에 적합하게 상응하는 독립적인 기준이 필요하다. 이러한 관점에서 차별의 인식과 평가는 어떤 사람이 받은 대우를 다른 사람이 받은 대우가 아닌 별도로 마련되어 있는 기준과 비교하는 것이다. 예를 들어 X가 차별을 받았는지 평가하기 위해 받은 대우를 자신이 받아야만 하는 별도의 기준과 비교하는 것이다.¹⁴⁹⁾ ‘독립적 차별’¹⁵⁰⁾이라고 부를 수 있는 이러한 차별 개념에 따르면 자신이 받은 대우나 처한

148) Deborah Hellman, “Two Concepts of Discrimination”, *Virginia Law Review* 102(4), 2016, pp. 895~952: p. 901.

149) Deborah Hellman, “Two Concepts of Discrimination”, *Virginia Law Review* 102(4), 2016, pp. 895~952: p. 899.

150) 헬먼(D. Hellman)은 “비교적(comparative)” 차별에 대응하는 개념을 만들기 위해 “비(非)비교적(noncomparative)” 또는 “독립적(independent)”이라는 용어를 번갈아 가며 사용한다. 한글 번역에서는 ‘비’가 연달아 중복되는 ‘비비교적’이라는 조어는 부자연스러울 뿐만 아니라 엄밀한 의미에서 ‘비비교적’ 차별 역시 별도의 기준으로 비교가 실행된다는 점에서 보다 자연스럽고 의미에 충실한 ‘독립적’이라는 단어를 사용하기로 한다. “차별의 비비교적 개념(the noncomparative



상태를 다른 사람이 받은 대우나 처한 상태와 비교하는 것은 자신이 받은 대우나 처한 상태에서 무엇이 부족한지 그 결핍 상태를 중요하게 부각시켜주지만 그것이 대우나 상태 자체를 부당하게 만들지는 않는다. 수익 또는 부담에 관한 내용의 어떤 법이나 정책이 누군가를 부당하게 차별한다고 하려면 단지 그 법이나 정책의 대상에 다른 사람들은 포함하고 어떤 사람들은 배제함으로써 발생하는 결핍 상태 만으로는 부족하고 그 결핍을 초래한 처우가 그 대상에서 배제된 사람들이 마땅히 받아야 하는 처우의 기준을 벗어나야 한다.

그런데 이러한 표준으로서 기준이 정해지게 되면 비교는 더 이상 작동을 멈추고 그 어떤 실제 작용을 하지 못한다.¹⁵¹⁾ 예를 들어 대학 입학 전형에서 불합격 처분을 받은 지원자는 합격 처분을 받은 백인 지원자처럼 대학 입학에 대해 관련성이 있지만, 합격 기준이 목표하는 바가 학업을 가장 잘 수행하는 것이라면 중간등급의 지원자는 상위등급의 지원자와 관련성이 없다. 중간등급의 지원자와 상위등급의 지원자는 다르기 때문이다. 예를 들어 중간등급과 상위등급을 가르는 기준이 정해지게 되면 중간등급에 해당하여 입학이 거절된 흑인은 자신이 받은 불합격 처분을 상위등급에 해당하여 입학이 허락된 백인의 합격처분과 비교하는 것이 아니라 상위등급과 중간등급을 가르는 기준이자 합격여부를 결정하는 기준으로서 표준에 따라 받아야 할 처분을 따져봐야 한다. 물론 엄밀히 분석한다면 이때에도 비교를 하지 않는 것은 아니다. 자신이 실제로 받은 처분과 입학 기준에 따라 자신이 받아야 할 처분을 비교하기 때문이다.

만약 입학이 거절된 흑인 역시 상위등급일 경우라고 할지라도 비교의 기준으로서 표준은 달라지지 않는다. 언뜻 보면 이 경우에 같은 상위등급인 흑인의 불합격 처분에 적용된 기준을 백인의 합격 처분에 적용된 기준과 비교하는 것처럼 보인다. 그러나 독립적 차별 개념에 따르면 이러한 경우에도 비교의 기준이 되는 것은 같은 상위등급에서도 합격처분을 받은 백인에게 적용된 기준이 아니라 자기에게 적용되어야 하는 기준으로서 중간등급과 상위등급을 구분하는 기준이자 합격여부를 판단하는 일반적 기준이다. 상위등급의 백인이 받은 합격처분은 이러한 일반적 기준이 적용된 또 다른 예에 불과한 것이 된다. 상위등급의 백인 합격자가 없는 경우에도 일반적 입학 기준에 따라 상위등급인 경우 합격 처분을 받아야 한다면 상위등급에 해당하지만 불합격 처분을 받은 흑인은 그러한 별도의 독립적 기준에 따라 이의를 제기할 수 있는 것이다.

conception of discrimination)”에 관한 헬먼의 논의는 Deborah Hellman, “Two Concepts of Discrimination”, *Virginia Law Review* 102(4), 2016, pp. 895-952: pp. 909-922 참조.

151) Deborah Hellman, “Two Concepts of Discrimination”, *Virginia Law Review* 102(4), 2016, pp. 895-952: p. 912.



이렇게 차별에 대해 독립적 기준을 적용하는 것은 차별에 관한 권리를 자유 또는 자율성에 관한 권리와 구별하기 어렵게 한다. 차별 여부를 판단하기 위한 실체적 기준을 인간으로서 가져야 할 권리에서 찾는 것처럼 보이기 때문이다. 그러나 차별을 독립된 관점으로 접근하는 것이 전혀 무의미한 것만은 아닌데, 독립적 차별 개념은 비교적 차별 개념의 문제로 제기되는 하향평준화(levelling down)에 대한 개방성에 적절히 대응할 수 있게 해준다.¹⁵²⁾

비교적 차별 개념은 다른 사람에게 적용된 기준이 확정되고 확인되기 전까지 차별 여부를 판단할 수 있는 실체적 기준이 마련되지 않을 뿐만 아니라 차별 여부의 평가는 다른 사람에게 적용된 기준에 의존한다. 이러한 개념은 평등에 관한 홈즈(E. Holmes)의 구분에 따르면 엄격한 의미의 평등 개념과도 연결된다.¹⁵³⁾ 따라서 다른 사람이 받은 대우에만 관심이 있지 그것이 상향이건 하향이건 문제 삼지 않는다. 대학 입학 전형에 관한 앞의 예에서 상위등급의 흑인 지원자가 불합격 처분을 받은 경우에 비교적 차별 또는 엄격한 평등 개념에 따르면 같은 상위등급의 백인 지원자도 불합격 처분을 받았다면 차별이라고 인식되거나 평가되지 않는다. 그러나 상위등급에 해당할 경우 합격할 수 있는 기준이 정해져 있다면 입학이 거절된 상위등급의 흑인 지원자뿐만 아니라 입학이 거절된 상위등급의 백인 지원자 모두 차별을 주장할 수 있게 되어 기준의 수준이 하향하는 것을 막을 수 있다.

하지만 이 경우에도 하향을 제한하는 기준이 최저 수준인지 최고 수준인지 아니면 그 사이 어디쯤인지 결정하는 문제가 남게 된다. 나아가 이 문제는 사회적 급부의 기준을 어느 수준에 맞추어야 하는 것인지에 관한 복지법의 문제와 중첩되거나 구별하기 어렵다는 별도의 문제점도 남긴다. 만약 사실적 급부에 관한 권리를 중심으로 하는 사회권이 비례성원칙에 의해 과소보호금지를 명령하는 것으로 보는 기본권이론에 따른다면 독립적 차별 개념으로 구성된 차별 받지 않을 권리는 사회권과 중첩되거나 구별되기 어려운 불필요한 중복이 될 수도 있다. 그러나 독립적 차별 개념으로 구성된 차별에 관한 권리가 어떤 기준을 설정하고 있다면 오히려 그 기준이 사회권에서 정하는 급부의 기준에 영향을 미칠 수 있다는 구성 역시 양립 가능하다. 독립적 차별 개념이 대우나 상태의 기준이 되는 수준의 하향화를 제어하는 기능을 갖는다는 것은 사회권의 대상인 급부 수준을 결정할 때 고려해야하는 중요한 조건으로 작용할 수 있다는 함의를 갖는다.

152) 상향평준화에 대한 계기로 차별과 존엄성의 연결에 관해서 본 논문 「제3장 제3절 IV. 3. 인간의 존엄성과 차별」 참조.

153) 본 논문 「제3장 제3절 II. 2. 형식적 평등에서 ‘같게’와 ‘다르게’의 구별」 참조.



제4장

머신러닝 알고리즘의 결정과 차별의 판단

지능적 에이전트로서 머신러닝 알고리즘은 인간의 고유 영역이라고 생각되어 왔던 인식과 판단 작용에 단순히 영향을 미치는 수준을 넘어 대체하는 수준을 향해 발전해 가고 있다. 그럼에도 불구하고 비전문가로서 일반인의 관점에서 머신러닝 알고리즘 또는 인공지능은 어디까지나 기계 또는 기술일 뿐 인간은 아니라는 범주적 구분과 기계 또는 기술은 인간을 이롭게 하는 도구에 불과하다는 기술 철학적 고정 관념에 기대어 그 실재적 능력이 과소평가되는 경향이 있다. 반면에 머신러닝 알고리즘과 인공지능 연구의 선두에 서서 각각의 기술적 한계를 직접 마주하는 전문가들은 그 잠재적 능력이 과학기술에 대한 막연한 환상에 기대어 과대평가되는 것을 경계한다. 특히 이러한 경계는 기술의 능력을 과대평가하여 그로 인한 우려가 정책에 반영되거나 법에 의한 기술의 통제가 이루어질 경우 기술의 진보를 가로막을 수 있다는 자기정당화 논거로 그 타당성을 획득하기도 한다. 그러나 기술은 사회로부터 구성될 수 있을 뿐만 아니라 역으로 사회를 구성할 수도 있다는 점을 강조하면 기술의 중립성은 이와 양립 불가능하고 기술 진보의 필요성만으로 법적 규제 필요성을 선형적으로 극복할 수는 없다.

기술이 사회와 맺는 관련성은 법이 사회와 맺는 관련성과 유사한 측면이 있고 사회의 규제 메커니즘으로 교합될 수 있다. 이러한 사회의 규제 메커니즘이 오프라인 시대에는 인간이 물리적 구조를 설계하고 그 구조물을 현실 세계에 실현하는 방식으로 작동했다면 온라인 시대를 넘어 온라인 시대에는 머신러닝 알고리즘이 사이버-물리적 구조를 설계하고 실현하는 방식으로 작동한다. 이는 법의 지배가 알고리즘의 지배로 대체될 수 있는 환경이 구축되고 있다는 것이기도 하다. 이러한 환경에서 추론에 기초한 머신러닝 알고리즘은 차별을 쉽게 우회할 수 있는 간극을 만들 수 있다. 이러한 간극은 기존 차별금지법의 구조 및 이론과의 사이에서 생기는 것일 수도 있지만 최적화에 지향된 법적 결정 그 자체와의 사이에서 발생하는 것일 수도 있다. 그러나 이러한 간극은 머신러닝 알고리즘의 작동원리와 차별금지법의 작동원리를 실제로 알 수 있을 때 확인할 수 있을 것이다. 머신러닝 알고리즘을 무지의 베일 안에 가려주는 법적 장치들이 헌법 차원의 긴장 관계를 형성하고 유지하고 있다면 간극의 확인은 가설적 수준에서 쉽게 벗어나기 어려울 것이다.



제1절 알고리즘의 지배와 법의 지배

인류의 유전자나 문화를 통해 전승되었을 기술에 관한 사유의 한 조각을 찾아가 보면 서양에서 기술은 일종의 기예(art)로 여겨졌다. 그리고 이런 기술은 기예로만 남지 않고 국가를 주조할 수 있는 것으로 보았다. 플라톤(Platon)에 따르면 기술들은 분명 그것들이 관여하는 대상을 지배하고 제어한다. 국가를 배에 비유했던 플라톤은 배를 만드는 기예처럼 기술을 통해 국가를 만들 수 있다고 믿었다.¹⁾ 동양에서 기술은 재주(技)이며, 눈으로 보는 것(目視), 즉 감각할 수 있는 기관으로 아는 것(官知)이다. 장자(莊子)는 잘 살아가는 본성(養生主)에 관해 논하면서 끝이 있는(有涯) 삶(生)을 살면서 끝이 없는(無涯) 앎(知)을 추구하는 것을 위태로운 것으로 본다. 기(技)를 넘어 선 것이 도(道)이며 도에 따라 사는 것은 자연의 결(天理)에 따라서 사는 것이다.²⁾ 대상을 지배하고 제어하는 기술의 특성과 지식을 추구하는 기술의 특성은 불가분이다. 관여 대상을 지배하고 제어하기 위해 그에 관한 지식은 필수적 이기 때문이다.

I. 사회의 규제 메커니즘으로서 법과 기술

1. 기술적 설계를 통한 사회의 구조화

기술의 가치는 효율성과 창의성에 있다. 기술은 같은 결과를 산출하는데 들어 가는 힘과 시간을 줄여줄 뿐만 아니라 이전에는 나올 수 없었던 새로운 결과를 산출해 낸다. 기계로 상징되는 기술은 생산력을 높여주는 데에 그치지 않고 산업의 구조까지 바꾼다. 몇 세기 전의 용어이지만 증기 기술과 전기 기술은 산업에 혁명을 일으켰다. 기계의 생산력은 인간의 육체적 노동에 의한 생산력을 압도하면서 인간을 대체해 왔다. 그리고 “표명된 사회의 복잡성이 은폐된 전자공학의 복잡성으로 대체”³⁾되는 가운데 전자공학의 기술로서 알고리즘은 사회 질서로서 법을 대체해 가고 있다. 통치(governance), 지배(ruling), 규제(regulation), 통제(control) 등의 관점

1) Platon, 국가[Politeia], 천병희(역), 숲, 2013, 56~66쪽.

2) 莊子, 장자(莊子), 조현숙(역), 책세상, 2016, 양생주 편: 73~78쪽.

3) Langdon Winner, *Autonomous Technology*, MIT Press, 1977, p. 285: “Manifest social complexity is replaced by concealed electronic complexity.”



에서 보자면 ‘알고리즘의 지배’⁴⁾가 ‘법의 지배’⁵⁾를 대체해 가고 있다고도 할 수 있다. 오프라인 환경에서는 물론 온라인 환경에서는 정보통신기술에 힘입어 그 위력은 더 강해져 왔다.

위너(L. Winner)에 따르면 “기술적 구조에는 권력, 권위, 자유와 사회정의의 조건들이 깊이 각인”⁶⁾될 수 있다. 예를 들어 도로 위를 지나는 다리를 설치하면서 다리의 높이를 결정하려고 할 때 교량 건축 기술에서 높이 설정 문제는 단순히 어떤 숫자를 결정하기만 하면 되는 것처럼 보인다. 그런데 이렇게 중립적으로 보이는 문제를 해결하기 위해 높이가 2미터 안팎인 다리를 설치하기로 결정했다고 하더라도 물리적으로 2미터를 넘는 차량은 그 다리가 설치된 도로를 이용해 다리 아래를 통과할 수 없다. 이러한 높이의 다리가 먼저 설치되고 그 이후에 차량을 제조한다면 이 차량이 2미터 높이의 다리가 설치된 도로를 무사히 통과할 수 있는 필요조건은 차체의 높이를 2미터로 제한하는 것이다. 그러나 이미 여러 가지 사양으로 제조되어 운행되고 있는 차량이 있는 상태에서 도로 위에 다리를 설치하면서 그 높이를 결정할 때에는 다리 아래를 통과하는 도로를 이용하는 차량의 사양이 결정의 조건이 된다. 그리고 이미 트럭이나 버스처럼 2미터가 넘는 차량이 있음에도 불구하고 다리의 높이를 2미터로 설계하는 것은 그러한 종류의 차량이 다리 아래를 지나는 도로를 이용하지 못하게 하는 효과를 낳는다.

-
- 4) 알고리즘의 지배(통치) 또는 알고리즘에 의한 지배(통치)에 관해서 Florian Saurwein · Natascha Just · Michael Latzer, “Governance of Algorithms: Options and Limitations”, *Info* 17(6), 2015, pp. 35~49; Danilo Doneda · Virgilio A. F. Almeida, “What Is Algorithm Governance?”, *IEEE Internet Computing* 20(4), 2016, pp. 60~63; Natascha Just · Michael Latzer, “Governance by Algorithms: Reality Construction by Algorithmic Selection on the Internet”, *Media, Culture & Society* 39(2), 2017, pp. 238~258; Franco Zambonelli · Flora Salim · Seng W. Loke · Wolfgang De Meuter · Salil Kanhere, “Algorithmic Governance in Smart Cities: The Conundrum and the Potential of Pervasive Computing Solutions”, *IEEE Technology and Society Magazine* 37(2), 2018, pp. 80~87 참조.
- 5) ‘법의 지배(the rule of law)’라는 표현은 영어권에서 다이시(A. V. Dicey)의 ‘헌법학입문’에서 처음 사용된 것으로 알려져 있다. 다이시는 ‘법의 지배’의 의미를 보통법원이 통상의 절차에 따라 확립한 판결에 의하지 아니하고는 누구도 처벌이나 신체의 훼손, 재산의 박탈을 당하지 않는다는 점, 누구도 법 위에 있지 않고 법 앞에 평등하므로 보통법과 보통법원의 재판권에 복속된다는 점, 그리고 영국 제도의 특별한 속성으로서 법의 지배 또는 우위는 개인의 권리가 헌법의 일반원리로부터 도출되는 것이 아니라 구체적 사건에 대한 법원의 사법적 결정에서 비롯되는 특수성을 갖는다는 점에서 찾는다. 이에 관한 자세한 내용은 Albert Venn Dicey, *Introduction to the Study of the Law of the Constitution*, E. C. S. Wade(Ed.), 10th ed., 1959[1st, 1885], pp. 187~203 참조; 헌법의 법치주의 모델에 대한 분석은 남중권, “헌법의 몇 가지 법치주의 모델 -개념과 구조-”, *법학연구* 59(3), 2018, pp. 1~34쪽 참조.
- 6) Langdon Winner, “Techne and Politeia”, *The Whale and the Reactor: A Search for Limits in an Age of High Technology*, University of Chicago Press, 1986, pp. 40~58: p. 40.



그리고 이러한 효과는 차량의 통과 여부에 그치지 않는다. 다양한 사회적 맥락과 결합하여 여러 가지 효과를 양산할 수 있기 때문이다. 2미터 높이의 다리가 설치된 도로가 도심으로 이어지는 주요 도로라면 트럭이나 버스 등 다수의 화물과 여객을 운송하는 차량은 도심에 진입할 수 없다. 게다가 트럭이나 버스의 운전자 그리고 버스의 승객들이 대부분 도심 밖에 거주지를 두고 있거나 경제적 소득이 낮거나 특정 인종이나 성으로 구성된 경우 이러한 특성을 가진 사람들 역시 버스를 이용해만 들어갈 수 있는 도심에 진입할 수 없다. 반론을 위해 우회 도로의 이용이나 도보를 상정할 수도 있겠으나 이것이 더 많은 시간과 비용을 들게 하여 높이 2미터의 조건을 충족하지 못하는 트럭이나 버스의 운전자와 이용자에게 불리하고 불편한 상황을 만들어 도심 진입에 제한을 가한다는 점에는 변함이 없다. 교량 건축 기술에서 높이 설정 문제가 단순히 어떤 숫자를 결정하기만 하면 되는 중립적인 문제라고만 볼 수 없는 이유가 여기에 있다.

위너는 이러한 기획이 실현된 실제 사례를 뉴욕시에 있는 롱아일랜드의 고가도로에서 찾는다.⁷⁾ 1920년대부터 1970년대까지 뉴욕의 도로, 공원, 교량 건설 및 그 밖에 다른 공공사업을 맡았던 건축가 모지스(R. Moses)는 자신이 설계한 공원도로에 버스가 통행할 수 없는 사양에 따라 고가도로를 설계했고, 그러한 설계에는 모지스가 갖는 사회의 계급에 대한 편향과 인종에 대한 편견이 반영됐다는 것이다.⁸⁾ 모지스는 그가 “상류 및 안락한 중간(‘upper’ and ‘comfortable middle’)”⁹⁾ 이라고 부른 계급으로서 자동차를 소유한 백인은 기분전환 또는 출퇴근을 위해 공원도로를 자유롭게 이용할 수 있지만, 통상 대중교통을 이용하는 가난한 사람과 흑인은 공원도로에 가까이 올 수 없도록 하려고 9피트(약 2m 74cm) 높이로 고가도로를 설계하여 건축했다. 그 당시 버스의 높이는 12피트(약 3m 66cm)였는데, 3피트(약 91cm)만 높이를 낮추는 매우 간단한 기술적 고려를 통해 중대한 사회적 효과를 발생시키는 기술-사회적(techno-social) 구조를 만들어 낼 수 있었던 것이다.

7) Langdon Winner, “Do Artifacts Have Politics?”, *The Whale and the Reactor: A Search for Limits in an Age of High Technology*, University of Chicago Press, 1986, pp. 19~39: p. 22 이하 참조.

8) 모지스(R. Moses)의 전기 작가 카로(R. A. Caro)에 따르면 모지스는 흑인이 “본질적으로 불결하다 (inherently ‘dirty’)”는 생각을 가지고 있었다. Robert A. Caro, *The Power Broker: Robert Moses and the Fall of New York*, Vintage Books, 1975[1974], p. 318.

9) Robert A. Caro, *The Power Broker: Robert Moses and the Fall of New York*, Vintage Books, 1975[1974], p. 480.



2. 기술적 조치를 통한 물리적·심리적 제한

물론 의도나 효과는 다를 수 있지만 비슷한 기술적 조치들은 현재의 일상에서도 흔히 볼 수 있다. 백화점이나 대형마트의 주차장 진입로에 일정한 높이의 장애물을 설치하여 진입 가능한 차량의 종류를 제한할 수 있고, 주차장 곳곳에 과속 방지턱을 설치해서 주차장 내부에서 차량의 운행 속도를 제한할 수 있다. 또한 아파트 단지 내부에 차로를 만들지 않고 우회도로를 건설하여 보행 주민의 안전을 확보할 수도 있지만, 아파트 건물 내부로 이어지는 주차장의 높이가 제한되어 있을 경우 적재함이 있는 차량의 진입이 사실상 제한될 수도 있다.¹⁰⁾ 이 경우 화물 운반자는 차량 적재함의 높이를 낮추거나 아파트 단지 밖에서 물건을 내려 직접 이동시켜야 하는 불이익을 받는다.¹¹⁾ 고속도로나 자동차 전용도로를 비롯해 도로 곳곳에 설치된 과속차량 단속 카메라는 실제로 과속차량을 적발하는 데에도 사용되지만 카메라가 설치된 지점에서 차량 운전자가 단속을 피하기 위해 감속하도록 유도함으로써 차량 속도를 실제로 제한하는 효과도 있다. 특히 단속 카메라 같은 감시 장치에 사용되는 기술은 물리적 제한을 심리적 제한으로까지 확장시킨다. 누군가 지켜보고 있는 것 같다는 인식과 경험으로 형성된 심리적 사실은 위축된 행동으로 나타나는 데에 영향을 미칠 수 있다.¹²⁾

법적 조치가 취해질 때 이를 실현할 수 있는 기술적 조치가 마련되어 있는 경우도 있지만 없는 경우도 있다. 예를 들어 국외에 거주하는 국민에게 선거권이 보장된다고 해도 거주 국가에서 투표할 수 있는 기술적 조치가 마련되어 있지 않다면 헌법적 권리는 유명무실하다.¹³⁾ 시민의 자유를 제한하는 여러 가지 법적 조치가 취해진다고

10) 주택법[법률 제15356호, 2018. 1. 16. 개정] 제35조의 위임 및 주택건설기준 등에 관한 규정 [대통령령 제28628호, 2018. 2. 9. 개정] 제27조 제5항의 위임을 받은 주택건설기준 등에 관한 규칙 [국토교통부령 제471호, 2017. 12. 26. 개정] 제6조의2 제2호에 따라 준용되는 주차장법 시행규칙 [국토교통부령 제498호, 2018. 3. 21. 개정]에 따르면 지하식 또는 건축물식 노외주차장의 차로의 “높이는 주차바닥면으로부터 2.3미터 이상으로 하여야 한다.”(동 규칙 제6조 제5호 가목) 그런데 주차장 높이가 시행규칙의 기준인 2.3미터는 넘지만 대부분의 택배용 화물 차량의 높이에 해당하는 2.5미터에서 2.6미터를 넘지 않을 경우 이러한 화물 차량은 아파트단지의 지상으로도 지하로도 통행할 수 없게 된다.

11) 정부는 주택건설기준 등에 관한 규칙을 개정하여 지상공원형 아파트에 대해 지하주차장 층고를 ‘2.7미터 이상’으로 상향 조정하는 안을 제시함으로써 이 문제에 대처했다. 주택건설기준 등에 관한 규칙 일부개정령(안) 입법예고(2018년 6월 20일, 국토교통부공고 제2018-790호) 참조.

12) Neil M. Richard, “The Dangers of Surveillance”, *Harvard Law Review* 126, 2013, pp. 1934~1965 참조.

13) 법률에서 선거권에 관하여 거주요건을 두지 않는 경우 국외 거주 국민도 선거를 할 수 있지만 그렇게 하는 것은 “선거기술상 불가능”하다는 이유로 거주요건을 두는 것이 합헌적(가각)이라고 보았던 결정(헌재 1999. 1. 28. 97헌마253 등, 판례집 11-1, 54: 62)은 비록 “선거기술상의



하더라도 이를 집행하고 실행할 수 있는 기술적 수단이 없다면 시민의 자유는 사실상 제약되지 않는 효과가 있다. 반대로 어떤 기술적 수단에 대한 법적 규제가 없거나 허용되어서 법적으로 자유가 보장된다고 하더라도 일상생활에 밀접한 영향을 미치는 기술적 수단이 규제적 성격을 가질 경우 법적으로 보장된 자유는 사실상 제약된다. 레식(L. Lessig)은 사이버공간의 주권을 논하면서 이와 유사한 논의를 펼친다.¹⁴⁾ 공산주의 국가의 예로 베트남을 거론하며 베트남 사회가 미국 사회보다 법의 측면에서 더 강압적이고 광범위한 규제를 갖고 있을 수도 있고 그렇지 않을 수도 있지만, 생활구조상 국가에 의한 실제적인 규제는 불가능하다는 것이다. 그 이유로 통제의 하부구조가 존재하지 않는다는 점을 든다. 국가의 규제규범이 있을지라도 이를 실행할 구조가 없다면 실질적인 사실상의 자유가 있는 것이다. 결국 구조는 규제를 억제할 뿐만 아니라 가능하게도 한다.

3. 규제 메커니즘의 실현 조건

레식은 통제 받는 사람을 중심으로 할 때 규제가 법, 사회규범, 시장, 구조에 의해 실현되는 것으로 본다.¹⁵⁾ 흡연을 예로 들면, 법은 금연구역을 지정하여 일정 장소에서 흡연하는 것을 금지하고,¹⁶⁾ 간접흡연을 방지하도록 의무를 부과할 수 있다.¹⁷⁾ 그러나 실제로 흡연을 적발하거나 이를 법정에서 다루는 경우는 드물다. 오히려 법보다는 길거리에서 담배를 피우지 않거나 식사 중에 담배를 피우지 않아야 한다는 사회규범의 규제를 받는다. 시장에서 담배가격이 상승하면 동일한 소득에서 담배를 피울 수 있는 능력은 상대적으로 규제되고, 담배를 제조하는 기술이 담배에 니코틴 함유량을 강하게 하거나 독한 향이 나도록 그 구조를 설계하면 건강상의 제약 또는 흡연 장소의 제약으로 사실상 흡연이 규제된다. 각각의 규제에는 형벌이나 과태료, 사회적 비난, 가격의 부담, 그리고 불편함이나 육체적 부담 같은 장치가 동원된다.

어려움"이 있다고 하더라도 그러한 기술상의 문제가 재외국민의 선거권 행사를 전면적으로 박탈하기 위한 합당한 사유라 보기 어려워 거주요건을 두는 것이 위헌적(헌법불합치)이라는 결정(헌재 2007. 6. 28. 2004헌마644 등, 판례집 19-1, 859: 877)으로 변경되기도 했다.

14) Lawrence Lessig, *Code: And Other Laws of Cyberspace, Version 2.0*, Basic Books, 2006, pp. 281~282.

15) Lawrence Lessig, *Code: And Other Laws of Cyberspace, Version 2.0*, Basic Books, 2006, pp. 121~125.

16) 국민건강증진법은 공중이 이용하는 시설의 소유자·점유자 또는 관리자에게 해당 시설의 전체를 금연구역으로 지정하고 금연구역을 알리는 표지를 설치하도록 의무를 부과한다(동법 제9조 제4항).

17) 공동주택관리법 제20조의2: “공동주택의 입주자등은 발코니, 화장실 등 세대 내에서의 흡연으로 인하여 다른 입주자등에게 피해를 주지 아니하도록 노력하여야 한다.”(동조 제1항)



규제의 여러 요소들이 상호 지지하거나 반대하는 논거로 작용하면서 전체적인 규제를 강화하거나 약화하는 형태로 영향을 미치는데 이러한 규제 방식은 사이버 공간에도 적용된다. 그리고 사이버공간의 강력한 규제자로서 광범위하게 자유를 위협하는 것은 코드이다. 레식은 19세기 중반부터 20세기 중반까지 자유를 위협했던 규제자가 사회규범, 국가 권력, 시장이라고 한다면, 20세기 말에서 21세기에 이르는 시기에 주목해야 할 새로운 규제자는 코드라고 주장한다.¹⁸⁾ 코드(code)는 법률을 의미하기도 하지만 사이버공간을 구성하는 소프트웨어나 하드웨어에 삽입되어 있는 명령어를 의미하기도 한다.¹⁹⁾ 코드는 네트워크의 연결로 창출된 사이버공간에 형성되어 있는 환경이자 구조다. 알고리즘은 이러한 명령어로서 코드가 진행하는 절차이자 순서도로서 작동의 연속성을 보장함으로써 코드로 구성된 환경적 구조를 체계화함으로써 규제적 기능을 수행한다. 오늘날 알고리즘의 규제는 단순한 은유에 머물러 있지 않고 실제로 사이버-물리 세계를 규제한다.²⁰⁾

II. 데이터 알고리즘 사회와 알고리즘의 지배

1. 데이터 알고리즘 사회

볼킨(J. Balkin)은 “어떤 결정을 하고 그 결정을 실행하기도 하는 알고리즘, 로봇, 인공지능 에이전트가 만들어 내는 사회적·경제적 결정에 따라 조직되는 사회”²¹⁾를 ‘알고리즘의 사회(algorithmic society)’라고 하면서, 우리 시대가 인터넷 사회에서 알고리즘 사회로 급속히 이동하고 있다고 진단한다. 그리고 그 이면에 있는 빅데

18) Lawrence Lessig, *Code: And Other Laws of Cyberspace, Version 2.0*, Basic Books, 2006, pp. 120~121.

19) 코드의 중의적 의미를 구별하기 위해 레식은 법률을 제정하는 미국 의회의사당(United States Capitol)이 있는 워싱턴 D. C.(Washington, District of Columbia.)가 동부 연안에 위치해 있고, 첨단기술을 연구하고 개발하는 연구단지로서 실리콘밸리(Silicon Valley)로 유명한 샌프란시스코(San Francisco)가 서부 연안에 있다는 점에 착안하여 법률로서 코드를 ‘동부 연안 코드(East Coast Code)’로 기술적 명령어로서 코드를 ‘서부 연안 코드(West Coast Code)’로 구별해서 부른다. Lawrence Lessig, *Code: And Other Laws of Cyberspace, Version 2.0*, Basic Books, 2006, pp. 72~74 참조.

20) Tim O'Reilly, “Open Data and Algorithmic Regulation”, in *Beyond Transparency: Open Data and the Future of Civic Innovation*, Brett Goldstein · Lauren Dyson(Eds.), Code for America Press, 2013, pp. 289~300: p. 291; 심우민, “인공지능의 발전과 알고리즘의 규제적 속성”, 법과 사회 53, 2016, 41~70쪽 참조.

21) Jack M. Balkin, “The Three Laws of Robotics in the Age of Big Data”, *Ohio State Law Journal* 78(5), 2017, pp. 1217~1241: p. 1219.



이터는 알고리즘 사회를 가동시키는 연료²²⁾이면서 또한 알고리즘 사회의 산물로서 알고리즘 사회의 한 축을 담당하는 것으로 본다. 이른바 ‘데이터 알고리즘 사회’라고 부를 수 있는 이러한 사회의 기반구조는 데이터와 알고리즘으로 형성된다. 데이터 알고리즘 사회에서는 어떤 활동을 하더라도 데이터가 발생하고, 어떤 정보도 데이터로 환원될 수 있다. 그리고 머신러닝이나 데이터 마이닝 같이 알고리즘을 활용하는 기술은 이러한 데이터를 구조화한다. 데이터를 구조화하는 것은 데이터를 구조 속에 편입시키는 것이기도 하지만 데이터의 관계를 설정할 구조를 만드는 것이기도 하다. 또한 머신러닝 알고리즘은 그 데이터를 분석하여 다양한 결정을 내리고 이를 직접 실행함으로써 기존의 시스템을 변경하고 재구축한다.

2. 온라인 생활양식의 확장

기계와 인간의 역사에서 인간의 육체노동을 대체하는 기계가 개발됨으로써 노동자로서 인간은 기계를 조작하거나 기계가 할 수 없는 부분을 메우는 영역으로 이동하거나 밀려났다. 기계는 물리적인 결과물을 생산하는 것에서 이제는 마치 지능을 가진 것처럼 정보와 지식을 생산해 내고 있다. 기계는 인간의 물리적 환경을 만들어 내던 것에서 나아가 정신적 환경까지 창조해 내고 있다. 이에 따라 기업 활동에서 컴퓨터와 인터넷, 그리고 데이터가 차지하는 비중은 매우 높아졌다. 온라인과 오프라인 생활의 구별이 더 이상 무의미해지는 생활은 사이버-물리 시스템(cyber-physical system)을 환경으로 삼는다. 플로리디(L. Floridi)는 21세기 초반을 살아가는 세대가 “온라이프와 온라인의 명확한 차이를 경험하는 마지막 세대”²³⁾일 것이라고 하면서 “온라인(online) 상태에 있는지 오프라인(offline) 상태에 있는지 묻는 것이 더 이상 사리에 맞지 않는 초연결 현실(hyperconnected reality)의 새로운 경험을 언급하기 위해”²⁴⁾ 신조어로 ‘온라이프(onlife)’를 고안하여 개념화한다.²⁵⁾

22) “Fuel of the Future; The Data Economy”, *The Economist*, London, 423(9039), 6 May 2017, pp. 17-20.

23) Luciano Floridi, “A Look into the Future Impact of ICT on Our Lives”, *The Information Society* 23(1), 2007, pp. 59-64: p. 62.

24) Floridi, Luciano, “Introduction”, in *The Onlife Manifesto: Being Human in a Hyperconnected Era*, Luciano Floridi(Ed.), Springer, 2015, pp. 1-3: p. 1.

25) 플로리디(L. Floridi)의 정보철학에 관한 자세한 내용은 Massimo Durante, *Ethics, Law and the Politics of Information: A Guide to the Philosophy of Luciano Floridi*, Springer Berlin Heidelberg, 2017 참조.



3. 사이버-물리 시스템 환경에서 알고리즘의 지배

사이버-물리 시스템으로 이루어진 생활환경에서 인간은 자발적으로 이동식 감지기를 소지하거나 부착하고 다니면서, 그것도 아니면 주거 공간에 감지 장치를 설치해 놓음으로써 일상적이고 지속적으로 자신의 신체, 위치, 의식 정보를 특정 기업에게 전송하게 된다. 기업은 빅데이터라는 그 크기와 형태를 알기 어려운 정보를 알고리즘으로 분석하여 사용하면서 고려할 변수를 선택하고 정의한다. 그러나 이러한 결정은 첨단 기술의 형식을 갖추고 있다는 점에서 차이가 있을 수 있지만, 구조적인 측면에서 모지스가 다리를 설계하기 위해 다리의 높이라는 변수를 선택하여 얼마로 정의할 것인지 결정하는 것과 다르지 않다.²⁶⁾ 그렇기 때문에 이러한 알고리즘 기술 역시 막연하게 중립적이라고만 볼 수 없는 편향된 결과를 산출할 수 있다는 점은 새로운 사실이 아니다. 다만, 사이버-물리적 시스템이라는 전방위의 사이버네틱(cybernetic) 환경을 지배하는 알고리즘에 의한 결정이 좀 더 세련되고 섬세한 형태로 광범위한 사회적 효과를 가져 온다는 것은 새로운 사실이 될 수 있을지도 모른다.

III. 데이터 알고리즘 사회에서 전달되는 법과 지시하는 법

1. 정보의 두 가지 의미와 정보로서 법

힐데브란트(M. Hildebrandt)는 우리의 환경이 급증하는 데이터 기반 행위로 포화 상태에 이를 경우 법의 힘을 지탱하는 정보와 커뮤니케이션의 기반구조가 변형되어 법의 문법과 법의 지배에 심각한 위협을 야기할 것이라고 주장한다.²⁷⁾ 그리고 머신러닝 알고리즘이 인간의 머리로 포착할 수 없는 정확성을 갖고 있다는 것을 인정해야 하며 2020년대가 지나기 전에 대부분의 법률가들이 관련 입법, 판례, 법적 논변을 찾을 때 머신러닝에 의지할 것이라고 예견한다.²⁸⁾ 나아가 정보의 힘은 계산의 힘에 밀려나고, 법적 논증(legal argumentation)은 통계적 계산(statistical calculation)으로 대체되고 있다고 주장한다.²⁹⁾

26) 모지스((R. Moses))의 도시 설계에 관해서 본 논문 「제4장 제1절 I. 1. 기술적 설계를 통한 사회의 구조화」 참조.

27) Mireille Hildebrandt, “Law as Information in the Era of Data-Driven Agency”, *The Modern Law Review* 79(1), 2016, pp. 1~30: pp. 7~8.

28) Mireille Hildebrandt, “Law as Information in the Era of Data-Driven Agency”, *The Modern Law Review* 79(1), 2016, pp. 1~30: p. 10.



이러한 주장을 위해 힐데브란트는 먼저 정보의 두 가지 측면을 구분하며, 정보의 성격을 갖는 법 역시 그러한 두 가지 측면을 갖고 있는 것으로 본다. 정보는 단순히 대상을 서술하는 측면도 있지만, 그 대상에게 영향을 미치는 형성적인 측면을 갖고 있다는 것이다. 정보는 일정한 내용을 알려주기만(inform) 하는 것이 아니라 그 내용 자체를 형성하는(in-form) 것이다. 이러한 의미에서 정보에는 힘이 있다. 그리고 그 힘은 기호를 구성하는 표시 자체로부터 나오는 것이 아니라 기호에 부여된 의미로부터 나온다.

고전적인 언어학의 용어를 빌려 좀 더 섬세하게 표현하자면 정보로서 기호의 힘은 시니피앙(signifiant), 즉 기표(記表)의 체계에서 나오는 것이 아니라 시니피에(signifié), 즉 기의(記意)의 체계에서 나오는 것이다.³⁰⁾ 예를 들어 ‘트리(tree)’나 ‘목(木)’이라는 소리-이미지를 나타내는 기표의 체계가 ‘나무’라는 개념을 형성하는 기의의 체계와 결합되어 있지 않다면 기호의 힘은 없는 것이다.

기계가 인간을 상대로 바둑 게임에서 이기기 위해 인간이 바둑 게임을 하면서 기록해 둔 기보(碁譜)를 학습한다고 해보자. 기계의 입장에서 기보는 일련번호를 매길 수 있는 좌표에 검정색 점과 흰색 점을 찍는 순서를 표시해 둔 것으로서 게임에서 승리 또는 패배한 결과와 연동된 일종의 알고리즘이다. 그런데 인간이 실제로 수행한 게임의 기보 이외에 수학적으로 가능한 기보를 모두 학습한다면 그 개수는 점의 색깔 수 2개와 가로와 세로 각각 19개 줄을 그어 만들어지는 361개의 좌표수에 의해 결정된다. 그 개수를 구하기 위한 방식을 수학적으로 표현하는 것은 ‘느낌표(!)’ 하나로 해결된다. ‘361!’은 자연수 361부터 1까지 하나씩 연이어 곱하는 것을 의미하며 ‘ $361 \times 360 \times 359 \times \dots \times 3 \times 2 \times 1$ ’로 표시한다.

일상적인 언어의 문법 체계와 수리학 언어의 공식 체계에서 단순한 점 위에 선 하나를 그은 ‘!’ 표시는 그 의미도 ‘계승(階乘)’으로 달라지고, 읽는 법도 ‘factorial (팩토리얼)’로 달라진다. 이때 그 의미를 결정하는 것은 문법체계와 수리체계 또는 문학적 맥락과 수학적 맥락이다. 마찬가지로 법이 정보로서 힘이 있다면 그 힘은 법적 표현 자체로부터가 아니라 그 법적 의미를 구성하는 법체계 또는 법적 맥락과 결합됨으로써 발생하는 것이라 할 수 있다.

29) Mireille Hildebrandt, “Law as Computation in the Era of Artificial Legal Intelligence: Speaking Law to the Power of Statistics”, *University of Toronto Law Journal* 68(suppl. 1), 2018, pp. 12~35: p. 30.

30) 소쉬르(F. Saussure)의 언어학에서 기호는 별개로 분리된 기의(signifié, signified)와 기표(signifiant, signifier)로 구성된다. 기의는 일반적으로 보다 추상적인 개념(concept)을 대체하고, 기표는 단순히 물리적이거나 물질적인 것이 아닌 심리적으로 각인된 감각적 소리-이미지(sound-image)를 대체하기 위해 고안한 것이다. Ferdinand de Saussure, *Course in General Linguistics*[*Cours de linguistique generale*, 1916], Perry Meisel · Haun Saussy(Eds.), Wade Baskin(Trans.), Columbia University Press, 2011, pp. 65~67 참조.



2. 전달되는 정보와 조종하는 정보

정보를 영향력이나 형성력의 측면이 아니라 전달이나 전송의 측면에서 접근하면 정보의 성격은 사뭇 달라진다. 새넨(C. Shannon)은 1948년 발표한 “수학적 커뮤니케이션 이론”³¹⁾에서 정보의 존재 양상을 의미와 분리된 표시, 다시 말해 내용과 분리된 신호로만 제한한다. 새넨은 커뮤니케이션의 근본문제를 “한 지점에서 선택된 메시지를 다른 지점에서 정확하게 또는 근접하게 재생산하는 것”³²⁾으로 설정하고, 공학적 측면에서 커뮤니케이션의 의미론적 측면은 제거했다. 중요한 것은 시스템을 설계할 당시에는 나중에 실제 선택될 메시지가 무엇인지 알지 못하기 때문에 어떤 메시지가 선택되더라도 작동할 수 있도록 시스템을 설계하는 것이다. 메시지가 의미를 담고 있다고 하더라도 공학적 커뮤니케이션의 기능이 그 의미까지 전달하는 것은 아니라고 보았기 때문에, 어떤 내용을 신호로 바꾸어 보다 빠르고 저렴하게 내용의 동일성을 훼손시키지 않고 완전하게 비밀을 유지하면서 전송할 수 있는지에 초점을 맞춘 것이다. 즉, 정보 그 자체가 중요한 것이 아니라 분리된 별개의 정보를 권한 없는 제3자가 볼 수 없도록 A지점에서 B지점으로 완전하게 전달하는 것이 중요한 것이다. 여기에서 의미는 애매하거나 불필요하거나 관련 없는 것이다. 왜냐하면 의미는 송신자가 의도한 맥락, 메시지를 전달 받은 수신자의 배경과 맥락, 그리고 그것들의 차이에 의존할 뿐만 아니라 그러한 맥락의 차이는 서로 다른 시간에 존재하고 다른 언어로 표현되기 때문이다.

전달의 관점에서 보면 정보의 내용은 수량화할 수 있어야 하고, 확률로 계산될 수 있어야 한다. 새넨의 정보에 대한 이해는 통상의 직관에 반하는데, 정보를 무질서의 정도에 관한 엔트로피(entropy)를 기준으로 측정될 수 있는 것으로 이해하면서 엔트로피의 증가와 연계시킨다. 정보는 불확실성과 우연성의 측면에서 측정되는 것으로서 엔트로피가 낮아 불확실성이나 우연성이 없으면 역설적으로 정보는 없는 것이 된다. 다시 말해 어떤 결과의 발생 확률이 높으면 정보량은 적은 것으로 측정된다. 이러한 이론에서 정보는 “‘비트(bits)’로 측정될 수 있는 양이고, 상징이 발생할 확률의 측면에서 정의된다.”³³⁾ 이러한 개념을 사용하면 정보를 양화시킬 수 있지만, 그 의미와 질을 포기해야만 한다.

31) Claude E. Shannon, “A Mathematical Theory of Communication”, *The Bell System Technical Journal* 27, 1948, pp. 379~423(July) and pp. 623~656(October); Claude E. Shannon · Warren Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, 1963[1st, 1949] 참조.

32) Claude E. Shannon, “A Mathematical Theory of Communication”, *The Bell System Technical Journal* 27, 1948, pp. 379~423(July) and pp. 623~656(October): Introduction.

33) Frank Webster, *Theories of the Information Society*, 4th ed., Routledge, 2014, p. 30.



사이버네틱스(cybernetics)의 제창자인 위너(N. Wiener)는 새년의 이론을 뒤집어 정보를 감소하는 엔트로피로 보았다. 더 많은 구조를 발견할수록 더 많은 정보를 갖는다는 것이다. 지식과 동일시되는 이러한 정보 개념에는 메시지의 도움을 받아 환경을 통제하는 것에 대한 위너의 관심이 반영되어 있다. 이러한 정보는 행위와 결정에 연결된다. 본래 키잡이 또는 선박조종사를 뜻하는 그리스어 ‘kybernetes’에서 유래한 사이버네틱스에는 조종 또는 통치의 사고가 담겨 있다. 위너가 저술한 책 “사이버네틱스(Cybernetics)”의 부제가 “동물과 기계에서 통제와 커뮤니케이션”³⁴⁾이라는 것은 이러한 생각을 잘 표현해준다.

3. 사실을 구성하는 법과 알고리즘 시스템

정보로서 형성적 기능을 갖는 것이 법만 있는 것은 아니다. 정책, 기준, 관행, 도덕도 그런 기능을 갖는다. 형성적 기능을 갖는 정보는 일종의 표준(standard)으로서 새로운 사실을 만들어 내기도 한다. 예를 들어 대학의 입학 기준은 지원자에게 입학을 위해 충족해야 할 사실적 조건을 알려주기도 하지만 그 조건의 충족여부에 따라 합격과 불합격이라는 새로운 사실을 만들어내기도 한다.³⁵⁾ 이러한 표준이 사실을 구성하는 관계는 지시하기와 진술하기의 구별에 기초한다. 지시하기(denoting)는 사실들의 관계를 만들어 주는 것이고, 진술하기(stating)는 사실들의 관계를 있는 그대로 보여주는 것이다. 지시하기를 통해 형성되는 관계는 새로운 사실을 만들어 내고, 진술하기를 통해 드러나는 사실들의 관계는 발견되거나 재현되는 것이다.

인공지능과 법(AI & Law) 연구에서 하그(J. Hage)는 원칙 기반의 추론(principle-based reasoning)을 설명하면서 두 가지 유형의 사실을 원칙으로 공식화 할 경우 원칙의 공식화가 사실들 간의 관계를 지시할 뿐 진술하지 않는다고 하여 지시하기와 진술하기의 구별을 시도한다.³⁶⁾ 예를 들어 ‘어떤 유형의 사실 R이 있다면, 어떤 유형의 사실 C가 있다.’는 것을 원칙으로 공식화하는 것은 어떤 유형의 사실 R를 어떤 유형의 사실 C와 관계 맺도록 지시하는 것이지, 어떤 유형의 사실 R이 어떤 유형의 사실 C와 맺고 있는 관계를 진술하는 것이 아니라는 것이다.

34) Norbert Wiener, *Cybernetics: or Control and Communication in the Animal and the Machine*, 2nd ed., MIT Press, 1961[1st, 1948].

35) 독립적 차별 개념에서 표준으로 기능하는 기준에 관해선 본 논문 「제3장 제4절 IV. 2. 독립적 차별」 참조.

36) Jaap Hage, “A Theory of Legal Reasoning and a Logic to Match”, *Artificial Intelligence and Law* 4(3-4), 1996, pp. 199~273: p. 203 이하.



하그에 따르면 지시하기와 진술하기의 이러한 구별은 세계(world)와 말(word)의 관계를 통해 좀 더 명확히 설명될 수 있다.³⁷⁾ 세계와 말 사이의 관계는 설(J. Searle)이 언어행위 이론에 관한 연구에서 앤스콤(G. E. M. Anscombe)으로부터 ‘장보기 목록(shopping list)에 관한 사례’³⁸⁾를 차용하여 제시한 “맞춤의 방향(directions of fit)”³⁹⁾이란 개념으로 설명된다. 이때 장보기 목록이 가리키는 방향은 각각 다르다. 장바구니에 담긴 물건은 남편이 장보기 목록에 따라 담아 넣은 것이다. 장바구니에 물건을 담는 행위 또는 장바구니에 담긴 물건을 ‘세계’라고 하고, 장보기 목록을 ‘말’이라고 할 때 남편은 ‘말에 세계를 맞춘 것’이다. 반면에 형사의 장보기 목록은 장바구니에 물건을 담는 행위 또는 장바구니에 담긴 물건에 따라 작성된 것이다. 즉, 형사는 ‘세계에 말을 맞춘 것’이다. 이때 형사의 장보기 목록이 설명서(description)라면 남편의 장보기 목록은 지침서(directive)인 셈이다. 남편의 장보기 목록은 형성적 기능을 갖는 정보이자 표준으로서 새로운 사실을 만들어 내는 역할을 한 것이고 이러한 기능은 세계를 말에 맞추는 방향으로 작동한다.

말에 세계를 맞추는 지시하기는 군인을 겁쟁이로 만들 수도 있다. 만약 ‘군인이 다가오는 적군을 피해 도망가면 겁쟁이다.’라는 표준이 설정돼 있다면 다가오는 적군을 피해 도망가는 군인은 단순히 적군이 다가오는 것을 보고 도망가는 군인으로 남지 않는다. 이 군인은 다가오는 적군을 피해 도망가는 군인이면서 동시에 겁쟁이가 된다. 표준에 따라 도망가는 행위를 겁먹은 행위라고 지시하고, 그러한 행위를 하는 사람을 겁쟁이라고 지시한다. 도망가는 행위를 하는 군인이라는 사실에 겁먹은 행위를 하는 겁쟁이라는 새로운 사실이 추가적으로 구성된다.

이러한 지시하기의 구성적(constitutive) 성격은 해석 도식으로서 규범적 질서에 준거하여 사건을 해석함으로써 새로운 제도적 사실이 만들어질 수 있다는 주장의

37) Jaap Hage, “A Theory of Legal Reasoning and a Logic to Match”, *Artificial Intelligence and Law* 4(3-4), 1996, pp. 199~273: p. 204 이하.

38) 장보기 목록에 관한 사례는 다음과 같다. 어떤 사람이 가게에서 구매할 물건의 목록이 적힌 종이를 가지고 장을 본다. 이 사람은 목록을 보면서 물건을 하나씩 장바구니에 담아 넣는다. 그리고 그 사람을 뒤따르는 또 다른 사람이 있다. 앞에 가는 사람은 아내로부터 장보기 목록을 전달받아 그 임무를 수행하는 남편이고, 그에 뒤따르는 사람은 고객이 장바구니에 어떤 물건을 담는지 하나도 빠짐없이 종이에 적어 구매 목록을 만드는 형사이다. 남편이 장을 다 보고 나면 장바구니에 담긴 물건을 놓고 두 개의 장보기 목록이 남게 된다. 이에 관해서 Gertrude Elizabeth Margaret Anscombe, *Intention*, 2nd ed., Harvard University Press, 2000[1st, Basil Blackwell, 1957], pp. 56~57 참조.

39) John R. Searle, *Expression and Meaning: Studies in the Theory of Speech Acts*, Cambridge University Press, 1979: pp. 3~4.



근거가 될 수 있다. 법을 제도적 사실의 하나로 볼 수 있다는 논의⁴⁰⁾ 역시 이러한 맥락에서 크게 벗어나지 않는다. 법의 지시적 성격은 새로운 규범적 사실들을 만들어낸다. 법이 그 효력이 미치는 대상을 일정한 기준으로 구별하거나 분리해서 분류할 경우에 분류에 따라 포함되거나 배제되는 집단은 기준에 사용된 특성을 가졌다는 사실이 진술되는 것에 그치지 않고 그러한 특성에 따라 규정되는 새로운 법적 사실에 구속되도록 지시되고 구성되는 것이기도 하다.

이와 같은 맥락은 머신러닝 알고리즘에 대해서도 적용해 볼 수 있다. 머신러닝 알고리즘은 데이터 간의 관계를 설명할 수 있는 규칙을 추론함으로써 데이터로 구성된 세계를 구조화한다. 머신러닝 알고리즘의 데이터 학습을 지도하기 위해 클래스 레이블이나 표본의 레이블을 정의하는 경우⁴¹⁾ 레이블을 설정하고 그 레이블에 데이터를 지정하는 것은 레이블을 이용해 데이터로 구성된 세계를 재구성하는 것이라고도 볼 수 있다. 레이블에 연결된 데이터는 머신러닝 알고리즘의 세계에서 있는 것이 되지만 레이블에 연결되어 있지 않은 데이터는 없는 것으로 취급받기 때문이다.

앞에서 살펴봤듯이 머신러닝 알고리즘 기반의 이미지 검색 시스템에 대체 가능한 표현처럼 보이는 ‘장보기’와 ‘쇼핑’을 입력했을 때 머신러닝 알고리즘 시스템이 보여준 이미지 세계는 다르게 구성될 수 있다.⁴²⁾ 이는 머신러닝 알고리즘이 산출하는 지식이 새로운 사실을 지시하는 것을 의미하며, 차별의 맥락에서 새로운 특성의 추론에 의한 간접차별 영역의 확장 또는 우회적 문제로 나타날 수 있다.⁴³⁾ 그리고 세계를 재구성하는 힘을 가진 머신러닝 알고리즘이 데이터의 편향을 그대로 학습할 뿐만 아니라 확대시킬 수도 있다는 점⁴⁴⁾은 머신러닝 알고리즘 시스템이 그 편향을 반복해서 산출하여 세계를 재구성할 수 있음을 시사한다.

40) Neil MacCormick, “Law as Institutional Fact”, in Neil MacCormick · Ota Weinberger, *An Institutional Theory of Law: New Approaches to Legal Positivism*, Kluwer Academic Publishers, 1986, pp. 49~76.

41) 머신러닝 알고리즘의 지도학습방식과 작동원리에 관해서 본 논문 「제2장 제2절 III. 1. 지도학습과 분류 및 예측」 및 「제3장 제1절 머신러닝 알고리즘의 작동원리」 참조.

42) 같은 의미로 대체할 수 있는 표현처럼 보이는 ‘장보기’와 ‘쇼핑’도 머신러닝 알고리즘의 세계에서는 다른 레이블이기도 하다. 머신러닝 알고리즘이 이용되는 이미지 검색기에서 ‘장보기’와 ‘쇼핑’에 대한 검색 결과가 다르다는 점은 본 논문 「제2장 제3절 IV. 2. 이미지 검색 실험」 참조.

43) 이에 관해서 본 논문 「제4장 제2절 III. 2. 머신러닝 알고리즘의 특성 추론과 간접차별」 참조.

44) 이미지 데이터를 통해 학습한 머신러닝 알고리즘이 ‘shopping’이라는 레이블을 여성 행위자에게 연결시키는 편향을 드러낼 수 있다는 점은 Jieyu Zhao · Tianlu Wang · Mark Yatskar · Vicente Ordonez · Kai-Wei Chang, “Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints”, *Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2979~2989 및 이에 관한 본 논문 「제2장 제3절 III. 2. 데이터세트의 성별 편향과 모델에 의한 증폭」 참조.



IV. 예측하는 알고리즘과 예측되는 법

1. 법의 체계성

알고리즘 사회에서 법은 예측된다. 체계로서 법이 작동하기 위한 내부의 위계를 엄두에 둔다면 추상적이고 일반적인 수준의 헌법과 법률에서 개별 사건에 대한 구체적이고 특수한 판결에 이르기까지 법체계는 법적 결정을 동력으로 작동한다. 체계성은 법적 결정의 산출물이자 결과물인 법규범명제와 그 의미내용을 그때그때의 임시변통적인 것(ad hoc)에서 벗어나 보다 예측 가능한 것으로 만들어준다. 여기에는 법의 누적적(cumulative) 성격과 자기 조직적 성격이 반영되어 있다. 법은 법 그 자체에 기반을 둬으로써⁴⁵⁾ 법 위에 세워지고, 법은 법으로부터 도출됨으로써 이러한 체계성을 드러낸다. 체계성을 법 개념의 핵심 요소로 본다면 최초의 명령권자가 있어서 첫 날 어떤 명령을 공포하고, 다음 날 또 다른 명령을 공포하는 식으로 명령들이 쌓여 형성된 명령 더미가 곧 법인 것은 아니고, 합법이면서 동시에 불법인 규칙 또한 법이 아니다.

2. 머신러닝 알고리즘에 의한 법적 결정의 예측

개별 사건에 대해 어떤 법적 판단을 받을 것인지 예측할 수 있다면 개별 사건의 행위자들은 자신의 행위를 그에 비추어 선택할 수 있다. 이때 법적 판단을 예측하는 것은 개별 사건을 재판하는 재판부 구성원으로서 개별 법관이 어떤 결정을 내릴 것인지 예측하는 것이다. 예측의 문제는 확률의 문제이기도 하고 확률의 문제는 계산의 문제이기도 하다. 특히 법적 판단의 최종 결과값은 우선적으로 이진법에 따른다. 기본적으로 인용되거나 기각되고, 유죄이거나 무죄이고, 불법이거나 합법이고, 위헌이거나 합헌이다. 이진법에 따르는 결과 값을 예측하는 것은 50퍼센트의 확률로 출발한다. 자신의 사건이 또는 의뢰인의 사건이 기각되거나 무죄이거나 합법이거나 합헌일 확률은 절반인 것이다. 그러나 이러한 확률이 통계적 자료에 기초할 경우 그 확률은 더 높아질 수도 있고 낮아질 수도 있다.

법적 판단은 유사한 사건의 결과가 통상 어땠는지, 현재 사건의 담당 재판부의 법관이 유사한 사건에서 어떤 결정을 내렸는지에 관한 데이터뿐만 아니라 현재 사건의

45) Jeremy Waldron, "The Concept and the Rule of Law", *Georgia Law Review* 43(1), 2008, pp. 1~61: pp. 32~36.



소송을 진행할 때 보이는 법관의 감정적 태도가 결정에 어떤 영향을 미치는지에 관한 데이터에 이르기까지 법적 결정에 영향을 미칠 수 있는 가능한 모든 데이터를 기반으로 머신러닝 알고리즘이 통계적 예측 분석을 실행한다면 정확성은 향상되고 그 비용은 절감될 수 있다. 의사결정 트리⁴⁶⁾를 이용한 분류트리(classification tree) 알고리즘은 법률 전문가보다 대법원의 결정에 대한 대법관의 투표 결과를 더 잘 맞출 수 있고,⁴⁷⁾ 음성 분석 알고리즘은 재판을 진행하는 법관의 목소리를 분석하는 것만으로 대법관과 재판연구관이 논의해서 판결문을 작성하는 것보다 훨씬 싸고 정확하게 대법관의 투표 결과를 예측할 수 있다.⁴⁸⁾ 물론 이러한 실용적인 접근이 법을 오해하게 하거나 오용하게 한다는 이유로 못마땅하게 여겨질 수도 있지만 성능이 향상되고 있는 머신러닝 알고리즘이 얼마든지 이러한 과제를 수행할 수 있는 잠재력을 가지고 있다는 점만큼은 인정하지 않을 수 없다.

3. 머신러닝 알고리즘에 의한 법적 결정의 대체

법관이 어떤 결정을 할 것인지 예측하는 단계에서 사용되는 알고리즘은 적어도 최종적으로 법적 효력을 갖는 결정에 관한 한 외부의 관찰자로 남아 있다. 결정을 예측할 뿐 직접 법적 효력을 갖는 결정을 하지는 못하는 것이다. 그러나 이러한 상황은 예측 분석 알고리즘의 사용자가 바뀌면 쉽게 반전될 수 있다. 법관 즉, 법적 효력을 갖는 결정을 최종적으로 승인하는 권한을 갖는 기관이 직접 이러한 알고리즘을 사용하는 경우이다. 법관이 국가기관으로서 내렸던 결정은 모두 흔적을 남긴다. 판결, 결정, 명령의 어떤 형태도 기록에 남아 있다. 예측 분석 알고리즘이 갖는 통계적 계산의 힘을 빌린다면 법관 스스로 자기 자신이 어떤 판단을 내릴 것인지 알 수 있다. 이때 법관이 할 수 있는 일은 알고리즘의 예측에 따라 자신의 과거 결정 패턴과 일치하게 결정을 하거나 그 예측에 항변하는 것이다. 만약 법관이 알고리즘의 예측에 항변하는 것을 포기하는 순간 알고리즘이 법관의 판단을 지원 하던 조력의 차원은 직접 법관의 판단을 대신하는 대체의 차원으로 이행한다.⁴⁹⁾

46) 의사결정 트리(decision tree)에 관해서 본 논문 「제2장 제2절 IV. 1. 기호주의와 논리 및 규칙 기반의 역연역법」 참조.

47) Andrew D. Martin · Kevin M. Quinn · Theodore W. Ruger · Puline T. Kim, “Competing Approaches to Predicting Supreme Court Decision Making”, *Perspectives on Politics* 2(4), 2004, pp. 761~767. 이 연구는 연방 대법원 판사인 샌드라 데이 오코너(Sandra Day O’Conner)의 판결 방식을 결정트리 알고리즘으로 보여준다.

48) Bryce J. Dietrich · Ryan D. Enos · Maya Sen, “Emotional Arousal Predicts Voting on the U. S. Supreme Court”, 2017, https://scholar.harvard.edu/files/msen/files/scotus_audio.pdf, pp. 1~9.



법관의 결정을 국가의 사무를 관장하는 국가기관이 내리는 여러 가지 결정 중의 한 가지 예라고 한다면 이러한 알고리즘은 다른 공무원의 결정에 대해서도 조력하는 수준을 넘어서 대체할 수 있는 가능성은 열려 있다. 비록 각 국가기관과 공무원이 관장하는 고유한 사무와 권한이 별개로 분리되어 있다고 해도 알고리즘의 확장성과 범용성은 이를 단지 사용자가 달라지는 문제를 다루는 것에 불과하게 만들 수 있다. 이러한 경향에 제동을 걸 수 있는 답변은 국가 사무는 인간 에이전트만 담당할 수 있고 실제로 알고리즘 에이전트는 담당할 수 없다고 사무의 기술적 대체 불가능성을 주장함으로써 기술의 발전과 능력에 관한 시대착오적 인식을 드러내는 것에 머물거나, 대체 가능성은 인정하되 국가 사무는 인간 에이전트의 최종 승인을 거쳐야 한다고 주장하거나 알고리즘 에이전트가 단독으로 처리하는 사무에 제한을 두어야 한다고 주장함으로써 인간 에이전트 중심의 정책이 담긴 목표규범을 제시하는 논변으로 전개될 수 있을 것이다. 한 발 더 나아가간다면 지능적 에이전트인 경우 인간 에이전트이건 알고리즘 에이전트이건 구별하지 않고 동일한 권한을 부여해야 한다고 주장함으로써 지능적 에이전트 간에 처리할 수 있는 사무를 차별적으로 제한하지 않도록 하는 논변이 제시될 수 있을 것이다.

49) ‘데이터 알고리즘 사회’의 주요 현상으로서 “인간의 이성적 사고나 추론, 직관을 알고리즘 명령어와 인공지능 분석으로 대체하려는 흐름”을 제시하는 경우로 이광석, 데이터 사회 비판, 책읽는수요일, 69쪽 참조.



제2절 머신러닝 알고리즘의 분류 및 추론과 간접차별

방대한 규모의 데이터에 노출되어 훈련된 머신러닝 알고리즘은 데이터를 관찰함으로써 발견한 패턴을 설명할 수 있는 규칙을 추론한다. 이렇게 추론된 규칙은 향후 새로운 데이터가 입력될 때 적용되어 데이터를 패턴에 맞게 분류하고 정렬하여 결과로 보여준다. 일련의 결정 절차로서 추론된 규칙은 미래의 결과에 대한 예측을 확률의 형태로 보여주기도 한다. 그런데 이때 규칙을 추론하는 알고리즘 에이전트의 작동에 고의 또는 의도가 있다고 볼 수 있을까? 차별 개념을 행위 중심의 의무론에 따라 이해하고 이를 기초로 차별금지법을 구성하거나 해석할 때 고의 또는 의도의 유무는 핵심적인 요소가 된다. 특히 법적 의미를 갖는 행위를 인간의 행동에 한정하려고 하는 입장에서 기계의 행동과 인간의 행동을 구별하기 위해 내심의 의도나 동기를 부각하는 것은 기계와 인간의 차이를 마음(minds)이 있는지 또는 생각할 수 있는지 여부에서 찾는 분석철학의 입장과 맥락상 연결된다.

I. 인공물에 의한 분류와 인간에 의한 분류

1. 인식 모델과 인지 작용에 의한 지능적 분류

차별의 전제로서 구별은 그 자체 인식과 밀접한 관련을 맺고 있으며, 집단을 분류하는 토대가 된다.⁵⁰⁾ 그리고 인지는 본질적으로 일반성 또는 보편성을 포함한다는 고대의 전통적 주장에 따르면 특수한 것은 직접적으로 그 자체 지능적이지 않은 것이다. 이러한 이해에 따르면 머신러닝 알고리즘에 의해 작동하는 시스템 같은 인공물이 지능적이라고 평가받기 위해서는 일반성 또는 보편성에 부합하는 분류를 실행해야 한다. 브랜덤(R. Brandom)은 ‘직관을 개념에 따라 분류하기 시작하는 것이 인지(cognition)’라는 칸트(I. Kant)의 설명⁵¹⁾을 이런 전통이 특별히 잘 개발된 대표적인 경우로 본다.⁵²⁾ 믿거나 판단하는 것의 기본적인 형식은 특수한 것을 보편적인 것에 포함시키는 것이다. 그러므로 “인식은 특수한 것을 포괄하는 반복적인 개념을

50) 이에 관해서 본 논문 「제3장 제2절 I. 차별의 전제로서 구별, 분리, 분류」 참조.

51) Immanuel Kant, *Critique of Practical Reason*[*Kritik der praktischen Vernunft*, 1st, 1788], Mary J. Gregor(Ed. and Trans.), Rev. ed., Cambridge University Press, 2015, Sec. 7: pp. 28~30.

52) Robert Brandom, *Making It Explicit: Reasoning, Representing and Discursive Commitment*, Harvard University Press, 1994, p. 85.



드러내는 것”⁵³⁾이다. 분류되는 사물의 인식과 그 사물이 분류되는 방법의 인식은 분류하기를 구성하는 인식 즉, 그 사물을 포괄하는 반복적인 개념을 드러내는 것으로부터 나온다. 칸트는 초기 경험주의자들이 개념적 이해가 부족한 의식적 이해를 용인하는 것에 반대하며 어떤 종류의 인식에서든 개념적 분류가 있어야 한다고 주장한다.

브랜덤은 이러한 칸트식 인식 모델의 실용주의 버전을 헤겔(G. Hegel)에게서 찾는다.⁵⁴⁾ 개념적 분류의 근원은 실상에서 어떤 사물을 어떤 종류로 대우하는 데에서 발견되는데, 이는 특수한 어떤 것을 보편적인 어떤 것으로 대우하는 것이다. 헤겔은 인식의 기원에 관한 욕구 이론의 형식(the form of erotic theory)을 빌려 이러한 분류의 원천을 동물의 욕구(Begierde)로 본다.⁵⁵⁾ 동물은 무엇인가 땅에 떨어졌을 때 그것을 먹을 것 즉, 음식으로 분류하고 먹는다. 다시 말해 먹는 것은 그것을 대우하는 것이고, 그것에 반응하는 것이며, 그것을 실상에서 음식으로 분류하는 것이다. 유기체의 일부에 대한 그러한 반복적 활동은 그런 활동을 이끌어내는 경향이 있는 사물들 사이의 유사성에 대한 반복적 관심을 유발한다.

이와 같이 특수한 자극의 분류는 예를 들어 ‘자신의 먹이가 될 것’과 ‘자신을 먹이로 할 것’ 즉, 자신의 먹이와 포식자 같은 일정한 종류에 서로 다른 반복적 원형 개념을 일치시키는 것으로 이루어지고 이러한 분류는 그에 대한 유기체의 반응에 함축 또는 암시되어 있다. 이렇게 실상에서 함축적 또는 묵시적으로 분류하는 인식이나 이해가 어떤 종류의 명시적 인식도 상정하지 않는다는 것은 묵시적 인식이나 이해에 필요한 것이 “신뢰할 수 있게 구별하는 반응적 성질(a reliable differential responsive disposition)”이라는 점에서 명확해진다.

2. 반응적 분류와 개념적 분류

모든 물리적 체계(physical system)의 작용에 나타나는 반응의 규칙성(the responsive regularities)에 따라 분류를 무차별적으로 파악하는 개념 사용 모델에 따르게 되면 개념 적용에 필요하다는 의미에서 종(sapience), 인식, 의식으로 분류 개념을 구분하는

53) Robert Brandom, *Making It Explicit: Reasoning, Representing and Discursive Commitment*, Harvard University Press, 1994, p. 86.

54) Robert Brandom, *Making It Explicit: Reasoning, Representing and Discursive Commitment*, Harvard University Press, 1994, p. 87.

55) Georg W. F. Hegel, *Phänomenologie des Geistes*[1st, 1807], Eva Moldenhauer · Karl Markus Michel(Eds.), 2nd ed., Suhrkamp, 1989, S. 90-91.



칸트식 합리주의 전략은 사소한 것이 되어버린다. 철근은 용해나 부식으로 어떤 환경에 반응한다. 이는 그 어떤 환경으로서 높은 열기와 낮은 열기를 분류하는 것이며 많은 양의 물기와 적은 양의 물기를 분류하는 것이다. 그러나 브랜덤은 철근이 수행하는 분류에서도 나타나는 ‘반복적으로 구별하고 반응하는 성질’은 개념 사용에 필수적인 조건이지만 충분조건은 아니라고 보면서 추론주의(inferentialism) 관점에서 “반응적 분류(responsive classification)”와 “개념적 분류(conceptual classification)”를 구별하여 인식한다.⁵⁶⁾ 개념의 올바른 적용과 그릇된 적용의 차이에 동의하는 규범적 차원(normative dimension)이 필요하기 때문이다.⁵⁷⁾ 그럼에도 불구하고 개념은 합의된 규범적 차원의 기준에 따라 적절하고 정확하게 사용될 수도 있고, 부적절하고 부정확하게 사용될 수 있다.

인공도구(artificial instrument)가 구별 가능한 일정한 반응적 경향에 따라서 분할 가능한 자극의 부분 집합을 등가의 클래스에 속하게 하고, 그러한 분류가 인공도구의 이용자에게 실천적 또는 이론적 의미 있는 구별과 일치하도록 구축된 경우 인공도구는 더 이상 철근 덩어리와 같은 물리적 체계와는 다른 분류를 하는 것으로 볼 수 있다. 온도계는 수은의 팽창과 수축 반응을 분할된 클래스에 연결시키고 이를 이용자에게 실천적 또는 이론적 의미가 있는 온도의 높낮이 구별에 일치되도록 구축된 인공도구의 대표적인 예이다.

브랜덤은 간단한 사고실험을 설계하는 데 철근 같은 물리적 체계와 다른 예로 측광기(spectrophotometer)를 제시하는데, 이 측광기는 음향 장치에 연결돼서 적절한 빈도의 빛을 비추는 경우에만 ‘저것은 빨강다.’라는 소리를 출력할 수 있다. 그리고 이러한 측광기를 적절한 빈도의 빛을 비추는 경우에만 ‘빨강’을 외치도록 상정된 광적인 인간 보고자(reporter)와 비교한다. 여기서 보고자는 “비추론적으로 주변의 온도나 색에 대해 확신을 얻고 주장을 펼치는 관찰자(observer)”⁵⁸⁾이다. 이 보고자는 다시 같은 자극 아래에서 같은 소리를 읊조리게 훈련된 앵무새와 비교된다.

56) Robert Brandom, *Making It Explicit: Reasoning, Representing and Discursive Commitment*, Harvard University Press, 1994, pp. 86~87.

57) Robert Brandom, *Making It Explicit: Reasoning, Representing and Discursive Commitment*, Harvard University Press, 1994, pp. 26~30.

58) 여기에 ‘온도’가 포함된 것은 철근 덩어리와 같은 물리적 체계와 다른 인공도구의 예로 온도계를 측광기와 함께 제시했기 때문이다. Robert Brandom, *Making It Explicit: Reasoning, Representing and Discursive Commitment*, Harvard University Press, 1994, p. 87.



이러한 비교는 결국 보고자를 측광기 같은 인공도구나 앵무새와 구별하게 하는 인간의 실천적 능력은 무엇인지라는 질문으로 귀착되고 브랜덤은 이를 이해(understanding)에 관한 문제로 본다.⁵⁹⁾ 보고자의 반응은 타자에게 의미가 있을 뿐만 아니라 반응하는 보고자 자신에게도 의미가 있다는 것이다. 물론 측광기나 앵무새의 반응도 타자에게 의미가 있지만 측광기와 앵무새는 자신의 반응을 이해하지 못한다. 그리고 이러한 차이를 구성하는 이해와 관련된 실천적 능력을 셀러스(W. Sellars)에 따라 “이유를 제시하고 묻는(giving and asking for reasons)”⁶⁰⁾ 능력에서 찾는다. 이유를 제시하고 묻는 반응은 “확신과 주장을 정당화하는(justifying beliefs and claims)”⁶¹⁾ 역할을 수행하고, 그럼으로써 반응에 “의미(significance)를 부여”⁶²⁾하기 때문이다.

3. 기계와 인간의 구별 기준으로서 이해와 그 한계

브랜덤이 개념적 분류를 추론주의 입장에서 설명하기 위해 동일한 상황에 대해 같은 반응을 보이는 앵무새 및 측광기와 인간의 분류를 비교한 것과 다르게 설(J. Searle)은 튜링테스트⁶³⁾에 담긴 인간의 정신에 대한 기능주의적 접근을 반박하기 위해⁶⁴⁾ ‘중국어 방(chinese room)’이라고 불리는 사고실험을 제안했다.⁶⁵⁾ 설 역시 튜링테스트에서처럼 응답자와 질문자 사이를 가리는 격리 공간을 조성한다. 중국어를 전혀 알지 못하고 영어만 알고 있는 사람을 방 안에 두고 밖에서 중국어 기호를 넣는다. 그 사람은 어떻게 중국어 기호들을 산출해야 하는지에 관한 영어 지침서를 가지고 문법 규칙들을 활용해 중국어 기호로 답을 내 놓는다. 그리고 시간이 흘러 그 사람이 중국어 기호들을 조작하는 일에 대한 지침서를 따르는 데

59) Robert Brandom, *Making It Explicit: Reasoning, Representing and Discursive Commitment*, Harvard University Press, 1994, p. 88.

60) Robert Brandom, *Making It Explicit: Reasoning, Representing and Discursive Commitment*, Harvard University Press, 1994, p. 89.

61) Robert Brandom, *Making It Explicit: Reasoning, Representing and Discursive Commitment*, Harvard University Press, 1994, p. 89.

62) Robert Brandom, *Making It Explicit: Reasoning, Representing and Discursive Commitment*, Harvard University Press, 1994, p. 89.

63) Alan M. Turing, “Computing Machinery and Intelligence”, *Mind* 59(236), 1950, pp. 433~460 및 이에 관한 본 논문 「제2장 제1절 II. 2. 튜링의 모방게임과 학습하는 기계」 참조.

64) Steve Schwartz, *A Brief History of Analytic Philosophy: From Russell to Rawls*, Wiley-Blackwell, 2012, pp. 183~192.

65) John R. Searle, “Minds, Brains and Programs”, *Behavioral and Brain Sciences* 3(3), 1980, pp. 417~424.



아주 익숙해진다. 만약 어떤 프로그램 설계자가 방 바깥에 있는 누군가의 관점에서 원어인 중국인 화자들의 답과 전혀 구별할 수 없는 답을 출력하는 프로그램을 자유 자재로 사용한다고 할 경우, 지침서에 따라 중국어 답을 내놓는 사람 역시 외부자의 관점에서 중국어를 사용하지 않는다고 말할 수 없다. 결국 중국어 방에 있는 그 사람도 튜링테스트를 통과할 수 있지만 설은 이 경우에 그 사람이 중국어를 이해하지는 못한 것이라고 주장한다.

이러한 사고실험을 통해 설이 증명하려고 한 것은 두 가지이다. 그중에 하나는 튜링테스트가 생각, 지능, 이해 또는 그 비슷한 어떤 것에 대한 시험이 아니라는 것이다. 중국어 방에 있던 그 사람이 중국어 기호를 이해하지 못한다 하더라도 튜링테스트를 통과할 수 있는 것처럼 컴퓨터도 튜링테스트를 통과할 수 있지만 아무것도 이해하지 못한다는 것이다. 다른 하나는 순수하게 기능적인 어떤 설명만으로 생각, 이해, 지능의 본성을 포착할 수 없다는 것이다. 그런데 앞서 브랜덤처럼 이해를 이유 제시 및 질문 능력이라는 실천적 능력과 연관시키는 관점에 따른다면 튜링테스트에서 질문 시간이 5분으로 제한되지 않았다면 컴퓨터가 이유를 제시하고 묻는 능력에 어떤 한계를 드러냄으로써 이해 능력을 결여하고 있다는 점을 증명할 수 있다고 비판할 수도 있다. 그런데 문제는 인간과 비교되는 인공물이 색을 말하는 측광기의 수준에 계속 머물러 있지 않다는 것이다.⁶⁶⁾ 인간의 어떤 능력을 인공물이 갖고 있지 않거나 인공물의 어떤 능력이 인간의 그것에 미치지 못한다는 식의 인간 고유의 능력이나 인간의 우월한 능력에 기초한 논변은 인공물의 과제 수행능력이 향상되어 지능적 분류 및 예측을 수월하게 해내는 상황에 직면하면 그 토대를 상실하여 한계를 맞이할 수 있다.

II. 간접차별과 직접차별

일정한 종류에 원형 개념을 반복적으로 일치시키는 작업은 외관상 중립적인 분류로 보일 수 있다. 그런 분류가 차별의 전제가 될 경우 차별 개념은 한층 복잡해진다. 간접차별(차별효과)은 비록 앞에서 차별 형식의 하나로 다루었지만,⁶⁷⁾ 미국의 경우 간접차별이론의 형성은 1971년 그릭스(Griggs)와 듀크 전력회사(Duke Power Co.)

66) 인공물의 지능적 판단 수준에 관해서 본 논문 「제2장 제1절 II. 인공지능 연구와 학습하는 기계」, 「제2장 제1절 III. 4. 21세기의 인공지능 로봇과 인류 미래에 관한 논의」 및 「제2장 제3절 머신러닝 알고리즘의 오류와 편향」 참조.

67) 이에 관해서 본 논문 「제3장 제2절 II. 4. 차별효과와 간접차별」 참조.



간의 소송사건에서 비롯된다.⁶⁸⁾ 이 사건에 대한 판결⁶⁹⁾의 주심 법관은 이솝 우화(Aesop's Fables)에 나오는 '여우와 황새(the fox and the stork)' 이야기⁷⁰⁾를 원용하여 간접차별의 문제를 다룬다.⁷¹⁾

듀크 전력회사의 고용주는 채용절차에 응시한 지원자들에게 필기시험을 보도록 했다. 필기시험의 통과 여부는 외관상 중립적인 채용 기준처럼 보인다. 그러나 주심 법관은 흑백 분리 정책에 따라 낮은 수준의 교육을 받아 온 흑인은 상대적으로 필기 시험을 통과하기 어렵다는 점을 지적한다. 그러면서 겉으로는 중립적인 것처럼 보이는 사유이지만 필기시험을 통과할 수 있는 능력과 직업을 수행하는 능력 사이에 관련이 있다고 보기 어려울 뿐만 아니라, 그러한 낮은 관련성을 지원자가 입증할 수도 없어 지원자를 간접적으로 차별한다고 밝힌다. 외관상 인종 간에 중립적으로 보이는 규칙이나 관행이라고 할지라도 어떤 한 인종 집단에게 다른 인종 집단과 비교해 불균형적으로 불리한 영향(disparate adverse impact)이 있고 그러한 규칙이나 관행의 필요성이 정당화되지 않으면 차별적이라고 판명될 수 있다는 것이다.

아울러 이러한 차별 유형의 부당성은 행위자의 의도나 동기에 대한 언급 없이도 정립될 수 있다고 밝힌다. 우화 속의 여우는 사악한 의도를 가지고 있었던 반면, 고용주에게는 그러한 의도가 없거나 무의식적 편향이 작용했을 뿐이지만 고용주의 의식이나 무의식 같은 심리적 사실이 집단에 대한 불리한 효과에 대한 부당성을 좌우할 수 없다는 것이다. 통상 '차별효과'나 '불평등 효과'로 번역되는 'disparate impact'는 이러한 맥락 속에 등장한 용어이고 이에 대응하는 일반적인 용어는 '간접차별'로 번역되는 'indirect discrimination'이다. 그런데 이러한 간접차별을 별도의 차별 유형으로 구분하는 기준은 각양각색이고 심지어 구분의 시도가 포기되기도 한다.

68) 차별효과이론의 간략한 역사는 Michael Selmi, "Indirect Discrimination and the Anti-Discrimination Mandate", in *Philosophical Foundations of Discrimination Law*, Deborah Hellman · Sophia Reibetanz Moreau(Eds.), Oxford University Press, 2013, pp. 250~268: pp. 252~255 참조.

69) *Griggs v. Duke Power Co.*, 401 U. S. 424, Supreme Court of the United States (1971); 이 사건에 대한 미국연방대법원의 판결이 선고된 1971년은 공고롭게도 알고리즘의 NP-완전 문제가 처음으로 증명된 해이기도 하다.

70) 여우는 황새를 초대해 음식을 대접하면서 수프를 납작한 그릇에 담아 내놓는다. 주둥이가 길고 뾰족한 황새는 부리 끝을 적시는데 만족할 수밖에 없었다. 여우가 악의를 가지고 자신을 불리하게 대접했다고 생각한 황새는 다음에 여우를 자신의 집에 초대하여 음식을 대접하면서 입구가 얇고 긴 관 모양으로 된 호리병에 내놓았다. 여우의 주둥이는 호리병을 닮지 않았기 때문에 음식을 먹을 수 없었다.

71) Hugh Collins · Tarunabh Khaitan, "Indirect Discrimination Law: Controversies and Critical Questions", in *Foundations of Indirect Discrimination Law*, Hugh Collins · Tarunabh Khaitan(Eds.), Hart Publishing, an imprint of Bloomsbury Publishing Plc, 2018, pp. 1~30 참조.



1. 차별로 인한 개인의 불이익과 집단의 불이익

차별의 유형을 직접차별과 간접차별로 구분하는 첫 번째 방식은 개인의 불이익을 주장하는지 혹은 집단의 불이익을 주장하는지에 따르는 것이다.⁷²⁾ 이러한 구분 기준은 반드시 그런 것은 아니지만 직접차별은 보통 한 개인이 또 다른 개인에게 특정 행위를 하는 것과 관련된다는 점에 근거를 둔다. 이런 방식의 직접차별 유형에 관한 예는 어떤 고용주가 특정 여성을 고용하지 않기로 결정하는 것이다. 반면에 간접차별은 항상 집단에 관한 것이다. 간접차별의 인식은 어떤 집단이 행위, 규칙 또는 관행에 의해서 같은 기준이 적용되는 다른 비교 집단에 비해 불균형적인 불이익을 입었는지 여부를 판별하는 것으로부터 출발하기 때문이다. 이러한 구분 방식은 같은 상황을 다른 관점에서 보는 것이기 때문에 특정 사례가 직접차별로 다루어질 수도 있고 간접차별로 다루어질 수도 있다. 특정 여성에 대한 결정이 개인에 대한 것이면서 동시에 여성 집단의 구성원에 대한 것이기 때문이다. 그러므로 직접차별인지 간접차별 인지는 우선적으로 차별을 주장할 때 원용하는 불이익이 개인의 불이익인지 집단의 불이익인지에 따라 결정된다. 그렇다면 앞의 그릭스(Griggs) 사건의 경우 원고가 개인의 불이익을 주장했다면 직접차별 사안으로도 다룰 수 있게 된다.

2. 차별의 의도 또는 동기

두 번째 방식은 어떤 일관된 기준을 가지고 그에 따라 직접차별과 간접차별을 구분한다.⁷³⁾ 그러나 그 기준은 법체계마다 달라 차별의 의도 또는 결과를 기준으로 구분할 수 있다.⁷⁴⁾ 차별의 의도에 따라 구분하는 대표적인 예로 미국은 성별이나 인종 같이 법적으로 보호되는 특징을 결정의 근거로 사용하여 차별하려는 의도에 따라 직접차별과 간접차별을 구분한다. 그래서 그러한 의도에 대한 명확한 증거가 없거나 단지 의심스러운 분류인 경우 법적 청구는 그 집단에 대한 차별 효과로 제기되어야 한다. 이때 차별효과원칙은 의도원칙만으로 구성된 차별금지원칙을 쉽게 우회하는 것을 방지하는 데 필수적인 것으로 여겨지기도 한다.⁷⁵⁾

72) Hugh Collins · Tarunabh Khaitan, “Indirect Discrimination Law: Controversies and Critical Questions”, in *Foundations of Indirect Discrimination Law*, Hugh Collins · Tarunabh Khaitan(Eds.), Hart Publishing, an imprint of Bloomsbury Publishing Plc, 2018, pp. 1~30: p. 19.

73) Hugh Collins · Tarunabh Khaitan, “Indirect Discrimination Law: Controversies and Critical Questions”, in *Foundations of Indirect Discrimination Law*, Hugh Collins · Tarunabh Khaitan(Eds.), Hart Publishing, an imprint of Bloomsbury Publishing Plc, 2018, pp. 1~30: p. 20.

74) 차별 개념의 의무론적 구성과 결과론적 구성에 관해서 본 논문 「제3장 제2절 III. 행위 중심의 차별과 결과 중심의 차별」 참조.



3. 차별의 결과 또는 효과

반면에 차별로 인한 결과나 효과에 따라 구분하는 예로 영국은 의도의 증명 여부를 직접차별의 기준으로 삼지 않고, 법으로 보호되는 집단을 완전히, 즉 백퍼센트(100%) 배제하는 결정을 수용했는지 여부를 직접차별과 간접차별을 가르는 기준으로 삼는다. 따라서 보호집단을 완전히 배제하면서 그와 비교 관계에 있는 집단은 아무도 배제되지 않는 경우를 직접차별로 보고, 보호집단에 대한 배제 결과의 비율이 백퍼센트 미만이지만 비교 집단에서 배제된 결과의 비율이 그와 비례적이지 않은 경우 간접차별로 본다.⁷⁶⁾

성별을 이유로 어떤 결정을 내리는 것에 대해 법적으로 금지된 상황에서 한쪽 성별이 모두 배제되고 다른 쪽 성별이 모두 포함된 경우 직접차별로 보는 반면, 한쪽 성별에서 일부 배제된 비율이 다른 쪽 성별에서 일부 배제된 비율과 같지 않은 경우, 예를 들어 한쪽 성별 X에서 5분의 2(40%)만 배제된 반면 다른 쪽 성별 Y에서는 5분의 3(60%)이 배제된 경우 성별에 따른 간접차별이 있는 것으로 보는 것이다. 한쪽 성별이 모두(100%) 배제되고 다른 쪽 성별은 극히 일부만 배제된 경우, 예컨대 100분의 1(1%)만 배제된 경우 역시 직접차별이 아니라 간접차별로 보게 된다. 이러한 구분은 결정의 근거로 보호 집단의 특징을 언급하지 않는 경우에도 비율의 차이만으로도 직접차별이 될 수 있다는 점이 특별하다.

영국은 직접차별과 간접차별 모두 규칙이나 관행의 불리한 효과에만 집중한다. 이 경우 법의 공통된 목적은 보호 집단에 불리한 영향이 발생하는 것을 방지하는 것이라는 점에서 차별 개념과 형식에 통일된 기반을 마련할 수 있다. 이 관점에서 직접차별과 간접차별 개념은 법으로 규율하는 공정한 기회의 평등이나 사회적 통합 같은 사회적 목표를 달성하기 위해 사용되고, 편견이나 고정관념의 희생자에게 보상을 제공할 수도 있다.

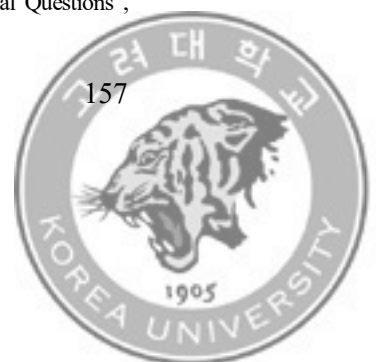
4. 차별의 정당화 가능성

세 번째 방식은 차별의 정당화 가능성을 기준으로 직접차별과 간접차별을 구분하는 것이다.⁷⁷⁾ 이는 부당성과 정당성이라는 차별에 관한 규범적 평가 결과를 직접

75) Larry Alexander · Kevin Cole, “Discrimination by Proxy”, *Constitutional Commentary* 14(3), 1997, pp. 453~463: p. 455.

76) James v. Eastleigh Borough Council (1990) 2 AC 751 참조.

77) Hugh Collins · Tarunabh Khaitan, “Indirect Discrimination Law: Controversies and Critical Questions”,



차별과 간접차별의 구분 기준으로 사용한다. 일종의 귀납적 구분 방식이다. ‘정당화 가능성’이라는 기준을 조금 더 엄밀하게 구성하면 ‘정당화 가능성의 여부’를 기준으로 하는 경우와 ‘정당화 가능성의 정도’를 기준으로 하는 경우로 나눌 수 있다.

(1) 정당화 가능성의 여부

먼저 정당화 가능성의 여부를 기준으로 하는 경우 예외를 인정하지 않음으로써 정당화가 불가능한 것은 직접차별, 예외를 인정함으로써 정당화가 가능한 것은 간접차별이 된다. 이러한 구분에 따르면 정당화 가능성이 없다는 점에서 직접차별 유형에 해당되는 것은 곧바로 부당하고 예외 없이 금지되는 절대적 차별이다. 반면에 간접차별은 정당화 가능성이 있어, 즉 정당화에 개방되어 있어 차별적 행위나 상황의 정당성을 주장할 수 있고 그에 따라 차별적 행위나 상황이 허용될 수 있는 예외가 있다는 의미에서 상대적 차별이다. 절대적 성격의 직접차별은 차별 금지 사유 또는 그에 따른 보호집단을 강하게 보호할 수 있다. 특히 국가에 의해 발생하는 차별을 억제하는 데 기여할 수 있을 것이다. 그런데 차별금지법의 적용 범위에 사적 영역도 포함되어 사인에게 효력을 미칠 경우 그러한 절대적 성격은 차별금지법의 효력으로 의무 부담을 지게 되는 사인이 어떤 법익이 충돌 관계에 있을 때 이를 절대적으로 부인하는 것이 될 수 있다.

(2) 정당화의 난이도

정당화 가능성의 ‘정도’를 기준으로 하는 것은 대부분의 법체계에서 간접차별의 예보다는 직접차별의 예에 대해서 정당화를 더 어렵게 한다는 점을 토대로 한다. 차별에 대한 평가 단계에서 정당화를 어렵게 하는 방법 중 하나는 예외를 인정하거나 차별 상황을 허용할 수 있는 심사 방법을 선택하되 그 강도를 세게 하는 것이다. 따라서 예외를 인정하거나 법익에 대한 제한을 허용하는 경영상 필요성 심사나 비례성 심사를 선택하는 경우 이러한 심사 방법이 매우 엄격하게 적용되는 것을 직접차별로 보고, 심사의 강도가 느슨하게 적용되는 것을 간접차별로 보게 된다. 물론 심사 방법은 다르게 구성될 수도 있다. 심사 방법을 단계적으로 구성하여 부당성 또는 불법성이 강하게 추정되어 반증이 어려운 것과 부당성 또는 불법성이 그만큼 강하게 추정되지는 않아 상대적으로 반증이 쉬운 것으로 나누는 것이다. 미국 연방대법원에서 사용하는 이른바 ‘합리성 심사’는 이러한 단계적 구성의

in *Foundations of Indirect Discrimination Law*, Hugh Collins · Tarunabh Khaitan(Eds.), Hart Publishing, an imprint of Bloomsbury Publishing Plc, 2018, pp. 1~30: pp. 21~23.



대표적인 예이다.⁷⁸⁾ 이와 같은 구성의 심사 방법에 따르면 반증 가능성이 낮은 심사 단계가 적용되는 것, 즉 부당성 또는 불법성이 강하게 추정되는 것을 직접차별로 보고, 반증 가능성이 높은 심사 단계가 적용되는 것, 즉 부당성 또는 불법성이 상대적으로 약하게 추정되는 것을 간접차별로 보게 된다.

5. 차별의 해소 방법

차별 문제에 대한 해결 방법을 기준으로 구분하는 것이다.⁷⁹⁾ 차별 문제를 해결한다는 것은 차별의 부당성 또는 불법성을 시정할 수 있는 조치를 취한다는 것이다. 따라서 부당성 또는 불법성의 근거를 어디에서 찾느냐에 따라 해결 방법도 달라진다. 행위자의 악의적인 의도에 부당성이 있다고 보면 단지 그런 의도를 갖지 말라고 권고하는 것이 유일한 해결책이 된다. 행위에서 부당성의 근거를 찾으면 문제되는 행위를 취소하고 다음부터 동일한 행위를 하지 못하도록 하면 된다. 행위의 결과에서 부당성 또는 불법성을 찾으면 결과로서 발생한 피해를 복구하는 보상 조치가 필요하다.⁸⁰⁾

그런데 차별에 대한 해결 방법의 차이는 직접차별과 간접차별을 구분하는 근거가 될 수도 있지만 오히려 직접차별과 간접차별을 구분한 결과가 해결 방법의 차이로 나타난 것일 수도 있다. 예를 들어 차별의 결과에 대해 재정적 보상 지원까지 요청하는 것을 직접차별로 보고, 차별적 규정의 사용을 중단하도록 하는 데 그치는 것을 간접차별로 보는 것이라고 할 수도 있지만, 오히려 다른 방식으로 구분한 직접차별과 간접차별에 대해 그에 상응하는 해결 방법으로 취하던 조치가 어느 정도 유형화할 정도로 구별된 결과로 보는 것이 타당할 수 있다. 행위에서 부당성의 근거를 찾으면서 문제되는 행위를 취소한 경우에도 이에 그치지 않고 취소한 행위를 대체할 수 있는 새로운 행위를 권고하여 보다 적극적인 해결책을 제시할 수도 있기 때문이다. 예를 들어 고용주에게 고용주의 정당한 목적을 고려해 비례성 심사를 충족할 수 있는지 필요성의 관점에서 보다 적합한 대안을 제시하는 것이다. 이는 고용주에게 보호되는 집단에게 공정한 기회를 제공할 수 있는 규정, 기준, 관행을

78) 미연방대법원의 차별에 대한 단계 심사를 차별 발견에 관한 휴리스틱(Heuristics)으로 보는 경우로 Deborah Hellman, "Two Concepts of Discrimination", *Virginia Law Review* 102(4), 2016, pp. 895~952: pp. 906~909 참조.

79) Hugh Collins · Tarunabh Khaitan, "Indirect Discrimination Law: Controversies and Critical Questions", in *Foundations of Indirect Discrimination Law*, Hugh Collins · Tarunabh Khaitan(Eds.), Hart Publishing, an imprint of Bloomsbury Publishing Plc, 2018, pp. 1~30: pp. 23~24.

80) 차별시정조치와 관련된 차별금지의무의 내용은 본 논문 「제5장 제1절 III. 차별금지의무의 성격과 내용」 참조.



정립할 것을 고용주에게 요구함으로써 소수자 집단에게 우호적인 적극적 조치 명령에 근접하게 된다.⁸¹⁾ 이때 적절한 권고는 의무적으로 제시할 수도 있고 선택적으로 제시할 수도 있을 것이다.

6. 구분의 포기

직접차별과 간접차별의 차이를 명확하게 관념화하는 것은 종종 난관에 봉착하지만, 차별금지법의 기초를 이해하는 데에 도움을 준다. 그렇지만 이러한 구분이 명확히 정의된다고 해도 실제 법원이나 재판소에서 다른 방식으로 그 의미가 조작되는 경향이 있다는 점도 인정해야 한다. 유럽연합(EU) 법에서 재판소는 어떤 집단에게 불균형 또는 불비례적인 불이익이 존재한다는 점으로부터 중립적인 규칙이나 관행이 실제로는 은폐된 의도적 차별의 예라는 것을 추론한다.⁸²⁾ 직접차별과 간접차별의 차이를 의도의 존재 여부로 가르는 구분에 따르면 그러한 중립적 규칙이나 관행은 간접차별이면서 직접차별인 것이다. 특정 집단에 대한 실질적이고 지속적인 차별을 간접차별로 보면서 직접차별로 간주하는 이러한 증명 방식에서 의도의 추론은 비록 보호되는 집단의 특성과 관련 없는 특성이지만 보호되는 집단이 불평등하게 겪고 있는 불이익에 대해 직접차별을 청구할 수 있는 근거를 제공한다. 그러나 EU법은 간접차별의 역할이 차별금지법의 독립적 형식으로서 해악의 의도와 상관없이 적용될 수 있다는 데에 있다는 것도 수용한다. 실제로 간접차별의 기능을 직접차별의 증명 도구로 약화시키는 것은 모두에게 공정한 기회를 제공하는 데 방해가 되는 구조적 불평등과 비가시적 장벽을 제거하는 목적에 사용될 수 있는 법적 메커니즘을 간접차별에 관한 법에서 발견하려는 시도를 좌절시키는 것이기도 하다.

이렇게 간접차별 개념의 기능적 사용은 구분의 기초가 되는 도덕적 원칙 없이 직접차별과 간접차별을 구분하는 것은 불안정할 수밖에 없다는 것을 증명하는 것처럼 보이게 한다. 그럼에도 불구하고 구분을 조작하고 각 개념을 많은 사실 관계에 적용하여 절차적, 실체적 또는 구제의 이익을 얻도록 하는 능력은 직접차별과

81) 이러한 접근을 적극적 시정조치와 간접차별을 연결하는 흥미로운 시도로 보기도 한다. Kasper Lippert-Rasmussen, "Indirect Discrimination, Affirmative Action and Relational Egalitarianism", in *Foundations of Indirect Discrimination Law*, Hugh Collins · Tarunabh Khaitan(Eds.), Hart Publishing, an imprint of Bloomsbury Publishing Plc, 2018, pp. 173~196 참조.

82) CHEZ Razpredelenie Bulgaria AD v Komisia za Zashtita ot Diskriminatsia, C-83/14, Judgment of the Court of 16 July 2015, para. 94: "직접차별과 달리 간접차별은 비록 중립적 용어로 표현되었지만, 다시 말해 보호되는 특성과 관련되지 않은 다른 기준에 의한 것이지만 특별히 그 특징을 가진 사람들이 불이익을 받는 결과를 초래하는 조치에 기인한다."



간접차별의 구별을 포기하고 차별에 대한 통일된 법으로 대체되어야 한다는 결론으로 이끌기도 한다.⁸³⁾ 예컨대 캐나다 대법원은 구분의 인위성(artificiality), 차별 방법에 따라 달라지는 구체책, 불리한 영향을 받는 집단은 항상 숫자상 소수자라는 의심스러운 가정, 고용주의 방어에 실제로 적용할 때 어려움, 시스템의 차별에 대한 정당화, 전통적 분석과 인권법의 표현목적 및 조건 간의 불일치, 인권 분석과 현장 분석의 불일치 등을 이유로 차별 개념에 대한 종래 분석에 대한 변경의 필요성을 피력했다.⁸⁴⁾

III. 머신러닝 알고리즘의 이용과 차별의 우회

1. 머신러닝 알고리즘의 의도 숨기기와 직접차별

간접차별 개념이 처음으로 문제된 미국에서 차별은 행위 중심으로 구성하여 행위자의 차별에 대한 의도가 있었는지 여부에 따라 의도가 있는 경우 직접차별로 보아 그에 대한 심사를 엄격하게 한다. 이때 직접성은 차별금지 사유에 대한 것이며 동시에 그러한 사유를 기준으로 형성된 집단에 대한 것이기도 하다. 그래서 법으로 차별을 금지하는 사유를 결정이나 판단의 기초로 삼는 것은 동일한 특성을 공유하는 집단에 대한 직접적 차별로 이해하는 것이다. 그렇기 때문에 일단 법으로 금지된 사유를 기준으로 분류가 실행되면 그 자체로 차별의 의심을 받는다. 그런데 목적 변수나 클래스 레이블, 표본 레이블을 정의하고, 훈련용 데이터를 수집하면서 특성을 선택하는 과정에서 대용물을 사용하여 만들어진 모델을 잘 활용하는 결정자는 그 의도를 쉽게 가릴 수 있다.⁸⁵⁾ 편견적인 관점을 가진 결정자들의 의도를 머신러닝 알고리즘으로 쉽게 가릴 수 있는 것이다.

그런데 결과만을 놓고 차별을 의심할 뿐 그 의도를 알 수 없는 상황도 실제로는 세심하게 조직된 것일 수 있다. 예를 들어 결정자들은 어떤 목적을 가지고 모델을 훈련시키는 데이터 수집을 편향적으로 할 수 있다. 그렇다면 학습을 통해 생성되는 알고리즘은 특정 집단의 특수한 자격을 가진 구성원에게 비우호적인 규칙으로

83) British Columbia (Public Service Employee Relations Commission) v British Columbia Government Service Employees' Union, Case No. 26374, [1999] 3 S.C.R. 3 참조.

84) British Columbia (Public Service Employee Relations Commission) v British Columbia Government Service Employees' Union, Case No. 26374, [1999] 3 S.C.R. 3: pp. 17~30(para. 27~49).

85) Solon Barocas · Andrew D. Selbst, "Big Data's Disparate Impact", *California Law Review* 104, 2016, pp. 671~732: pp. 692~693 참조.



모델을 형성하도록 제안할 수 있다. 그 과정이 복잡하고 길다고 해도 얼마든지 시도할 수 있는 일이다. 왜냐하면 결정자의 결정을 지원하는 모델의 불편부당성을 주장하거나, 훈련용 데이터에 담겨 있는 이전의 관행이 반복된 것일 뿐이라고 주장하거나 그렇지 않으면 모델의 단순 오류를 지적함으로써 얼마든지 결정의 결과에 대한 책임을 회피할 수 있기 때문이다. 또한 정교하고 세밀한 차이보다는 굵고 성긴 차이를 만들어 내는 특성을 선택하면 결정의 결과에 나타나는 특정 집단의 불비례적인 구성⁸⁶⁾에 대해서 얼마든지 정당화할 수 있는 여지를 남길 수 있다. 그리고 하위 수준의 여러 가지 변수들을 포착하는 데 실패한 분석임에도 불구하고 이러한 분석이 추상화된 모델을 사용해서 구직자를 선별한 결과 어떤 특정 집단에 속한 사람들을 유망하지 않은 후보자로 평가하게 되는 경우에 설령 그러한 분석 과정에 그 어떤 의도가 담겨 있다고 하더라도 상위 수준으로 추상화되는 과정 속에서 그 의도는 얼마든지 희석되고 사라질 수 있다.

2. 머신러닝 알고리즘의 특성 추론과 간접차별

결정자는 어떤 합당한 경영상의 이유보다는 편견 때문에 발생하는 결정의 특수성을 제한하기 위해 의도적으로 보다 일반적인 특성을 선택할 수도 있다. 이는 마치 캔버스의 특정 부분을 작은 붓으로 칠하여 색의 편중을 드러내는 것보다 캔버스 전반을 큰 붓으로 넓게 칠하여 색을 분산시킴으로써 편중의 정도를 가리는 것과 유사한 태도라고 할 수 있다. 똑같은 편향을 가진 색칠이라고 하더라도 작은 붓으로 한 곳에 집중해서 칠하는 것보다는 넓은 붓으로 여러 곳을 칠하는 것이 편향의 중심을 흐트러뜨리는 착시 효과를 발생시키기 때문이다.⁸⁷⁾

그러나 이러한 선택은 불비례적인 구성의 특정 집단을 결정짓는 특성 즉, 금지된 차별 사유를 가림으로써 그러한 특성을 가진 특정 집단의 구성원을 체계적으로 (systematically) 열악한 위치에 가져다 놓을 수 있다. 머신러닝 알고리즘은 잠재적으로 보이지 않는 속성들을 추론하기 때문에 민감한 개인정보⁸⁸⁾를 포함해 결정의 근거로

86) 차별 인식에서 불비례성의 역할은 본 논문 「제4장 제3절 II. 차별의 인식과 불비례성」 참조.

87) 편향과 분산에 관해서 본 논문 「제2장 제2절 IV. 5. 유추주의와 서포트 벡터 및 사례 기반의 조건부 최적화」 참조.

88) 개인정보 보호법[법률 제14839호, 2017. 7. 26. 개정] 제23조 제1항: “개인정보처리자는 사상·신념, 노동조합·정당의 가입·탈퇴, 정치적 견해, 건강, 성생활 등에 관한 정보, 그 밖에 정보주체의 사생활을 현저히 침해할 우려가 있는 개인정보로서 대통령령으로 정하는 정보(이하 “민감정보”라 한다)를 처리하여서는 아니 된다.(단서 생략)”



삼지 않도록 법으로 정해진 특성들까지 추론해 낼 수 있다.⁸⁹⁾ 예를 들어 마케팅 기업이 사용하는 머신러닝 알고리즘은 여성 고객이 신생아용 기저귀나 무향 비누, 엽산제 등을 인터넷 장바구니에 넣어 둔 사실로부터 고객의 임신 사실을 추론하여 임신 및 출산 용품에 관한 맞춤형 광고를 보여줄 수 있다.⁹⁰⁾

이로써 개인들을 특정 집단에 연결시킬 수 있는 특성들을 찾아내어 그 집단의 구성원에게 귀속시킬 수 있고, 해당 집단이 사회적으로 또는 법적으로 비우호적인 집단일 경우 이를 매개로 더 쉽게 비하하고 처벌하고 배제할 수 있다. 머신러닝 알고리즘은 특정 집단의 구성원을 구별하여 결정자들에게 이익을 줄 수 있고, 결정자들이 개인의 집단 자격에 관한 명확한 정보에 접근할 수 없는 경우에도 그들을 구별할 수 있는 능력을 갖게 해준다. 특히 머신러닝 알고리즘은 그런 집단의 자격이 되는 특성의 대용물을 찾아주는데, 이것은 직접적으로 개인들을 특정 집단으로 분류한다는 외부의 시선을 차단하는 보호막을 설치하는 것과 같다. 게다가 훨씬 연결고리가 멀리 떨어져 있고 복잡한 대용물들을 발견하는 경우 그 보호막은 더욱 강력해진다. 이러한 머신러닝 알고리즘을 활용하는 결정자들은 그런 사실을 직접 배우지 않고도 개인들을 각각의 집단으로 자동적으로 분류할 수 있는 위치에 서게 되면서, 동시에 개인의 특정 집단 자격을 규정하는 특성을 고려하지 못하도록 만들어 놓은 법적 제한을 쉽게 우회할 수 있는 길 위에 들어서게 된다. 그리고 무엇보다 이런 머신러닝 알고리즘은 값비싼 것으로 머물러 있지 않고 보다 흔해졌고 그래서 간과하기도 더 쉬워졌다.⁹¹⁾

3. 머신러닝 알고리즘의 통계적 추론과 합리적 차별

머신러닝 알고리즘이 추론해낸 특성들은 데이터에 기반을 두고 있다. 머신러닝 알고리즘은 데이터셋으로 훈련되기 때문에 그 자체로 통계적 기반을 갖고 있으며 자의적인 추론이 아니라 통계적 근거가 있는 추론인 것이다.⁹²⁾ 통계적 특성은 종종

89) 머신러닝 알고리즘을 이용해 사람에 관한 데이터를 처리할 때 중첩적으로 발생하는 개인정보 보호의 문제는 본 논문 「제5장 제3절 머신러닝 알고리즘의 차별로부터 보호와 개인정보의 보호」 참조.

90) Kashmir Hill, “How Target Figured Out a Teen Girl Was Pregnant Before Her Father Did”, *Forbes*, 16 February 2012, <https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did>, 접속일: 2018년 7월 3일.

91) Solon Barocas · Andrew D. Selbst, “Big Data’s Disparate Impact”, *California Law Review* 104, 2016, pp. 671~732: pp. 692~693 참조.

92) 통계적 근거가 합리적인 이유로 제시되는 경우에 관해서 본 논문 「제5장 제3절 II. 3. 통계적 분석에 대한 차별금지와 개인정보보호의 접근방식」 참조.



객관성과 합리성을 담보하는 것처럼 보이게 할 수 있다. 그렇기 때문에 합리성의 외피를 두르고 있는 머신러닝 알고리즘의 결정은 합리적일 것이라는 믿음 속에 차별의 의심에서 제외되거나 통계와 확률의 뒷받침을 받는 인과성과 합리성을 토대로 쉽게 정당화될 수 있다. 합리성 심사를 통해 차별 여부를 판단하게 될 경우 합리적 차별은 더 이상 법적으로 금지된 차별이 아닌 것이다.

머신러닝 알고리즘은 분명 차별의 합리화를 통해 기존에 차별이 금지된 사유를 무력화하는 것은 물론 비가시적인 새로운 차별 사유를 생산해낼 수 있는 능력을 가지고 있다. 차별금지법의 역사에서 장애를 차별금지사유로 채택한 것은 합리적이라고 여겨지는 이유들에 근거한 차별에 대해서도 제한을 가할 수 있는 여지를 만들어 놓은 것이기도 하다. 따라서 차별을 합리적인 것이 무엇인지 찾기 위한 개념이 아니라 “합리성의 자기반성적 개념(die Selbstreflexion der Rationalität)”⁹³⁾으로 보는 것처럼 합리성과 차별의 관계는 지속적으로 재정립될 필요가 있다.

예를 들어 기업이 머신러닝 알고리즘을 이용하여 직원을 채용하는 모델을 생성할 때 과거에 해당 기업에 축적된 채용 평가 데이터를 사용하는 경우⁹⁴⁾ 이러한 통계적 자료를 활용하는 것은 사적 자율성의 행사로서 보장 받고 그 결과 역시 통계적 근거를 가진 합리적인 것으로 평가될 수 있다. 이 경우 차별 개념은 채용 모델을 생성할 때 보다 정확하고 공정한 결과가 나올 수 있도록 조치해야 한다는 규범의 근거가 될 수 있다. 그러나 문제는 여기서 끝나지 않을 수 있다.

합리성의 대용물로 사용되는 대표적인 가치는 효율성이다. 설령 채용 모델이 부정확한 데이터나 집단에 대한 편향이 심한 데이터를 기반으로 생성된 것이라고 할지라도 정확한 데이터나 편향이 낮은 데이터를 수집하는 데에 비용이 드는 경우 기업은 경영상의 효율성을 이유로 채용 모델의 적용에 대한 합리성을 또 다시 주장할 수 있다. 그렇다면 차별을 통해 궁극적으로 제기하는 반성적 고려는 효율성을 합리성의 대용물로 사용하는 고정관념에 관한 것이 될 수 있다.⁹⁵⁾

93) Alexander Somek, *Rationalität und Diskriminierung:: Zur Bindung der Gesetzgebung an das Gleichheitsrecht*, Springer, 2001, S. 37~39 참조.

94) 이와 관련된 실제 사례와 문제점은 Jeffrey Dastin, “Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women”, Reuters, 10 October 2018, <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G> 및 본 논문 「제2장 제3절 III. 1. 고용 알고리즘과 편향의 학습」과 「제3장 제1절 IV. 1. 집단에 대한 또 다른 특성의 귀속」 참조.

95) 차별의 형식으로서 고정관념과 합리적 차별의 관계에 대해서 본 논문 「제3장 제2절 II. 3. 통계적 고정관념과 합리적 차별」 참조.



또 다른 예로 머신러닝 알고리즘과 관련해서 차별금지사유로 제정될 수 있는 대표적인 특성은 지능이다. 지능을 사유로 사람을 구별하는 것은 지극히 합리적이고 당연한 것으로 여겨져 법체계에 편입되어 있다. 대한민국에서 지능은 범죄를 반복할 위험성을 평가하는 한 요소가 된다.⁹⁶⁾ 재범 위험성을 평가하는 기준으로 지능을 그 판단의 근거로 삼는 것에 대해 합리적이라는 사회적 합의는 더 이상의 자기반성적 판단으로 나아가는 것을 멈추게 함으로써 그 합의를 확정적이고 안정적인 것으로 만든다. 그러나 차별의 렌즈를 통할 경우 지능을 이유로 인간을 구별하는 것의 합리성은 재고의 대상이 된다.

인간의 재범 위험성을 지능에 따라 평가해온 역사를 표현하는 축적된 데이터를 기반으로 훈련된 머신러닝 알고리즘에 따라 작동하는 인공지능 에이전트를 가정해 보면, 인공지능 에이전트가 지능이 낮은 인간에 대해 범죄를 지을 확률이 높은 위험 대상으로 평가하고 그에 따른 행위를 하더라도 기존의 합리성 기준에만 의지하는 사회에서는 지속적으로 합리적이라고 판단될 수밖에 없다. 또한 인간(human)과 비인간(nonhuman)의 관계를 지능에 따라 위계적으로 구성할 경우 인간은 그보다 지능이 낮은 동물보다는 우위를 점할 수 있기 때문에 차별이 정당화될 수 있을지 모르지만 인간보다 지능이 높은 인공지능 에이전트가 출현할 경우 인간이 동물을 차별하는 바로 그 합리적 근거는 인간 종(species)이 차별받는 것을 정당화하는 유력한 논거가 될 수 있다.⁹⁷⁾ 이와 같은 경우 차별 개념은 합리적이라고 판단된 것에 대해 외부적 시선으로 다시금 자기를 관찰하는 계기를 마련해 줄 수 있다.

96) 형법 제51조는 “형을 정함에 있어서는 다음 사항을 참작하여야 한다.”고 한 후 제1호에서 “범인의 연령, 성행, 지능과 환경”을 규정함으로써 지능을 형사재판의 양형 참작사유로 고려하도록 한다. 머신러닝 알고리즘이 재범의 위험성 평가에 사용되는 경우에 관해서 본 논문 「제2장 제3절 I. 2. 루미스 사건」 및 「제5장 제3절 III. 1. 개인정보 수집제한의 한계와 이용단계의 중요성」 참조.

97) 싱어(P. Singer)는 지능을 기준으로 범주를 구분하고 분류하여 정렬함으로써 형성된 위계를 “지능의 위계(hierarchy of intelligence)”로 표현하면서 이러한 종류의 차별주의를 인종이나 성별과 같은 기준에 따른 노골적인(blatant) 차별주의보다 세련된(sophisticated) 것으로 본다. 또한 평등한 이익고려의 원칙(the principle of equal consideration of interests)을 인간의 평등을 위해 가능한 최선의 기초로 보면서 이러한 평등의 원칙이 적용되는 범위를 인간에게 한정시키지 않고 인간이 아닌 동물(nonhuman animals)에게도 확장시킨다. 이에 관해서 Peter Singer, *Practical Ethics*, 3rd ed., Cambridge University Press, 2011, pp. 19~20 및 pp. 48~70 참조.



제3절 머신러닝 알고리즘의 최적화와 차별의 정당화

머신러닝 알고리즘의 중요한 연구 목표 중에 하나는 최적화이다. 즉, 문제를 해결하는 최적의 방법을 찾는 것이다.⁹⁸⁾ 가장 효율적인 알고리즘을 찾는 과정은 비용을 최소화하는 것이고 이때 비용은 계산의 복잡도와 소요 시간으로 나타난다. 특히 복잡도는 이론상 해결할 수 없는 문제 또는 해결 가능하더라도 시간이 필요 이상으로 많이 소요되어 실천의 영역에서 다루는 것이 비효율적인 문제를 구별하는 기준이 된다. 사회의 문제를 해결하는 방법으로서 법 역시 해결할 수 있는 문제와 해결하기에 너무 많은 시간이 소요되는 문제를 구별해 줄 수 있다. 또한 추상적 수준에서 법의 복잡도가 너무 높아지면 이를 사례에 구체화하는 과정에서 복잡도가 더 상승하게 된다. 복잡도의 상승과 시간의 지연을 막기 위해서는 단순화한 모델을 만들어 적용하거나 어느 지점에선가 세부적인 논증을 중단해야 한다. 그리고 가치나 이익 간 비교의 과정은 이와 같은 단순화나 논증의 중단을 통해 비로소 결론에 도달할 수 있고 이러한 단순화나 논증의 중단 절차에서 또 다른 차별이 발생하는 것은 아닌지 문제된다.

I. 머신러닝 알고리즘과 법적 추론에서 최적화

1. 법적 특이점으로서 반성적 평형

인공지능 알고리즘이 법 실무에 어떤 영향을 미칠 것인지에 대해서 누구도 확실한 답을 내놓을 수는 없다. 다만 기술의 발전과 실무에 적용되는 종류와 속도를 바탕으로 여러 가지 전망과 예견을 할 수 있을 뿐이다. 알레어(B. Alaire)의 경우 미래의 법을 예견하면서 인공의 법 지능 에이전트가 활동함으로써 더 이상 법의 불확실성을 걱정하지 않아도 되는 순간을 “법적 특이점(legal singularity)”⁹⁹⁾이라고 부르기도 한다.

98) 알고리즘의 연구 목표로 제시될 수 있는 문제들 중에는 최적화 문제 이외에도 두 가지 가능성 중 하나를 선택하는 결정 문제, 어떤 대상을 찾는 문제로서 검색 문제 등 여러 종류가 있을 뿐만 아니라 알고리즘이 존재하지 않는 해결 불가능한 문제나 이론상 알고리즘이 존재하지만 지나치게 비효율적인 것만 존재하는 경우 등 문제에 대한 정답보다 근사값을 찾는 것이 중요한 문제도 있다.

99) Benjamin Alarie, “The Path of the Law: Towards Legal Singularity”, *University of Toronto Law Journal* 66(4), 2016, pp. 443~455: p. 445.



본래 특이점(singularity) 개념의 기원은 천체 물리학의 특이점 이론¹⁰⁰⁾에 두고 있다. 특이점은 일반적으로 알려진 물리학 법칙과 시간이 그 작동을 멈추는 극히 작으면서 극히 뜨거운 고밀도의 에너지 덩어리이다. 이 특이점 개념은 인공지능이 자율성을 획득하고, 인간의 자연지능을 따라잡는 것은 물론이고 이를 초월함으로써 더 이상 인간의 지능으로 이해할 수 없는 순간이 다가올 것이라고 예견하기 위해 미래학자 커즈와일(R. Kurzweil)이 차용하면서 사회경제학적 의미를 갖게 된다. 그는 특이점을 “기술의 변화 속도가 너무 빠르고 그 영향이 너무 깊어서 인간 생활이 돌이킬 수 없게 변화하는 미래 시기”¹⁰¹⁾라는 의미로 사용한다.

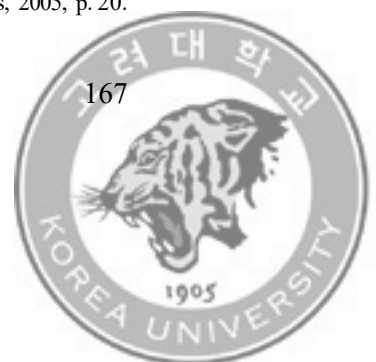
알레어는 이런 용어를 법의 영역에 적용한 것이다. 법이 특이점에 이를 것이라는 주장은 두 가지 추세에 근거를 두고 있다. 하나는 세계의 관찰 가능한 현상들의 정량화가 방대해지고 있다는 것이고 다른 하나는 패턴을 인식하는 새로운 기술과 방법이 훨씬 정교해지고 있다는 것이다. 이것은 달리 말해 처리할 수 있는 데이터가 방대해지고 있다는 것과 그러한 방대한 데이터에 근거한 규칙의 추론이 더욱 정교해진다는 것이다. 이러한 추세가 정점에 이르렀을 때 법적 특이점이 도래하고 이 시기에 법적 불확실성은 시대에 뒤떨어진 유물로 전락하게 된다고 본다.

아울러 법적 특이점은 한편으로는 그러한 시기가 도래할 것이라는 전망에 대한 회의론자들이 옳다는 것이 경험적으로 잘못됐다고 증명되는 시기이면서 다른 한편으로는 롤즈(J. Rawls)가 ‘정의론(a theory of justice)’에서 제시한 “반성적 평형(reflective equilibrium)”¹⁰²⁾ 상태에 도달한 시기이기도 하다. 롤즈의 정식화에 따르면 정의의 원칙들로부터 나온 판단들이 충돌할 경우, 충돌하는 다양한 신념들을 안정적인 평형 상태에 도달할 때까지 조정함으로써 그 충돌은 극복된다. 롤즈의 반성적 평형 상태에서 정의의 원칙들과 그 결과인 법들은 안정적이고 충돌하지 않으며 법의 지배를 받는 사람들에게 실천적인 지침을 제공한다. 구체적 상황 속에 있는 개인에게 꼭 맞는 최적화된 법(optimized law)이 안내되는 것이다.

100) 특이점 이론은 1966년 천체 물리학자 스티븐 호킹(Stephen Hawking)이 수학자 로저 펜로즈(Roger Penrose)와 협동해 블랙홀(black hole)에 특이한 한 점이 존재한다는 것을 수학적으로 증명하면서, 우주가 빛까지 집어 삼키는 블랙홀의 한 점, 즉 특이점에서 출발해 시작되었다고 주장한 이론이다.

101) Ray Kurzweil, *The Singularity Is Near: When Humans Transcend Biology*, Viking, 2005, pp. 7~8.

102) John Rawls, *A Theory of Justice*, Original ed., Belknap Press of Harvard University Press, 2005, p. 20.



2. 머신러닝 알고리즘에서 최적화

최적화는 머신러닝 알고리즘의 학습 개념에서 중요한 요소 중에 하나로서¹⁰³⁾ 함수에서 최고의 출력값을 산출하는 입력값을 발견하는 것과 관련된 수학의 분과이기도 하다.¹⁰⁴⁾ 머신러닝 알고리즘이 학습을 위해 논리, 신경망, 유전자 프로그래밍, 그래픽 모형, 서포트 벡터 등으로 표현하고, 이를 정확도, 제공 오차,¹⁰⁵⁾ 적합성, 사후 확률, 마진 등으로 평가했을 때 최적화는 가장 높은 점수로 평가되는 모델을 찾아 주는 것에 관한 문제인 것이다.¹⁰⁶⁾ 예를 들어 총 수입을 최대화하도록 주식을 매수하고 매도하는 순서를 발견하는 것, 임의의 데이터를 정확하게 분류하기 위해 판별식의 마진이나 예시들의 가중치를 변경하는 것 등은 최적화 문제이다.

또한 최적화는 복잡하고 변화무쌍한 환경에 가장 잘 적응할 수 있는 상태를 만드는 것이기도 하다. 잘 적응한다는 것은 되도록 적은 에너지를 들여 빠른 시간 안에 설정된 목표에 도달할 수 있는 방법을 고안한다는 것이다. 머신러닝 알고리즘은 병, 죽음, 굶이, 길이, 높이, 온도, 시간 등 사실적 대상을 최적화할 수 있기 때문에 최적의 알고리즘을 발견하거나 설계하는 것은 머신러닝 알고리즘을 연구하는 목표 중에 하나가 되기도 한다.

3. 법적 추론에서 최적화

최적화는 법적 추론에 관한 이론 구성에 응용되기도 하는데,¹⁰⁷⁾ 규범을 원칙(Prinzipien)과 규칙(Regeln)으로 모델화하는 기본 구성에서 원칙 모델을 최적화요구(Optimierungsgebote)¹⁰⁸⁾로 보는 것이다.¹⁰⁹⁾ 규범 모델로서 최적화는 출력값인 목적이

103) Pedro Domingos, “A Few Useful Things to Know About Machine Learning”, *Communications of the ACM* 55(10), 2012, pp. 78~87: pp. 79~80 및 본 논문 「제2장 제2절 II. 2. 머신러닝 알고리즘의 학습과 데이터」 참조.

104) Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, New York: Basic Books, a member of the Perseus Books Group, 2015, p. 30.

105) 제공 오차는 예측 값과 실제 값의 차이를 제공하여 더한 것이다.

106) Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, New York: Basic Books, a member of the Perseus Books Group, 2015, pp. 239~241 참조.

107) Robert Alexy, *Theorie der Grundrechte*, 1. Aufl., Suhrkamp, 1994, S. 75~77 참조.

108) 알렉시(R. Alexy)는 최적화의 맥락에서는 명령(Gebot)을 허용(Erlaubnisse)과 금지(Verbote)를 포괄하는 넓은 의미로 사용한다. Robert Alexy, *Theorie der Grundrechte*, 1. Aufl., Suhrkamp, 1994, S. 76, Fn. 23 참조.

109) 이준일, “법학에서 최적화”, 법철학연구 3(1), 2000, 101~130쪽 참조.



최고값에 도달할 수 있도록 하는 조건들을 입력값으로 찾아야 한다는 것이다. 드워킨(R. Dworkin)이 도입한 법원칙 모델을 구체적으로 전개시킨 알렉시(R. Alexy)의 원칙이론¹¹⁰⁾에 따르면 최적화를 요구하는 원칙의 성격으로부터 적합성, 필요성, 좁은 의미의 비례성의 세부 원칙을 갖는 비례성원칙이 논리적으로 도출되며, 반대로 비례성원칙으로부터도 원칙의 최적화요구로서 성격이 논리적으로 귀결된다.¹¹¹⁾ 이렇게 함으로써 최적화를 요구하는 이론으로서 원칙이론을 비례성원칙과 필연적 관계로 결합시킨다.¹¹²⁾ 이러한 이론 구성을 기본권이론에 적용하여 기본권에 원칙 규범의 성격을 부여하는 것은 기본권 실현을 최대화할 수 있는 조건들을 입력값으로 찾는 과제가 기본권을 체계 내에 편입한 헌법의 과제가 될 수 있도록 하려는 기획으로도 볼 수 있다.¹¹³⁾

이때 최적화의 조건으로 법적인 가능성과 사실적 가능성이라는 제약을 가하면 복잡한 차원에 접어드는 것으로부터 조금이라도 비켜가는 효과를 기대할 수 있다. 최적화에서는 단순한 함수들이 종종 복잡한 해법을 제시하기 때문에 복잡성을 감축시키기 위해 제약 사항을 부가하여 최적화에 일정 수준의 제한을 가하는 일종의 조건부 최적화(constrained optimization)인 셈이다. 조건부 최적화는 제한적 조건들에 종속된 함수를 최대화하거나 최소화하는 문제로 머신러닝에 관한 유추주의의 최상 알고리즘인 서포트 벡터 머신에 사용되는 기법이기도 하다.¹¹⁴⁾ 실제로 알렉시는 법적 추론의 기초가 되는 논변 형식으로 포섭 공식(the Subsumption Formula)만을

110) 알렉시의 원칙이론이 고전 논리(classical logic)에 따르고 있다고 보며, 규칙과 원칙 개념에 가폐논리(defeasible logic)를 적용하는 것이 적절하다는 주장으로 Bartosz Brożek, “Legal Rules and Principles: A Theory Revisited”, *I-Lex* 7(17), 2012, pp. 205-226 참조; 법적 논증에서 인공지능의 논리 기반 추론에서 사용된 비단조논리(Non-monotonic Logic)의 활용 가능성을 다룬 것으로 유승익, “인공지능의 추론방식을 활용한 법적 논증 - 폐기가능성, 비단조논리, 형량 -”, 원광법학 32(2), 2016, pp. 299-323 참조.

111) Robert Alexy, *A Theory of Constitutional Rights*, Julian Rivers(Trans.), Oxford University Press, 2002, p. 66.

112) Robert Alexy, “Constitutional Rights and Proportionality”, *Revus: Journal for Constitutional Theory and Philosophy of Law* 22, 2014, pp. 51-65: p. 57.

113) 규범을 규칙과 원칙으로 구분하면서 다른 구조를 가진 것으로 보는 법존재론(legal ontology)의 성격이 원칙이론의 특성이자 전략이라는 지적으로 Ralf Poscher, “The Principles Theory: How Many Theories and What Is Their Merit?”, in *Institutionalized Reason: The Jurisprudence of Robert Alexy*, Matthias Klatt(Ed.), Oxford University Press, 2012, pp. 218-247, 특히 p. 220: “법존재론의 이론적 주장을 토대로 형성된 원리적(doctrinal) 이론에 대해 원리적 논변으로 반박할 수 없다.”

114) 서포트 벡터 머신에 관해서 Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, New York: Basic Books, a member of the Perseus Books Group, 2015, p. 193 이하 및 이에 관한 본 논문 「제2장 제2절 IV. 5. 유추주의와 서포트 벡터 및 사례 기반의 조건부 최적화」 참조.



고려했다가,¹¹⁵⁾ 머신러닝 관련 용어를 사용해 ‘가중치 공식’이라고도 부를 수 있는 중요도 공식(the Weight Formula)¹¹⁶⁾을 추가하여 각각 규칙(rule) 개념과 원칙(principle) 개념에 연결시키고, 마지막 세 번째 공식으로 사례 간의 유추(analogy) 또는 비교(comparison)를 추가하여 사례(case) 개념에 연결시킴으로써 시스템을 완성한다.¹¹⁷⁾

이를 머신러닝 알고리즘의 방법론에 비유해 보면 마치 기호주의자의 포섭 공식과 베이지주의자의 중요도 공식에 유추주의자의 사례 비교를 결합시킨 것과 유사하다.¹¹⁸⁾ 원리로부터 출발하는 기호주의와 베이지주의 방식을 취합하고, 거기에 모델 없이 사례로부터 유추할 수 있는 방식을 추가함으로써 법적 논증 모델을 완결하려는 것이다. 이러한 3차원의 조합은 여전히 자연으로부터 학습하는 진화주의와 연결주의가 갖는 함의를 미지의 것으로 남겨둔다. 각각의 접근방식에 따라 법적 추론 알고리즘이 서로 연결되어 스스로 오류를 수정해 갈 뿐만 아니라, 서로 교차함으로써 진화할 수 있다는 가설이 반영된다면 법적 추론의 마스터 알고리즘도 탄생할 수 있을지 모른다. 물론 이러한 5차원의 시스템을 인간이 이해할 수 있을지는 의문이지만, 고차원 단계에서부터는 컴퓨터 알고리즘의 도움이 필요하다는 점을 역설적으로 말해주는 것일 수도 있다.

115) Robert Alexy, *A Theory of Legal Argumentation*, Ruth Adler · Neil MacCormick(Trans.), Clarendon Press, 1989, pp. 221~230.

116) 중요도 공식은 형량의 구조를 수학적 모델을 사용해 설명해 보려고 한 알렉시의 시도로 다음과 같이 표현된다.

$$W_{i,j} = \frac{I_i \cdot W_i \cdot R_i}{I_j \cdot W_j \cdot R_j}$$

알렉시에 따르면 형량은 사건을 경쟁하는 두 개의 원칙(principles: P_i, P_j)에 포섭시키는 것에서 출발하여 두 원칙에 대한 제약의 강도(the intensity of interferences: I_i, I_j), 두 원칙의 추상적 비중(abstract weights: W_i, W_j), 문제되는 수단이 사건의 구체적 상황에서 한 원칙의 비실현과 또 다른 한 원칙의 실현에 대해 갖는 의미 관한 경험적 가정의 신뢰도(the reliability of the empirical assumptions: R_i, R_j)에 각각 가치를 할당함으로써 구체적 비중($W_{i,j}$)을 산출한다. 이때 각 변수에 숫자를 할당하면 원칙 P_i 의 구체적 비중을 계산하는 것은 단순한 연역에 불과한 것이 된다. 이에 관해서 Robert Alexy, “On Balancing and Subsumption”, *Ratio Juris* 16, 2003, pp. 433~449: pp. 443~448 및 Robert Alexy, “Two or Three?”, in *On the Nature of Legal Principles*, Martin Borowski(Ed.), Franz Steiner and Nomos, 2010, pp. 9~18: pp. 10~11 참조.

117) Robert Alexy, “Two or Three?”, in *On the Nature of Legal Principles*, Martin Borowski(Ed.), Franz Steiner and Nomos, 2010, pp. 9~18: pp. 17~18.

118) 기호주의, 베이지주의, 유추주의에 대해서 Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, New York: Basic Books, a member of the Perseus Books Group, 2015, pp. 57~92, pp. 143~176, pp. 177~202 및 이에 관한 본 논문 「제2장 제2절 IV. 머신러닝 알고리즘과 인공지능」 참조.



II. 차별의 인식과 불비례성

1. 차별의 인식 도구로서 불비례성

차별을 인식하는 방법은 차이를 발견하는 것이다. 차이를 발견한다는 것은 차이를 가시화하는 것이기도 하다. 상대적인 성격을 갖는 차이를 집단 수준에서 찾기 위해 비율(proportion)이 활용될 수 있다. 예를 들어 집단 간에 이익을 얻거나 손해를 입은 구성원의 비율에 차이가 있다면 비례성(proportionality) 관념은 집단 간 차이를 발견하는 인식 도구로 사용될 수 있다. 그런데 이때 비례성은 차이가 비례적이지 않아 정당하지 않거나 중요하지 않다는 소극적 평가의 도구라기 보다 통계적 차원에서 불비례적인 차이를 적극적으로 인식할 수 있는 도구로서 불비례성(disproportionality)이라고 할 수 있다. 어떤 특성을 이유로 한 구별에 토대를 둔 차별은 같은 특성을 공유하는 사람들을 집단화한다. 그러므로 어떤 특성을 갖는 집단과 그와 다른 특성을 갖는 비교 집단 간의 차이를 차별로 포착하기 위한 지표를 어떻게 구성할 것인지는 차별을 인식하고 평가하는 데 이론적으로도 실천적으로도 매우 중요한 요소가 된다. 특히 이러한 지표는 차별에 관한 분석의 단계에서 각별한 관심을 받는다.¹¹⁹⁾ 더욱이 법의 효과에서 불비례성은 절차가 편견으로 오염됐다는 실패의 신호일 수 있기 때문이다.¹²⁰⁾

2. 법적 차별의 측정 지표

차별을 측정하는 다양한 방법은 세계에 널리 수용되어 있는데, 차별을 분석하기 위해 영국 법률은 ‘위험차(risk difference, RD)’를, 유럽연합사법재판소는 ‘위험비(risk ratio, RR)’를, 미국 법률과 법원은 주로 ‘상대적 기회(relative chance, RC)’를 언급한다. 그 밖에 오즈비(odds rate, OR)도 있다. 우선 각 지표 개념을 이해하기 위해 아래의 ‘자료 20’ 같은 간단한 통계 분할표를 그려볼 필요가 있다.¹²¹⁾ 두 집단에

119) Andrea Romei · Salvatore Ruggieri, “A Multidisciplinary Survey on Discrimination Analysis”, *Knowledge Engineering Review* 29(5), 2014, pp. 582~638 참조.

120) Vicki C. Jackson, “Constitutional Law in an Age of Proportionality”, *Yale Law Journal* 124(8), 2015, pp. 3094~3196: p. 3175.

121) Dino Pedreschi · Salvatore Ruggieri · Franco Turini, “The Discovery of Discrimination”, in *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, Bart Custers · Toon Calders · Bart Schermer · Tal Z. Zarsky(Eds.), Springer, 2013, pp. 91~108; 본문의 분할표는 p. 98에 있는 ‘Fig. 5.1’의 분할표를 약간 수정한 것이다.



대한 이익의 제공여부를 기준으로 하면 기본적으로 ‘2×2 분할’이지만 구별되는 집단의 수가 더 많아지면 ‘k×2 분할’의 표가 만들어질 수 있다.

자료 20. 보호집단과 비보호집단 중에 이익이 거부 또는 부여된 인원

집단	이익		합
	거부	부여	
보호되는	a	b	n1
보호되지 않는	c	d	n2
합	m1	m2	n

위 표는 어떤 특성을 이유로 차별 받지 않도록 보호되는 집단의 전체 인원(n1)과 그러한 특성을 갖지 않아 보호되지 않는 집단의 전체 인원(n2) 중에 각각 이익이 거부된 인원(a, c)과 이익이 부여된 인원(b, d)을 나눈 것이다. 이익이 거부된 인원(m1)과 이익이 부여된 인원(m2)을 합한 총 인원(n)은 보호집단의 전체 인원(n1)과 비보호집단의 전체 인원(n2)을 합한 총 인원(n)과 같다.

통계학에서 위험차(RD)는 절대위험감소(absolute risk reduction)로도 불리며 보호집단의 전체 인원(n1) 중에 이익이 거부된 인원(a)의 비율(a/n1)과 비보호집단의 전체 인원(n2) 중에 이익이 거부된 인원(c)의 비율(c/n2) 사이의 절대적 차이를 말한다. 이를 공식화하면 ‘ $RD = a/n1 - c/n2$ ’가 된다. ‘a/n1’을 ‘p1’로, ‘c/n2’를 ‘p2’로 대체 하면, ‘ $RD = p1 - p2$ ’가 된다.

위험비(RR) 또는 상대적 위험(relative risk)은 보호집단의 전체 인원 중에 이익이 거부된 사람의 비율(p1)과 비보호집단의 전체 인원 중에 이익이 거부된 사람의 비율(p2) 사이의 상대적 차이를 말한다. 이를 공식화하면 ‘ $RR = p1 / p2$ ’가 된다.

선택률(selection rate)로도 불리는 상대적 기회(RC)는 고용 차별에 관한 문헌에서 유래한 것으로 각 집단의 전체 인원에서 이익이 거부된 인원(a 또는 c)이 아니라 이익을 받은 인원(b 또는 d)의 비율을 변수로 삼는다. 즉, 상대적 기회(RC)는 보호집단의 전체 인원(n1) 중에 이익을 받은 인원(b)의 비율(1-p1)과 비보호집단의 전체 인원(n2) 중에 이익을 받은 인원(d)의 비율(1-p2) 사이의 상대적 차이를 측정하는 것이다. 이를 공식화하면 ‘ $RC = (1 - p1) / (1 - p2)$ ’가 된다.

교차비 또는 승산비로 불리는 오즈비(OR)는 보호집단의 전체 인원에서 이익이 거부된 인원(a)과 이익을 받은 인원의 상대적 차이(a/b)와 비보호집단의 전체 인원에서 이익이 거부된 인원(c)과 이익을 받은 인원(d)의 상대적 차이(c/d) 사이의 상대적



차이를 측정한다. 이를 공식화하면 ' $OR = (a / b) / (c / d) = (a \times d) / (b \times c) = p1(1 - p2) / p2(1 - p1)$ '로 된다.

3. 확장된 차별 측정 지표와 머신러닝 알고리즘의 차별

위험차(RD), 위험비(RR), 상대적 기회(RC) 모두 보호집단을 비보호집단과 비교함으로써 차별을 측정하는 것이라면, 이익이 거부된 비율($p1$) 또는 이익이 부여된 비율($1 - p1$)을 집단을 불문한 총 인원(n) 중 이익이 거부된 총 인원($m1$)의 평균 비율(p)과 비교하여 측정할 수도 있다. 확장된 기준으로서 이와 같은 지표는 데이터 마이닝의 차별 발견에 도입된다.¹²²⁾ 먼저 확장된 차이(extended difference, ED)는 보호집단의 전체 인원에서 이익이 거부된 인원의 비율($p1$)과 전체적으로 이익이 거부된 평균 비율(p)의 절대적 차이를 측정하는 것으로, ' $p1 - p$ '가 된다. 확장된 비율(extended ratio, ER)은 보호집단의 전체 인원에서 이익이 거부된 인원의 비율($p1$)을 이익이 거부된 평균 비율(p)과 상대적으로 비교하는 것으로 ' $p1 / p$ '가 된다. 마지막으로 확장된 기회(extended chance, EC)는 보호집단에서 이익을 부여받은 인원의 비율($1 - p1$)과 이익을 부여 받은 평균 비율($1 - p$)의 상대적 차이를 비교하는 것으로 ' $(1 - p1) / (1 - p)$ '로 공식화할 수 있다.

4. 소결

이상의 차별 인식 기준을 모으면 다음과 같다.

- ▶ 위험차(risk difference, RD) = $p1 - p2$
- ▶ 위험비(risk ratio, RR) = $p1 / p2$
- ▶ 상대적 기회(relative chance, RC) = $(1 - p1) / (1 - p2)$
- ▶ 오즈비(odds ratio, OR) = $p1(1 - p2) / p2(1 - p1)$
- ▶ 확장된 차이(extended difference, ED) = $p1 - p$
- ▶ 확장된 비율(extended ratio, ER) = $p1 / p$
- ▶ 확장된 기회(extended chance, EC) = $(1 - p1) / (1 - p)$

122) Dino Pedreschi · Salvatore Ruggieri · Franco Turini, "Measuring Discrimination in Socially-Sensitive Decision Records", *Proceedings of the SIAM International Conference on Data Mining*, SIAM, 2009, pp. 581~592.



차별에 관한 인식에서 보호집단이 비보호집단 또는 평균보다 이익이 더 많이 거절되거나 더 조금 부여되는 것에 초점을 맞춘다면 위험비(RR), 오즈비(OR), 확장된 비율(ER)의 값이 1보다 큰 경우, 위험차(RD), 확장된 차이(ED)의 값이 0보다 큰 경우, 그리고 상대적 기회(RC)와 확장된 기회(EC)의 값은 1보다 작은 경우에는 이익의 분배가 불비례적인 경우로서 의심스러운 차별 또는 잠정적인 차별 등으로 차별 평가의 대상이 될 수 있을 것이다.

III. 비례성 및 합리성과 차별의 평가

1. 차별의 평가 지표로서 비례성

비례성이 차별의 평가 도구로 사용될 때 두 가지 방식으로 사용될 수 있다. 그 중에 하나는 인식 도구로 사용된 불비례성에 대해 허용 가능한 정도를 정하는 것이다. 다시 말해 측정된 지표에 대해 어느 범위까지를 비례적이라고 볼 것인지 한계의 폭을 정하는 것이다. 예를 들어 위험비(RR), 오즈비(OR), 확장된 비율(ER) 처럼 비율 간 상대적 차이를 측정한 지표에 대해 1보다 큰 경우에도 1을 초과하기만 하면 곧바로 비례성을 충족하기 못하는 차별이라고 볼 것인지 2까지는 비례성을 충족한다고 볼 것인지 비례성의 내용을 결정하는 것이다. 평가적 관점에서 비례성의 용도는 인식된 차별적 상황을 수정할 것인지 여부를 결정하는 기준으로 사용된다. 차별을 인식하기 위해 적극적으로 사용되는 불비례성의 지표는 통계적 분석에 의존하지만 차별을 평가하기 위해 소극적으로 사용되는 비례성의 기준은 규범적이다.

예를 들면 어떤 정당에서 4명의 대의원을 선출하는 선거에 정강에 따라 여성에게 1석이 할당되고, 여성 2명과 남성 5명이 후보자로 나서 투표 결과 여성 1명, 남성 3명으로 대의원이 구성된 경우를 생각해 볼 수 있다. 대의원이 되지 못한 것 즉, 낙선을 불이익으로 보고, 여성을 보호집단으로 보는 조건에서 위험비(RR)로 차별을 측정하면 보호집단의 전체 인원 2명 중에 이익이 거부된 인원 1명의 비율 0.5와 비보호집단의 전체 인원 5명 중에 이익이 거부된 인원 2명의 비율 0.4 사이에 상대적 차이($0.5 / 0.4$)는 1.25이다. 이때 불비례성의 기준인 1은 통계적 분석에 의존하여 차별을 인식하는 데에 사용되지만 차별의 평가 도구로서 비례성 기준을 1.5로 정할 경우 규범적 차원에서 차별로 평가되지 않을 수 있다.



인식 기준과 평가 기준을 구별하는 관념은 보다 단순한 방식으로 선거구 획정에 관한 사례에서 확인할 수 있다. 선거구 A의 인구가 500명, 선거구 B의 인구가 400명인 상황에서 각각 1명의 대표를 뽑는다고 할 경우, A선거구와 B선거구의 상대적 기회(RC)는 A선거구에서 당선되어 대표가 되는 비율($1/500$)과 B선거구에서 당선되어 대표가 되는 비율($1/400$) 사이의 상대적 차이로 0.8이 된다. 이는 불비례성 기준인 1 미만이므로 차별로 인식될 수 있지만, 평가 도구로서 비례성 기준을 0.5로 정할 경우 규범적 차원에서 차별로 평가되지 않을 수 있다. 평등원칙의 기준으로 선거구 사이에 인구비례의 차이가 1:2를 넘지 말아야 한다는 결정¹²³도 1:1을 불비례적 평등 인식의 기준으로 하고, 평가의 기준으로 비례적 평등의 기준을 1:2로 설정한 것이라고 볼 수 있다.

미국의 고용평등기회위원회(EEOC)에서는 어느 정도의 불비례성이 법적으로 제재를 가하기에 의미 있는 것인지를 판단하는 기준으로 이른바 ‘5분의 4 규칙(Four-Fifths Rules)’을 제시하기도 한다.¹²⁴ 이때 각 집단에서 고용이 거부된 인원의 비율이 아니라 고용이 인정된 인원의 비율을 변수로 삼는다는 점에서 상대적 기회(relative chance)를 차별에 관한 측정 지표로 사용하는 것으로 볼 수 있다. 그리고 비율의 상대적 차이가 ‘4:5’를 넘지 않는 범위에서 법적으로 허용되는 비례성의 범위를 별도로 설정한 것이다. 다시 말해 20%를 초과하는 비율 차이가 발생할 경우 법적으로 허용되는 비례성의 범위를 벗어나는 것으로 평가된다.

2. 차별의 평가 도식으로서 비례성 심사와 합리성

또 다른 의미의 비례성은 차별적 조치가 공적 또는 사적 목적을 달성하기 위해 시행된 것일 경우 그러한 조치가 목적에 상응하는 수단인지 평가하는 용도로 사용된다. 이때 비례성은 일반적으로 목적과 수단의 타당한 관계를 평가하기 위해 비례성 심사(proportionality test)라는 구조화된 형식을 갖춘다. 첫째, 수단이 목적을 달성하기 위해 합리적인 연관성(rational connection)을 가져서 적합해야(suitable) 한다. 둘째, 목적을 달성하는 과정에서 피해를 최소화(minimal impairment)할 수 있는 다른

123) 헌법재판소는 1995년부터 선거구획정에 관해서 인구편차 문제를 다루기 시작하여 국회의원 지역선거구 인구편차의 허용한계를 1:4(헌재 1995. 12. 27. 95헌마224 등, 판례집 7-2, 760: 761), 1:3(헌재 2001. 10. 25. 2000헌마92 등, 판례집 13-2, 502: 503), 1:2(헌재 2014. 10. 30. 2012헌마192 등, 판례집 26-2상, 668: 669~670쪽)로 시간차를 두면서 단계적으로 줄여 나갔다. 그 이상적 기준은 1:1에 지향되어 있다.

124) ‘5분의 4 규칙’에 관한 자세한 내용은 조순경·한승희·정형욱·정경아·김선욱, 간접차별의 이론과 여성노동의 현실, 푸른사상, 2007, 73~79쪽 참조.



수단이 없어서 필요한(necessary) 것이어야 한다. 마지막 셋째, 목적 달성에 적합하고 필요한 수단이라고 할지라도 그 효과(effects)가 목적의 중요성과 비례해야(proportionate) 한다. 수단의 효과와 목적의 중요성 사이의 심사 요소로서 비례성 그 자체(proportionality as such)는 좁은 의미의 비례성이라고 불리기도 한다. 수단의 정당성을 평가하기 위한 전체 과정에서 목적의 중요성은 별도의 독립된 평가 대상이 될 수 있고,¹²⁵⁾ 그러한 목적의 중요성은 비례성 심사에서 수단의 효과와 비교되고 형량되는(balanced) 대상으로 포함되어 평가될 수 있다.¹²⁶⁾

그런데 차별적 조치가 정당한 것인지 평가할 때 비례성 심사의 모든 부분 원칙이 적용되는 것인지는 차별 개념을 어떻게 구성하느냐에 따라 달라진다. 특히 헌법적 권리로서 기본권에 관한 논증에서 차별 개념에 대한 구성 방식은 구체적 사례에서 문제를 해결하기 위해 비례성 심사를 진행할 때 묵시적으로 표현된다. 헌법재판소의 초기 결정에서 다루어졌던 한 가지 사례를 들어보자. 문제가 되는 법률 조항은 담보공탁에 관한 것으로 연체대출금에 관한 경매절차에서 경락허가결정에 대해 항고하려는 사람에게 담보로서 경락대금의 10분의 5에 해당하는 현금 또는 자기앞 수표 및 유가증권을 공탁하지 않은 경우 항고장이 접수된 원심법원에게 각하하도록 하고, 이 각하 결정에 대해 즉시항고도 할 수 없도록 규정하고 있다.¹²⁷⁾ 이의가

125) R. v. Oakes(Sa Majesté La Reine c. David Edwin Oakes), Case No. 17550, [1986] 1 S. C. R. 103: 138~140.

126) 헌법재판소는 초기 결정에서부터 “차별의 목적의 정당성과 필요성에 있어서나 그 수단의 적정성에 있어서 합리적인 근거가 있다고 보기 어렵다.”는 식으로 차별의 목적과 수단의 관계뿐만 아니라 목적의 정당성도 고려하고 있다(헌재 1989. 5. 24. 89헌가37 등, 판례집 1, 48: 56~58); 헌법재판소는 차별의 판단 기준으로 합리성을 의미하는 자의금지원칙 외에 엄격한 심사기준으로 비례성원칙을 적용하면서(헌재 1999. 12. 23. 98헌마363, 판례집 11-2, 770: 787), 차별의 근거로 헌법에 명시된 사유를 제시하는 경우에도 헌법에서 특별히 평등을 요구하고 있는 경우가 아니라면 반드시 비례성 심사를 해야 하는 것은 아니라고 보기도 한다(헌재 2010. 11. 25. 2006헌마328, 판례집 22-2하, 446: 453~454); 평등권이 보호범위를 가지고 있지 않다는 이유로 평등심사에 비례성원칙을 적용할 수 없다고 보는 견해로 한수웅, “헌법 제37조 제2항의 과잉금지 원칙의 의미와 적용범위”, 저스티스 (95), 2006, 5~28쪽: 27쪽 참조; 또한 헌법재판소가 비례성원칙의 적용에서 목적의 정당성을 고려하는 것과 달리 헌법상 비례성원칙의 부분원칙은 적합성원칙, 필요성원칙, 좁은 의미의 비례성원칙으로만 구성된다고 보는 견해로 이준일, “헌법상 비례성원칙”, 공법연구 37(4), 2009, 25~44쪽 참조; 유럽사법재판소에서 적용되는 비례성원칙의 내용에 관해서는 김대환, “유럽연합법원(EuGH) 판결에 나타난 비례성원칙의 내용과 심사강도”, 세계헌법연구 17(3), 2011, 1~27쪽: 10쪽 참조.

127) 구 금융기관의연체대출금에관한특별조치법(법률 제1808호, 1966. 8. 3. 제정; 법률 제2153호, 1970. 1. 1. 개정; 법률 제2570호, 1973. 3. 3. 개정) 제5조의2: “연체대출금에 관한 경매절차에 있어서 경락허가결정에 대한 항고를 하고자 하는 자는 담보로서 경락대금의 10분의 5에 해당하는 현금 또는 금융기관이 발행한 자기앞수표 및 대통령령으로 정하는 유가증권을 공탁하여야 한다(제1항). 항고의 제기에 있어서 그 항고장에 제1항의 규정에 의한 담보의 공탁이 있는 것을



제기된 이유는 “일반 경매법에 의한 통상의 경매절차에 있어서와 달리 유독 금융기관의 연체대출금에 관한 경매절차에 있어서만 경락허가결정에 대한 항고를 하고자 하는 자에게 공탁을 하도록” 하여 “합리적 이유없이 금융기관에게 우월의 지위를 부여한 것”으로 보았기 때문이다.¹²⁸⁾ 비교적 차별 개념에 따르면 문제의 핵심은 해당 법률조항이 금융기관을 특별히 우대했느냐이다. 다른 채권자의 채권과 달리 금융기관의 연체대출금 채권에 대해서만 항고인에게 담보 공탁을 항고 조건으로 부과함으로써 신속한 회수를 보장하는 것이므로 비교 대상은 경매절차에 참여하는 다른 일반 채권자이다. 그리고 불특정다수에게 채무를 부담함으로써 획득한 자금을 대출하여 발생한 채권 및 이러한 대출사무를 규칙적이고 조직적으로 하는 법인이 갖는 채권에 대해서만 신속하게 회수할 수 있도록 보장해 주는 것은 그렇지 않은 채권자 및 채권을 다르게 대우하는 것일 뿐만 아니라 불리하게 대우하는 것이다. 따라서 특별한 사정이 없는 한 금융기관을 다르게 또 유리하게 대우하는 규정은 차별적이다.

하지만 논증은 여기에서 끝나지 않는다. 특별한 사정이 있다는 것이 증명되면 차별적이라는 평가는 정당한 차별이 될 수 있다. 그리고 특별한 사정은 ‘합리적 이유’로 대체된다. 합리적 이유의 증명 여부에 따라 대우에 대한 평가의 결론이 달라지는 것이다. 그렇기 때문에 차별을 판단하는 기준을 비례성과 구별하여 합리성으로 보기도 한다.¹²⁹⁾ 그런데 합리적 이유는 비례성 심사의 한 부분원칙 즉, 수단이 목적을 달성하기 위해 합리적 연관성(rational connection)을 가져야 한다는 원칙으로서 적합성 원칙을 연상시킨다. 이러한 적합성 원칙은 마치 합리성과 비례성을 연결시켜주는 원칙처럼 보이기도 하고, 합리성 원칙이 비례성 원칙에 편입된 것처럼 보이기도 한다. ‘합리적’이라는 표현에 연연하지 않는다면 가능한 한 피해를 최소화해야 한다는 필요성 원칙에도 합리적 성격을 부여할 수 있다. 예상할 수 있는 피해를 고려하지 않는 것보다 고려하는 것이 보다 합리적이라고 볼 수 있기 때문이다. 다만, 차별을 순수하게 의무론적으로 구성한다면 결과로서 피해를 고려하지 않기 때문에 분류와 목표 간 합리적 연관성만이 판단 기준이 될 것이다.

증명하는 서류를 첨부하지 아니한 때에는 원심법원은 그 항고장을 접수한 날로부터 7일내에 결정으로 이를 각하하여야 한다(제2항). 제2항의 결정에 대하여는 즉시항고를 할 수 없다(제3항).”

128) 현재 1989. 5. 24. 89헌가37 등, 판례집 1, 48: 50~51.

129) 차별의 문제가 실제로 자유침해와 중첩되는 사례가 많다는 점을 지적하며 차별의 판단기준으로서 ‘합리성’과 자유침해의 판단기준으로서 ‘비례성’을 엄격하게 구분하여 별도로 판단이 수행되어야 한다는 견해로 이준일, “차별, 소수자, 국가인권위원회”, 헌법학연구 18(2), 2012, 177~222쪽: 211~214쪽 참조.



3. 비례성의 합리성 문제와 차별의 정당화 논증에서 잠재적 차별 가능성

비례성에 대한 내적 비판 중 하나는 비례성에 합리성이 부족하다는 것이다.¹³⁰⁾ 알렉시(R. Alexy)에 따르면 합리성의 형식적 차원이 실체적 차원과 연결되게 하는 방법은 “합리적 법적 논증 이론에 중요도 공식을 끼워 넣는 것”¹³¹⁾이다. 알렉시는 자신의 원칙이론이 형식 이론이라는 주장에 대해 반박하기 위해 프레게(G. Frege)가 “개념표기(Begriffsschrift)”에서 ‘형식 언어’와 ‘일상 언어’를 각각 ‘현미경’과 ‘눈’에 비교한 것¹³²⁾을 예로 든다. 합리성의 형식적 차원은 현미경과 같고, 합리성의 실체적 차원은 눈과 같다는 것이다. 그리고 형식적 측면의 기본 요소는 숫자(number)이고, 실체적 측면의 기본 요소는 논변(argument)이라고 주장한다. 숫자는 분류(classification)하기에 적절한데, 분류명제에 따르면 제약의 강도와 중요성의 정도, 비보호의 강도와 중요성의 정도는 숫자화 할 수 있다. 그리고 논변은 숫자화한 분류를 정당화하는 것이다.

형량으로 환원되는 비례성에 중요도 공식을 활용함으로써 그 기본 요소인 숫자로 변수를 대체하여 분류하는 것이 합리적이지 않다는 주장은 하버마스(J. Habermas)의 자의성 논변이나, 슐링크(B. Schlink)의 결정주의,¹³³⁾ 그리고 포셔(R. Poscher)의 직관주의 등을 통해 제기된다. 하버마스는 이행 질서에 어떤 가치를 편입시켜야 하는지에 관한 합리적 기준이 없기 때문에 형량은 자의적이거나 비성찰적으로 관습적 표준과 위계질서에 따라 실시될 것이라고 보고,¹³⁴⁾ 슐링크는 좁은 의미의 비례성 심사에서 궁극적으로는 심사자의 주관성만이 효력을 갖게 돼 그 주관성이 다소간의 가치로서 기본권 적용에 대한 사례별 판단의 결과로 나타나게 된다고 본다. 그리고 포셔는 형량 프로세스 자체가 충돌하는 원칙들의 상대적 비중에 관한 판단자의 직관에 의존하기 때문에 원칙들의 형량이 도덕적 직관 같은 다른 실천적 지식원보다 더 나은 점이 무엇인지 명확하지 않다고 주장한다.¹³⁵⁾

130) Aharon Barak, *Proportionality: Constitutional Rights and Their Limitations*, Cambridge University Press, 2012, p. 484.

131) Robert Alexy, “Proportionality and Rationality”, in *Proportionality: New Frontiers, New Challenges*, Vicki C. Jackson · Mark V. Tushnet(Eds.), Cambridge University Press, 2017, pp. 13~29: p. 22.

132) Gottlob Frege, *Begriffsschrift und Andere Aufsätze*[1., 1879], Ignacio Angelelli(Hrsg.), 2. Aufl., Olms, 1993, V.

133) Bernhard Schlink, “Freiheit durch Eingriffsabwehr – Rekonstruktion der klassischen Grundrechtsfunktion”, *Europäische GRUNDRECHTE-Zeitschrift* 11, 1984, S. 457~468: S. 462.

134) Jürgen Habermas, *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*[*Faktizität und Geltung*, 1992], William Rehg(Trans.), Polity Press, 1996, p. 259.



분류명제의 정당화 가능성에 대한 회의적인 관점을 뒷받침하는 사례로 2006년 독일연방헌법재판소의 전자 데이터 감시(electronic data-screening)에 관한 결정¹³⁶⁾이 제시된다. 이 사건은 잠재적 테러리스트를 발견하기 위해 남성, 18세에서 40세의 나이, 학생이거나 학생이었을 것, 이슬람교, 출생국가라는 기준¹³⁷⁾에 따라 사람들의 신원 정보를 자동화된 데이터 처리 시스템에 전송한 것이 문제가 됐다. 이때 개인정보자기 결정권(informationelle Selbstbestimmung)에 대한 제약은 상당히 중대한 반면, 신체 안전에 대한 위험은 미국의 2001년 9·11 테러 이후 일반적으로 위협적인 상황으로서 추상적 수준이라고 판단했다.¹³⁸⁾ 그런데 재판관 하스(Haas)는 별개의견에서 개인정보 자기결정권에 대한 제약은 가벼운 비중을 갖는다고 보았다.¹³⁹⁾ 형량 회의론자는 개인정보 자기결정권에 관한 이 사례에서 판단을 좌우한 것은 결정의 다수성에 있다고 본다.

이에 대해 알렉시는 다수를 향한 논변이 아니라 법과 연결되어 있는 정당성을 향한 논변이었다고 반박한다.¹⁴⁰⁾ 또한 부동의가 반드시 비합리적인 것은 아니며 합리적인 부동의가 있다는 것이다. 합리적인 부동의는 권위 있는 다수의 결정이 이루어진 다음에도 여전히 남아 있어 미래의 논증에서 의해 다수가 동의할 수 있는 변경 가능성이 남는 것이라고 한다. 또한 논증대화 규칙을 준수했는지 여부도 기준이 된다고 한다. 그런데 이에 더해 알렉시는 다수 견해와 소수 견해가 논증에 할애한 쪽수를 거론한다. 다수 의견은 11쪽을 소수 의견은 4쪽을 할애함으로써 논변의 양적 차이가 있다는 점도 강조한다.

이러한 해명에도 불구하고 합리성의 형식 차원에서 제약의 강도와 중요성의 정도, 비보호의 강도와 중요성의 정도를 숫자화 하는 과정은 분류에 적절한 숫자를 대응물로 선택하는 것에 그치지 않는다. 숫자가 중요도를 판단하기 위해 순위를 갖기 때문이다. 숫자는 그 숫자가 레이블로 지정된 클래스를 만드는 것에 그치는

135) Ralf Poscher, “The Principles Theory: How Many Theories and What is Their Merit?”, in *Institutionalized Reason: The Jurisprudence of Robert Alexy*, Matthias Klatt(Ed.), Oxford University Press, 2012, pp. 218~247: p. 241.

136) BVerfG, Beschluss des Ersten Senats vom 04. April 2006 - 1 BvR 518/02 - Rn. [1-184], http://www.bverfg.de/e/rs20060404_1bvr051802.html 참조.

137) BVerfG, Beschluss des Ersten Senats vom 04. April 2006 - 1 BvR 518/02 - Rn. [1-184], Rn. 8.

138) 미국의 2001년 9·11 테러 이후 선포된 ‘테러와의 전쟁’이 ‘데이터와의 전쟁’으로 귀결됐다는 점은 본 논문 「제1장 제1절 II. 1. 테러와의 전쟁과 데이터와의 전쟁」 참조.

139) BVerfG, Beschluss des Ersten Senats vom 04. April 2006 - 1 BvR 518/02 - Rn. [1-184], Rn. 169.

140) Robert Alexy, “Proportionality and Rationality”, in *Proportionality: New Frontiers, New Challenges*, Vicki C. Jackson · Mark V. Tushnet(Eds.), Cambridge University Press, 2017, pp. 13~29: p. 24~25 참조..



것이 아니라 클래스 간에 순위와도 관련을 맺고 있다. 그러므로 숫자를 대용물로 삼아 숫자와 연계된 클래스에 제약의 강도와 중요성의 정도 그리고 비보호의 강도와 중요성을 지정할 때, 다시 말해 각각의 특성에 순위가 있는 숫자를 레이블로 정의할 때 이미 정의하는 자, 즉 판단자의 편향이나 고정관념이 반영될 수 있다는 것이다. 비록 그 특성이 가치에 관한 것이라고 하더라도 그 가치를 대용물로 하는 특성을 가진 어떤 집단과 잠재적으로 연관성을 가질 수 있다. 그렇기 때문에 어떤 가치나 특성에 어느 정도 비중을 두고 판단했는지 그 과정을 아는 것은 분명 결정에 대해 검증하는 데에 유용한 토대가 될 수 있다. 그러한 연관성을 발견하는 것은 마치 간접차별의 대용물에 연결된 특성을 가진 집단을 발견하는 머신러닝 알고리즘을 이해하는 것만큼 어려운 일이지는 않지만, 그럼에도 불구하고 판단 과정을 공개하고 그 과정에 접근할 수 있도록 하는 것은 결정에 대한 검증의 최소 요건이 될 수 있다.



제4절 머신러닝 알고리즘의 불투명성과 차별의 은폐

알고리즘의 설계는 직관에 따른 문제 해결 방법을 구체적이고 명시적으로 표현한다는 측면에서 객관성의 확보와 밀접한 관련을 맺는다. 그러나 이러한 알고리즘을 알고리즘이 생성하게 되면서 그 객관성에도 불구하고 알고리즘은 인간이 이해하기 어렵거나 이해할 수 없는 영역으로 옮겨가고 있다. 또한 알고리즘의 객관성은 누구든 접근할 수 있는 공개성과 투명성을 통해 담보될 수 있지만 알고리즘이 사적 소유에 맡겨지게 되면 비밀의 영역으로 넘겨져 비공개성과 불투명성 속에서 알고리즘이 내놓은 결과에 대해 그 원인을 분석하거나 시정할 수 있는 기회조차 갖기 어려워지게 된다. 특히 인간이 축적해 온 차별의 역사와 경험이 담긴 사회의 데이터를 통해 훈련된 알고리즘으로 차별이 재생산되고 확대될 수 있다는 점은 문제를 더욱 심각하게 만든다. 그런데 머신러닝 알고리즘의 불투명성은 알고리즘 자체의 내부 작동방식에 기인하는 측면도 있지만 이미 사회의 법규범이 일방적으로 투명성을 옹호하거나 불투명성을 옹호하는 방식으로 정립되어 있지 않고 불투명성의 완전한 구축도, 해체도 어려운 상호 지양적 관계로 설정되어 있기 때문에 아닌지 검토가 필요하다.

I. 불투명성의 세 가지 차원

투명성에 대한 장벽으로서 불투명성이 만들어지는 방식은 크게 세 가지이다. 첫째는 기업이나 국가의 의도적인 비밀주의에 의한 불투명성이다. 이러한 비밀주의는 기술적 조치로 인해 실제로 그 내용에 접근할 수 없는 방식과 법적 조치로 그 내용을 공개하지 못하도록 담당자에게 의무를 부과하는 방식이 적절히 혼합되어 유지된다.¹⁴¹⁾ 둘째는 기술적 문맹에 의한 불투명성이다. 비록 기술적 조치나 법적 조치가 제거 돼 그 내용이 공개된다고 해도 접근할 수 있는 사람이나 기관이 기술의 작동방식과 그 의미에 대해 이해할 수 있는 능력이 부족한 경우 여전히 불투명성은 유지된다. 셋째는 머신러닝 알고리즘의 특성과 알고리즘을 적용하기 위해 필요한 계층 구조로부터 발생하는 불투명성이다.¹⁴²⁾

141) 비밀주의(secretcy)를 현실적인 것(real secrecy)과 법적인 것(legal secrecy)으로 구분하는 경우로 Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*, Harvard University Press, 2015, p. 6 참조.



1. 비밀주의와 불투명성

기업이나 정부가 자기보호나 은폐를 위해 만들어내는 알고리즘의 비밀주의는 알고리즘에 의해 정확히 무슨 일이 벌어지는지 파악하기 불가능하게 한다.¹⁴³⁾ 설령 기업이나 정부에 대한 감사(audit)를 통해 알고리즘 방법론에 의도적인 차별이 작용하지 않았다고 확신한다 해도, 알고리즘의 결함은 여전히 심각한 문제로 남는다. 스위니의 연구¹⁴⁴⁾에서 이미 문제가 제기됐듯이 알고리즘은 꼭 ‘유추되는 인종’이 없더라도 각각의 이름을 추적하여 별개의 온라인 광고 범주와 연결시킬 수 있다. 이러한 작동은 어떤 이름이나 그와 유사한 이름을 가진 사람들에 대한 과거의 평가를 기계적으로 추론한 결과일 수도 있다. 프로그램의 코드와 기초 데이터, 알고리즘 모델 등에 접근하지 못하는 한 어떤 종류의 추적이 진행되는지, 그리고 ‘현실의 삶’에서 오랫동안 문제시되어온 차별이 이제는 어떻게 사이버공간으로까지 확대되었고 어떻게 다시 현실의 삶에 영향을 미치는지 무성한 추측만이 가능할 뿐 그 원인과 과정을 구체적으로 알 수 없다.

점수와 등급에 기초해 구조화된 사회에서 머신러닝 알고리즘은 평점을 부여하고 그에 따라 순위를 정렬하는 규칙을 만들기 때문에 머신러닝 알고리즘이 생성한 규칙으로서 알고리즘은 그 자체로 법이 되어 그 평가 대상이 되는 사람들에게 특정 기준을 내면화하도록 조장하고 실패에 대해 처벌하는 구조를 만들어 낼 수 있다.¹⁴⁵⁾ 이처럼 알고리즘을 이용함으로써 지배 권력을 획득하는 세력이 등장할 수 있다는 우려는 시스템의 투명성과 객관성을 강조함으로써 불식시킬 수 있는 것처럼 보이기도 한다. 알고리즘 기반의 컴퓨터는 같은 사례들을 같게 취급함으로써 공정하게 작동할 것이라고 주장하는 것이다.

이러한 논지에 따르면 비공개 결정과정을 거치는 판사나 검사 또는 배심원보다는 알고리즘에 의할 경우 결과를 의심하는 사람들도 ‘내막을 살펴보고’ 시스템이 어떻게

142) Jenna Burrell, “How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms”, *Big Data & Society* 3(1), 2016, pp. 1~12 참조.

143) Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*, Harvard University Press, 2015, p. 39.

144) Latanya Sweeney, “Discrimination in Online Ad Delivery”, *Communications of the ACM* 56(5), 2013, pp. 44~54 및 이에 관한 본 논문 「제2장 제3절 I. 1. 스위니의 연구」 참조.

145) Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*, Harvard University Press, 2015, p. 191.



작동하는지 직접 파악할 수 있기 때문에 컴퓨터 계산이 법체계의 모델이 될 수 있다는 주장도 가능하다. 그러나 이와 같이 비교적 개방적인 접근법은 알고리즘에 관해 스파머, 해커, 사기꾼, 조작자, 경쟁자 및 일반 대중에게 알려진 바가 적을수록 더 효과적이라는 새로운 접근법에 쉽게 잠식된다. 투명성은 기술적으로든 법적으로든 견고한 비밀주의로 대체되고 정당화의 문제는 유보된다. 영업 비밀 보호는 공시가 필요하지 않은 알고리즘에서 효과적으로 재산권을 창출하고, 국가 기밀 규정은 국가 안보가 관련된 사안에서 더욱더 가공할 만한 각종 법적 기제를 제공한다.¹⁴⁶⁾ 이렇게 투명성을 통한 정당화보다 비밀주의를 통한 보호를 중시하는 태도는 이른바 ‘블랙박스 사회(black box society)’¹⁴⁷⁾가 탄생하는 토양이 되고, 그와 더불어 정보화 시대의 수많은 사회적 위험을 야기한다.¹⁴⁸⁾

2. 기술적 문맹 상태와 불투명성

기술적 문맹의 문제는 단순히 알고리즘의 소스 코드나 설계 이력을 공개하는 것만으로 해결될 수 없는 차원의 불투명성이다. 전문지식이 집약된 고차원의 알고리즘의 설계에 사용된 코드에 대한 문식 능력(literacy), 즉 기술적 코드(code)를 읽고 작성할 줄 아는 능력은 전문 기술자와 일반 대중 사이에 현격한 차이가 나타나는 부분이다. 이러한 능력은 알고리즘 기술을 이해하고 활용하는 것과 직결되어 있다. 아날로그 기술을 체득한 세대와 디지털 기술을 체득한 세대가 공존하는 현 세대에서 이러한 차이는 극명하다. 컴퓨터 장치를 위한 코드의 작성은 인간의 자연어에서와는 다르게 정확성, 형식성, 완결성을 요구하고, 이를 위해서는 다양한 수준의 추상화가 필요하다.¹⁴⁹⁾

자신의 생활환경을 조성하는 기술을 체득해야만 기본적인 생활이 가능해지는 상황에서 복잡한 기술의 코드를 이해하고 활용할 줄 모른다는 것은 커뮤니케이션을 위해 필수적인 기호를 읽고 쓸 줄 모르는 문맹 상태에 놓이는 것이기도 하다. 이는 교육의 문제¹⁵⁰⁾와 좀 더 밀접하게 관련되어 있기 때문에 단순히 알고리즘에 관한

146) Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*, Harvard University Press, 2015, p. 193.

147) ‘블랙박스 사회’에 관한 설명은 본 논문 「제4장 제4절 II. 1. 블랙박스 사회의 딜레마」 참조.

148) Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*, Harvard University Press, 2015, p. 193.

149) Jeannette M. Wing, “Computational Thinking”, *Communications of the ACM* 49(3), 2006, pp. 33~35: p. 34.

150) Irene Lee · Fred Martin · Jill Denner · Bob Coulter · Walter Allan · Jeri Erickson · Joyce Malyn-Smith



정보를 공개(disclose)하거나 고지(notice)하는 것만으로 불투명성을 제거하고 투명성을 제고하는 충분한 조건이 될 수 없다는 점을 방증한다.

3. 복잡한 머신러닝 알고리즘의 작동방식과 불투명성

알고리즘의 불투명성을 야기하는 것이 관련 기술에 대한 일반인의 기본적인 이해도가 낮기 때문일 수도 있지만 그 내용이 복잡한 경우에는 기술에 대한 이해도가 높은 전문가도 이해하지 못할 수 있다. 이때에는 일반인과 전문가를 구별하는 것의 의미가 상실된다. 만약 정부나 기업이 사용하는 시스템의 알고리즘에 대해 정보 공개 명령이 이루어진 경우 누구도 이해하기 어려운 내용의 복잡한 자료를 공개한다면 이에 대한 불투명성은 공개 여부나 자료의 이해능력과는 무관해진다. 알고리즘은 변수가 많아질수록 차원이 증가하여 해결할 수 없거나 어려워질 수 있어 알고리즘의 복잡도(complexity)는 종종 ‘차원의 저주(curse of dimensions)’¹⁵¹⁾라고 표현되기도 한다.

머신러닝 알고리즘이 데이터를 처리할 때 작동하는 내부적 방식이 인간의 사고 수준에서 근본적으로 이해할 수 없는 것이라면 이때 발생하는 불투명성은 근원적인 의미를 갖게 된다. 버렐(J. Burrell)은 머신러닝 알고리즘의 특징인 고차원의 수학적 최적화(mathematical optimization in high-dimensionality)와 인간 수준의 추론 및 의미론적 해석 양식에 대한 요구 사이에서 발생하는 부정합성이 그와 같은 불투명성의 원인이 된다고 주장한다.¹⁵²⁾ 이처럼 알고리즘의 실제 데이터 처리 방식과 인간적 차원의 사고방식 간 부정합성이 불투명성의 원인이라면 단순히 코드의 분량이 얼마나 되는지, 알고리즘 설계를 위해 투입된 프로그래머의 인원이 얼마나 되는지, 모듈 또는 서브루틴 간 연결 개수가 얼마나 되는지 살펴보는 것만으로 그 원인이 제거되기는 어렵게 된다. 더구나 머신러닝 알고리즘 개발 과정에서 예측 불가능성에서 유래하는 어느 정도의 불투명성은 성공적인 알고리즘 전략에 불가피한 부수 효과로 여겨지기도 한다.

· Linda Werner, “Computational Thinking for Youth in Practice”, ACM Inroads 2(1), 2011, pp. 32~37 참조.

151) Richard E. Bellman, Adaptive Control Processes, Princeton University Press, 1961 및 Trevor Hastie · Robert Tibshirani · J. H. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd ed., Springer, 2009, p. 14 참조.

152) Jenna Burrell, “How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms”, Big Data & Society 3(1), 2016, pp. 1~12: pp. 4~5.



II. 투명성과 불투명성 사이의 헌법적 긴장 관계

1. 블랙박스 사회의 딜레마

블랙박스(black box)는 비행기, 자동차, 기차 등에 부착되어 기계의 작동에 관한 모든 데이터를 기록하는 장치로서 종종 그 내막을 알 수 없는 불가사의한 방식으로 작동하는 시스템을 가리키는 데 사용된다. 파스칼레(F. Pasquale)는 이러한 블랙박스를 법과 기술이 교차로 작동하여 만들어 내는 지배의 알고리즘이 사회에서 점점 더 불투명해지는 상황을 포착하기 위한 개념으로 제시한다.¹⁵³⁾ 온라인에서 하는 모든 활동은 기록되고 철두철미한 감시 카메라의 시선을 벗어난 사각지대를 찾기 어려워진 데다 자발적으로 위치 추적기(스마트폰)를 몸에 지니고 다니는데 그치지 않고 여기 저기 센서(sensor)가 내장되어 있는 정보수집용 감식장치(스마트스피커)를 집 안에 들여 놓아 내밀한 일상의 적나라한 정보까지 하나도 남김없이 내보냄으로써¹⁵⁴⁾ 그 정보를 처리하는 기업과 국가로부터 더 면밀히 추적당할 수 있음에도 불구하고 정작 그렇게 얻어간 또는 내보낸 정보가 얼마나 광범위하게 유통되고 누구의 이익으로 어떻게 활용되는지 그 내막을 알지 못하기 때문이다. 이러한 메커니즘에 따라 작동하는 사회는 분명 블랙박스 사회이다. 하지만 모순에 봉착하는 것이 정보의 수집과 이용을 규제할 수 있는 법과 제도가 없기 때문인 것은 아니다. 기술의 발전 속도와 어느 정도 시차가 있지만 이러한 상황에 대응하기 위한 법과 제도들은 속속 갖춰지게 마련이다.

2. 정보 관련 입법에 함의된 투명성과 불투명성

컴퓨터 보급이 일반화되고 네트워크 간의 네트워크 즉, 인터넷이 구축되는 시기를 분기점으로 정보와 관련된 법과 제도는 지속적으로 도입되었고 정비되고 있다. 대한민국에서 대표적인 법률은 그 제정 순서에 따라 정보통신망 이용촉진 및 정보

153) Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*, Harvard University Press, 2015, p. 3.

154) 예를 들어 2018년 5월, 미국 포틀랜드(Portland)의 한 가정에서 주고받은 사적 대화는 음성으로 조종하는 스마트스피커인 아마존의 알렉사(Alexa)에 녹음되어 임의로 시애틀(Seattle)에 거주하는 연락처 목록에 있는 사람에게 전송되었다. 이에 관해서 Gary Horcher, "Woman Says Her Amazon Device Recorded Private Conversation, Sent It out to Random Contact", KIRO7, 25 May 2018, <https://www.kiro7.com/news/local/woman-says-her-amazon-device-recorded-private-conversation-sent-it-out-to-random-contact/755507974> 참조, 접속일: 2018년 5월 25일.



보호 등에 관한 법률,¹⁵⁵⁾ 신용정보의 이용 및 보호에 관한 법률,¹⁵⁶⁾ 국가정보화 기본법,¹⁵⁷⁾ 공공기관의 정보공개에 관한 법률,¹⁵⁸⁾ 전자정부법,¹⁵⁹⁾ 위치정보의 보호 및 이용 등에 관한 법률,¹⁶⁰⁾ 국가공간정보 기본법,¹⁶¹⁾ 디엔에이신원확인정보의 이용 및 보호에 관한 법률,¹⁶²⁾ 개인정보 보호법¹⁶³⁾으로 나열해 볼 수 있다. 각 법률의

-
- 155) 정보통신망 이용촉진 및 정보보호 등에 관한 법률[법률 제15628호, 2018. 6. 12. 개정]은 1986. 5. 12. ‘전산망보급확장과이용촉진에관한법률’로 제정된 것이 1999. 2. 8. ‘정보통신망이용촉진등에관한법률’로, 2001. 12. 31. ‘정보통신망이용촉진및정보보호등에관한법률’로 변경되어 현재에 이르게 된 것이다. ‘전산망보급확장과이용촉진에관한법률’의 입법 목적은 “전산망의 개발보급과 이용등을 촉진하여 정보화사회의 기반을 조성함으로써 국민생활의 향상과 공공복리의 증진에 이바지”(동법 제1조)하는 것이다.
- 156) 신용정보의 이용 및 보호에 관한 법률[법률 제15748호, 2018. 8. 14. 개정]은 1995. 1. 5. “신용정보업을 건전하게 육성하고 신용정보의 효율적 이용과 체계적 관리를 도모하며 신용정보의 오용·남용으로부터 사생활의 비밀 등을 적절히 보호함으로써 건전한 신용질서의 확립에 이바지”(동법 제1조)하려는 목적으로 제정되었다.
- 157) 국가정보화 기본법[법률 제15369호, 2018. 2. 21. 개정]은 1995. 8. 4. ‘정보화촉진기본법’으로 제정된 것이 2009. 5. 22. ‘국가정보화 기본법’으로 변경되어 현재에 이르고 있다. 구 정보화촉진 기본법의 입법 목적은 “정보화를 촉진하고 정보통신산업의 기반을 조성하며 정보통신기반의 고도화를 실현함으로써 국민생활의 질을 향상하고 국민경제의 발전에 이바지”(동법 제1조)하는 것이었는데, 법명을 변경하면서 입법 목적도 “국가정보화의 기본 방향과 관련 정책의 수립·추진에 필요한 사항을 규정함으로써 지속가능한 지식정보사회의 실현에 이바지하고 국민의 삶의 질을 높이는 것”(현행법 제1조)으로 ‘지식정보사회’에 초점을 맞추어 수정하였다.
- 158) 1996. 12. 31. 제정된 공공기관의 정보공개에 관한 법률[법률 제14839호, 2017. 7. 26. 개정]의 입법 목적은 “공공기관이 보유·관리하는 정보의 공개의무 및 국민의 정보공개청구에 관하여 필요한 사항을 정함으로써 국민의 알권리를 보장하고 국정에 대한 국민의 참여와 국정운영의 투명성을 확보”(동법 제1조)하는 것이다.
- 159) 전자정부법[법률 제14914호, 2017. 10. 24. 개정]은 2001. 3. 28. ‘전자정부구현을위한행정업무등의 전자화촉진에관한법률’로 제정된 것이 2007. 1. 3. ‘전자정부법’을 법명을 바꾸어 현재에 이르고 있다. ‘전자정부를 구현하고, 행정의 생산성, 투명성 및 민주성’을 높이려는 목적(동법 제1조)은 세부적인 자구의 변경에도 불구하고 변함이 없다.
- 160) 2005. 1. 27. 제정된 위치정보의 보호 및 이용 등에 관한 법률[법률 제15608호, 2018. 4. 17. 개정]은 “위치정보의 유출·오용 및 남용으로부터 사생활의 비밀 등을 보호하고 위치정보의 안전한 이용환경을 조성하여 위치정보의 이용을 활성화함으로써 국민생활의 향상과 공공복리의 증진에 이바지”(동법 제1조)하려는 목적을 갖는다.
- 161) 국가공간정보 기본법[법률 제12736호, 2014. 6. 3. 개정]은 2009. 2. 6. ‘국가공간정보에 관한 법률’로 제정된 것이 2014. 6. 3. ‘국가공간정보 기본법’으로 명칭이 변경되었지만, “국가공간정보 체계의 효율적인 구축과 종합적 활용 및 관리에 관한 사항을 규정함으로써 국토 및 자원을 합리적으로 이용하여 국민경제의 발전에 이바지”(동법 제1조)하겠다는 목적은 그대로이다.
- 162) 2010. 1. 25. 제정된 디엔에이신원확인정보의 이용 및 보호에 관한 법률[법률 제13722호, 2016. 1. 6. 개정]은 “디엔에이신원확인정보의 수집·이용 및 보호에 필요한 사항을 정함으로써 범죄수사 및 범죄예방에 이바지하고 국민의 권익을 보호”(동법 제1조)하려는 목적을 갖는다.
- 163) 데이터 주체에 관한 정보를 다룬 법률의 결정판이라고 할 수 있는 개인정보 보호법[법률 제14839호, 2017. 7. 26. 개정]은 2011. 3. 29. “개인정보의 수집·유출·오용·남용으로부터 사생활의 비밀 등을 보호함으로써 국민의 권리와 이익을 증진하고, 나아가 개인의 존엄과 가치를 구현하기 위하여 개인정보 처리에 관한 사항을 규정”(제정법 제1조)하는 목적으로



제목과 제정된 순서만 놓고 보더라도 1980년대 후반부터 2010년대 초반까지 전개된 입법 정책에 관한 일련의 흐름을 읽을 수 있다. 사회 저변에 전산망을 보급·확장하고 이용을 촉진하는 것으로 시작해 기업과 국가의 정보화에 무방비로 노출되는 정보의 주체를 보호하는 방안을 강구하는 것이다. ‘국가공간정보 기본법’에서도 나타나듯이 국가는 국민에만 한정하지 않고 국가의 토대가 되는 모든 것에 대해 알고 싶어 한다.

정보에 관한 개별 법률들에서 고려되는 이익이나 가치로 한편에는 지식정보사회를 상정하여 국민의 생활수준을 향상시키거나 공공복리를 증진시키고, 국민경제를 활성화하려는 목적이 놓여 있다. 헌법에 따라 “국가는 과학기술의 혁신과 정보 및 인력의 개발을 통하여 국민경제의 발전에 노력하여야 한다.”(제127조 제1항)는 측면을 감안하면 국가에게 정보는 국민경제의 발전과 긴밀하게 연결되어 있는 개발 대상으로 자리매김하게 된다. 그리고 다른 한편에는 국민의 알 권리, 사생활의 비밀 등 자유와 권리를 보호하는 목적도 세워져 있다. 이를 위해 정보를 수집하고 이용하고 공개하는 절차와 그 내용에 관한 사항, 그리고 그에 대한 권한과 책임 있는 기관 또는 조직을 규정한다. 특히 국가의 행정 업무는 직접적으로 국민에게 영향을 미치므로 “국정운영의 투명성”이나 “행정의 생산성, 투명성 및 민주성”을 고려하도록 하고 있다.¹⁶⁴⁾ 기본권의 관점에서 보자면 이러한 입법은 알 권리를 구체화한 것이기도 하다.

또한 국민 개인의 정보를 국가가 처리할 수 있다는 것은 국민의 삶을 일일이 들여다 볼 수 있다는 것이므로 사생활의 비밀과 자유 또는 프라이버시(privacy)에 관한 권리는 그에 대항할 수 있는 유력한 헌법적 근거이자 무기가 될 수 있다. 사생활의 비밀과 관련하여 헌법은 “모든 국민은 사생활의 비밀과 자유를 침해받지 아니한다.”(제17조), “모든 국민은 통신의 비밀을 침해받지 아니한다.”(제18조), 나아가 “모든 국민은 주거의 자유를 침해받지 아니한다. 주거에 대한 압수나 수색을 할 때에는 검사의 신청에 의하여 법관이 발부한 영장을 제시하여야 한다.”(제16조)고 규정한다. 이를 구체화한 대표적인 입법으로 부정경쟁방지 및 영업비밀보호에 관한

제정되었다. 이러한 입법 목적은 카드사 등에서 빈번하게 발생하던 개인정보 유출 사고가 계기가 된 2014. 3. 24. 개정[법률 제12504호]에서 “수집·유출·오용·남용으로부터 사생활의 비밀 등을 보호함으로써 국민의 권리와 이익을 증진”이 “처리 및 보호에 관한 사항을 정함으로써 개인의 자유와 권리를 보호”로, “구현하기 위하여 개인정보 처리에 관한 사항을 규정함”이 “구현함”으로 변경되었다(동법 개정이유 및 제1조).

164) 앞의 ‘공공기관의 정보공개에 관한 법률’ 및 ‘전자정부법’에서 입법 목적(각 법률 제1조) 참조.



법률,¹⁶⁵⁾ 통신비밀보호법,¹⁶⁶⁾ 금융실명 거래 및 비밀보장에 관한 법률¹⁶⁷⁾을 꼽을 수 있다. 이러한 법률은 영업활동에 관련된 기술상 또는 경영상의 정보나 금융거래 정보 같은 사업이나 경제활동에 관한 정보를 국가가 임의로 직접 들여다보거나 그런 정보의 전달 자체에 개입하는 것을 제한하는 방어벽의 기능을 한다.

영업활동이나 경제활동, 창작활동 등에 관한 정보는 정보 생산자 또는 소유자의 권리로 보호 받을 수도 있다. 대표적인 입법으로 발명가에게 독점적 권리를 부여하여 창작으로서 발명을 보호하려는 특허법¹⁶⁸⁾, 저작자와 그 저작물을 저작권권과 저작재산권으로 보호하는 저작권법¹⁶⁹⁾을 들 수 있다. 이러한 법률 역시 헌법에 근거를 두고 있다. 헌법은 “저작자·발명가·과학기술자와 예술가의 권리는 법률로써 보호한다.”(제22조 제2항)고 하여 정보를 저작자·발명가·과학기술자·예술가의 권리 아래 귀속시킬 수 있는 근거를 마련해 두고 있고, 나아가 “모든 국민의 재산권은 보장된다. 그 내용과 한계는 법률로 정한다.”고 하여 일반적으로 보장되는 재산권의 대상으로서 정보를 재산 개념에 편입시켜 재산권에 귀속시킬 수 있는 해석 가능성을 제공하고 있다.

165) 부정경쟁방지 및 영업비밀보호에 관한 법률[법률 제15580호, 2018. 4. 17. 개정]은 1961. 12. 30. “부정한 수단에 의한 상업상의 경쟁을 방지하여 건전한 상거래의 질서를 유지”하려는 목적을 가지고 제정된 ‘부정경쟁방지법’에 1991. 12. 31. 개정을 통해 영업비밀 조항이 추가된 후, 1998. 12. 31.에 현행 법령으로 변경된 것이다. 1991년 부정경쟁방지법에 영업비밀 조항이 신설될 당시의 정책적 문제인식은 “과학기술투자의 확대와 기술혁신에 따라 산출되는 기술상·경영상 유용한 정보(營業秘密)의 중요성이 높아지고” 있다는 것이다(동법 개정이유). 현행법의 목적은 “국내에 널리 알려진 타인의 상표·상호(商號) 등을 부정하게 사용하는 등의 부정경쟁행위와 타인의 영업비밀을 침해하는 행위를 방지하여 건전한 거래질서를 유지”(동법 제1조)하는 것이다.

166) 1993. 12. 27. 제정된 통신비밀보호법[법률 제15493호, 2018. 3. 20. 개정]은 “통신 및 대화의 비밀과 자유에 대한 제한은 그 대상을 한정하고 엄격한 법적 절차를 거치도록 함으로써 통신비밀을 보호하고 통신의 자유를 신장”(동법 제1조)하는 것을 목적으로 한다.

167) 1997. 12. 31. 제정되어 현행에 이르는 금융실명거래 및 비밀보장에 관한 법률[법률 제14242호, 2016. 5. 29. 개정]은 “실지명의(實地名義)에 의한 금융거래를 실시하고 그 비밀을 보장하여 금융거래의 정상화를 꾀함으로써 경제정의를 실현하고 국민경제의 건전한 발전을 도모”(동법 제1조)하려는 목적을 갖는다.

168) 1952. 4. 23. 별도의 입법 목적을 규정하지 않고 ‘1946년 특허법’을 개정하면서 ‘발명’을 “신규하고 유용한 기술, 방법, 기계, 생산품, 물질의 합성 급 식물의 변종, 기타 신규유용한 개량을 포함”(동법 제2조 제7호)하는 것으로 정의했다. 1961. 12. 31. 전부개정을 통해 “발명을 장려, 보호·육성 함으로써 기술의 진보발전을 도모하고 국가산업의 발달에 기여하게”(동법 제1조)한다는 목적이 제시되었다. 현행 특허법[법률 제15582호, 2018. 4. 17. 개정]은 발명을 “자연법칙을 이용한 기술적 사상의 창작으로서 고도(高度)한 것”(동법 제2조 제1호)으로 정의한다.

169) 1957. 1. 28. “학문적 또는 예술적저작물의 저작자를 보호하여 민족문화의 향상발전을 도모”(동법 제1조)하려는 목적으로 제정된 저작권법[법률 제14634호, 2017. 3. 21. 개정]은 내용에 일부 변경이 가해져 “저작자의 권리와 이에 인접하는 권리를 보호하고 저작물의 공정한 이용을 도모함으로써 문화 및 관련 산업의 향상발전에 이바지”할 목적으로 시행되고 있다.



III. 민주주의 국가의 디폴트(default)로서 투명성

투명성과 공개성을 옹호하는 가장 단순하면서 가장 직관적인 이론적 설명은 자유민주주의 국가는 개념적으로 가능한 한 투명해야 한다는 것이다.¹⁷⁰⁾ 즉 민주주의 개념으로부터 투명성과 공개성이 나온다. 투명성은 잘 아는 상태에서 공적 토론을 할 수 있게 하고 국가에 대한 신뢰와 정당성을 생성하고 또한 선거일에 개인의 결정에 영향을 미친다.¹⁷¹⁾ 그러므로 국가작용의 기본설정(default)은 투명성이다. 샤우어(F. Schauer)에 따르면 투명성의 정도는 정보 보유자, 투명해져야 하는 정보, 그리고 정보에 대한 접근권자의 세 가지 변수 간 함수로 결정된다.¹⁷²⁾ 투명성 개념은 매우 넓고 다양한데, 그 중에 중심이 되는 이론은 네 가지로 정리할 수 있다.¹⁷³⁾ 첫 번째는 공정성과 효율성에 관한 논변, 두 번째는 인기 있는 혁신과 클라우드소싱에서 이익을 취하려는 정책에 관한 논변, 세 번째는 정보 프라이버시 보호에 관한 논변, 네 번째는 개인의 자율성과 관련된 두 가지 형태의 권리에 관한 논변이다.

1. 공정하고 효율적인 정책을 위한 유인책으로서 투명성¹⁷⁴⁾

국가행위는 결함을 수반하고, 편향되고, 비효과적이며 비효율적일 수 있다. 관련 공무원들은 그들 자신의 극심한 편견에 입각해서 또는 사적 관심에 과도하게 영향을 받아서 부적절하게 권리와 이익을 형량할 수 있다. 정부 관련 부처의 기술부족이나 무능, 부패, 부주의, 단순 오류, 수용할 수 없는 관점의 영향을 받을 수도 있고, 권한을 다른 목적을 충족하는 데로 확장시키려고 할 수 있다. 이러한 가능성을 고려할 때 투명성은 국가 행위자에게 그들의 행위와 그 결과에 대해 책임감을 부여할 수 있다. 그래서 종종 투명성은 책임성과 같은 의미로 사용되기도 한다. 그러나 엄밀하게 봤을 때 투명성과 책임성은 동일한 개념이 아니다.¹⁷⁵⁾ 책임성은 국가의

170) Adam M. Samaha, "Government Secrecy, Constitutional Law, and Platforms for Judicial Intervention", *UCLA Law Review* 53(4), 2006, pp. 909~976: p. 970.

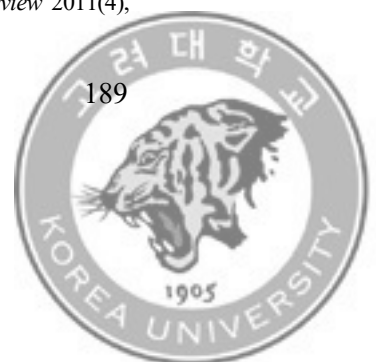
171) Mark Fenster, "The Opacity of Transparency", *IOWA Law Review* 91, 2006, pp. 885~949: p. 898.

172) Frederick Schauer, "Transparency in Three Dimensions", *University of Illinois Law Review* 2011(4), 2011, pp. 1339~1357 참조.

173) Tal Z. Zarsky, "Transparent Predictions", *University of Illinois Law Review* 2013(4), 2013, pp. 1503~1570: pp. 1533~1553.

174) Tal Z. Zarsky, "Transparent Predictions", *University of Illinois Law Review* 2013(4), 2013, pp. 1503~1570: pp. 1533~1538.

175) Frederick Schauer, "Transparency in Three Dimensions", *University of Illinois Law Review* 2011(4),



개별 공무원들에게 그들의 행위, 가능한 결함, 부정행위와 관련된 윤리적 의무를 지시하는 것인 반면, 투명성은 그러한 책임성을 가능하게 하는 필수적 도구이다.

그런데 정보공유의 범위를 넓힌다고 해서 공정성과 효율성이 증진할 것이라는 가정이 당연한 것은 아니다. 브랜다이스(L. Brandeis)는 공정함을 증진시키는 방법으로 투명성을 이용하는 것을 옹호한 것으로 유명한데, 레식(L. Lessig)은 그 기초를 제공하는 두 가지 논거로 부끄러움과 시장 또는 민주적 힘의 효과를 제시한다.¹⁷⁶⁾ 여기에는 두 가지 가정이 있는데, 하나는 공공 일반이 관심을 갖는다는 것이고 다른 하나는 그러한 관심에 대해 공무원이 반응한다는 것이다. 하지만 자동화된 알고리즘의 결정이라는 맥락에서 두 가정이 그대로 들어맞지 않을 수 있다. 자동화된 알고리즘의 결정은 공공의 관심보다는 소수의 전문가의 제한적 관심을 받는다는 점에서 부끄러움은 따르지 않거나 제한적 효과만을 가져온다. 다만, 투명성은 공무원의 행위가 잘 정립된 규범 또는 현행법과 충돌한다든지 인종 또는 종교 같은 민감한 요인에 의지할 때 그 사실을 드러내 준다. 공공의 맥락에서 볼 때 투명성은 정치적 결과로 선출된 공무원이 관료들을 압박할 수 있게 해 준다. 정보를 가능한 한 넓게 분배하는 것은 정치적 동역학을 불러일으킨다. 그러나 이 역시 수치심과 관련된 맥락에서 했던 가정을 묵시적으로 반복하기 때문에 공공 일반의 관심과 정치인의 주목을 이유로 투명성을 정당화하기는 어렵다. 또한 국가행위의 투명성은 자칫 정치인을 보수적으로 만들거나 인기영합주의로 이끌 수 있고, 비효율적이고 불공정할 수 있는 결정을 내리게 할 수 있어 단기적이거나 근시안적인 대응으로 이어질 수도 있다는 반론에 부딪히기도 한다.¹⁷⁷⁾

2. 투명성의 확장 범위와 클라우드소싱

투명성은 예측 모델의 정확성과 공정성을 증진시킬 수 있는데, 최종 결과의 개선에 관해서 정부의 외부에서 지식을 편입시킴으로써 의미 있는 피드백이 이루어질 수 있다.¹⁷⁸⁾ 비슷한 맥락에서 정부의 자동화된 결정 시스템의 투명성을 확장해야 한다고 주장하기도 한다.¹⁷⁹⁾ 그러나 외부 지식에 의존하는 것은 정부를 위해 소프트웨어를

2011, pp. 1339~1357: p. 1346.

176) Lawrence Lessig, "Against Transparency", *New Republic*, 9 October 2009, <https://newrepublic.com/article/70097/against-transparency>, 접속일: 2017년 12월 6일.

177) Frederick Schauer, "Transparency in Three Dimensions", *University of Illinois Law Review* 2011(4), 2011, pp. 1339~1357: p. 1353.

178) Mark Fenster, "The Opacity of Transparency", *IOWA Law Review* 91, 2006, pp. 885~949: p. 885.

179) Danielle Keats Citron, "Technological Due Process", *Washington University Law Review* 85,



개발한 계약자가 영업 비밀을 누설하는 경우처럼 상당한 취약성을 창출한다. 이를 고려해 정보를 선택된 기관이나 전문가 집단에게만 제공하고 정보가 더 공개되지 않도록 할 수도 있다. 부끄러움과 정치적 힘에 기댄 첫 번째 논변이 기술적 문제와 결정을 위한 투명성을 정당화하는 것에 어려움이 있었다면 제한된 범위의 외부 지식에 의존하는 크라우드소싱(crowd sourcing)에 근거한 투명성의 정당화 논변에서 기술적 문제와 결정은 그 핵심이 된다.¹⁸⁰⁾

3. 투명성 통제와 프라이버시

개인정보를 누가 통제(control)할 것이냐의 문제로 투명성에 접근하면 투명성은 개인의 자율성 신장과 자기 정보를 통제할 능력으로 이해된다.¹⁸¹⁾ 자기정보에 관한 통제능력을 보장하는 조건은 선택(choice)과 고지(notice)이다. 선택은 동의(content)의 형식을 빌린다. 그런데 동의의 형식이 일차적으로 정보 수집과 이용을 통제하지만 동의의 형식을 우회하는 추론된 정보 수집을 막을 길이 없고, 동의의 형식으로 이용에 대한 통제를 가하는 것 역시 제한적이다. 고지의 조건은 개방성 또는 투명성을 높일 것처럼 보이지만 실제로는 정보의 수집 단계에서 주로 적용되고 고지사항에 대해 일반적인 방식으로 공공의 접근이 허용된다거나 고지를 통해 분석이 어떻게 수행되고, 그 결과가 어떻게 적용되는지에 관한 내용까지 제공하는 것을 기대하기 어렵다. 특히 자동화된 알고리즘의 결정에 대해서 데이터의 수집 단계에서조차도 데이터 베이스가 어떻게 형성됐는지 또는 유사한 분야가 어떻게 나중에 사용될 하나의 데이터 세트로 집적됐는지 이해할 권리가 포함된 것으로 인식되지도 않는다.

프라이버시와 자율성에 근거한 논변은 대량 데이터분석이 실행될 때 전체적인 투명성에 관한 권리에 대한 주장으로 변형될 수 있다. 그러나 투명성에 관한 권리에 근거하여 투명성을 확장시키는 것이 데이터를 수집하고 처리하는 모든 과정에서 옹호될 수 있는지 즉, 수집 이후의 단계인 분석과 이용 단계에 모두 투명성이 적용될 수 있는 것인지에 대해서는 투명성을 제한하는 논거가 그 한계를 설정해 준다.¹⁸²⁾

2007, pp. 1249~1313: pp. 1308~1313.

180) Tal Z. Zarsky, "Transparent Predictions", *University of Illinois Law Review* 2013(4), 2013, pp. 1503~1570: p. 1538.

181) Tal Z. Zarsky, "Transparent Predictions", *University of Illinois Law Review* 2013(4), 2013, pp. 1503~1570: p. 1541.

182) 머신러닝 알고리즘이 데이터를 수집하고 분석하고 이용하는 단계에서 개인정보 보호의 문제와 차별금지 문제가 중첩될 수 있다는 점은 본 논문 「제5장 제3절 머신러닝 알고리즘의 차별로부터



4. 투명성과 개인의 자율성

자율성에 근거한 논변은 매우 다른 측면에서 투명성에 대한 정당화를 창출한다.¹⁸³⁾ 개인들이 예측 프로세스에서 산출된 결정에 의해 부정적인 영향을 받는다면 그들은 그 이유를 이해할 권리를 갖는다. 결정 기준에 대한 설명을 받아야 하고, 이러한 행위의 이면에 숨겨진 논리에 대한 설명을 받아야 한다. 물리적 환경의 작동을 지배하는 기본적인 물리학 법칙, 화학 법칙, 생물학 법칙에 대해 접근하지 못하게 제한하는 것을 받아들이지 않는 것처럼 말이다.¹⁸⁴⁾

개인들은 자신에게 영향을 미치는 사건의 이유를 학습할 권리가 있다. 이런 개념은 유럽연합과 그 회원국들에게 깊이 각인되어 있다.¹⁸⁵⁾ 예측 모델에서 해악은 부가적 비용이나 시간의 소요 또는 재정적 부담이지 헌법의 보호를 받는 생명, 자유, 재산 같은 것이라고 볼 경우 예측 모델의 효과는 헌법의 적법절차원칙에 따라 보호될 수 있는 범위 밖에서 나타날 수도 있다. 예를 들어 공항검색과 금융조사에 사용되는 예측은 헌법의 적법절차상 요건을 갖추지 않고도 가능한 기법이 될 수 있다. 그러나 표적으로 선택된 것에 대해 그 이유를 알 권리가 언급되기도 한다.¹⁸⁶⁾ 이러한 권리는 이용 단계에 좀 더 관련이 있고, 그 사유가 부정확하다면 반대할 수 있을 것이다. 그러나 선택된 결과에 도달한 내부 작동을 이해하지 못한다면 결과는 여전히 자의적이고 부당하게 보일 것이다. 심지어 존엄성 관련 이익은 단지 상관성이 아니라 인과성이 발견될 것을 요구한다고 주장하기도 한다. 또한 관련 개인에게 이미 정보가 제공되었다면 전체인구에 대한 투명성이 정부의 이익에 해를 입힌다고 반론하기 어렵다.

자동화된 알고리즘의 결정은 그 영향을 받는 개인의 자율성을 방해한다. 그러나 단순히 이러한 프로세스가 고도의 자동화를 특징으로 한다는 이유만으로 투명성이 증대되어야 한다는 주장은 실제로 유럽연합의 입법으로 실현되었지만

보호와 개인정보의 보호」 참조.

183) Tal Z. Zarsky, "Transparent Predictions", *University of Illinois Law Review* 2013(4), 2013, pp. 1503~1570: p. 1544.

184) Julie E. Cohen, *Configuring the Networked Self: Law, Code, and the Play of Everyday Practice*, Yale University Press, 2012, p. 235.

185) 이에 관해서 본 논문 「제5장 제4절 IV. 머신러닝 알고리즘에 대한 설명의무와 설명청구권」 참조.

186) Sherry F. Colb, "Innocence, Privacy, and Targeting in Fourth Amendment Jurisprudence", *Columbia Law Review* 96, 1996, pp. 1456~1525: p. 1464, pp. 1489~1493.



이러한 논변에 대해 설득력이 떨어지며 심지어 컴퓨터 검색의 열등함에 관한 신러다이트(neo-Luddite) 신념에서 도출된 것이라고 보기도 한다.¹⁸⁷⁾ 물론 러다이트(Luddite)¹⁸⁸⁾ 운동이 항상 반동적 보수주의에 머무르는 것이 아니고 새로운 대안을 제시한다는 측면이 없는 것은 아니다.¹⁸⁹⁾

그렇지만 자동화를 존엄성의 훼손과 연결 짓는 것은 21세기에는 시대착오적인 것으로 보일 수 있다. 컴퓨터화한 프로세스가 공정하고 효율적인 결과를 제시한다면 존엄성이 훼손됐다고 보기 어렵고, 자동이건 수동이건 개인의 권리들을 위태롭게 하고 자의적으로 보이는 모든 단계에 보호 장치가 필요한 것이지 자동화된 영역에 대해서만 특별히 높은 수준의 투명성이 필요한 것은 아니기 때문이다. 또한 자동화된 알고리즘의 결정이 오류를 생성하므로 더 많이 공개해야 한다는 것¹⁹⁰⁾도 인간의 결정 역시 은폐된 내부의 오류와 심한 편견에 근거한다는 점¹⁹¹⁾을 고려하면 특별한 의미를 가질 수 없다.

또 다른 주장으로 가능한 것은 자동화된 결정은 덜 의심받는 경향이 있다는 것이다. 컴퓨터화한 자동화는 흠 없는 완벽한 결정 능력을 갖고 있는 분위기를 연출한다는 점에 문제의 심각성이 있긴 하지만,¹⁹²⁾ 자동화의 본질에 대해 대중과 관련 결정자를 교육하는 것으로 해결될 수 있는 문제라면 투명성이 과연 충분한 답이 될 것인지는 여전히 의문이다.

187) Tal Z. Zarsky, “Transparent Predictions”, *University of Illinois Law Review* 2013(4), 2013, pp. 1503~1570: p. 1550.

188) 러다이트(Luddite)는 보통 반기술철학(antitechnology philosophy)을 부를 때 사용되는 용어인데, 역사적으로는 1811년부터 1817년 사이 영국의 직물 공장의 노동자로 구성된 단원을 가리키며, 이들은 산업혁명을 초래한 기계를 파괴함으로써 실업의 위기에서 벗어나고자 했다. 러다이트는 이들 단원의 지도자가 ‘제너럴 네드 러드(General Ned Ludd)’라는 데에 기원을 둔 것으로 전해진다. Steven E. Jones, *Against Technology: From the Luddites to Neo-Luddism*, Taylor and Francis, 2013, p. 3 참조.

189) Julie E. Cohen, *Configuring the Networked Self: Law, Code, and the Play of Everyday Practice*, Yale University Press, 2012, p. 270.

190) Danielle Keats Citron, “Technological Due Process”, *Washington University Law Review* 85, 2007, pp. 1249~1313: p. 1278.

191) Tal Z. Zarsky, “Governmental Data Mining and its Alternatives”, *Pennsylvania State Law Review* 116, 2012, pp. 285~330 참조.

192) Kenneth A. Bamberger, “Technologies of Compliance: Risk and Regulation in a Digital Age”, *Texas Law Review* 88, 2010, pp. 669~740: p. 675.



IV. 투명성 제한의 근거로서 비밀과 차별

1. 공공안전과 국가안보를 위한 기밀과 영업비밀

앞서 루미스 사건¹⁹³⁾에서도 나타나듯이 국가는 불법적인 행위를 예견할 수 있는 자동화되고 개별화된 예측에 특별한 관심을 갖는다. 자동화된 개별적 예측 모델을 만들기 위해서는 불법행위와 연결되는 행위자의 특성을 포착하는 것이 중요하다. 이러한 특성은 개인을 식별할 수 있는 특성일 수도 있지만 행동 패턴을 아는 것만으로도 사전적인 예측은 가능하다. 그런데 이러한 예측 모델이 공공 일반에 투명하게 공개될 경우 이를 악용하려는 행위자는 예측 모델의 표적이 되는 것으로부터 벗어날 수 있다. 즉, 투명성이 예측 모델로 달성하려고 하는 목표를 약화시키는 것이다.¹⁹⁴⁾

특히 법집행과 국가안보에 관한 예측 모델의 적용에서 투명성은 오히려 심각한 피해를 가져올 수 있다. 예를 들어 세법을 준수하고 세입을 증가시키려는 예측 모델이 있다고 할 때 그 모델에 관한 투명성은 조세를 회피할 수 있게 하고 궁극적으로 세입을 감소시킬 수 있다. 항공 보안을 위해 공항 검색대에서 테러리스트를 식별하기 위해 고안된 예측 모델의 알고리즘이 공개되는 경우에도 투명성은 예측 모델이 제 기능을 하는 것을 방해한다.

이러한 강력한 논거는 현행법에 반영되어 정보 공개와 관련된 법률의 비공개에 관한 예외조항으로 규정되곤 한다. 또한 예측 모델의 투명성을 확보하기 위해 소프트웨어의 소스코드를 공개하는 경우 관련 소프트웨어를 개발한 업체는 경쟁업체에게 영업비밀이 노출되는 것을 우려할 수밖에 없다. 그렇기 때문에 이러한 비밀주의 역시 감사 목적에 필요한 범위에서 투명성을 제한하도록 요구하는 논거가 된다. 기업과 국가의 의도적 비밀주의는 투명성보다는 불투명성을 지지하는 핵심 논거가 된다.

2. 대용물로 사용된 특성에 의한 낙인

예측 모델에는 여러 가지 특성이 대용물로 사용된다. 앞서 살펴보듯이 공공안전이나 국가안보를 위한 목적으로 고안된 예측 모델에 따라 데이터를 처리할 때 사용된

193) State v. Loomis, 2016 WI 68, 881 N. W. 2d 749 (2016) 및 이에 관한 본 논문 「제2장 제3절 I. 2. 루미스 사건」 참조.

194) Tal Z. Zarsky, "Transparent Predictions", *University of Illinois Law Review* 2013(4), 2013, pp. 1503~1570: pp. 1553~1560.



대용물의 목록을 열람할 수 있게 될 경우 대부분의 열람자는 반사회적이거나 범죄적인 행동을 할 위험과 연결된 대용물로 사용된 특성과 동일한 특성을 가진 개인들의 집단을 부정적으로 가정하고 인식하게 될 수 있다.¹⁹⁵⁾ 즉, 투명한 공개를 통해 위험 예측에 사용된 특성을 가진 개인 일반에 대한 고정관념을 형성시키는 것이다. 그러한 가정이 잘못된 것이라고 할지라도 다수의 인식 속에 형성된 고정관념은 사회적으로 문제 있는 결과를 양산해 낼 수 있다.

이런 일이 가능한 이유 중 하나는 예측 모델이 찾아내는 상관성의 맥락을 오해하기 때문이다. 예를 들어 위험 요소로 사용된 대용물과 연결된 특성을 가진 사람의 동료로서 그 사람과 같은 특성을 공유한다는 것만으로도 얼마든지 같은 위험 집단에 묶일 수 있다는 맥락을 이해하지 못할 경우 예측 모델에 사용된 대용물은 쉽게 공공 일반에게 고정관념을 심어줄 수 있다. 또 다른 이유는 예측 모델에 기초한 일반화에 있다. 예를 들어 어떤 예측 모델이 A 도시 거주자의 소득 신고 비율이 저조하다는 것을 A 도시 거주자는 신뢰할 수 없거나 진실하지 못하다고 일반화하는 것이다. 상관성을 잘못 이해하는 경우인데 국가가 공개하는 차원에서 특정 요인들이 상관성을 띤 형태로 나타나게 되면 특정 집단에 대한 낙인화가 뒤따를 수 있다. 그렇게 되면 낙인화된 집단의 구성원으로서 개인들은 고착된 프로파일로부터 벗어날 수 없고 부정적인 고정관념에 갇힐 수 있다.

이러한 논의는 여러 해 동안 잘못된 대우와 극심한 편견에 시달려온 특정 보호 집단에 초점이 맞춰져 있다. 물론 자동화된 예측 분석이 이러한 집단을 또 다시 차별하여 차별을 재생산할 것이라는 우려에 대해서 예측 모델에 사용되는 패턴들이 과거에 차별을 받은 개인들의 집단에 대한 대용물이 아니라고 하여 안심시킬 수도 있다. 실제로 그러한 보장이 가능하다 해도 낙인화와 관련된 우려는 완전히 해소되지 않는다. 예측 모델은 새로운 집단을 만들어 내기 때문이다. 사회적으로 충분하게 이해되고 공유되지 않은 조롱과 불신의 대상이 되는 사회의 하위집단이 생성될 수 있는 것이다. 그러한 집단은 인종, 성별, 국적처럼 잘 알려져 있는 기존의 집단으로 분리되지 않을 것이고, 사회적으로도 지리적으로도 분산되어 있을 수 있다. 또한 예측 모델에 따른 알고리즘의 프로세스는 자동적이라는 고유의 특성을 갖는다는 사실을 고려하면 특정 요인이 대용물로 선택된 이유에 관한 조사가 내재되어 있지

195) Tal Z. Zarsky, "Transparent Predictions", *University of Illinois Law Review* 2013(4), 2013, pp. 1503~1570: pp. 1560~1563.



않을 수도 있다. 게다가 상관관계를 인과관계로 오인함으로써 단순한 고정관념에 과학성을 부여할 수도 있다. 따라서 새로운 집단을 낙인화한다는 논거에 따르면 예측 모델에 의한 특성 간 관계의 발견을 불투명하게 할 것이 요청된다.¹⁹⁶⁾

196) 민감한 정보와 같은 특수한 정보를 이용함으로써 어떤 특성을 공유하는 특정한 집단에 대한 차별을 정확하게 드러내는 것이 차별을 시정하는 데에 도움이 된다는 관점을 취할 경우 알고리즘의 정확성에서 기인하는 낙인에 의한 차별은 알고리즘의 투명성을 제한하는 논거가 아니라 오히려 알고리즘의 불투명성을 제한하고 투명성을 확대하는 논거로 사용될 수도 있다. 이에 관해서 본 논문 「제5장 제4절 III. 특정범주의 개인정보와 자동화된 차별적 결정」 참조.



제5장

머신러닝 알고리즘의 차별에 관한 책임과 개인정보 보호

머신러닝 알고리즘이 생성한 결정 규칙에 따른 행위가 산출하는 결과는 기존의 차별에 관한 법명제의 의미체계에 포착될 수 없거나 그 체계적 틀을 용이하게 벗어난다. 그렇기 때문에 새로운 사회적 실재를 포착할 수 있는 법체계를 구성하는 것은 정책의 과제로 남는다.¹⁾ 과학기술로서 머신러닝 알고리즘은 사회의 무수한 데이터로부터 학습하고 다시 사회의 데이터를 처리하여 새로운 데이터를 사회의 커뮤니케이션에 유통시킨다. 적어도 커뮤니케이션의 측면에서 머신러닝 알고리즘은 사회의 행위자로서 역할을 담당하고 있다. 그렇기 때문에 차별적 결정을 사회의 해악으로 여기고 이를 방지하고 교정하려고 했던 차별금지법체계를 알고리즘 에이전트가 우회할 수 있다는 것은 차별금지법에 대한 심각한 도전이 아닐 수 없다.

환경에 반응하고 환경을 조종할 수 있는 에이전트의 특성을 갖는 머신러닝 알고리즘은 데이터를 구별하고 분류함으로써 환경을 지각한다. 머신러닝 알고리즘 에이전트에게 인간의 데이터는 그저 지각의 대상이 되는 환경에 관한 데이터이다. 인간에 관한 데이터로부터 일정한 패턴을 발견하여 그에 따라 목표한 과제를 수행한 알고리즘이 사회의 차별을 재생산하고 확대시킨다면 그에 대한 사회의 규제는 필수적이다. 그런데 구체적 차별사유에 개별적으로 대응하여 정립되어 온 차별 금지법의 규제 방식이 알고리즘의 차별에 대한 대응 방식으로 적절한지는 독자적으로 논의되지 않고 단지 그 복잡성과 난해함에 대한 언급 수준의 문제제기가 있을 뿐이고, 머신러닝 알고리즘처럼 자동화된 결정 시스템이 데이터를 처리하는 것에 대해 자연인인 개인의 데이터 보호 관점에서 정립되어 온 개인정보 관련 법제가 있을 뿐이다.

1) 정책(policy)은 보다 면밀한 탐구와 계획이 필요하고, 중국에는 법으로 귀결될 수 있을 뿐만 아니라 규제적 개입의 찬반에도 불구하고 공익을 우선적으로 고려할 수 있다는 점에서 인공지능에 관한 논의를 윤리(ethics)나 통치(governance)의 틀보다 정책의 틀로 접근하는 것이 필요하다는 주장으로 Ryan Calo, "Artificial Intelligence Policy: A Primer and Roadmap", *U. C. Davis Law Review* 51(2), 2017, pp. 399~436: pp. 407~410 참조.



데이터 프라이버시 중심의 개인정보보호법 논의에서 차별의 문제는 불공정 또는 프라이버시에 대한 해악의 문제로 언급되곤 한다. 자유와 평등의 관계라는 헌법학의 전통적 주제는 데이터를 기반으로 한 알고리즘 중심의 사회에서 데이터 프라이버시와 알고리즘을 이용한 차별 사이의 관계라는 사이버-물리 세계의 주제로 재탄생할 수 있다. 이때 공통적으로 제기될 수 있는 법적 대응 방법은 우선적으로 불투명한 알고리즘의 결정에 대해 그 결정의 상대방 즉, 인간이 알 수 있도록 설명을 강제하는 것이다. 그러나 불투명한 알고리즘에 대한 설명은 알고리즘의 차별이라는 구체적 문제에 대한 궁극적 해결 방법이 될 수 있는지에 대해서는 보다 진전된 논의가 필요하다.



제1절 차별에 관한 책임의 구조

알고리즘으로 자동화된 결정이 사회의 사이버-물리적 환경을 조성하고 조종하는 세계에서 알고리즘의 오류와 편향, 과대적합, 복잡성 등은 아날로그 세계를 전제로 개발되고 발전되어 온 차별금지법체계에 혼란을 야기한다. 기존의 법체계에서도 차별은 매우 다루기 어려운 복잡한 현상이었다. 차별은 완전한 평등, 완전한 자유, 완전한 존엄, 완전한 통합에 도달하는 것을 방해하는 장애물이자 언제나 불완전한 평등, 불완전한 자유, 불완전한 존엄, 불완전한 통합의 상태를 창출할 수 있는 사회의 위험 요소이기도 하다. 복잡한 사회 현상으로서 차별이 인간의 이해 수준을 넘어설 정도로 복잡한 알고리즘 에이전트를 만나 자동화되고 정교해지고 합리적인 모습으로 사회에 재생산된다는 것은 알고리즘이 지배하는 사회에서 상당한 위험이 된다. 이러한 경향은 머신러닝 알고리즘이라는 특수한 과학기술에 정향되어 있다기보다는 과학기술 발전의 일반적 맥락 속에 있고, 법체계는 이러한 위험의 불확정성을 예측 가능성이라는 자체적 성격으로 감소시켜 왔다. 그런데 법을 미세규범으로 기능하는 알고리즘이 대체할 경우 불투명한 알고리즘의 작동 결과에 대한 책임 소재가 명확하게 밝혀지기 어렵다면 그 위험은 분산되어야 한다.

I. 과학기술의 발전과 책임 구조의 변화

1. 위험에 대응하기 위한 규칙 시스템의 진화 과정

어떤 시대의 사회라도 그 사회를 지배하는 일정한 책임구조가 있다. 마찬가지로 책임이 현대사회만의 특수한 현상은 아니지만 과학기술의 발전이 책임구조에 변화를 불러일으켰다는 점만은 틀림이 없다.²⁾ 그리고 이러한 변화가 책임에 대한 새로운 해석도식과 새로운 개념구성에 반영되어 왔다는 점에서 19세기와 20세기에 전개된 현대사회 제도의 정치사를 “인간의 결정에서 비롯된 산업적 불안과 위험에 대응하기 위해 갈등으로 점철된 규칙 시스템이 진화하는 과정”³⁾이라고 본 벡(U. Beck)의 통찰은 여전히 유효하다.

2) Ulfrid Neumann, “Zur Veränderung von Verantwortungsstrukturen unter den Bedingungen des wissenschaftlich-technischen Fortschritts”, in: *Recht als Struktur und Argumentation: Beiträge zur Theorie des Rechts und zur Wissenschaftstheorie der Rechtswissenschaft*, Nomos-Verl.-Ges, 2008, S. 188-202 참조.

3) Ulrich Beck, *World at Risk*[*Weltrisikogesellschaft*, 2007], Ciaran Cronin(Trans.), Polity Press, 2009, p. 7.



책임은 귀속의 문제이다. 결과로서 발생한 손해를 어디에 귀속시킬 것인지가 곧 책임의 문제인 것이다. 전통적인 책임 모델은 ‘불운’과 ‘불법’의 구별을 토대로 삼는다. 규범에 따라 행동한 사람은 그 결과가 다른 사람에게 손해를 발생시켰을지라도 그에 대한 책임을 부담하지 않는다. 규범에 반하여 행동한 사람은 그 결과로 발생한 다른 사람의 손해에 대해 책임을 부담한다. 규범을 준수한 행위, 즉 합법인 행위에 대해서 그에 대한 책임을 면제시켜 주지만 규범을 준수하지 않는 행위, 즉 불법인 행위에 대하여 그 책임을 면제시켜 주지 않는다. 손해를 입은 사람의 입장에서 다른 사람의 합법인 행위에 따른 결과로 자신에게 발생한 손해는 감내해야 하는 불운인 것이다.

2. 사고와 위험 책임

불운과 불법을 기초로 책임을 부담시키는 책임 모델은 과학기술이 발전하는 조건에서 발생하는 위험의 문제를 다루기 어렵다. 예를 들어 원자력 발전 기술을 사용하는 사람은 법률과 규칙에 따르기만 하면 다른 사람에게 어떤 손해가 발생해도 전통적인 책임 모델에서는 어떤 책임도 지지 않는다. 이를 해소하기 위해 전통적 책임 모델의 불운과 불법 사이에 ‘사고(Unfall)’ 개념이 자리 잡게 된다. 사고는 “사회적 효용 때문에 일반적으로 인정되는 위험으로 인해 도저히 피할 수 없는 구체적 결과를 낳는 현상”⁴⁾을 뜻한다. 사고는 합법적인 행위라도 그로 인해 발생한 손해에 책임을 져야 하는 영역을 창출한다. 그러므로써 사고에 대한 위험 책임이 전통적 책임 모델의 구조를 변형시킨다. 위험 책임 개념을 법적 책임으로 받아들이면서 책임이 확장되는 경향을 갖게 된다. 예를 들어 형법의 영역에서 침해범 위주의 구성요건은 위험범을 흡수하는 쪽으로 확대되고 있다. 이러한 경향은 민주적 헌법 국가에서 이른바 ‘책임의 인플레이션’ 현상이 나타나고 있다는 인식⁵⁾을 불러일으키는 주요 논거가 된다.

위험 책임 법리는 책임을 귀속시킬 수 있는 불법적 위험과 책임을 귀속시킬 수 없는 허용된 위험을 구별한다. 위험을 책임의 영역에 끌어들임으로써 책임의 구조에 어떤 변경이 발생하는지 살펴보기 위해 기본적인 책임의 관계를 이해할 필요가 있다. 책임은 세 가지 부분으로 구성된 관계인데, 누군가가 다른 사람에

4) Ulfrid Neumann, “Zur Veränderung von Verantwortungsstrukturen unter den Bedingungen des wissenschaftlich-technischen Fortschritts”, in: *Recht als Struktur und Argumentation: Beiträge zur Theorie des Rechts und zur Wissenschaftstheorie der Rechtswissenschaft*, Nomos-Verl.-Ges, 2008, S. 188-202: S. 191.

5) Horst Dreier, “Verantwortung im demokratischen Verfassungsstaat”, in: *Verantwortung in Recht und Moral*, Ulfrid Neumann · Lorenz Schultz(Hrsg.), Steiner, 2000, S. 9-38 참조.



대해 무엇인가에 관해 책임을 부담한다. 책임의 관계는 누군가가 자신의 행위로 인해 발생한 어떤 결과에 관해 다른 사람에게 부담하는 것이다. 다시 말해 책임의 주체와 책임의 대상 그리고 책임의 상대방 사이의 관계로 이루어지는 것이다.

3. 책임의 주체와 상대의 확장

책임의 주체는 고전적인 관점에 따르면 인격으로서 개인이다. 칸트는 인격을 “책임이 귀속되는 행위를 할 능력이 있는 주체”⁶⁾로 본다. 민사법 영역에서는 일찍이 법적 인격을 자연인에게만 부여하던 골레를 벗어났다. 그럼으로써 법인, 즉 조직이나 기관에게 책임을 귀속시킬 수 있도록 했다. 이는 달리 말하면 책임의 주체가 개인에서 집단으로 확장됐다는 것을 의미한다. 의사결정은 조직 내부에서 이루어지는데 그 결정에 개인이 얼마만큼 관여하고 영향력을 행사했는지는 쉽게 파악되지 않을 뿐만 아니라 쉽게 감추어질 수도 있다. 요나스(H. Jonas)는 기술시대의 윤리학으로 책임의 원칙을 제시하면서 책임의 주체를 인류 전체로까지 확대시킨다.⁷⁾ 이때 제기할 수 있는 것이 책임의 상대방 문제이다. 상대방 없는 책임은 무의미하기 때문이다. 인류가 책임의 주체라면 그 상대방은 환경이거나 과거 또는 미래의 세대일 것이다.⁸⁾

4. 책임 대상의 확대

책임의 대상은 전통적으로 자신의 행위로 인한 결과이다. 그것도 규범을 준수하지 않은 행위로 인한 결과이다. 이는 규범주의 모델에 따르는 것이기도 하다. 합법적인 자신의 행위로 인한 결과도 책임의 대상에 포함시키는 것은 결과주의 모델에 해당한다. 결과주의 모델은 어떤 행위라도 결과와 관련이 있으면 책임을 지도록 구성할 수도 있지만, 그 행위에 따른 결과가 어느 정도 규칙성 또는 반복성을 가질 때 비로소 책임을 부담하도록 구성할 수도 있다. 결과주의 모델에서 행위와 결과의 관련성은 책임의 주체와 책임의 상대방 간의 거리를 암시한다. 행위로 인한 결과의 인과관계를 무한히 확장하여 책임 대상 범위를 넓히면 그만큼 책임의 상대방 또한 무한히 확장된다. 인과관계의 연결고리에서 멀리 떨어져 있는 상대방도 책임의 상대방이 되는 것이다. 이는 책임을 보편화시킨다. 또한 위험을 책임의 영역으로

6) Immanuel Kant, *Metaphysik der Sitten*, AB 22.

7) 요나스가 제시하는 책임의 원칙에 관해서 Jonas, Hans, 책임의 원칙: 기술 시대의 생태학적 윤리[*Das Prinzip Verantwortung: Versuch Einer Ethik Für Die Technologische Zivilisation*, 1979], 이진우(역), 서광사, 1994, 84~94쪽 및 Jonas, Hans, 기술 의학 윤리: 책임 원칙의 실천[*Technik, Medizin und Ethik: zur Praxis des Prinzips Verantwortung*, 1987], 이유태(역), 숲, 2005 참조.

8) 대한민국헌법은 전문에서 “우리들의 자손의 안전과 자유와 행복을 영원히 확보할 것을 다짐”한다고 하여 책임의 상대에 미래 세대를 포함하고 있다.



끌어들이면 책임의 대상은 자신의 행위에 한정되지 않고 부작위로 확장된다. 자신이 보호하는 아동의 행위로 발생한 손해의 결과도 자신이 아동을 보호하지 않은 부작위의 결과로 보아 책임의 대상이 된다.

II. 차별에 관한 국가와 사인의 책임

1. 민주주의 사회에서 자동화된 차별의 위험

정부가 특정 정책을 수행하기 위해 시민을 분류하고 예측하거나 기업이 사업의 목표를 달성하기 위해 고객을 분류하고 예측하는 것에는 항상 차별의 위험이 담겨 있다. 더구나 정부나 기업이 이러한 분류 및 예측을 위해 자동화된 결정 시스템을 도입했을 때 의도를 확인하기 어려운 머신러닝 알고리즘이 추론과 통계적 근거에 기초하여 수행한 분류와 예측은 차별행위를 수행한 주체를 찾아 그 의도를 확인하여 책임을 묻는 방식으로 차별을 금지하는 접근법에 한계를 가져온다. 따라서 차별로부터 보호되어야 하는 시민 또는 고객에게 인간의 결정을 대체하는 머신러닝 알고리즘 시스템의 결정이 차별에 관한 규범적 조건을 우회하거나 정당화하거나 은폐할 수 있게 된다는 것 역시 법의 보호로부터 벗어난 영역에서 상시적인 차별에 노출될 수 있는 위험이 된다.

민주주의 사회에서 나타나는 위계에 대해 경제적 위계(economic hierarchy)와 신분 위계(status hierarchy)로 나누기도 하는데,⁹⁾ 마샬(T. H. Marshall)이 이론화했던 영국 사회의 경우를 살펴보면 경제적 위계는 지주 귀족에서부터 상인 및 산업 자본 엘리트, 그리고 전문직, 정신노동자를 거쳐 숙련 노동자와 비숙련 노동자에 이르는 식의 구조를 갖는다.¹⁰⁾ 이러한 경제적 위계 구조에서 한 사람의 지위는 그 사람이 시장 또는 생산 수단과 맺고 있는 관계에 따라 결정된다. 경제적 위계 구조에 대한 투쟁의 형태는 재분배의 정치(politics of redistribution)로 나타난다. 따라서 재분배 정치의 목표는 기회나 문화에 대해 계급적 차이를 축소시키는 것이다. 반면 신분 위계는 아일랜드인 보다 영국인이, 유태인이거나 이슬람교도보다 기독교도가, 구교도보다 신교도가, 흑인이나 황인보다 백인이, 여성보다 남성이, 동성애자보다 이성애자가, 장애인보다

9) Will Kymlicka, *Contemporary Political Philosophy: An Introduction*, 2nd ed., Oxford University Press, 2002, pp. 331~333 참조.

10) 마샬(T. H. Marshall)은 시민권(citizenship)을 권리로 이해하여 18세기, 19세기, 20세기를 각각 시민적 권리(civil rights), 정치적 권리(political rights), 사회적 권리(social rights)의 시대로 본다. Thomas H. Marshall, *Citizenship and Social Class*, Cambridge University Press, 1950, p. 21.



‘정상인’¹¹⁾이 더 낫다는 식의 구조를 갖는다. 차별받는 하층 신분 집단이 신분 위계 구조에 대해 벌이는 투쟁은 인정의 정치(politics of recognition)¹²⁾ 또는 차이의 정치(politics of difference)¹³⁾로 나타난다. 인정의 정치는 표현, 해석, 소통의 사회적 양식에 자리 잡고 있는 문화적 부정의(cultural injustices)¹⁴⁾에 초점을 맞춘다.

민주주의 사회를 표방하는 국가라고 할지라도 오랜 세월 형성되어 온 위계 구조는 그 구조 자체를 드러내지 않고도 끊임없이 방대한 데이터를 생산해 낼 수 있다. 그런데 다시금 그 데이터의 관계가 머신러닝을 통해 분석되어 알고리즘으로 생성된다면 누구도 이 알고리즘이 학습의 토대가 된 위계 구조를 그대로 답습하지 않을 것이라고 보장할 수는 없을 것이다. 또한 차별이 개인에 대해서뿐만 아니라 사회 전체에 대해 미치는 해악을 고려할 때 차별의 재생산 및 확대의 자동화는 한 국가뿐만 아니라 인류 사회 전체에 대한 심각한 위협이 된다는 점도 부정하기 어렵다.¹⁵⁾

그리고 인간의 주관적 레이블링이 개입하지 않는 비지도학습의 경우¹⁶⁾ 훈련용 데이터의 원천이 사회라고 할지라도 인간이 사회의 집단을 구별하는 방식과 다른 관점에서 머신러닝 알고리즘이 집단을 구별하는 방식을 통해 인간이 이해하기 어렵거나 이해할 수 없는 차별의 영역이 발생할 수 있다는 점은 머신러닝 알고리즘에 의한 결정에 잠재되어 있는 차별의 위험이기도 하다. 차별금지법체제에서 차별금지 사유로 다루지 않는 특성이라고 할지라도 머신러닝 알고리즘은 얼마든지 통계적 자료에 기초해서 정교한 집단화를 수행하여 개인에게 법적·사실적으로 중요한 영향을 미치는 결정을 형성할 수 있기 때문이다.

11) 신분적 위계 구조에서 비장애인은 인습에 따라 ‘정상인’으로 표현되기도 한다.

12) Charles Taylor, “The Politics of Recognition”, in *Multiculturalism: Examining the Politics of Recognition*, Amy Gutmann(Ed.), Princeton University Press, 1994, pp. 25~73.

13) Iris Marion Young, *Justice and the Politics of Difference*, Princeton University Press, 1990, p. 25: “권리는 소유개념으로는 적절히 파악될 수 없다. 권리는 사물이 아니라 관계이다. 권리는 타인과의 관계에서 무엇을 할 수 있는지를 정하는 제도적으로 정의된 규칙들이다. 정의란 분배뿐 아니라, 개인의 능력과 집합적 의사소통 및 협동을 발전시키고 행사하는 데 필수적인 제도적 조건을 일컫는다. 이와 같은 정의 개념에서 본다면, 부정의란 무엇보다도 두 가지 형태의 구속, 억압, 지배를 의미한다. 이 구속 중에는 분배적 유형도 있지만, 분배논리와는 잘 맞지 않는 다른 요소들도 있다. 의사결정절차와 분업 그리고 문화가 그런 것이다.”

14) Nancy Fraser, “Social Justice in the Age of Identity Politics: Redistribution, Recognition and Participation”, in *Culture and Economy after the Cultural Turn*, Larry J. Ray · R. Andrew Sayer(Eds.), SAGE, 1999, pp. 25~52: p. 29.

15) Christian Delacampagne, 인종차별의 역사[*Une Histoire du Racisme*, 2000], 하정희(역), 예지, 2013 참조.

16) 머신러닝 알고리즘의 비지도학습방식에 관해서 본 논문 「제2장 제2절 III. 2. 비지도학습과 군집」 참조.



예를 들어 기업이 이용하는 머신러닝 알고리즘이 맞춤형 서비스를 제공하기 위해 고객들의 물품 구매 기록으로부터 채소 위주의 소비 패턴을 보이는 고객들만을 추출할 경우 채식주의자 집단을 분류하게 된다.¹⁷⁾ 이러한 분류 자체로도 잠재적인 차별의 위험이 발생할 수 있지만, 보험 회사에서 완전채식주의라는 구별 항목을 넣어 별도로 요금을 책정하는 경우에도 잠재적인 차별의 위험이 발생한다.¹⁸⁾ 그러나 채식 위주의 식사를 하는 특성이 차별금지법체계로 포착되지 않는 상황에서는 그러한 위험은 허용된 위험으로 남게 된다.

결국 머신러닝 알고리즘의 차별적 결정이 만들어 낸 차별적 상황을 책임의 대상으로 삼을 것인지 그리고 그 책임의 주체와 상대방을 어떤 범위로 설정할 것인지 책임을 분배하는 문제가 남게 된다. 다만 이러한 책임 분배의 문제는 사인에게 선별적인 방식으로 책임을 부담시킬 경우 사인 간에 또 다른 차별의 가능성을 잠재적으로 포함할 수 있다.

2. 차별에 관한 국가의 책임과 시스템의 차별

차별에 대한 책임의 주체는 기본적으로 국가와 사인으로 구분할 수 있다.¹⁹⁾ 차별 행위의 주체를 국가와 사인으로 나눌 경우 국가와 사인의 어떤 행위가 법으로 금지된 차별에 해당한다면 법에 정한 효과에 따라 행위자에게 법적 책임이 귀속된다. 가해자와 피해자의 도식으로 구성된 차별 문제는 이러한 전통적인 행위와 책임의 구성 방식과 잘 부합한다. 그리고 평등원칙을 논증부담 또는 입증부담의 원칙으로 재해석하는 절차주의적 이해에 따르면 불평등으로서 차별을 정당화하는 측이 논증책임 또는 입증책임을 부담하게 된다.²⁰⁾

17) 마케팅 기업이 사용하는 머신러닝 알고리즘은 신생아용 기저귀나 무향 비누, 엽산제 등을 인터넷 장바구니에 넣어 둔 여성 고객의 온라인 활동 흔적만으로도 고객의 임신 사실을 추론하여 맞춤형 광고로 임신 및 출산 관련 용품을 보여 줄 수 있다. 이에 관해서 Kashmir Hill, "How Target Figured Out a Teen Girl Was Pregnant Before Her Father Did", *Forbes*, 16 February 2012, <https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did> 및 본 논문 「제4장 제2절 III. 2. 머신러닝 알고리즘의 특성 추론과 간접차별」 참조.

18) 보험회사가 완전채식주의(veganism)를 구별 항목으로 넣는 경우의 예는 Tal Z. Zarsky, "An Analytic Challenge: Discrimination Theory in the Age of Predictive Analytics", *I/S: A Journal of Law and Policy for the Information Society* 14(1), 2018, pp. 11~35: p. 33 참조.

19) 차별의 문제를 가해자와 피해자의 구도 속에서 가해자 측, 차별의 주체를 기준으로 국가에 의한 차별과 사인에 의한 차별을 유형화하는 경우로 이준일, *차별없는 세상과 법*, 홍문사, 2012, 41~44쪽 참조.

20) 평등원칙을 절차주의의 관점에서 논증부담원칙으로 재해석한 경우로 이준일, "소수자(Minority)와 평등원칙", *헌법학연구* 8(4), 2002, 219~243쪽 참조.



차별을 위험으로 구성하는 경우에는 위험을 발생시킨 행위를 한 주체에게 책임을 귀속시킬 수 있고, 가해자로서 차별의 주체를 국가와 사인으로 구분하여 각각의 책임 여부나 정도를 분담시킬 수 있다. 그런데 차별의 결과는 있지만 그 결과와 결부된 원인 행위의 제공자를 찾을 수 없는 경우, 예를 들어 누적적인 편견과 고정 관념에 따라 체계적으로 구축된 시스템에 의해 차별적 결과가 발생한 경우에 대해 행위 중심의 책임 법리 구성은 대응에 한계가 있다.²¹⁾ 위험의 관점에서 보면 이러한 차별의 유형은 행위에 결부시킬 수 없는 위험 즉, 허용된 위험으로 합법의 영역에 든 법정책적 차원의 결정이라고 볼 수도 있다.²²⁾

그렇다면 차별에 대한 국가의 책임은 직접 차별적 결정을 했기 때문일 수도 있고, 차별적 상황이 발생하는 것을 방지하지 못했거나 발생한 차별적 상황을 해소하기 위한 구제책을 마련하지 못했기 때문일 수도 있다. 이러한 근거가 타당하기 위해서는 국가에게 차별적 결정을 하지 않을 의무 또는 차별적 상황을 발생시키지 않거나 발생한 차별적 상황을 해결할 의무가 있어야 한다. 대한민국헌법에 따르면 국가는 “누구든지 성별·종교 또는 사회적 신분에 의하여 정치적·경제적·생활의 모든 영역에 있어서 차별을 받지 아니”(제11조 제1항)하는 것을 “보장할 의무를 진다”(제10조). 이는 차별에 대한 책임을 헌법의 차원에서 국가에게 귀속시키는 하나의 논거가 될 수 있다. 허용된 위험에 대해 그 발생 확률을 감소시키는 것은 헌법으로 보장된 누구도 차별을 받지 않을 권리를 사회의 시스템으로 구현하는 것에 대한 책임을 누구에게 귀속시킬 것인지에 관한 다른 차원의 문제가 된다. 일차적으로 그 책임이 국가에게 있다면 국가는 그 책임을 이행하는 다양한 방법을 모색해야 한다.

3. 차별에 관한 사인의 책임과 차별금지법의 특수한 책임 주체

문제는 차별에 대한 책임을 사인에게도 부담시킬 수 있는 것인지 그 근거를 찾거나 논거를 구성하는 것이다. 차별금지법의 역사적 전개는 크게 두 가지 흐름을 보이는데, 하나는 미국의 인종, 영국의 성별처럼 어떤 한 가지 사유에 관한 역사적 차별의 경험에 관한 법적 대응으로 시작되어 차츰 그 적용 사유를 확대하는 경향으로 전개

21) 차별의 문제를 대립적 구도로 접근하는 방식의 한계를 지적하며 개인주의의 대립적 시스템이 집단 기반 모델에 의해 구조를 개선하는 기관의 역할이 강조되는 방식으로 보완될 것을 주장하는 견해로 Sandra Fredman, *Discrimination Law*, 2nd ed., Oxford University Press, 2011, pp. 280~295, 특히 pp. 294~295 참조.

22) 허용된 위험(Risiko)과 위법한 위험(Gefährdung)을 용어상 구별하는 경우는 Werner Krawietz, *Theorie der Verantwortung – neu oder alt?: Zur normativen Verantwortungsattribution mit Mitteln des Rechts*, in: *Verantwortung: Prinzip oder Problem?*, Kurt Bayertz(Hrsg.), WBG, 1995, S. 184~216 참조.



되어 오고 있다는 점이고, 또 다른 하나는 차별금지법이 규정하는 의무를 부담하는 주체가 국가에 사인으로 확대되어 왔다는 점이다. 이는 차별에 대한 책임을 귀속시킬 주체에 관한 문제이면서 동시에 차별금지법을 적용할 영역을 공적 영역에 한정할 것인지 사적 영역으로 확대할 것인지에 관한 문제이기도 하다.

헌법 차원에서 가능한 논거 중 하나는 차별에 관한 법조문을 근거로 ‘누구든지 차별을 받지 아니한다.’는 점을 강조하는 구성이다. 차별을 ‘받는’ 입장에서 차별을 ‘누가 가하는지’에 따라 차별의 유무가 달라지는 것은 아니기 때문이다. 차별의 주체는 국가에 의한 차별과 사인에 의한 차별이라는 종류를 구분하는 기준이 될 수 있을 뿐이다. 이러한 구성은 사인이 차별을 일방적으로 ‘당하기만 하는’ 것에서 ‘주고’ 받을 수 있는 것으로 보는 관점에도 열려 있다. 물론 다른 방식으로 논거를 구성할 수도 있다. 누구든지 차별을 받지 않는 것을 ‘보장할 국가의 의무’를 강조하는 것이다. 차별에 대한 책임을 국가에게 귀속시킨다는 것은 국가에게 차별로부터 모든 사람을 보호할 의무를 부과함과 동시에 그러한 의무를 실현할 권한을 국가에게 부여하는 것이기도 하다. 이러한 권한에 근거해서 국가는 사인에게도 다른 사람을 차별로부터 보호할 의무를 부담시킬 수 있는 것이다. 다만, 헌법의 차원에서 차별 그 자체에 대해 말해주는 바가 없다면 차별 개념에 대한 이론적 구성과 해석에 따라 의무의 내용은 달라질 것이다.

차별금지법의 적용 범위를 사적 영역으로 확대하는 경우에도 재화와 용역의 분배 기능을 수행하는 경우로 국한할 것인지 분배와 상관없는 대우도 포함할 것인지 문제된다. 즉, 차별로부터 보호할 의무를 사인에게 부과할 것인지 여부 그 자체가 아니라 그러한 의무를 모든 사인에게 부과할 것인지 아니면 어떤 사인에게만 부과할 것인지 여부가 문제되고, 특정한 사인으로 범위를 제한한다면 그 범위를 결정하는 기준이 무엇인지가 중요한 의미를 갖는다. 국가인권위원회법의 “평등권 침해의 차별행위”(제2조 제3호)²³⁾는 차별행위의 주체를 특별히 제한하지 않고 ‘모집, 채용, 교육, 배치, 승진, 임금 및 임금 외의 금품 지급, 자금의 융자, 정년, 퇴직, 해고 등을 포함하는 고용 관계, 재화·용역·교통수단·상업시설·토지·주거시설의 공급이나 이용 관계, 교육시설이나 직업훈련기관에서 실시되는 교육·훈련이나 그 이용 관계’로 적용 분야를 제한함으로써 차별에 대한 책임의 주체로 사인을 포함시키면서 그 범위를 제한한다. 이는 어디에 거주하면서, 무슨 직업을 갖고,

23) 이에 관해서 본 논문 「제3절 제3절 I. 2. 법률의 차별 관련 규정과 ‘평등권 침해의 차별행위’」 참조.



어떤 물건이나 서비스를 구입해서 생활을 할 것인지는 생활의 기초가 되는 자원의 분배와 관련된 업무로서 고용이나 직업, 교육이나 훈련, 주거시설이나 토지, 재화나 서비스의 공급과 이용에 관련된 업무의 담당자인 공급자나 관리자에게 차별금지의무를 부담시켜 책임을 귀속시키는 것으로 볼 수도 있다.

그런데 가해자와 피해자의 구도로 차별의 문제를 구성하면 차별금지법이 생활의 기초가 되는 자원의 분배에 관한 업무를 담당하는 공급자나 관리자에게 한정해서 차별금지의무를 부과하는 것에 대한 일관된 설명이 어렵다. 또한 차별의 문제를 역사적이고 경험적으로 불리한 대우를 받아 온 집단을 보호하기 위한 것으로 구성하는 경우에는 과거에 직접 또는 간접적으로 차별에 관여한 적이 없음에도 불구하고, 즉 현재의 차별에 대한 과거의 가해자가 아님에도 불구하고 현재 직원의 고용, 주거시설의 임대, 재화나 서비스의 공급을 담당하는 자에게 차별금지의무를 부담시키는 것에 대해서도 적절한 설명을 찾기 어렵게 된다. 차별의 문제는 가해자를 찾아 가해자에게 손해를 배상하도록 하는 관점도 중요하지만 그에 못지않게 누구든 차별의 대상이 될 수 있다는 공동의 위험에 대한 책임을 어떻게 분산시킬 것인지에 대한 관점도 고려되어야 하는 이유이다.

III. 차별금지의무의 성격과 내용

차별금지법에 따라 부과되는 의무의 면면을 들여다보면 단순히 적극적 의무와 소극적 의무라는 이분법의 구분 도식을 적용하여 이론으로 구성하는 것이 큰 의미를 갖지 못할 수 있다. 실제로 미국, 캐나다 등 여러 국가의 차별금지법제에서 채택하고 있는 의무의 내용은 직·간접차별에 대한 금지뿐만 아니라 합당한 배려, 적극적 조치의 이행, 보복 또는 괴롭힘의 금지 등 다양하다.²⁴⁾ 이렇게 차별에 관한 책임을 구체화하는 의무의 종류와 내용이 개별적이고 산발적인 것은 차별 개념으로 포착되는 차별 유형이 정치적으로 포착돼 그에 대한 구제방법으로 구상한 내용이 차별금지법체계에 편입되어 온 차별금지법의 역사적 전개 양상 때문이기도 하다. 그래서 이러한 다양한 의무들을 적극적 의무와 소극적 의무의 구분 틀로 이론화하는 것을 지양하는 차별금지법이론을 구성하기도 한다.

24) 미국과 캐나다, 호주와 뉴질랜드, 인도와 남아프리카공화국, 그리스와 아일랜드 등 각국의 차별금지법제에 관한 내용은 이준일, *차별없는 세상과 법*, 홍문사, 2012, 145~161쪽 참조; 유럽의 차별금지법제에 관해서는 Evelyn Ellis · Philippa Watson, *EU Anti-Discrimination Law*, 2nd ed., Oxford University Press, 2012 참조.



1. 차별금지에 관한 소극적 의무와 적극적 의무 도식

알고리즘의 차별에 대한 책임은 책임의 주체가 부담하는 의무의 형식과 내용에 따라 다르게 구체화될 수 있다. 차별을 행위 중심으로 구성하면서 차별행위의 불법성에 대한 책임을 차별금지의무 위반에 둘 경우 차별에 대한 책임은 차별금지의무를 이행하는 것으로 귀결된다. 이때 소극적 의무와 적극적 의무라는 의무 형식의 이분법적 구분 도식을 그대로 적용하면 차별금지의무는 소극적 의무, 즉 부작위의무가 된다. 이런 구성은 차별을 오로지 적극적 행위로만 구성할 경우에 타당하다. 따라서 차별 개념이 적극적 행위뿐만 아니라 소극적 행위로도 구성된다는 입장을 취하면 차별금지의무를 소극적 의무로 보는 것은 차별 개념을 좁게 구성하고 있거나, 차별의 원인을 적극적 행위에서만 찾고 그러한 행위를 금지하는 것을 차별금지라고 함으로써 ‘차별금지(anti-discrimination)’의 의미를 오해한 것으로 보일 수 있다. 차별을 결과 중심으로 구성하여 행위와 연결시키는 경우에도 적극적 행위만 차별 결과에 영향을 미치는 것은 아니다. 불균형적인 시스템을 그대로 유지하는 소극적 행위도 불균형의 정도를 더 심화시킬 수 있기 때문이다. 이러한 관점에서도 역시 차별금지의무를 곧장 소극적 의무와 연결시키는 것은 마찬가지로 오해를 불러일으킬 수 있다.

차별 형식은 구별 또는 분류로부터 출발한다.²⁵⁾ 구별이나 분류가 선행 행위일 수도 있고 어떤 결정에 의해 비로소 구별과 분류의 결과가 나타날 수도 있다. 하지 말아야 할 구별이나 분류를 한 경우 우선적으로 문제가 된다. 이는 다른 점을 고려하지 말고 무시하라는 부작위 요구 형태로 나타난다. 이는 소극적 의무라 할 수 있다. 해야 할 구별이나 분류를 하지 않은 경우도 차별의 형식에 포함된다. 차이를 무시하지 말라는, 즉 다른 것을 다르게 대우하라는 작위 요구 형태로 나타난다. 소극적 의무 위반에 대한 책임은 작위를 금지하는 형태로 나타난다. 어떤 특성을 판단의 사유로 사용하려고 할 경우 이를 판단의 사유로 사용하지 못하도록 하는 것이다. 판단이 규범적 결정이면 그러한 사유를 판단 근거에서 제외하고 다시 결정이 이루어져야 한다. 적극적 의무 위반에 대한 책임은 부작위를 금지하는 형태로 나타난다. 부작위를 금지하는 것은 작위를 명령하는 것이기도 하다. 다만 그 구체적인 행위 내용의 정도나 수준은 달라질 수 있다.

25) 이에 관해서 본 논문 「제3장 제2절 I. 차별의 전제로서 구별, 분리, 분류」 참조.



2. 합당한 배려의무의 부차적 성격

차별금지법의 의무들을 적극적 의무와 소극적 의무의 이분법에 따라 구분하는 것이 합당한 배려의무에는 적절하게 들어맞지 않을 수 있다.²⁶⁾ 의무를 적극적 의무와 소극적 의무의 이분법에 따라 구분하면 합당한 배려의무는 적극적 의무의 유형에 속한다. 예를 들어 고용주가 장애인을 고용한 경우나 직원으로 고용한 여성이 임신한 경우, 각각 장애 유형에 따른 근무여건을 조성하는 것과 임신 상태에 맞게 근무 시간을 조정하거나 유급휴가를 부여하는 것이 합당한 배려(또는 합당한 조정)의 문제다. 그런데 고용주가 장애 유형에 따른 근무여건을 조성하지 않거나 임신 상태에 맞는 조정을 하지 않는 것이 비용 때문이 아니라 장애 여부 자체와 여성의 임신 여부 자체를 근거로 하는 것이라면 이때는 합당한 배려의 문제가 아니라 직접차별의 문제가 된다.²⁷⁾ 합당한 배려 개념을 직접차별과 관련시켜 구성하면 합당한 배려 의무는 차별 금지의무를 위반하지 않은 경우에 비로소 부과되는 부차적 의무의 성격을 갖는다. 이러한 개념 구성에 따르면 장애인이나 임신한 여성 집단에 속한다는 이유로 의도적으로 비용을 감당하지 않은 경우 합당한 배려의 문제는 발생하지 않는다. 이러한 입장에서 적극적 의무와 소극적 의무 도식은 합당한 배려 같은 차별금지법상 의무의 우선적(primary) 성격과 부차적(secondary) 성격을 보여주는 데 적합하지 않은 것으로 보이게 한다.

그런데 의도를 고려하지 않고 차별 개념을 구성할 경우 앞의 예에서 장애인 또는 임신한 여성이 다른 사실적 조건에 부합하는 환경을 조성하는 데 드는 비용을 고용주에게 부담시키는 것이 의무의 내용이자 차별에 대한 구제방법이라는 점에서 의도적으로 장애 또는 임신 사유를 고려해 비용을 부담하지 않은 것과 비의도적으로 비용을 부담하지 않은 것은 부작위로써 적극적 의무를 위반한 것이라는 점에서 그 규범적 차이는 상쇄된다. 오히려 합당한 배려를 직접차별에 대해 부차적인 것으로 보는 구성은 합당한 배려를 간접차별의 유형으로 취급함으로써 차별에 관한 권리의 내용에 포함시키는 것과 깊게 연결된다. 그리고 이러한 구성을 통해 합당한 배려와 적극적 조치는 동일하게 적극적 의무의 성격을 갖지만 구분될 수 있다. 차별에 관한 권리 구성의 측면에서 합당한 배려는 차별 받지 않을 권리의 내용이 되지만 적극적 조치는 권리의 내용에는 포함되지 않는다. 그렇다고 해서 차별에 관한 책임의 주체에게

26) Tarunabh Khaitan, *A Theory of Discrimination Law*, Oxford University Press, 2016, pp. 86~87.

27) Christine Jolls, "Antidiscrimination and Accommodation", *Harvard Law Review* 115(2), 2001, pp. 642~699: pp. 646~651 참조.



적극적 조치 의무를 부과할 수 없는 것은 아니다. 모든 권리에는 상응하는 의무가 있지만, 모든 의무에 모든 권리가 상응하는 것은 아니라는 규범 이론적 명제에 따르면 적극적 조치 의무는 권리에 상응하지 않는 의무로 구성할 수 있기 때문이다. 그렇다면 합당한 배려와 적극적 조치의 구분은 작위 및 부작위라는 행위 양식과 금지 및 명령이라는 당위 양식을 각각 결합한 형식에 따라 구분하는 것과는 다른 기준을 사용한다. 의무의 양식이나 형식에 따른 것이 아니라 의무의 수준 또는 정도에 따른 구분이라고 하는 것이 보다 정확할 것이다. 의무의 주체가 마땅히 받아들여야 하는 수준과 이를 넘어서는 수준을 구분하는 것이다. 이러한 관점은 차별의 인식 및 평가 방법에서 독립적 기준으로서 표준에 따른 차별에서 이미 살펴보았다.²⁸⁾

3. 적극적 조치와 머신러닝 알고리즘의 차별

적극적 조치(affirmative action)는 교정적인 것과 비교정적인 것으로 나눌 수 있다.²⁹⁾ 교정적 조치는 특정인의 특정 행위를 교정하는 것이다. 교정적 조치를 좁게 이해하면 과거의 차별로 인해 고통 받은 특정인에게 이익을 주는 것이다. 예를 들어 리치(Ricci) 사건³⁰⁾에서 시는 흑인이 아무도 선발되지 않았다고 하여 승진시험을 취소하고, 새로운 시험으로 대체한 것에 대해 패소했다. 최초의 승진시험이 정당하지 않은 간접차별이라는 점을 증명하지 못했기 때문이다. 비교정적 조치는 집단에 대해 일반적으로 행해진 과거의 차별에 대해 교정하려는 것임에도 불구하고 특정인의 행위를 교정하기 위해 설계되지 않는다. 비교정적 조치의 수혜자는 개별적으로 차별에 의한 고통을 받았을 필요가 없고, 비교정적 조치를 이행해야 하는 사람이나 단체는 과거의 차별에 대한 책임이 없을 수도 있다.

특히 비교정적 조치는 규제 설계 방식에 따라 세 가지 유형으로 나눌 수 있다.³¹⁾ 첫 번째 유형은 수혜자에게 재화에 대한 접근을 용이하게 할 것인지 아니면 재화를 직접 지급할 것인지에 따라 촉진적 조치와 분배적 조치로 나누는 것이다. 촉진적 조치 중에 대표적인 것이 투명성(transparency) 조치이다. 작업장의 노동자나 대학 입학 지원자의 성별, 인종별, 종교별 구성이나 남녀 간 임금 격차 등에 관한 정보를 공개하는 것은 기존에 배제된 집단이 훨씬 더 많이 대표되고 표현될 수 있게 해준다. 투명성은 보다 근원적인 차별을 적발하고 조사할 수 있도록 이끌 수 있다. 홍보, 훈련,

28) 비교적 차별과 독립적 차별의 구분은 본 논문 「제3장 제4절 IV. 비교적 차별과 독립적 차별」 참조.

29) Tarunabh Khaitan, *A Theory of Discrimination Law*, Oxford University Press, 2016, pp. 81~82.

30) Ricci v. DeStefano 557 US 557 (2009).

31) Tarunabh Khaitan, *A Theory of Discrimination Law*, Oxford University Press, 2016, pp. 83~86.



장학, 지원 등을 통해 희소한 자원에 접근할 가능성을 높일 수 있도록 수혜자의 행동에 영향을 미치는 조치 역시 촉진적 조치라고 할 수 있다. 희소한 자원의 관리자라고 할 수 있는 고용주, 대학, 임대업자 등을 규제하는 조치는 보통 분배적이다. 이러한 관리자는 공적 주체일 수도 있고 사적 주체일 수도 있다. 분배적 조치에는 수혜자에게 우선권을 부여하거나 할당제를 적용하고, 더 많은 자격을 갖추 수 있도록 기준을 다시 정의하는 것 등을 포함한다. 이때 수혜자에게 불리했던 불균형을 깨뜨리는 것은 약한 의미를 갖지만, 별도의 할당을 마련하는 것은 강한 의미를 갖는다.

두 번째 유형은 보호집단을 구별하는 특성을 기준으로 조치를 취하느냐에 따라 직접적 조치와 간접적 조치로 나눌 수 있다. 예를 들어 인종처럼 수혜자 집단이 갖는 특성을 직접인 이유로 삼아 조치를 취하는 것은 직접적 조치이다. 간접적 조치는 수혜자 집단이 보호집단으로서 구별되는 민감한 특성을 조치의 근거로 직접 사용하지 않으면서 수혜자에게 불비례적인 이익을 얻을 수 있도록 한다. 이러한 구분은 적극적 조치로 인한 역차별 문제에서 직접적 조치보다는 간접적 조치가 역차별을 정당화하는 데 보다 용이할 수 있다는 전망에 기초한 것이다. 소수 인종이라는 특성을 입학 기준으로 정하여 직접적으로 인종의 다양성을 조작하는 것보다,³²⁾ 성적의 상한을 입학 기준으로 정함으로써 간접적으로 인종의 다양성을 확보하는 것이 위헌적 상황을 비켜갈 수 있다는 것이다.

세 번째 유형은 희소한 자원의 관리자가 시행하는 조치가 법적 허용 상태에서 취해지는 것인지 법률상 또는 계약상 의무에 따른 것인지 여부에 따라 자발적 조치와 의무적 조치로 구분할 수 있다. 대부분의 헌법 구조에서 적극적 조치는 일정한 조건에서 허용하는 경향이 있다. 이러한 허용은 적극적 조치에 대해서 그로 인한 반사적 차별에 대한 책임을 면제해 주는 방식으로 나타난다. 자발적 적극적 조치를 허용하는 것은 적극적 조치의 종류와 형식을 특정하지 않고 맥락 의존적 설계를 가능하게 함으로써 입법자나 계약 당사자의 재량을 남겨 놓는 데에 기여한다.

머신러닝 알고리즘에 의해 차별이 발생하는 경우 이에 대해 적극적 조치를 취할 경우 직접적인 대상은 차별의 발생 원인이 되는 머신러닝 알고리즘 시스템이 될 것이다. 그리고 차별적 결정을 반복하여 확대시키지 않는 데에 우선적인 목적이 있다면 그 방식은 시스템을 교정하는 방식이 될 것이다. 그러나 이미 머신러닝 알고리즘에 의한 시스템이 작동하여 사회적으로 의미 있는 결과들을 산출하고

32) 인종적 소수자를 위한 직접적인 적극적 조치에 대해 위헌이라고 판단한 예로 *Hopwood v. Texas* 78 F 3d 932 (1996) 참조.



그것이 차별적 효과를 양산한다는 평가가 내려진 후에 이루어지는 조치는 사실상 무의미할 수 있다. 특히 머신러닝 알고리즘이 인간이 이해할 수 없는 방식으로 알고리즘을 생성한 경우 그 원인을 찾아 시정하는 것이 불가능할 수도 있기 때문이다. 따라서 머신러닝 알고리즘의 차별적 결정에 대한 적극적 조치는 사후적인 것보다는 사전적인 것에 좀 더 비중을 두게 된다. 예를 들어 사전 예방적 접근 방식의 적극적 조치는 알고리즘의 설계 단계에서 차별효과에 대한 영향 평가를 하는 방식으로 이루어질 수 있다.³³⁾

33) 공공기관이 자체적으로 자동화된 결정 시스템에 대해 수행하는 영향 평가에 관해서 Dillon Reisman · Jason Schultz · Kate Crawford · Meredith Whittaker, “Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability”, AI NOW, 2018, <https://ainowinstitute.org/aiareport2018.pdf> 참조; 알고리즘에 대한 영향 평가의 근거를 권리로 구성할 경우 논의될 수 있는 설명청구권에 관해서는 본 논문 「제5장 제4절 IV. 머신러닝 알고리즘에 대한 설명의무와 설명청구권」 참조.



제2절 머신러닝 알고리즘의 차별에 대한 책임의 분산

머신러닝 알고리즘의 차별을 사회의 위험으로 인식할 경우 그 위험은 사회에서 분배되어야 한다. 차별에 관한 위험 분배의 책임은 헌법의 지시에 따라 일차적으로 국가에게 있다. 그리고 그 구체적인 내용은 차별의 발생 원인이 국가에게 있는 경우에만 국가가 책임을 지는 방식, 사인에게 차별 발생의 원인이 있는 경우에는 사인에게 책임을 부담지우는 방식, 사인의 차별에 대해서도 국가가 책임을 지는 방식 등 다양한 구성이 가능해 보인다. 그러나 차별 받지 않을 권리가 헌법상 권리로서 보장된다는 점은 책임의 대상을 차별의 원인이 국가에게 있는 경우로 한정하지 않고 사인에게로 확장하는 타당한 근거가 된다. 또한 그에 대한 책임을 국가가 모두 부담할 것인지 사인에게도 부담시킬 것인지는 입법의 영역에 개방되어 있다. 머신러닝 알고리즘의 규제적 속성을 고려할 때 ‘규제에 대한 규제’라는 견제의 방식과 ‘권한과 책임의 분산’이라는 균형의 방식은 머신러닝 알고리즘의 차별과 관련된 차별적 권력의 견제와 차별에 대한 책임의 분배적 균형을 모색하는 데에도 적용될 수 있을 것이다. 국가나 사인도 머신러닝 알고리즘이 설계되고 이용되는 과정의 일정한 단계에서 관련을 맺는 주체가 될 수 있다.

I. 알고리즘의 설계자, 감독자, 심사자

1. 알고리즘의 작성과 권력분립의 유추

차별금지법이 그 적용 범위를 제한적으로 설정한다고 하더라도 알고리즘 기술이 관련 업무의 결정을 지원하거나 대체하는 것에 대해서 별도로 제한을 하지는 않는다. 그런데 차별적 결정에 알고리즘 기술이 사용될 경우 그에 대한 책임 관계를 구성하는 후보군은 늘어난다. 이런 후보군을 설정하기 위해 법률의 제정 절차에서 활용되는 권력분립 원칙을 유추하기도 한다. 알고리즘은 절차적 성격을 가진다는 점에서, 그리고 머신러닝 알고리즘의 경우 데이터셋을 학습하여 알고리즘을 생성하는 중요한 목표가 일반화(*generalization*)³⁴⁾에 정향되어 있다는 점에서 알고리즘은

34) 새로운 데이터에 대해서 실행하는 분류나 예측이 부정확하거나 편향적인 경우 그 알고리즘을 일반적으로 사용하기는 어렵다. 이에 관해서 Pedro Domingos, “A Few Useful Things to Know About Machine Learning”, *Communications of the ACM* 55(10), 2012, pp. 78~87 및 본 논문 「제2장



법과 유사한 속성을 갖는다.³⁵⁾

절차적 성격을 가진다는 점에서 알고리즘과 유사한 속성을 공유하는 법률의 제정 절차에서 활용되는 권력분립 원칙을 유추하기도 한다. 이러한 접근방식의 미국식 모델은 적법절차(due process)로부터 출발한다.³⁶⁾ 법률 코드의 작성자를 법률 코드의 이용자로부터 분리시키려는 것이 적법절차의 핵심 기능이라는 것이다.³⁷⁾ 이때 적법절차는 일반적인 법률을 제정하는 역할을 구체적인 상황에 그 법률을 집행하는 역할 및 그 집행이 효력 있거나 가치 있는 것인지 판단하는 역할로부터 분리하는 것을 보장한다.

이러한 접근방식은 적법절차에서 출발하지만 그 발상의 원천은 권력분립 사상에 자리 잡고 있다. 권력분립의 관념은 모든 범위의 국가 구조에 적용되지만 용어는 균형, 통제, 분할, 분리 등 일관되지 않은 용어에서 증명되듯이 완전히 다른 조직 원리와 관련된 개념이다. 이러한 권력분립의 의미는 세 가지로 나눌 수 있는데, 첫째 정치 조직체의 다른 부분들을 조직적으로 분할할 것을 요구한다는 의미, 둘째 모든 기관이나 부처에 대한 교차적인 견제와 균형의 일반적 규칙을 밝힌다는 의미, 셋째 권력에 특별한 업무나 기능을 할당하고 다른 권력들을 통해 이 기능을 행사하는 것을 방지한다는 의미이다.³⁸⁾ 이는 권력분립의 관념이 조직 분할, 견제와 균형, 기능 분할의 원리를 모두 담고 있다고도 볼 수 있지만 권력분립의 관념이 각각의 원리로 분화된 것이라고 볼 수도 있다.

2. 알고리즘 설계의 중요성

권력분립에 관한 논의를 머신러닝 알고리즘의 결정에 대한 책임을 묻는 맥락으로 끌어들이는 주된 이유는 알고리즘 역시 누군가 즉, 어떤 에이전트에 의해서 형성되고 실행된다는 점에 있다. 형성 및 실행의 과정 또는 절차의 측면에서 접근하면 알고

제2절 II. 2. 머신러닝 알고리즘의 학습과 데이터」 참조.

35) 법의 절차화에 관해서 Saliger, Frank, “Prozeduralisierung im (Straf-)Recht”, in: *Einführung in Rechtsphilosophie und Rechtstheorie der Gegenwart*, Winfried Hassemer · Ulfrid Neumann · Frank Saliger(Hrsg.), 9. Aufl., C. F. Müller, 2016[1. 1976], S. 434-452 및 Tatjana Sheplyakova(Hrsg.), *Prozeduralisierung des Rechts*, Mohr Siebeck, 2018 참조.

36) Danielle Keats Citron, “Technological Due Process”, *Washington University Law Review* 85, 2007, pp. 1249-1313 참조.

37) Nathan S. Chapman · Michael W. McConnell, “Due Process as Separation of Powers”, *Yale Law Journal* 121(7), 2012, pp. 1672-1807.

38) Christoph Möllers, *The Three Branches: A Comparative Model of Separation of Powers*, Oxford University Press, 2013, pp. 43-49 참조.



리즘을 설계하는 단계에서부터 알고리즘을 이용하여 데이터를 분석하고 결정에 사용하는 모든 단계에서 감독과 심사가 개입될 수 있는 구조로 재구성할 수 있다.³⁹⁾

앞에서 살펴봤듯이 머신러닝 알고리즘 기반의 시스템이 일단 이용되기 시작하면 그 이후에 이를 시정하는 방법은 법적으로 제약될 수도 있고, 기술적으로도 제약될 수 있다.⁴⁰⁾ 만약 사후적으로 머신러닝 알고리즘의 결정에 대해 감독이나 심사를 실시하여 교정적 조치를 부과하려고 하는 경우에도 우선 그 내부를 들여다 볼 수 있어야 하고, 내부를 들여다 볼 수 있는 경우라 하더라도 어떤 과정으로 알고리즘이 실행됐는지 기록으로 남아있어야 한다.⁴¹⁾ 그렇지 않다면 사후적인 감독과 심사는 무용지물이 된다. 설계에 의한 책임(responsibility by design)이 강조되는 이유이다. 따라서 사후 감독과 심사의 전제 조건은 알고리즘의 생성 과정과 실행 과정이 기록으로 남아 있을 수 있는 설계 단계에서의 기술적 조치가 이루어지도록 하는 것이다. 예컨대, 머신러닝 알고리즘을 이용해 자동 번역을 수행할 때에 성 편향을 고려하는 기능을 지원하도록 설계하는 것만으로도 번역 알고리즘을 통해 산출되는 결과가 성차별적 경향을 덜 확률을 대폭 낮출 수 있다는 점은 이미 살펴보았다.⁴²⁾

3. 시스템의 차별을 규제할 견제와 균형 시스템

어쨌든 이러한 권력분립의 교훈을 알고리즘의 결정 절차에 적용하면 알고리즘의 기능은 알고리즘을 설계하는 역할, 알고리즘에 입력하는 질문을 감독하는 역할, 알고리즘의 출력을 심사하는 역할의 상호작용으로 볼 수 있다. 이러한 상호작용에 대한 규제가 차별금지법에 없다는 것은 시스템 속에서 편향이 자리 잡지 못하도록 하는 견제와 균형의 시스템이 부재하다는 것이기도 하다.⁴³⁾

39) 특정 가치 실현에 관한 평가프로그램을 개발하기 위한 컴퓨터 과학 분야의 연구 예로 조병훈·박성빈, “컴퓨터 과학 교육: 웹 접근성 분석”, 한국컴퓨터교육학회 학술발표대회논문집 13(2), 2009, 243-246쪽 참조.

40) 법적인 제약에 관해서 본 논문 「제4장 제4절 II. 투명성과 불투명성 사이의 헌법적 긴장 관계」 참조; 기술적 제약에 관해서 본 논문 「제4장 제4절 I. 3. 복잡한 머신러닝 알고리즘의 작동방식과 불투명성」 참조.

41) 이에 관해서 Frank A. Pasquale, “Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society”, *Ohio State Law Journal* 78(5), 2017, pp. 1243~1255 참조.

42) 이에 관해서 본 논문 「제2장 제3절 IV. 1. 번역 실험」 참조.

43) Kate Crawford·Jason Schultz, “Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms”, *Boston College Law Review* 55, 2014, pp. 93~128: p. 120.



머신러닝 알고리즘의 차별을 예방하고 관리, 감독하는 기관은 차별에 중점을 둘 것인지 아니면 머신러닝 알고리즘에 방점을 찍을 것인지에 따라 그 구성이 달라질 수 있다. 우선적으로 고려해 볼 수 있는 것은 기존의 차별시정기구에 별도의 전문 부서를 설치하는 것이다. 국가인권위원회에 머신러닝 알고리즘에 의해 작동하는 시스템을 비롯해 인간의 개입 없이 이루어지는 자동화된 결정 시스템이 야기하는 차별에 관한 문제를 다루는 전문위원회를 구성하는 것이다.

다른 한편으로 알고리즘을 활용한 기술 또는 알고리즘 자체에 대한 규제적 접근을 모색할 때 구체적인 작동 또는 작용에 대해 이해하기 어렵거나 복잡하다는 알고리즘의 특성은 그와 유사한 특성을 가진 기존의 규제 대상에 대한 축적된 규제 방식의 탐색으로부터 출발할 수도 있다. 예를 들어 머신러닝 알고리즘이 사회에 반드시 필요하지만 부정확성이나 편향성으로 인해 사회에 어떤 해로운 영향을 미치게 될 것인지 알기 어렵다는 측면이 식품이나 의약품에 비교될 수 있다는 점을 근거로 식품의약품안전처 같은 기관을 별도로 구성하는 것이다. 머신러닝 알고리즘에 의한 결정 시스템이 사회에 미치는 여러 가지 영향중에 하나로 머신러닝 알고리즘의 차별 문제를 다루게 하는 것이다.⁴⁴⁾

어떤 식의 구성이 됐든 머신러닝 알고리즘에 의한 결정 시스템을 과학기술적으로 이해할 수 있는 전문가와 그 의미를 법적·정치적·사회적·문화적으로 이해할 수 있는 전문가 그리고 무엇보다 관련 시스템이 일상생활의 환경을 구축한다는 점을 고려해 비전문가의 고른 참여는 필수적이라고 할 수 있다.⁴⁵⁾

II. 알고리즘 시스템 운영자

1. 데이터 프로세싱과 차별

정보의 처리 과정을 수집과 이용으로 엄밀하게 나눈다면 서술적 의미의 차별은 수집된 정보를 구별하고 분리하고 분류한다는 측면에서 정보의 이용 과정과 우선적으로 관련을 맺는다.⁴⁶⁾ 그러나 이러한 주장이 온전히 타당하려면 수집된 정보가

44) 복잡한 알고리즘을 식품 또는 의약품에 비유하여 미국의 식품의약품(FDA) 같은 정부기관의 설계가 필요하다는 구상은 Andrew Tutt, “An FDA for Algorithms”, *Administrative Law Review* 69(1), 2017, pp. 83~123 참조.

45) 비전문가인 일반 대중의 참여는 투명성과도 밀접한 관련이 있다. 이에 관해서 본 논문 「제4장 제4절 III. 2. 투명성의 확장 범위와 크라우드소싱」 참조.

46) Giovanni Sartor · Andrea Omicini, “The Autonomy of Technological Systems and Responsibilities for their Use”, in *Autonomous Weapons Systems: Law, Ethics, Policy*, Nehal Bhuta · Susanne Beck · Robin



‘모든 것’에 대한 정보라는 점이 전제되어야 한다. 따라서 현실 인식이 아직은 모든 것에 대한 정보체계가 구축되지 않았다는 점에 가닿아 있다면 그 주장은 차별이 정보의 수집과도 관련이 있다는 내용으로 수정 보충되어야 한다. 정보를 수집하기 위해서 어떤 정보를 수집할 것인지 구별하고 분류하는 과정이 필요하기 때문이다. 예를 들어 방법용 감시카메라로 수집한 영상정보를 이용할 때 그 정보를 구별하고 분류하는 것이 필수적이다. 하지만 그보다 앞서 어떤 지역에 방법용 카메라를 설치하여 정보를 수집할 것인지 결정하는 과정에도 정보 수집 장소를 구별하고 분류하는 것 역시 필요하며 이러한 결정 절차는 차별과 관련을 맺는다.

현실 인식이 모든 것에 대한 정보체계를 구축해 가고 있는 경향에 맞닿아 있다면 그 주장은 아직 유효하다. 특히 아무리 정보에 관한 법체계가 정보 수집에 일정한 제한을 가한다고 하더라도 이를 얼마든지 우회할 수 있는 기술적 수단을 강구하여 정보를 ‘무(無)’차별적으로 수집하는 상황이라면 적어도 정보 수집에 관한 한 차별은 ‘없는’ 것이기 때문이다. 실제로 사물 인터넷(internet of things) 즉, 사물에 감지기(sensor)를 부착해 전 방위에서 사물끼리 커뮤니케이션이 가능해지는 인터넷 체계가 구축되고 있으며⁴⁷⁾ 이를 기반으로 온라인상의 생활양식과 오프라인상의 생활양식을 구분하는 경계가 불명확해지고 더 이상 그 경계를 찾아 생활양식을 구분하는 것이 무의미해지는 이른바 “온라이프(onlife)”⁴⁸⁾ 양식에서 이루어지는 모든 활동은 수백 개 민간 기업의 데이터베이스에 로그 기록으로 남고 있다는 사실을 부인하기는 어렵다.

그럼에도 불구하고 정보 수집이 여전히 차별과 밀접한 관련을 맺는다는 주장은 가능하다. 온전한 정보의 체계가 구축되고 있는 상황을 염두에 둔 경우라도 정보 수집은 차원을 달리하는 형태로 차별과 관련을 맺을 수 있다. 수집된 정보 중에 필요한 정보를 또 다시 추출하여 조합하는 가공 과정에서 메타 차원의 정보 수집 양식이 가능하기 때문이다. 메타 정보 분석을 위해 정보를 수집할 때에도 이를 구별하고 분류하는 과정은 필수적이다. 예를 들어 일반 이용자가 검색 엔진에 질문을 입력하면 검색 알고리즘은 입력된 질문과 관련된 정보, 즉 연결된 웹페이지를 수집하여 그 중에 최적화된 정보를 일정 기준에 따라 정렬하여 답변 내용으로 출력한다. 이 과정에서 구별과 분류는 무수히 이루어진다. 다만 입력한 질문에 대한 답변의 출력 시간이 극히 짧아 수없이 많은 구별과 분류가 진행됐다는 사실을 쉽게 인지하지 못할 뿐이다.

Geiss · Claus Kress · Hin Yan Liu(Eds.), Cambridge University Press, 2016, pp. 39~74 참조.

47) 2000년대 초반 15년은 개인컴퓨터(PC) 사이의 인터넷 기술이 주도했다면, 이후 15년은 사물 인터넷(IoT) 기술이 주도할 것이라고 전망되기도 한다.

48) ‘온라이프’ 개념에 관한 설명은 본 논문 「제4장 제1절 II. 2. 온라이프 생활양식의 확장」 참조.



2. 정보처리자로서 알고리즘 시스템 운영자

알고리즘 시스템 운영자는 알고리즘을 사용해 정보를 처리한다. 처리되는 정보 중에는 의뢰인이나, 고객, 일반 이용자에 관한 정보도 있다. 그러므로 알고리즘을 사용해 개인 또는 집단에 관한 정보를 처리한다는 점에서 알고리즘 시스템 운영자는 이들의 정보를 처리하는 수탁자 또는 사무 관리자의 지위를 갖는 것으로 보아 책임의 주체로 구성될 수 있다. 수탁자와 사무 관리자의 구분은 정보의 주체로부터 정보 처리에 관한 사무를 위탁 받았는지 여부에 따른 것이지만 넓게 보면 다른 사람의 업무를 대신 처리하여 그 타인, 즉 본인에게 사무를 처리한 결과가 귀속되도록 한다는 의미에서 ‘대리인’으로 표현할 수도 있다.

3. 개인정보의 수집과 정보 수탁자

볼킨(J. Balkin)은 알고리즘이 처리하는 대상으로서 ‘개인정보’에 방점을 찍어 “정보 수탁자(information fiduciaries)”⁴⁹⁾ 개념을 구상하기도 한다. 수탁자는 보통 변호사나 의사, 부동산 관리인처럼 타인의 사무를 위탁받아 처리하면서 신탁자와 신뢰관계를 형성하는 데, 수탁자의 지위를 만드는 것은 타인, 즉 고객이나 의뢰인 등이 서비스의 공급을 그들에게 의지한다는 점 때문이다. 그러나 수탁자와 고객 사이에는 지식과 능력 면에서 중대한 비대칭의 간극이 있고, 이러한 비대칭성은 고객이 수탁자가 자신을 대신해 무엇을 하고 있는지 쉽게 살펴볼 수 없는 구조 속에 숨겨져 있다. 이와 같은 사정을 고려해 법으로 수탁자에게 고객이 신뢰할 수 있는 태도로 성실하게 고객에 대한 서비스를 제공하도록 요구할 수 있고, 수탁자의 이해관계가 고객과 충돌하지 않도록 요구할 수도 있다. 게다가 수탁자가 종종 그들의 고객에게 해를 입힐 수 있는 민감한 개인정보를 수집하게 된다는 점은 고객과 수탁자 사이의 비대칭적 구조 속에서 개인정보에 대한 보호와 통제의 필요성을 한층 강화한다. 따라서 법은 그들 고객의 개인정보를 보호할 것과 그들의 고객에게 해를 입히는 방식으로 정보를 공개하지 못하도록 요구할 수 있는 것이다. 이때 자신의 고객에 관한 정보를 수집하고 처리하는 수탁자의 지위를 ‘정보 수탁자’로 부르는 것이다.

49) Jack M. Balkin, “Information Fiduciaries and the First Amendment”, *U. C. Davis Law Review* 49(4), 2016, 1183~1234쪽 참조.



개인의 인격은 데이터베이스 안에서 데이터로 조직된 “디지털 인격(digital person)”⁵⁰⁾으로 재구성된다. 수량화할 수 있는 정보는 정확한 만큼 특정한 의사 결정에 기여할 수 있다. 그런데 데이터베이스에 있는 정보는 종종 우리 삶의 맥락이나 의미를 포착하지 못하며, 따라서 데이터베이스에 있는 어떤 정보를 사용하는 것이 차별의 맥락과 연결되는지 섬세하게 포착하지 못할 수도 있다. 정보 수탁자 개념을 사용하여 알고리즘 시스템 운영자에게 정보 수탁자의 지위가 부여되면 알고리즘을 사용해 그들 고객의 정보를 수집하고 처리할 때 차별이 발생할 경우 알고리즘 시스템 운영자를 차별에 대한 책임자로 볼 수 있다. 이때 개인정보를 처리하는 알고리즘 시스템 운영자는 개인정보 보호에 대한 책임자와 차별에 대한 책임자라는 중첩된 지위를 갖게 된다. 특히 개인정보를 데이터베이스에 불리하게 포함시키는 경우 각각 그 적용 대상과 범위를 달리하는 개인정보보호법과 차별금지법 사이에 교두보를 놓아 두 법제가 교차하는 계기가 된다.⁵¹⁾

4. 개인정보의 추론과 사무 관리자의 책임

그런데 알고리즘 시스템 운영자는 고객이나 의뢰인 또는 일반 이용자의 부탁을 받지 않은 경우에도 일반 공중을 대상으로 정보를 처리할 수도 있다. 더구나 일반 공중을 대상으로 특수한 개인정보를 추론해 낼 수 있다는 점은 정보 처리자로서 알고리즘 시스템 운영자를 사무 관리자로서 책임의 주체로 볼 수 있게 한다. 인간이 이해할 수 있는 정형의 정보를 넘어서는 비정형 정보를 포함하는 ‘빅데이터(big data)’ 환경에서 개인의 프로파일(profile)은 머신러닝 알고리즘 같은 시스템에 의한 추론에 의해 구성된다. 이는 마치 디지털 인격이 개인을 둘러싼 사물의 환경적 조건으로부터 ‘추론된 인격(inferred person)’으로 진화한 것처럼 보인다. 머신러닝 알고리즘을 통해 개인의 프로파일을 작성하려면 개인에 관한 수많은 정보들을 연결하고 조합하고 분류함으로써 여러 가지 특성을 추출할 수 있는 패턴을 발견하고 그에 부합하는 인격을 추론해야 한다. 특히 비정형 데이터를 분석할 경우 인간이 이해할 수 있는 정형의 데이

50) Daniel J. Solove, *The Digital Person: Technology and Privacy in the Information Age*, New York University Press, 2004, p. 49.

51) Raphaël Gellert · Katja de Vries · Paul de Hert · Serge Gutwirth, “A Comparative Analysis of Anti-Discrimination and Data Protection Legislations”, in *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, Bart Custers · Toon Calders · Bart Schermer · Tal Zarsky(Eds.), Springer, 2013, pp. 61~88.



터로 포착하려면 지도학습방식에 따라 인간이 이해할 수 있는 특성을 레이블로 정의하여 데이터를 분석하거나, 비지도학습방식에 따라 알고리즘이 데이터 세트에서 추출한 지배적인 특성에 의미를 부여해야 한다.⁵²⁾

이때 개인의 특성에 관한 일종의 판별식을 찾는 과정에서 경유하게 되는 집단화 또는 군집화의 과정은 인간에 대한 범주화(categorization) 기능을 수행한다. 이러한 범주화를 통해 추론된 인격의 프로파일에는 집단과 군집을 규정하는 속성이 반영된다. 그래서 머신러닝 알고리즘의 추론을 통한 개인의 구별은 어떤 속성이 개인을 과잉 대표함으로써 일반적인 범주에 가두는 결과를 낳을 수 있다. 예를 들어 머신러닝 알고리즘이 통계적 분석을 통해 범죄율을 인종에 따라 구분해서 특정 인종이 다른 인종보다 범죄율이 높다는 사실이 밝혀지면 범죄율이 높은 인종과 같은 특징을 가진 개인은 집단적으로 통계화한 확률만큼 범죄를 저지를 가능성이 높은 사람이 된다. 인종을 기준으로 구분하지 않았다면 평균 수준의 범죄 가능성의 의심을 받을 사람이 특정 인종을 기준으로 한 분류에 의해 높은 범죄 가능성을 의심받게 된다. 개인의 유일성과 정체성을 결정하는 특징을 사회적 구별의 기준으로 삼을 경우 개인을 집단의 영역으로 소환하여 그 유일성과 정체성을 훼손하는 방식으로 차별을 가하는 구조를 형성하게 된다.

5. 알고리즘 시스템 운영자의 사용자 책임

사용자 책임이론은 국가가 공무원의 국가사무에 관한 행위로 발생한 손해나 손실에 대해 배상 또는 보상하거나 고용주가 피고용인의 업무상 행위로 발생한 손해에 대해 배상할 수 있는 근거가 된다. 그런데 이런 사용자 책임이론은 피사용자의 인격을 전제한다. 불किन은 알고리즘 안에는 인격이 없기 때문에 사용자 책임 이론에 따라 알고리즘 운영자의 의무와 책임을 모델화하는 것은 무용한 것으로 본다. 알고리즘으로부터 고의나 과실, 범의를 찾아 사용자에게 대신 책임을 물을 수가 없다는 것이다.⁵³⁾ 이러한 결론은 피사용자로서 알고리즘 에이전트에게 법적 인격을 인정할 수 없다는 전제를 불변의 것으로 상정하고 있다. 알고리즘 에이전트에게 법적 인격이 인정된다면 고의나 의도가 없는 알고리즘이 발생시킨 차별의 사태를 제어

52) 머신러닝 알고리즘의 학습 방식에 대해서 본 논문 「제2장 제2절 III. 머신러닝의 지도학습과 비지도학습」 참조.

53) Jack M. Balkin, "The Three Laws of Robotics in the Age of Big Data", *Ohio State Law Journal* 78(5), 2017, pp. 1217~1241: p. 1234.



하지 못한 책임을 사용자에게 지을 수도 있다. 또한 알고리즘 에이전트에게 법적 인격이 인정될 경우 알고리즘 에이전트 자체도 법적 책임의 주체가 될 수 있으므로 별도로 논의할 필요가 있다.

III. 알고리즘 에이전트에 대한 책임 귀속 문제

1. 양 극단의 제한주의와 허용주의

차별적 결정에 대한 책임의 주체로서 마지막 후보자는 알고리즘 에이전트 자체이다. 차별적 결정을 도출한 알고리즘의 사용자에게 법적 책임을 묻기 위해 사용자 책임이론을 구성할 때 부딪히는 난관은 피사용자가 인격을 갖지 않는다는 전제에서 비롯된다. 그런데 피사용자로서 알고리즘 에이전트가 인격을 갖는다고 본다면 달라진 전제에서 사용자 책임은 이론적으로 성립될 수 있다. 또한 이와 같은 전제의 설정은 알고리즘 에이전트 자체를 책임의 주체로 보는 것도 가능하게 한다. 과연 인공지능 에이전트, 자율적 에이전트 등 사회의 행위자로서 기능하는 인공물에 대해 인격, 특히 법적 인격을 부여할 것인지 여부는 머신러닝 알고리즘이 관여한 차별에 대한 법적 판단뿐만 아니라 인간이 아닌 사회적 행위자에 대해 구축해 온 전통적인 법 이론적 구성과 실무적 판단에도 커다란 변화를 가져올 수 있는 근본적인 문제이기도 하다.

기본적으로 알고리즘 에이전트 자체에 책임을 귀속시킬 수 있을 것인지 여부에 대한 양 극단에는 제한주의와 허용주의 입장이 자리 잡고 있다.⁵⁴⁾ 제한주의는 알고리즘 에이전트 같은 인공물에게 법적 책임을 부여할 수 없다는 것이고, 허용주의는 입법자가 유연하게 법적 책임을 부여할 수 있다는 것이다. 제한주의 논변은 보통 얼마나 지능적인지 또는 자율적인지 상관없이 기계는 절대로 법적 인격을 가질 수 없고, 그들의 행위에 대해 법적으로 책임이 없다는 주장을 입증하기 위해 법적 책임의 전제조건을 구성하는 것처럼 보이는 의도성, 자유의지, 자율성 또는 의식 같은 인간의 속성을 제시하면서⁵⁵⁾ 이러한 속성이 기계에게는 부족하다는 점을 강조한다. 반면에 허용주의는 법을 사회공학의 탄력적 도구로 보면서 누구라도 또는 무엇이랄도 법적

54) Bartosz Brożek · Marek Jakubiec, “On the Legal Responsibility of Autonomous Machines”, *Artificial Intelligence and Law* 25(3), 2017, pp. 293~304 참조.

55) 도덕적 책임의 기초를 자유 관련(freedom-relevant) 또는 통제(control)에 두는 예로 John Martin Fischer · Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge University Press, 2000, pp. 1~27 참조.



인격을 갖도록 할 수 있다는 입장에서 출발한다. 따라서 자율적 인공 에이전트에게 그들의 행위에 대한 책임을 귀속시키는 것은 아무런 문제가 되지 않는다. 심지어 돌이나 숲, 강처럼 그 어떤 대상에도 법적 인격이 부여될 수 있다.⁵⁶⁾

자율적 에이전트의 법적 책임 문제에서 허용주의는 법체계에서 국가, 기업, 재단 같은 법인의 책임이 인정되는 점을 강조하며 자율 기계에도 유추할 수 있다고 주장하는 반면, 제한주의는 국가, 기업, 재단 같은 법인의 법적 책임과 자율 기계의 법적 책임은 완전히 다른 문제라고 주장한다. 허용주의는 법인으로 인정되는 개체들이 의도성이나 자유의지를 가지고 있다고 보기 어렵다는 점을 들어 그러한 요인이 법적 책임을 귀속시키기 위한 필수 요건이 아니라면 자율적 에이전트에 대해서도 그런 요인을 법적 책임의 요건으로 요구할 수 없다고 한다. 그러나 제한주의는 법인의 행위가 구체적으로는 결국 신체를 가진 인간에 의해 수행되는 것으로서 대표자나 피고용인과 연결되어 있다는 점을 강조한다. 그렇기 때문에 인간과 연결되어 있지 않고 인간의 개입 없이 행동하는 자율적 에이전트를 인간과 연결되어 있는 다른 법인과 유사하게 취급할 수 없다고 한다.⁵⁷⁾

2. 원칙적 허용과 공리주의적 제한

하그(J. Hage)는 자율적 에이전트의 행위에 관해서 에이전트의 제조자나 이용자 또는 소유자 외에 자율적 에이전트 자체도 책임이 있다고 주장한다.⁵⁸⁾ 이 주장은 심리학적 고려로서 의도성이 있거나 자유의지가 있다는 것이 법적 개념으로 사용되는 것에 대한 문제제기와 맞닿아 있다. 법적 책임에 의도성과 자유의지가 필요하다는 점은 법적 행위에 관한 일상 경험의 초석이 되지만 이를 마치 실재하는 그 무엇처럼 취급하는 경향은 현실주의자의 실수라고 본다. 그래서 만일 인간에게만 책임이 있고 기계에게는 책임이 없다는 주장의 근거가 의도성과 자유의지뿐이라면 의도성과 자유의지가 과연 존재하는 것인지에 대한 현대 과학의 의문에 따라 의도성과 자유의지가 근거로서 작용하는 논변들의 지지기반은 흔들리게 된다는 것이다. 허용

56) 실제로 2017년 3월 20일 뉴질랜드는 오랜 분쟁 끝에 ‘황가누이 강(Whanganui River)’에게 법인격을 부여하는 법률을 제정하였다. 그 구체적 내용은 다음과 같다. Te Awa Tupua (Whanganui River Claims Settlement) Act 2017, Sec. 14 (1): “Te Awa Tupua is a legal person and has all the rights, powers, duties, and liabilities of a legal person.”

57) Bartosz Brożek · Marek Jakubiec, “On the Legal Responsibility of Autonomous Machines”, *Artificial Intelligence and Law* 25(3), 2017, pp. 293~304: pp. 300~301.

58) Jaap Hage, “Theoretical Foundations for the Responsibility of Autonomous Agents”, *Artificial Intelligence and Law* 25(3), 2017, pp. 255~271 참조.



주의의 논변에 따르면 적어도 법적 책임에 관한 한 인간과 자율 시스템 사이의 차이 자체만으로 다른 취급을 정당화할 수는 없을 것이다. 다만, 개념적인 한계가 없다는 것과는 별개로 인공 에이전트에게 법적 책임을 귀속시키는 것은 그 결과가 바람직할 경우에만 정당화될 수 있기 때문에 법적 책임의 개념이 무차별적으로 남용되는 것은 제한된다고 한다.

3. 민속 심리학에 기초한 법 개념과 허용의 한계

브로젝(B. Brożek)은 위 논변을 기계가 행위에 대해 원칙적으로 법적인 책임을 질 수 있는지의 문제에 대한 허용주의를 외부에 있는 공리주의적 기준으로 완화시키는 것으로 보고, 상황이 보다 복잡하다는 것을 인간의 행동에 대한 서술적이고 개념적인 분석을 통해 보여주고자 한다.⁵⁹⁾ 이때 인간의 행동을 분석하는 개념적 도식은 크게 민속 심리학적(folk psychological)이거나 과학적(scientific)이다. 민속 심리학은 보통 마음을 읽는 능력으로 이해되고, 인간이 다른 사람의 행동을 설명하고, 그들의 행동을 예측 또는 기대하며, 인간의 행동에 속하는 일반화를 생성하기 위해 자신의 행동과 타인의 행동을 서술할 수 있는 근본적인 능력들을 말한다.⁶⁰⁾ 이러한 민속 심리학적 도식은 대개 인간의 행동을 문화 의존적이고 직관적이며 무의식적인 학습에서 비롯된 것으로 본다. 반면 과학적 도식은 인간의 행동을 민속 심리학의 현상적 수준보다 깊이 있는 구조적 수준에서 설명한다. 이때 브로젝은 법의 개념적 도식은 주로 민속 심리학에 기초한다는 점을 강조함으로써 자율적 기계가 법적 책임을 질 수 있는지 여부는 민속 심리학의 관문을 통과하느냐에 달려 있는 것으로 본다. 그러면서 지능적이고 인간과 닮은 모습을 하고 필요할 때 도움을 주기도 하고 새로운 친구와 대화하며 함께 시간을 보낼 준비가 되어 있는 A가 복잡한 안드로이드 즉, 인간의 모습을 한 로봇으로 밝혀졌을 때 우리는 아마도 A에 대한 태도를 바꿀 것이라고 진단한다. 물론 자율적 로봇이 충분히 도덕적이고 법적인 에이전트가 됐을 때 민속 심리학적 개념 도식이 진화하는 것이 불가능한 것은 아니라는 여지를 남겨두기는 하지만 적어도 아직은 그러한 지점에서 멀리 떨어져 있기 때문에 실천적 영역에서 자율적 에이전트의 법적 책임을 인정하기는 어렵다고 주장한다.

59) Bartosz Brożek · Marek Jakubiec, “On the Legal Responsibility of Autonomous Machines”, *Artificial Intelligence and Law* 25(3), 2017, pp. 293~304: pp. 296~300.

60) Stephen Stich · Ian Ravenscroft, “What Is Folk Psychology?”, *Cognition* 50(1), 1994, pp. 447~468.



제3절 머신러닝 알고리즘의 차별로부터 보호와 개인정보의 보호

머신러닝 알고리즘이 데이터셋으로부터 훈련되고, 새로운 데이터를 처리한다는 점에서 머신러닝 알고리즘과 데이터는 불가분의 관계에 있다. 머신러닝 알고리즘이 데이터를 처리(*processing*)할 때 발생하는 차별의 문제는 데이터셋으로부터 모델을 생성하여 실제 결정이나 판단에 적용하는 과정을 수집, 분석, 이용으로 단계적으로 구분할 때,⁶¹⁾ 세 가지 측면에서 접근할 수 있다. 먼저 데이터셋 즉, 데이터베이스의 구성 단계에서는 데이터베이스에 어떤 데이터가 포함되는지 아니면 배제되는지에 따라 발생하는 데이터베이스 간 차별 문제이다. 그 다음 분석 단계에서는 데이터베이스 내부에서 데이터 간 관계를 분석하는 과정에서 구별을 실행함으로써 발생하는 데이터 간 차별 문제이다. 마지막으로 데이터베이스 분석을 통해 추론된 규칙 즉, 모델을 실제 결정이나 판단의 기준으로 적용하는 단계에서는 구별 기준으로 사용되는 정보가 차별 사유로 작용할 때 발생하는 데이터 이용에 의한 차별 문제이다.

이러한 데이터 처리 절차에서 활용되는 데이터가 개인의 식별성과 관련을 맺을 때, 즉 개인정보로서의 자격을 갖출 때 개인에 관한 데이터 자체는 법적 보호 대상이 되어 처리를 위해서 일정한 법적 요건을 갖추어야 한다. 이는 헌법 차원의 권리로서 개인정보 자기결정권에 근거를 두고 있다. 그리고 처리되는 개인정보가 헌법상 차별의 근거로 삼아서는 안 되는 사유 또는 차별금지법에서 판단의 근거로 사용하는 것을 제한하는 차별 사유에 해당하는 정보일 경우 개인정보의 보호 문제와 차별 사유로서 개인정보의 특성을 공유하는 집단을 알고리즘의 차별로부터 보호하는 문제가 중첩 또는 교차하게 된다. 따라서 개인정보의 보호와 알고리즘의 차별로부터 보호라는 두 가지 문제를 해결하는 방법은 각각의 법제 구성의 근거와 방식에 따라 달라질 수 있다.

I. 데이터베이스 구성 단계에서 차별과 개인정보보호

1. 후버 사건

오스트리아 국적의 후버(H. Huber)는 1996년에 독일로 건너간다. 유럽연합(EU)의 시민으로서 다른 회원국에서 일하며 사는 데 장애물이 없었지만, 외국인중양 등록기

61) 데이터 마이닝을 통한 예측 분석 프로세스의 단계를 수집, 분석, 이용으로 구분하는 경우로 Tal Z. Zarsky, "Transparent Predictions", *University of Illinois Law Review* 2013(4), 2013, pp. 1503~1570: pp. 1521~1530 참조.



(Ausländerzentralregister, 이하 ‘AZR’)에 관한 법률에 따라 그의 개인정보는 자동화된 데이터베이스에서 처리되어야 했다. 2000년 후버는 외국인등록 데이터베이스에 자신의 데이터가 있는 것은 차별적이라고 이의제기하면서 자신의 데이터를 그 시스템에서 삭제할 것을 요청하며 독일 정부를 상대로 소를 제기한다(이하 ‘후버 사건’). 독일 국민의 데이터는 해당 시스템에 저장되어 있지 않은데다가, 저장된 데이터는 범죄수사나 인구 통계 조사 같은 부차적 목적으로 사용되는 경우도 있기 때문이다. 해당 사건의 재판부는 판결을 위해 예비적으로 유럽사법재판소(ECJ)에 다음의 세 가지 질문에 대해 판단을 구한다. 유럽연합의 외국 시민에 관한 개인정보를 외국인 중앙 등록 시스템에서 일반적으로 처리하는 것이 첫째 유럽연합의 시민에 대해 국적을 이유로 한 차별을 금지하는 법과 양립할 수 있는지 여부, 둘째 유럽연합의 다른 회원국의 영토에서 한 회원국의 국적을 확립할 자유에 대한 제한을 금지하는 법과 양립할 수 있는지 여부, 셋째 필요성 요건을 규정한 데이터 보호 지침(Data Protection Directive)과 양립할 수 있는지 여부이다.⁶²⁾

유럽사법재판소는 이에 대한 판단에서 먼저 데이터 보호 지침의 필요성 요건과 관련해서 외국인 중앙 등록 시스템(AZ) 같은 유럽연합 시민의 개인정보를 처리하는 시스템이 주거권에 관한 법률을 보다 효과적으로 적용하기 위한 경우라고 할지라도 데이터가 입법의 목적에 필수적일 경우에만, 그리고 주거권 관련 입법을 회원국의 국적이 아닌 유럽연합 시민에 관해 효과적으로 적용될 수 있게 하는 한에서만 지침의 필요성 요건을 충족시키는 것이고, 통계적 목적으로 외국인 중앙 등록기 같은 등록 시스템에 개인정보를 담은 데이터를 저장하고 처리하는 것은 필요성 요건에 부합하지 않는 것으로 보았다. 그리고 국적을 이유로 한 차별 금지 조항은 유럽연합의 한 회원국이 다른 국적을 가진 유럽연합의 시민에 대해서만 개인정보를 처리하는 시스템을 범죄 척결의 목적으로 설치하는 것이 불가능하다는 의미로 해석되어야 한다고 보았다.⁶³⁾ 이러한 판단을 토대로 후버 사건에 대한 재판을 진행한 결과, 북 라인-베스트팔렌 주 고등 행정법원은 2009년에 후버의 데이터를 외국인 등록 시스템에 저장한 것은 데이터 보호 지침과 차별 금지 조항의 정당성 요건을 충족하여 적법하다고 결정했다.⁶⁴⁾

62) Heinz Huber v Bundesrepublik Deutschland, C-524/06, Judgment of the Court (Grand Chamber) of 16 December 2008, para. 30-40.

63) Heinz Huber v Bundesrepublik Deutschland, C-524/06, Judgment of the Court (Grand Chamber) of 16 December 2008, para. 82.

64) Oberverwaltungsgericht NRW(Nordrhein-Westfalen), 17 A 805/03, 24. 6. 2009, Rn. 1~41.



2. 분류금지원칙과 수집제한원칙

자동화된 데이터 시스템에서 개인정보가 처리되는 경우가 증가하면서 1970년대에 공정 정보 규정(Fair Information Practices, FIPs)⁶⁵⁾이 처음 제안된 이후 경제협력개발기구(Organization for Economic Cooperation and Development)는 1980년에 향후 국제적 표준으로 자리 잡게 된 개인정보 보호에 관한 가이드라인(guideline)을 통해 8개 원칙을 제시했는데,⁶⁶⁾ OECD의 1980년 가이드라인은 1995년 유럽연합의 데이터 보호 지침에도 영향을 미쳤다. 이러한 지침은 후버 사건에서 법적 판단을 위한 근거가 됐는데, 이 사건이 특별한 이유는 자동화된 처리 장치의 데이터베이스에 개인정보가 담긴 데이터를 포함시키는 것이 단순히 개인정보 보호에 관한 문제만을 야기하는 것이 아니라 차별에 관한 문제도 함께 발생한다는 점을 서로 연결시켰다는 점 때문이다. 데이터 처리 과정을 엄밀히 구분하면 후버 사건은 개인정보를 수집하고 데이터베이스를 구성하는 단계에서 발생한 문제로서 수집 제한 원칙, 데이터 품질 원칙, 목적 명시 원칙 등이 적용된다. 즉, 개인정보는 명시한 목적에 필요한 범위 내에서 데이터 주체의 동의나 다른 법적 요건을 갖추어 제한적으로 수집되어야 하는 것이다. 이때 유럽사법재판소는 개인정보 수집의 목적이 주거권의 효과적 적용이라는 것을 정당한 것으로 볼 경우에도 그 목적 달성에 필요 없는 범위의 개인정보를 수집하는 것을 제한하는 것으로 지침의 의미를 파악한 것이다.

데이터의 수집 단계에서 차별로부터 보호하기 위해 적용할 수 있는 원칙은 ‘분류 금지 원칙’ 같이 데이터의 구별과 관련된 원칙이다. 데이터 수집 단계에서 집단을 구별하는 것만으로도 차별의 전제 조건은 갖춰지게 되는 것이다. 특히 국적을 이유로 한 차별을 금지하는 법규범이 시행되고 있는 경우 법적 의미의 차별 요건을 구성하게 된다. 그런데 유럽사법재판소의 논증 방식에 따르면 이러한 차별은 차별 대상이

65) 공정정보규정(Fair Information Practices)은 1973년 미국의 보건교육복지부(the Department of Health, Education and Welfare) 장관인 리차드슨(E. Richardson)이 개인에 관한 정보를 담은 자동화된 데이터 시스템 사용의 증가에 대응하기 위해 설립한 자문 위원회가 제출한 보고서 “기록, 컴퓨터, 그리고 시민권(Records, Computers and the Rights of Citizens)”에서 최초로 제안되고 명명되었다. 공정정보규정의 역사에 관해서 Robert Gellman, “Fair Information Practices: A Basic History”, Ver. 2.18, 2017, <https://bobgellman.com/rg-docs/rg-FIPshistory.pdf>, pp. 1~46 참조.

66) 1980년의 OECD 가이드라인(OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data)은 제7조부터 제14조까지 수집 제한(collection limitation), 데이터 품질(data quality), 목적 명시(purpose specification), 이용 제한(use limitation), 안전 보호(security safeguards), 공개(openness), 개인의 참여(individual participation), 관리자의 책임(accountability)에 관한 8개 기본 원칙을 규정하고 있으며, 2013년 개정된 가이드라인(OECD Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data)에서도 기본 원칙은 유지되고 있다. OECD, The OECD Privacy Framework, 2013, http://www.oecd.org/sti/economy/oecd_privacy_framework.pdf, pp. 14~15 참조.



되는 집단에 대한 주거권의 효과적 적용이라는 입법 목적에 따라 정당화된다. 그렇다면 범죄 예방 같은 목적으로 국적을 구별하여 개인정보를 수집하는 것은 국적을 이유로 한 차별로서 금지된다고 본 것은 범죄 예방이라는 목적이 국적을 이유로 한 차별을 정당화하지 못한다고 판단한 것으로 볼 수 있다. 이익과 불이익의 측면에서 구분하자면 주거권의 적용은 인권 또는 기본적 권리에 관한 것으로 이익이 되는 반면, 범죄 척결 또는 범죄 예방은 수집되어 데이터베이스로 구성되는 데이터에 연동된 개인을 잠재적 범죄자로서 분석 대상으로 삼는 것이어서 해당 개인에게 불이익이 될 수 있다.

3. 차별금지법과 개인정보보호법의 적용범위

개인정보를 수집하여 데이터베이스를 구성하는 단계에서 동일한 사실 관계에 대한 해석이 ‘개인정보의 보호’라는 관점에서 접근하는 방식과 ‘차별로부터 보호’라는 관점에서 접근하는 방식에 따라 달라질 수 있다는 점은 개인정보 보호법과 차별금지법의 적용 범위가 다르다는 규범적 함의를 묵시적으로 드러낸다.⁶⁷⁾ 이러한 차이점은 개인정보의 수집 목적이 변경될 경우 개인정보 보호법을 적용한 해석과 차별금지법을 적용한 해석이 달라질 수 있는지 살펴봄으로써 명시적으로 드러낼 수 있을 것이다. 이를 위해 후버 사건을 변형하여 범죄 예방을 목적으로 외국인에게 중앙 등록 시스템에 개인정보에 관한 데이터를 등록하는 법률이 시행되고 있다고 가정해 본다.

이때 개인정보 보호의 관점에서 접근할 경우 범죄 예방을 위해 사람을 식별하는 정보로서 국적은 중요한 데이터가 될 수 있다는 이유를 들어 범죄 예방을 위한 개인의 국적 정보의 필요성을 주장하면서 범죄 예방이라고 명시한 목적에 필요한 범위 내에서 데이터 주체의 동의나 다른 법적 요건을 갖추어 개인정보를 제한적으로 수집한다면 개인정보 수집의 필요성은 인정될 수 있다.

그런데 차별금지 또는 차별로부터 보호의 관점에서 접근하면 범죄 예방이라는 목적은 국적을 차별 금지 사유로 규정한 법적 의미를 상쇄할 수 있는 합리화 또는 정당화 기능을 수행해야 한다. 그런데 국적에 따라 내국인과 외국인을 구별하여 외국인만의 데이터베이스를 구성하면서 데이터베이스를 범죄 예방의 기초 자료로 삼는 것이 외국인에게 불리하게 작용한다는 점은 후버 사건과 아무런 차이를 갖지 않게 된다.

67) Raphaël Gellert · Katja de Vries · Paul de Hert · Serge Gutwirth, “A Comparative Analysis of Anti-Discrimination and Data Protection Legislations”, in *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, Bart Custers · Toon Calders · Bart Schermer · Tal Zarsky(Eds.), Springer, 2013, pp. 61~88.



4. 복잡한 차별금지과 안정화된 개인정보보호

여기서 데이터 수집의 목적과 수집되는 데이터의 내용이 보다 가치중립적인 것으로 여겨지는 사안에 이르면 “상당히 안정화된 개념(a fairly stabilized notion)”으로서 ‘개인정보 보호’와 “통일성을 찾아가는 개념(a concept in search of unity)”으로서 ‘차별’의 접근 방식은 분명한 차이를 드러낸다.⁶⁸⁾ 이러한 차이를 부각시키기 위해 이번에는 통계 작성에 활용할 목적으로 사람의 눈동자 색에 관한 정보를 수집하여 데이터베이스를 구성하는 경우를 상정해 본다. 사람의 눈동자 색이 사람을 식별하거나 식별할 수 있는 정보가 아니라고 하여 개인정보 보호법의 적용 대상이 되는 개인정보에서 제외되지 않는 한 통계 작성이라고 명시한 목적에 필요한 범위 내에서 데이터 주체의 동의나 다른 법적 요건을 갖추어 눈동자 색이라는 개인정보를 제한적으로 수집한다면 개인정보 수집의 필요성은 인정될 수 있다.

이 경우에 차별금지법의 접근은 몇 가지 다른 방식을 취할 수 있다. 접근 방식이 달라질 수 있는 것은 차별금지법을 구성하는 법적 차별 개념이 복잡한 것과 무관하지 않다.⁶⁹⁾ 첫 번째는 차별금지법을 차별 사유와 불가분의 관계를 맺고 있다고 볼 경우 눈동자 색은 차별금지법에서 직접적인 차별 금지 사유로 규정하고 있지 않기 때문에 차별의 문제가 발생하지 않는다고 볼 수 있다. 두 번째는 눈동자 색이 특별히 규정된 차별 사유를 대체하는 특성 즉, 대용물(proxy)로서 차별 사유와 간접적으로 연동되어 있다고 볼 경우 차별의 문제가 발생할 수 있다. 세 번째는 법으로 규정된 차별 사유에 해당하지 않지만 사람의 신체적 특징에 따라 사람을 구별하고 있으므로 차별의 전제가 되는 구별 기준으로 작용한다는 점을 들어 차별 문제와 연결시킬 수 있다. 그런데 차별의 문제와 연결시킨 두 번째와 세 번째 방식에서도 통계 작성이라는 목적만으로 눈동자 색으로 구별된 개인에게 어떤 불이익을 준다고 할 수 없다고 본다면 차별은 합리화 또는 정당화 된다. 이렇게 차별을 받는 사람의 불이익이 차별의 판단에서 차별의 정당화에 영향을 미치는 것은 차별의 판단에서 결과를 고려하도록 구성된 경우에 한한다. 의도를 고려하는 구성에 따르면 통계적 목적에 불이익을 가한다는 의도가 없는 한 불이익의 발생 여부와 상관없이 차별은 정당화 된다.

68) Raphaël Gellert · Katja de Vries · Paul de Hert · Serge Gutwirth, “A Comparative Analysis of Anti-Discrimination and Data Protection Legislations”, in *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, Bart Custers · Toon Calders · Bart Schermer · Tal Zarsky(Eds.), Springer, 2013, pp. 61~88: pp. 64~67.

69) 이에 관해서 본 논문 「제3장 제3절 차별의 복잡성과 평등」 참조.



II. 데이터 분석 단계에서 차별과 개인정보보호

1. 데이터 분석과 이용의 구별

데이터를 수집하여 데이터베이스를 구성하는 단계에서 차별 문제는 주로 데이터베이스에 포함되거나 배제되는 것 자체가 이후의 결정이나 판단에 유리 또는 불리하게 영향을 미치는 경우에 발생하는 것이라면, 데이터베이스를 분석하는 단계에서 차별 문제는 데이터베이스에 포함된 데이터 간에 구별을 실행하는 경우에 발생한다. 예를 들어 보험료를 50퍼센트(%) 상향 조정하는 정책의 대상자가 될 데이터베이스를 생성하는 것이 데이터베이스 구축 단계의 차별과 관련된다면, 일단 구축된 데이터베이스에서 보험료 인상의 대상자를 선별하는 것은 데이터베이스 분석 단계의 차별과 관련된다. 데이터 분석은 넓은 의미에서 수집된 데이터를 이용하는 방식 중에 하나일 수 있다. 예를 들어 고용 절차에서 지원자의 데이터를 수집하여 데이터베이스를 구축하고 분석한 후 그 결과를 향후 실제 고용 결정에 이용할 수도 있고 이용하지 않을 수 있다는 의미에서 단계가 구분된다는 것이다. 그리고 데이터 마이닝 같은 머신러닝 기법은 주로 데이터 분석을 통해 데이터베이스 내부에서 구별을 실행함으로써 규칙을 생성하는 단계와 이를 실제 모델로 이용하는 단계로 구분될 수 있다는 점에서도 데이터의 분석과 이용은 구분될 수 있고 데이터 분석 단계에서도 개인정보의 보호법과 차별금지법의 적용은 중첩될 수 있다.⁷⁰⁾

2. 테스트-아샤 사건

벨기에의 한 소비자 단체(Test-Achats)는 유럽연합의 “상품과 서비스의 공급과 접근에서 남녀평등대우원칙 시행 지침”⁷¹⁾ 제5조 제2항이 유럽인권보호협약이 보장하는 인권, 특히 남녀평등대우원칙에 위반하는 것은 아닌지 그 효력을 문제 삼았다(이하 ‘테스트-아샤 사건’).⁷²⁾ 동 지침 제5조 제2항은 동조 제1항에서 유럽연합 회원국이

70) Raphaël Gellert · Katja de Vries · Paul de Hert · Serge Gutwirth, “A Comparative Analysis of Anti-Discrimination and Data Protection Legislations”, in *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, Bart Custers · Toon Calders · Bart Schermer · Tal Zarsky(Eds.), Springer, 2013, pp. 61~88: p. 79.

71) “Council Directive 2004/113/EC of 13 December 2004 Implementing the Principle of Equal Treatment between Men and Women in the Access to and Supply of Goods and Services”, *Official Journal of the European Union*, 21 December 2004, L 373, pp. 37~43.

72) Association Belge des Consommateurs Test-Achats ASBL and Others v Conseil des ministres,



지침에 지정된 입법 개정 기한(2007년 12월 21일) 이후부터 체결되는 모든 신규 계약에 대해 보험 및 관련 금융 서비스 목적의 보험료 및 보험급여 산정 요인으로 성별을 사용하는 것이 개별적인 보험료 및 보험급여의 차이로 귀결되지 못하게 보장할 의무를 규정한 것에 대한 예외로, 그 기한 이전까지는 관련성 있는 정확한 보험 통계 및 통계 자료에 근거한 위험 평가에서 성별이 결정적 요인으로 사용되는 경우 개별적인 보험료 및 보험급여의 비례적 차이를 허용하도록 하고 있다. 이에 대해 유럽사법재판소는 동 지침 제5조 제2항은 2012년 12월 21일부터 효력을 상실한다고 판결했다.⁷³⁾

3. 통계적 분석에 대한 차별금지와 개인정보보호의 접근방식

보험 및 금융 분야는 방대한 양의 개인정보를 다룰 뿐만 아니라 그 중에서 위험이나 신용을 평가해야 하는 복잡한 업무가 많기 때문에 데이터 마이닝 같은 머신러닝 기술이 가장 활발하게 이용되는 곳이다. 그러므로 얼마든지 개인정보 보호의 관점에서 관련 지침을 적용하여 판단할 수 있을 것이다. 이 경우 보험료 산정을 목적으로 개인의 성별 정보를 처리하는 것이 개인정보 보호의 관점에서 정당화될 여지가 있다. 특히 생명 보험 같은 경우 성별에 따른 수명의 차이가 있다는 통계적 사실을 근거로 성별이 보험료 산정을 위한 위험 평가에서 고려될 필요성이 있다고 주장할 수 있기 때문이다. 또한 차별로부터 보호의 관점에서 접근하는 경우에 성별과 수명의 통계적 관련성이 머신러닝 알고리즘의 분석 결과로 나타날 경우 이러한 결과는 생명 보험의 위험 평가를 위한 데이터 분석에서 결정적인 평가요소로 고려될 수 있는 합리적 근거가 될 수 있다.

그렇다면 과연 통계적 자료가 합리적인 이유로 제시되는 경우 차별 금지 사유인 성별을 데이터 분석의 평가요소로 사용하는 것이 정당화되는 것인지 문제 삼을 수 있다. 차별금지법의 요청이 성별 이외에도 수명과 연관성이 있는 다른 지표가 있는 경우 성별을 데이터 분석의 지표로 삼지 말라는 것으로 해석된다면 통계적 자료가 근거로 제시되는 경우에도 차별을 정당화하지는 못할 것이다. 그렇다면 머신러닝 알고리즘은 성별을 대체할 수 있는 다른 특성을 대용물로 사용할 수 있고 이 경우에는 간접차별의 문제로 전환되거나⁷⁴⁾ 법이 규정하지 않은 새로운 차별 사유의 인정 문제로 접어들게 된다.

C-236/09, Judgment of the Court of 1 March 2011, para. 1, 12~14.

73) Association Belge des Consommateurs Test-Achats ASBL and Others v Conseil des ministres, C-236/09, Judgment of the Court of 1 March 2011, para. 36.

74) 이에 관해서 본 논문 「제4장 제2절 III. 2. 머신러닝 알고리즘의 특성 추론과 간접차별」 참조.



4. 집단의 차별에 대한 개인의 권리의 한계

이렇게 데이터베이스의 분석 단계에서 개인정보 보호의 관점과 차별로부터 보호의 관점을 문제로 설정하여 판단할 수 있음에도 불구하고 테스트—아사 사건은 차별에 관한 쟁점만을 다루고 있다.⁷⁵⁾ 여기서 개인정보 보호법의 접근 방식이 갖는 한계가 드러나는데, 개인정보 보호법은 정보주체로서 개인의 권리를 중심으로 설계되었기 때문에 자기정보가 직접 관련되어 있지 않는 한 소비자 단체가 청구하는 소송에서 주장될 수 있는 권리로서 개인정보 보호법상의 권리 또는 헌법상의 개인정보자기결정권(자기정보통제권)은 일정한 한계를 갖는다.

III. 분석 모델 이용 단계에서 차별과 개인정보보호

1. 개인정보 수집제한의 한계와 이용단계의 중요성

행정국가 및 복지국가로의 경향을 보이는 현대 국가는 지속적으로 방대한 개인 정보를 수집하고 보유해 왔고, 정보통신기술의 발전을 등에 업고 성장한 기업들은 개인의 일상적인 생활이 고스란히 담겨 있는 방대한 정보를 축적해 왔다. 게다가 정부와 기업의 협력 관계를 통해 데이터 경제를 활성화하려는 정책에 비추어볼 때 개인정보의 수집이 사실상 통제되기 어렵다고 전망하는 것이 현실에 부합한다. 그렇기 때문에 개인정보의 수집 자체를 제한하는 방식에 중점을 두어 개인정보를 보호하는 방식의 접근은 그 실효성에 한계를 가질 수밖에 없다.

머신러닝의 관점에서 데이터의 분석 단계에서 데이터베이스를 구성하는 개인 정보는 모델을 생성하기 위해 훈련용 및 검증용 데이터세트로 이용될 수도 있고, 생성된 모델을 시험하기 위해 평가용 데이터세트로 이용될 수도 있다. 그리고 구체적인 결정에서 사용되는 모델의 입력 값으로서 이용될 수도 있다. 데이터베이스의 분석은 이용 목적에 지향되어 있다. 일단 데이터베이스가 구축되고 주된 목적에 부합하는 알고리즘 모델을 생성해 내면 구체적인 이용 분야에 따라 모델의 규칙이 달라질 수 있겠지만 광범위로 이용될 수 있다. 예를 들어 위험 평가 모델은 보험료를 책정하기 위해 보험가입자의 사고 위험을 예측하고 그에 따라 보험가입자를 분류하는 데 사용

75) Raphaël Gellert · Katja de Vries · Paul de Hert · Serge Gutwirth, “A Comparative Analysis of Anti-Discrimination and Data Protection Legislations”, in *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, Bart Custers · Toon Calders · Bart Schermer · Tal Z. Zarsky(Eds.), Springer, 2013, pp. 61~88: p. 80.



되거나 금융거래자의 신용을 평가하고 대출금 상환 불이행의 위험을 관리하기 위해 사용될 수도 있지만,⁷⁶⁾ 국가의 경찰작용으로서 범죄행위나 불법행위를 예방하기 위해 우범자나 잠재적 테러리스트를 선별하여 예측하는 데 사용되거나⁷⁷⁾ 형사사법절차에서 검사가 공소를 제기하지 아니하는 처분을 할 때,⁷⁸⁾ 법관이 구속사유,⁷⁹⁾ 양형사유,⁸⁰⁾ 형의 선고유예 및 집행유예의 정상참작사유⁸¹⁾ 또는 보호관찰의 사유⁸²⁾ 등을 심사할 때 혹은 검사가 전자장치 부착명령을 청구할 때⁸³⁾ 재범의 위험성을 예측하는 데 사용될 수도 있다. 그리고 이 단계에서 비로소 개인정보의 처리 결과가 데이터 주체에게 법적 효과를 비롯해 일정한 영향을 미치게 되기 때문에 가장 중요한 단계이기도 하다. 머신러닝 알고리즘을 이용한 집단에 대한 평가 모델이 개인에 대한 법적 평가에 활용될 때 문제가 될 수 있다는 점은 앞에서 루미스 사건을 통해 살펴보았다.⁸⁴⁾

-
- 76) 신용정보의 이용 및 보호에 관한 법률 제4조 제1항: “신용정보업의 종류 및 그 업무는 다음 각 호와 같다. 1. 신용조회업: 신용조회업무 및 다음 각 목의 업무; 가. 본인인증 및 신용정보주체의 식별확인업무로서 금융위원회가 승인한 업무; 나. 신용평가모형 및 위험관리모형의 개발 및 판매 업무(이하 각 호 생략)”
- 77) 경찰관직무집행법 제2조: “경찰관은 다음 각 호의 직무를 수행한다. 1. 국민의 생명·신체 및 재산의 보호; 2. 범죄의 예방·진압 및 수사; 2의2. 범죄피해자 보호; 3. 경비, 주요 인사(人士) 경호 및 대간첩·대테러 작전 수행; 4. 치안정보의 수집·작성 및 배포; 5. 교통 단속과 교통 위해(危害)의 방지; 6. 외국 정부기관 및 국제기구와의 국제협력; 7. 그 밖에 공공의 안녕과 질서 유지”
- 78) 형사소송법 제247조: “검사는 「형법」 제51조의 사항을 참작하여 공소를 제기하지 아니할 수 있다.”
- 79) 형사소송법 제70조 제2항: “법원은 제1항의 구속사유를 심사함에 있어서 범죄의 중대성, 재범의 위험성, 피해자 및 중요 참고인 등에 대한 위해우려 등을 고려하여야 한다.”
- 80) 형법 제51조: “형을 정함에 있어서는 다음 사항을 참작하여야 한다. 1. 범인의 연령, 성행, 지능과 환경; 2. 피해자에 대한 관계; 3. 범행의 동기, 수단과 결과; 4. 범행 후의 정황”
- 81) 형법 제59조 제1항: “1년 이하의 징역이나 금고, 자격정지 또는 벌금의 형을 선고할 경우에 제51조의 사항을 참작하여 개전의 정상이 현저한 때에는 그 선고를 유예할 수 있다.(이하 단서 생략)”; 형법 제62조 제1항: “3년 이하의 징역이나 금고 또는 500만원 이하의 벌금의 형을 선고할 경우에 제51조의 사항을 참작하여 그 정상에 참작할 만한 사유가 있는 때에는 1년 이상 5년 이하의 기간 형의 집행을 유예할 수 있다.(이하 단서 생략)”
- 82) 형법 제59조의2 제1항: “형의 선고를 유예하는 경우에 재범방지를 위하여 지도 및 원호가 필요한 때에는 보호관찰을 받을 것을 명할 수 있다.”
- 83) 특정 범죄자에 대한 보호관찰 및 전자장치 부착 등에 관한 법률 제5조: “① 검사는 다음 각 호의 어느 하나에 해당하고, 성폭력범죄를 다시 범할 위험성이 있다고 인정되는 사람에 대하여 전자장치를 부착하도록 하는 명령(이하 “부착명령”이라 한다)을 법원에 청구할 수 있다.(이하 각 호 생략); ② 검사는 미성년자 대상 유괴범죄를 저지른 사람으로서 미성년자 대상 유괴범죄를 다시 범할 위험성이 있다고 인정되는 사람에 대하여 부착명령을 법원에 청구할 수 있다.(이하 단서 생략); ③ 검사는 살인범죄를 저지른 사람으로서 살인범죄를 다시 범할 위험성이 있다고 인정되는 사람에 대하여 부착명령을 법원에 청구할 수 있다.(이하 단서 생략); ④ 검사는 다음 각 호의 어느 하나에 해당하고 강도범죄를 다시 범할 위험성이 있다고 인정되는 사람에 대하여 부착명령을 법원에 청구할 수 있다.(이하 각 호 생략); ⑤ ~ ⑦ 생략”



2. 데이터 이용에 의한 차별자와 피차별자의 연결

머신러닝 알고리즘이 실제로 사용될 때 발생할 수 있는 차별의 문제를 입체적으로 부각시키기 위해 간단한 사고실험을 진행해 보도록 한다.⁸⁵⁾

[사고실험 1-1] 교사가 강의를 준비하면서 학생이 성별에 차이가 있다는 것을 인식한다. 그리고 인식한 차이에 따라 학생을 마음속 또는 머릿속으로 분류한다. 그 다음 강의실의 왼쪽에 여학생이 앉고 오른쪽에 남학생이 앉도록 분리하기로 한다. 그리고 이러한 분류 기준을 공지사항으로 작성하여 컴퓨터 프로그램을 이용해 강의실의 전자화면에 게시한다. 교사는 강의실에 설치된 카메라를 통해 자신이 생각한 대로 학생들이 대체로 분리된 것을 확인한 후 강의실의 스피커와 연결된 마이크를 이용해 강의를 시작한다.

그런데 이와 같은 상황은 학생의 시선에서 볼 때 다르게 구성될 수 있다.

[사고실험 1-2] 수업을 들으러 간 학생이 강의실 전면의 전자화면에서 여학생은 강의실 왼쪽에 남학생은 강의실 오른쪽에 앉아서 수강하라는 공지사항을 본다. 학생은 자신이 성별에 따라 여성인지 남성인지 인식한다. 그리고 그 인식에 따라 강의실의 왼쪽 또는 오른쪽으로 이동하여 자리에 앉는다. 이 와중에 건물 승강기와 가까운 쪽의 강의실 출입구에 마련된 장애인 지정석에 자리를 잡고 있던 학생은 그대로 자리를 지키고 있어야 하는지 자신의 성별을 인식해서 그에 따라 자리를 이동해야 하는지 고민에 빠진다. 또 요즘 들어 자신의 정체성에 대해 진지한 의문이 들기 시작한 어떤 학생은 교사가 들어오면 도대체 왜 이런 공지사항이 나왔는지 물어보고 항의해야 하는 것 아닌지 깊은 생각에 잠긴다.

이러한 사고실험에서 교사는 학생들이 성별에 차이가 있다는 것을 인식하고 그에 따라 학생을 분류한다. 그리고 강의실의 좌우로 여학생과 남학생을 관념상 분리한다. 여기까지는 분리를 통한 차별이 의심되는 상황이 교사의 시선에 머물러 있다. 그러한 분류 기준을 공지함으로써 적용될 때 성별에 따른 구별행위는 비로소 그 대상인 학생에게 효력을 미치며 학생의 시선이 교사의 시선과 교차하게 된다. 그에 따라 학생은 자신의

84) 이에 관해서 *State v. Loomis*, 2016 WI 68, 881 N. W. 2d 749 (2016) 및 본 논문 「제2장 제3절 I. 2. 루미스 사건」 참조.

85) 실증적으로 초등학교 학생들을 대상으로 분리 수업을 진행하여 차별 경험을 관찰한 내용에 대해 기록한 것으로 William Peters, *A Class Divided, Then and Now*, Expanded ed., Yale University Press, 1987 및 이에 관한 본 논문 「제3장 제2절 I. 4. 분리를 통한 분류와 집단의 비대칭성」 참조.



성별을 인식하고 교사가 제시한 기준에 따라 이동하여 자리에 앉음으로써 현실적 분리가 완성된다. 결국 이 상황을 차별의 관점에서 포착할 수 있는 계기는 차별적이라고 의심되는 조치로서 성별을 기준으로 한 공지사항이 게시됨으로써 발생한다. 이때 비합리적인 방식으로 분류 대상이 된 사람에게 어떤 불이익을 주거나 피해를 입힐 의도가 행위자에게 있는 경우에만 차별금지법이 적용된다고 보면 교사의 공시행위만으로도 차별 개념에 포착될 수 있으나, 교사의 의도가 증명되지 않으면 차별은 정당화된다.

위 사고실험 속 교사의 시선을 따라가 보면 교사는 “강의를 준비하면서 학생이 성별에 차이가 있다는 것을 인식한다. 그리고 인식한 차이에 따라 학생을 마음속 또는 머릿속으로 분류”했을 뿐 결코 학생들에게 어떤 정신적 손상을 가하려는 의도는 없다. 반면에 분류 대상이 된 사람이 불이익이나 피해를 받은 경우에만 차별금지법의 적용을 받는 것으로 보면 학생은 게시물을 보고 좌석을 이동하여 분리를 실현하는 과정에서 입은 불이익 또는 피해를 입증해야 한다. 그런데 위 사고실험 속 학생의 시선을 쫓아가 보면 학생은 “고민에 빠진” 또는 “생각에 잠긴” 수준에 머물러 있다. 여기까지는 교사가 인간 에이전트라는 암묵적 가정이 전제되어 있다. 사고실험의 내용을 조금 더 추가해 본다.

3. 인간에 의한 차별과 알고리즘 시스템에 의한 차별

[사고실험 2] 학생들은 수업이 끝나고 교사에게 자신들이 구상한 차별금지법 적용 모델에 따라 책임을 물으려고 한다. 그런데 수업을 진행한 교사가 새롭게 학교에 도입된 인공지능 교사였음이 밝혀졌다.⁸⁶⁾ 이러한 사실 앞에서 교사를 당연히 인간이라고 생각했던 학생들은 좌절한다. 그 중에 어떤 지적인 학생이 인간과 인간의 흉내를 내는 컴퓨터를 분간하기 어려워질 것이라는 튜링의 예측이 현실화됐다는 점을 인지하고 자신들이 구상한 차별 법 모델에서 행위의 주체였던 ‘인간’ 대신에 비인간(non-human)도 포함될 수 있는 ‘행위자’만을 요건에 남긴다. 왜냐하면 “우리는 정당한 이유에서 실제로 특정한 생물학적 구조를 가진 생물체만이 생각한다고 믿는다. 그러나 나의 친구가 (몇 년이 지나 결국) 실리콘으로 구성된 것으로 밝혀진다면, 나는 그가 사람이었는지에 관한 나의 판단이 아니라 사람이 어떤 물질로 구성될 수 있는지에 관한 나의 생각을 바꿀 것”⁸⁷⁾이라는 데이빗슨(D. Davidson)의 말이 떠올랐기 때문이다.

86) 인공지능은 실제로 교육 분야에 도입되고 있다. 이에 관해서 Carl Smith, “AI That Can Teach? It’s Already Happening”, ABC News, 16 June 2018, <http://www.abc.net.au/news/science/2018-06-16/artificial-intelligence-that-can-teach-is-already-happening/9863574> 참조, 접속일: 2018년 7월 23일.



‘[사고실험 2]’의 추가된 내용에서 ‘인간’ 학생들이 좌절한 근거에는 마치 튜링이 그의 논문 “계산기와 지능”⁸⁸⁾에서 ‘기계는 생각할 수 있을까?’라는 질문이 부딪힐 수밖에 없다고 했던 것과 유사한 종류의 고정관념이 자리 잡고 있다. ‘차별은 인간만이 할 수 있다.’ 또는 ‘차별을 이해할 수 없는 기계는 차별할 수 없다.’ 같은 가정을 예로 들 수 있는데, 이러한 가정에는 차별의 실행자와 인간을 개념적으로 연결시킨 ‘차별’ 그 자체에 대한 고정관념이 담겨 있다.

사고실험 속에 등장하는 분석철학자 데이빗슨(D. Davidson)의 태도는 앞서 알고리즘 에이전트의 책임 귀속 문제에서 민속 심리학에 기댄 법 개념을 근거로 알고리즘 에이전트 자체의 책임 허용에 대한 한계를 주장한 논증에서 브로젝(B. Brożek)이 보인 태도와 상반된다. 다시 말해 “지능적이고 인간과 다름없는 모습을 하고 필요할 때 도움을 주기도 하고 새로운 친구와 대화하며 함께 시간을 보낼 준비가 되어 있는 A가 복잡한 안드로이드 즉, 인간의 모습을 한 로봇으로 밝혀졌을 때 우리는 아마도 A에 대한 태도를 바꿀 것”⁸⁹⁾이라고 본 브로젝의 진단에 등장하는 ‘우리’에 적어도 “우리는 정당한 이유에서 실제로 특정한 생물학적 구조를 가진 생물체만이 생각한다고 믿는다. 그러나 나의 친구가 (몇 년이 지나 결국) 실리콘으로 구성된 것으로 밝혀진다면, 나는 그가 사람이었는지에 관한 나의 판단이 아니라 사람이 어떤 물질로 구성될 수 있는지에 관한 나의 생각을 바꿀 것”이라고 주장하는 데이빗슨 같은 사람들은 포함되지 않는다.

브로젝의 진단을 차별의 맥락에 연결시켜 보면 똑같은 차별적 상황이라도 그 원인이 인간에게 있을 경우와 알고리즘 시스템에게 있을 경우 그에 대한 평가가 달라질 수 있다는 것이기도 하다. 그러나 데이빗슨과 같은 태도는 똑같은 차별적 상황이라면 그 원인이 인간에게 있는 것과 시스템에 있는 것 사이에 평가의 기준이나 결과가 달라지지 않아야 한다는 규범적 요청으로 전환시킬 수 있다. 만약 이러한 고정관념이 극복될 수 있다면 이른바 ‘시스템에 의한 차별’이 갖는 규범적 함의를 ‘인간에 의한 차별’ 못지않게 중요하게 받아들이는 데에 일정 부분 기여 할 수 있을 것이다.

특히 책임 부담의 측면에서 인간에 의한 차별이 행위자인 개인에게 집중되는 것과 비교할 때 시스템에 의한 차별은 그 시스템의 작동에 관여하는 주체들에게 책임을 분산시키는 것이 보다 중요한 의의를 가질 수 있다.⁹⁰⁾ 책임 귀속을 특정

87) Donald Davidson, “Turing’s Test”[1990], in *Problems of Rationality*, Oxford University Press, 2004, pp. 77~86: p. 79.

88) 이에 관해서 Alan M. Turing, “Computing Machinery and Intelligence”, *Mind* 59(236), 1950, pp. 433~460 및 본 논문 「제2장 제1절 II. 2. 튜링의 모방게임과 학습하는 기계」 참조.

89) 이에 관해서 본 논문 「제5장 제2절 III. 3. 민속 심리학에 기초한 법 개념과 허용의 한계」 참조.

90) 이에 관해서 본 논문 「제5장 제2절 머신러닝 알고리즘의 차별에 대한 책임의 분산」 참조.



행위자에게 귀속시켜야 한다는 부담에서 벗어날 수 있다면 위 사고실험에서 머신러닝 알고리즘의 차별에 대한 책임의 주체는 알고리즘 에이전트로서 인공지능 교사, 이러한 인공지능 교사를 사용 또는 채용한 사용자로서 학교, 교사의 직무를 수행하는 인공지능 알고리즘을 설계한 개발자, 개발된 알고리즘에 관한 저작권 소유자 중에 일부 또는 모두가 될 수 있다.

4. 시스템의 차별에 대한 차별금지법과 개인정보보호법의 한계

선부른 의인화의 함정에만 빠지지 않을 수 있다면 추가된 내용을 포함하여 위 사고실험 전체 내용에서 교사를 머신러닝 알고리즘 시스템 또는 인공지능 에이전트라고 상정해 보는 것은 데이터로부터 학습한 머신러닝 알고리즘이 실제로 이용될 때에도 유사한 차별의 문제가 발생할 수 있다는 점을 발견하는 데 도움을 준다. 그리고 이 경우 개인정보 보호의 관점에서 접근하는 방식과의 차이점도 좀 더 선명해진다. 인공지능 교사 즉, 머신러닝 알고리즘 시스템이 분류⁹¹⁾ 기준으로 성별 또는 그 대용물을 지표로 삼을 경우 지표가 나타내는 특성을 가진 개인들은 집단(클래스)으로서 데이터 프로세싱과 관련을 맺는 것이지 식별된 또는 식별가능한 개인으로서 관련을 맺는 것이 아니기 때문이다. 다시 말해 인공지능 교사의 분류는 개인을 식별하지 않고 집단을 식별한다. 그러므로 집단 프라이버시를 보호법적으로 인정하여 집단에게 집단정보 보호에 관한 권리를 부여하지 않는 한 개인의 식별 관련성에 지향된 개인정보 보호법은 머신러닝 알고리즘의 작동에 의한 결정을 통제하는 법적 근거로 작용하는 데에 한계가 있다.

차별금지법의 접근 방식은 집단의 경계를 설정해주는 지표를 분류 기준으로 사용하는 것만으로 법을 적용할 수 있는 구조를 갖고 있다. 그러나 이러한 차별금지법도 가해자와 피해자의 구도 속에서 가해행위에 대한 불법의도와 그로 인해 발생한 피해결과에 천착할 경우 이른바 ‘시스템의 차별’⁹²⁾이 갖는 위험, 즉 의도를 입증하기 어려운 알고리즘 시스템의 작동방식에 대한 난해성, 피해자 개인이 직접 상대방이 되지 않는 집단을 매개로 한 피해의 간접성, 데이터세트로부터 훈련되어 통계적 근거를 갖춘 알고리즘 시스템의 차별적 결정이 갖춘 기본적 합리성이 주변적인 것으로 취급되는 한계에 직면하게 된다.

91) ‘개념적 분류(conceptual classification)’와 ‘반응적 분류(responsive classification)’의 구별에 관해서 Robert Brandom, *Making It Explicit: Reasoning, Representing and Discursive Commitment*, Harvard University Press, 1994, pp. 87~89 및 본 논문 「제4장 제2절 1. 2. 반응적 분류와 개념적 분류」 참조.

92) Marie Mercat-Bruns, “From Disparate Impact to Systemic Discrimination”, in *Discrimination at Work: Comparing European, French, and American Law*, Elaine Holt(Trans.), University of California Press, 2016, Chapter 4: pp. 82~144 참조.



제4절 자동화된 차별적 결정에 구속되지 않을 권리와 설명의 한계

머신러닝 알고리즘 시스템이 사회의 차별을 재생산하고 확대할 수 있다는 위험과 그 위험의 실현으로서 차별효과의 발생에 적절하게 대처하기 어려운 주된 이유 중에 하나는 머신러닝 알고리즘 자체가 불투명하다는 것이다.⁹³⁾ 결정 규칙으로서 알고리즘이 정확하게 확인되지 않는다면 차별의 의심을 받는 알고리즘의 결정에 대해 그 근거가 무엇인지 심사하기 어렵다. 그래서 알고리즘의 투명성을 제고하기 위한 법적 장치로서 결정의 효력이 미치는 관련 당사자들에게 알고리즘의 결정에 대해 설명 받을 권리를 부여하고 이를 보장하기 위한 설명의무를 알고리즘의 결정에 대한 책임자에게 부담시키는 방법 또는 알고리즘의 결정에 대한 책임자가 별도로 부담하는 공공의무로서 차별적 결정에 대해 설명 또는 해명할 의무를 책임자에게 부담시키는 방법이 고려될 수 있다.

I. 머신러닝 알고리즘 시스템의 자동화된 결정과 차별

1. 자동화된 개별적 결정과 GDPR의 제정

머신러닝 알고리즘이 적용되는 영역이 확대되면서 인간의 개입 없이 자동화된 알고리즘 시스템이 도출한 결과가 법적 효과를 비롯해 인간에게 중대한 효과를 미칠 수 있다는 점은 국제적 관심의 대상이 되었고, 그에 관한 법적 규제안에 관한 논의의 결과물이 입법이나 정책의 형태로 제시되고 있다. 그리고 이러한 입법이나 정책 속에 머신러닝 알고리즘 같은 자동화된 알고리즘 시스템에서 처리하는 데이터의 주체로서 인간이 알고리즘 자체에 대해 갖는 권리가 담겨 있는지 확인하거나 해석을 통해 그러한 권리를 도출해 내려는 시도는 권리를 중심으로 한 문제 접근 방식의 전형적인 태도이다. 특히 유럽연합(EU)의 1995년 데이터 보호지침⁹⁴⁾을 대체하는 개인정보 보호법으로서 일반데이터보호법(General Data Protection Regulation, 이하 ‘GDPR’)⁹⁵⁾이 2016년에 제정되면서 과연 GDPR에 그런

93) 이에 관해서본 논문 「제4장 제4절 I. 불투명성의 세 가지 차원」 참조.

94) “Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data”, *Official Journal of the European Union*, 23 November 1995, L 281.



권리로서 이른바 ‘설명에 관한 권리(right to explanation)’가 규정되어 있는지 또는 해석을 통해 그런 권리를 도출할 수 있는지 여부가 머신러닝 알고리즘 시스템이 불러일으킬 사회적 문제에 대한 법적 대응 논의로서 중심에 서게 되었다. 이러한 논의를 형성하는 데 결정적인 역할을 한 것은 ‘머신러닝에서 인간의 해석가능성’을 주제로 한 학술대회에 제출된 굿먼(B. Goodman)과 플랙스먼(S. Flaxman)의 “알고리즘의 결정형성⁹⁶⁾에 관한 EU법과 ‘설명에 관한 권리’”⁹⁷⁾라는 소논문이다. 이 논문의 배경이 되는 GDPR은 초안이 제시될 당시 종이 기반의 관료주의적 요구로부터 실제의 준수, 법의 조화, 개인에 대한 권리부여로 변화의 기준을 이동시켰다는 점에서 칸트가 외부 대상으로부터 개인의 인지능력으로 현실에 대한 이해를 이동시킨 것에 비견할 만한 “코페르니쿠스 혁명”이라 여겨지기도 했다.⁹⁸⁾

2. 차별의 문제와 설명에 관한 권리

굿먼과 플랙스먼은 당시에 유럽연합과 그 회원국의 데이터 프라이버시(data privacy)와 관련된 통일적인 보호를 위해 제정된 GDPR의 시행(2018년 5월 25일)⁹⁹⁾을 앞둔 상태에서 ‘프로파일링을 포함하는 자동화된 개별 의사결정’이란 제목이 붙여진

95) “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)”, *Official Journal of the European Union*, 4 May 2016, L 119; GDPR에 대한 자세한 해설로 Voigt, Paul · Axel von dem Bussche, *The EU General Data Protection Regulation (GDPR): A Practical Guide*, Springer, 2017 참조; GDPR의 전체 조문과 국내법의 차이에 대한 비교는 박노형 · 고환경 · 구태언 · 김경환 · 박영우 · 이상직 · 이창범 · 정명현 · 최주선, *EU 개인정보보호법 - GDPR을 중심으로 -*, 박영사, 2017 참조.

96) 통상 ‘의사결정’으로 번역하는 ‘decision-making’은 최종 결론으로서 결정(decision)뿐만 아니라 그런 결론에 도달하기 위해 결정을 형성 또는 생성하는(making) 절차(procedure) 또는 과정(process)도 포함한다. 특히 현대의 조직적 결정이 조직 내부의 결정 프로세스의 작동으로 이루어진다는 점을 고려하면 결정에 도달하는 프로세스는 필수적이다. 더구나 알고리즘에게 ‘의사’ 또는 ‘의지’가 있는지 불투명한 상황에서 인간의 결정에 사용되는 ‘의사결정’을 언어 사용의 관행상 그대로 기계의 결정에 사용하는 것은 무의식적 또는 무비판적 의인화를 감수하는 비용을 지불해야 한다. 반대로 결정에서 ‘의사’의 요소를 본질적인 것으로 보지 않는다면 굳이 결정 앞에 ‘의사’를 덧붙일 이유가 없어 보인다.

97) Bryce Goodman · Seth Flaxman, “EU Regulations on Algorithmic Decision-Making and a ‘right to explanation’”, ver. 1, ICML Workshop on Human Interpretability in Machine Learning, June 2016, pp. 26~30.

98) Christopher Kuner, “The European Commission’s Proposed Data Protection Regulation: A Copernican Revolution in European Data Protection Law”, Bloomberg BNA Privacy and Security Report, 2012, pp. 1~15.

99) Article 99(2) GDPR: “It shall apply from 25 May 2018.”



제22조를 중심에 두고 GDPR이 적용될 경우 발생할 수 있는 문제를 다룬다. 그런데 데이터 프라이버시를 보호법적으로 하는 ‘개인정보 보호법’이라는 법 분야와 ‘설명에 관한 권리’라는 데이터 주체의 권리를 내세운 논문 제목 때문에 자칫 간과하기 쉬운 중요한 내용은 ‘설명에 관한 권리’를 논의하는 이유가 차별의 문제와 밀접한 관련을 맺고 있기 때문이라는 점이다. 실제로 굿먼과 플렉스먼이 초점은 “차별에 대한 GDPR의 입장에서 발생하는 문제”와 “GDPR의 ‘설명에 관한 권리’로부터 발생하는 문제”에 맞추어져 있다.¹⁰⁰⁾ 그리고 차별의 문제를 포괄하고 있는 규정으로 GDPR 제22조 제4항을, ‘설명에 관한 권리’를 구체화하는 조항으로 제15조, 제22조, 전문(recital)¹⁰¹⁾ 제71항을 제시한다.

3. 차별에 비중을 둔 해석의 필요성

굿먼과 플렉스먼의 문제제기 이후에 그에 대한 반론의 형식으로 제기된 바흐터(S. Wachter) 및 그 동료들의 주장¹⁰²⁾을 국내에 소개하면서 ‘설명에 관한 권리’를 개인정보 자기결정권의 “현대적 변형” 또는 “새로운 발전적 형태”로 보고 기본권의 일종으로 인정할 필요가 있다는 논의¹⁰³⁾ 또는 GDPR의 프로파일링 규정에 대한 분석적 논의¹⁰⁴⁾ 등은 대체로 GDPR 제13조 제2항 (f), 제14조 제2항 (g) 및 제15조 제1항 (h)와 그 내용에서 언급되는 범위 안에서 제22조 제1항 및 제4항을 해석하거나 병렬적으로 제22조의 각 항을 분석하는 데에 초점이 맞추어져 있고, 제22조를 해석하는 경우에도 제4항에 특별히 비중을 두지는 않는다.¹⁰⁵⁾ 따라서 이하에서는 GDPR 제22조 제4항을 머신러닝 알고리즘의 차별적 결정의 대상이 되지 않을 권리, 즉 그 결정에 구속되지 않을 권리의 법적 근거로 보아 동 조항으로부터 출발하는 해석론을 전개해보도록 한다.

100) Bryce Goodman · Seth Flaxman, “EU Regulations on Algorithmic Decision-Making and a ‘right to explanation’”, ver. 1, ICML Workshop on Human Interpretability in Machine Learning, June 2016, pp. 26~30: p. 27.

101) ‘recital’은 국내에서 ‘전문해설’, ‘상설’ 등으로 번역되는데, 법적 구속력을 갖는 본문과 대비되는 의미를 부각하기 위한 것으로 보인다.

102) Sandra Wachter · Brent Mittelstadt · Luciano Floridi, “Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation”, *International Data Privacy Law* 7(2), 2017, pp. 76~99.

103) 박상돈, “헌법상 자동의사결정 알고리즘 설명요구권에 관한 개괄적 고찰”, *헌법학연구* 23(3), 2017, 185~218쪽: 205쪽 및 212쪽.

104) 박노형 · 정명현, “EU GDPR상 프로파일링 규정의 법적 분석”, *안암법학* 56, 2018, 283~315쪽.

105) 제22조를 전반적으로 언급하면서 “적절한 안전장치가 마련되지 않을 경우 프로파일링 및 자동화된 의사결정 기술이 기존에 형성된 고정관념과 사회적 구분을 영속화할 수 있다는 문제점을 지적”한 EU 제29조 작업반의 견해를 소개한 경우로 권건보 · 이한주 · 김일환, “EU GDPR 제정 과정 및 그 이후 입법동향에 관한 연구”, *미국헌법연구* 29(1), 2018, 1~38쪽: 13~15쪽.



II. 자동화된 결정에 구속되지 않을 권리와 그 예외

1. GDPR 제22조 제4항의 형식적 구조와 차별 관련성

GDPR 제22조 제4항은 “제2항에 언급된 결정은 제9조 제1항에 언급된 특정 범주의 개인데이터에 근거를 두어서는 안 된다. 다만, 제9조 제2항 (a) 또는 (g)가 적용되지 않고 데이터주체의 권리와 자유 및 정당한 이익을 보호하는 적합한 조치가 마련되어 있지 않은 경우에 한한다.”고 규정하고 있다. 규정의 맥락을 이해하기 위해서는 GDPR 제22조 제4항에 언급된 동조 제2항과 제9조 제1항 및 제2항 (a) 또는 (g)을 부가적으로 살펴보아야 하는데, 그 전에 우선 확인해 둘 것이 있다. 동 조항의 핵심 문장은 ‘결정이 특정 범주의 개인데이터에 근거해서는 안 된다(Decisions ... shall not be based on special categories of personal data ...).’라는 점이다. 개인을 지시하는 특정 범주의 데이터는 개인의 특성이면서 동시에 같은 특성을 공유하는 집단을 범주화한다는 점에서 이를 근거로 어떤 결정에 이르는 것을 금지하는 것은 차별금지법명제의 형식 구조¹⁰⁶⁾를 갖춘 것으로 볼 수 있다.

2. 자동화된 결정에 구속되지 않을 권리

GDPR 제22조 제2항은 동조 제1항의 결정이 적용되지 않는 예외를 규정하고 있어 제1항을 먼저 보면 “데이터 주체는 프로파일링을 포함해서 그 또는 그녀에게 법적 효력을 발생시키거나 이와 유사하게 중대한 영향을 미치는 자동화된 프로세싱에 오로지 근거를 둔 결정에 따르지 않을 권리를 가져야 한다.”¹⁰⁷⁾ 여기서 정보주체로서 데이터 주체(data subject)는 데이터와 관련 있는 “식별된 또는 식별가능한 자연인”이고,¹⁰⁸⁾ 프로파일링(profiling)은 “자연인과 관련 있는 어떤 개인적 측면을 평가하기

106) 이에 관해서 본 논문 「제3장 제3절 I. 3. 헌법과 법률의 차별 관련 규정과 그 구조적 유사성」 참조.

107) Article 22(1) GDPR: “The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.”

108) 다음에 추가한 밑줄 참조. Article 4(1) GDPR: “‘personal data’ means any information relating to an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.”



위해 개인정보를 이용하는 모든 형태의 자동화된 개인정보 프로세싱”을 말한다.¹⁰⁹⁾ 개인정보(personal data)는 “데이터 주체와 관련된 모든 정보”로서 데이터 주체의 개인정보를 의미하고,¹¹⁰⁾ 프로세싱(processing)은 “개인데이터 또는 개인정보 세트에 관해 실행되는 모든 형태의 작동 또는 작동세트”로서 “자동화된 수단에 의한 것인지 여부”는 묻지 않는다.¹¹¹⁾ 이러한 정의 규정에 따르면 제22조 제1항은 ‘식별된 또는 식별가능한 자연인(데이터 주체)에게 자기와 관련된 데이터 또는 데이터세트에 관해 실행되는 모든 형태의 작동 또는 작동세트(프로세싱) 중에 자기의 개인적 측면을 평가하기 위한 것(프로파일링)을 포함해 자기에게 법적 효력을 발생시키거나 이와 유사하게 중대한 영향을 미치는 자동화된 프로세싱에 오로지 근거를 둔 결정에 구속되지 않을 권리’를 부여한다.

이러한 법적 조건에서 ‘자연인(natural person)’이 아닌 법인이나 그 밖에 법적 인격을 획득하지 못하고 단지 환경일 뿐인 사물은 권리를 주장할 수 있는 데이터 주체에서 제외되고, ‘자동화된(automated)’ 프로세싱이 아닌 수동의(manual) 프로세싱은 데이터 주체가 권리를 주장하여 거부할 수 있는 결정 프로세싱에서 배제된다. 특히 ‘오로지(solely)’ 자동화된 프로세싱에 의한 결정에 대해서만 거부할 수 있도록 하는 것은 수동과 자동이 결합된 ‘반자동’ 같은 혼종(hybrid)의 프로세싱에 의한 결정에 대해서까지 구속되지 않을 권리를 보장하지는 못한다. 나아가 법적 효력을 발생시키는 프로세싱과 ‘유사하게 중대한 영향을 미치는(similarly significantly affects)’ 프로세싱으로 이분화한 것은 사실적 효력에 대해 유사성과 중대성에 관한 해석에 따라 권리가 적용되는 결정의 형성 수단으로서 자동화된 프로세싱의 범위를 폭넓게 제한할 수 있는 재량의 공간을 남긴다.

109) 다음에 추가한 밑줄 참조. Article 4(4) GDPR: “‘profiling’ means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.”

110) 앞의 Article 4(1) GDPR 참조.

111) 다음에 추가한 밑줄 참조. Article 4(2) GDPR: “‘processing’ means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.”



3. 자동화된 결정에 구속되는 예외

GDPR 제22조 제2항은 이에 더해 동조 제1항이 적용되지 않는 예외 즉, 데이터 주체에게 법적 효력을 발생시키거나 이와 유사하게 중대한 영향을 미치는 자동화된 프로세싱에 오로지 근거한 결정이지만 데이터 주체에게 이에 따르지 않을 권리가 부여되지 않는 예외를 세 가지로 규정한다. “그 결정이 데이터 주체와 데이터 컨트롤러 간에 계약을 체결하거나 이행하기 위해 필요한 경우”(a), “컨트롤러에게 적용되고 또 데이터 주체의 권리 및 자유와 정당한 이익을 보호하는 적합한 조치를 규정한 유럽연합 또는 회원국 법이 그 결정을 허용하는 경우”(b), “데이터 주체의 명시적 동의에 근거하는 경우”(c) 제22조 제1항은 이러한 결정에 적용되지 않는다.¹¹²⁾ 여기서 컨트롤러(controller)는 “자연인이나 법인, 공공기관(public authority), 정부 기관(agency), 그 밖에 단독으로 또는 타인과 공동으로 개인데이터 프로세싱의 목적 및 수단을 결정하는 자”¹¹³⁾로서 식별된 또는 식별가능한 자연인과 관련된 정보를 처리하는 목적과 수단을 결정하는 자라면 자연인에 한정되지 않는다. 제22조 제2항을 동조 제1항과 결합하면 ‘계약의 체결 및 이행에 필요’한 경우, ‘데이터 주체의 권리 및 자유와 정당한 이익을 보장하는 법이 허용’한 경우 또는 ‘데이터 주체가 명시적으로 동의’한 경우 데이터 주체는 자기에 대해 법적 효력을 발생시키거나 이와 유사하게 중대한 영향을 미치는 자동화된 프로세싱에 오로지 근거한 결정이라고 할지라도 이에 따르지 않을 권리를 갖지 않는다.

4. ‘설명에 관한 권리’ 존부 논쟁

그런데 GDPR 제22조 제2항에 따라 자동화된 결정에 구속되는 경우에도 데이터 컨트롤러는 제22조 제3항에 따라 계약의 체결 및 이행에 필요한 경우(a)와 데이터

112) Article 22(2) GDPR: “Paragraph 1 shall not apply if the decision: (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller; (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or (c) is based on the data subject's explicit consent.”

113) 다음에 추가한 밑줄 참조. Article 4(7) GDPR: “‘controller’ means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law.”



주체가 명시적으로 동의한 경우(c)에는 “데이터 주체의 권리 및 자유와 정당한 이익을 보호하기 위해 적합한 조치로서, 최소한 컨트롤러의 인적 개입을 획득하고 자신의 견해를 표명하며 결정에 이의를 제기할 권리를 보호하기 위한 조치를 이행해야 한다.”¹¹⁴⁾ 그런데 전문(recital) 제71항의 마지막 부분을 보면 프로세싱에 적용되는 ‘적합한 안전 장치’에는 “인간의 개입을 획득할 권리, 자기의 견해를 표현할 권리, 그러한 평가 이후에 도달한 결정의 설명을 획득할 권리, 그리고 결정에 이의제기할 권리”가 포함되어야 한다고 서술되어 있다. 법적 구속력이 있는 본문 제22조 제3조에 규정된 “최소한 컨트롤러의 인적 개입을 획득하고 자신의 견해를 표명하며 결정에 이의를 제기할 권리를 보호하기 위한 조치”와 비교할 때 본문에는 전문과 다르게 프로세싱의 “평가 이후에 도달한 결정의 설명을 획득할 권리”가 빠져 있다.

굿먼과 플렉스먼은 전문 제71항에 서술된 ‘설명을 획득할 권리’라는 표현에 입각해 ‘설명에 관한 권리’의 존재를 논증하려 하고,¹¹⁵⁾ 바흐터와 그 동료들은 전문에는 법적 효력이 없다는 점과 정작 본문에는 ‘설명을 획득할 권리’에 관한 표현만 특별히 빠졌다는 데에 규범적 의의를 두어 굿먼과 플렉스먼이 말하는 ‘설명에 관한 권리’가 GDPR에 규정되어 있지 않다는 주장을 펼친다.¹¹⁶⁾ 그러나 설명의 대상이 프로세싱 이후에 도출된 결과에만 한정되지 않는다면 ‘결정에 대한 설명의 획득’이라는 표현이 없다는 이유만으로 ‘설명에 관한 권리’ 자체가 없다고 하는 것은 자칫 권리 대상의 범위 설정 문제를 권리 자체의 존부 문제와 결합시켜 권리 이해에 혼동을 불러일으킬 수 있다. 바흐터와 그 동료들 역시 굿먼과 플렉스먼이 제시한 GDPR 제15조 제1항의 자동화된 프로세싱에 포함된 로직에 대한 의미 있는 정보에 대한 접근권을 중요하게 다루고 있기 때문이다. 이에 관한 논의는 항을 바꿔 마지막에서 다루기로 한다.¹¹⁷⁾

114) 다음에 추가한 밑줄 참조. Article 22(3) GDPR: “In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.”

115) Bryce Goodman · Seth Flaxman, “EU Regulations on Algorithmic Decision-Making and a ‘right to explanation’”, ver. 1, ICML Workshop on Human Interpretability in Machine Learning, June 2016, pp. 26~30.

116) Sandra Wachter · Brent Mittelstadt · Luciano Floridi, “Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation”, *International Data Privacy Law* 7(2), 2017, pp. 76~99.

117) 이에 관해서 본 논문 「제5장 제4절 IV. 머신러닝 알고리즘에 대한 설명의무와 설명청구권」 참조.



III. 특정 범주의 개인정보와 자동화된 차별적 결정

1. 특정 범주의 개인정보에 근거한 자동화된 결정에 구속되지 않을 권리

이제 다시 제22조 제4항으로 돌아가 보면 “제2항에 언급된 결정은 제9조 제1항에 언급된 특정 범주의 개인정보에 근거를 두어서는 안 된다. 다만, 제9조 제2항(a) 또는 (g)가 적용되지 않고 데이터 주체의 권리 및 자유와 정당한 이익을 보호하는 적합한 조치가 마련되어 있지 않은 경우에 한한다.” 제22조 제2항에 언급된 결정은 앞에서 펼친 해석에 의하면 데이터 주체가 ‘따라야 하는(shall be subject to)’ 자동화 프로세싱에 근거한 결정이다. 동조 제4항에 대한 해석을 완성하려면 동조 제2항에 언급된 결정이 근거로 삼아서는 안 되는 특정 범주의 개인정보를 확인해야 한다. 제9조 제1항은 “인종이나 민족적 출신, 정치적 견해, 종교나 철학적 신념, 노동조합원 자격을 드러내는 개인정보 프로세싱과 유전데이터, 자연인을 고유하게 식별할 목적의 생체인식데이터, 건강에 관한 데이터, 성생활 또는 성적지향에 관한 데이터 프로세싱은 금지된다.”¹¹⁸⁾고 규정한다.

그런데 이 조문은 금지되는 데이터 프로세싱을 둘로 나누고 있어 제22조 제4항에서 지시한 “제9조 제1항에 언급된 특정 범주의 개인정보”가 프로세싱 대상으로 양분된 데이터 중 어느 한 부분을 가리키는지 아니면 두 부분 모두를 가리키는지 양 갈래의 해석 가능성을 남긴다. 제9조 제1항의 규정 형식은 금지하는 프로세싱을 “인종이나 민족적 출신, 정치적 견해, 종교나 철학적 신념, 노동조합원 자격을 드러내는 개인정보”와 “유전데이터, 자연인을 고유하게 식별할 목적의 생체인식데이터, 건강에 관한 데이터, 또는 성생활 또는 성적지향에 관한 데이터”로 나눈다. “제9조 제1항에 언급된 특정 범주의 개인정보”를 엄밀하게 해석하면 전자만을 지시하여 “인종이나 민족적 출신, 정치적 견해, 종교나 철학적 신념, 노동조합원 자격을 드러내는 데이터”만이 해당된다고 해석할 수 있다. 그러나 개인정보가 GDPR에서 식별된 또는 식별 가능한 자연인에 관한 정보로 정의된다는 점과 제9조 제1항이 프로세싱을 금지하는 여러 종류의 데이터를 특정하고 있다는 점을 고려하면 전자의 데이터 종류뿐만

118) 특정 범주의 개인정보는 다음에 추가한 밑줄 참조. Article 9(1) GDPR: “Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited.”



아니라 자연인의 “유전데이터, 자연인을 고유하게 식별할 목적의 생체인식데이터, 건강에 관한 데이터, 또는 성생활 또는 성적지향에 관한 데이터”도 제22조 제4항에서 지시하는 특정 범주의 개인데이터에 포함된다. 그렇다면 제22조 제2항의 데이터 주체가 ‘따라야 하는’ 자동화 프로세싱에 근거한 결정은 제9조 제1항에 언급된 종류의 데이터를 근거로 할 경우 제22조 제4항에 의해 ‘구속되지 않을 권리(the right not to be subject to)’의 대상이 되고, 데이터 주체는 그 결정에 ‘따르지 않을’ 수 있다.

2. 자동화된 결정에 구속되는 예외

GDPR 제22조 제4항에도 예외를 규정한 단서가 있는데 “제9조 제2항 (a) 또는 (g)가 적용되고 데이터 주체의 권리 및 자유와 정당한 이익을 보호하는 적합한 조치가 마련되어 있는 경우”이다. 이를 위해서는 ‘제9조 제2항 (a) 또는 (g)가 적용될 것’과 ‘데이터 주체의 권리 및 자유와 정당한 이익을 보호하는 적합한 조치가 마련되어 있을 것’이라는 두 가지 요건을 모두 충족해야 한다. 후자의 요건은 제22조 제1항의 예외로서 동조 제2항의 조건 중 ‘(b)’를 갖춘 경우이다. 즉, 제22조 제2항 (b)의 조건을 만족시켜 데이터 주체의 권리 및 자유와 정당한 이익을 보호하는 적합한 조치가 마련되어 있다고 하더라도 그 결정이 제9조 제1항에 언급된 종류의 데이터에 근거를 두고 있다면 제22조 제4항의 예외적 단서 부분에 해당되지 않는 동시에 제22조 제1항의 예외에도 해당되지 않는다. 그러므로 제22조 제4항의 예외 조건 중 전자의 요건인 제9조 제2항 (a) 또는 (g)는 이러한 예외를 가능하게 해주는 제22조 제1항의 ‘예외 조건의 예외 조건의 예외 조건’으로서 제22조 제1항이 적용되지 않는 조건이 된다.

결국 제22조 제4항의 예외 조건은 동조 제1항 및 제2항과 결합하여 ‘데이터 주체에게 법적 효력을 발생시키거나 이와 유사하게 중대한 영향을 미치는 자동화된 프로세싱에 오로지 근거한 결정이 계약의 체결 및 이행에 필요한 경우, 데이터 주체의 권리 및 자유와 정당한 이익을 보장하는 법이 허용한 경우 또는 데이터 주체가 명시적으로 동의한 경우라고 할지라도 특정 범주의 개인데이터에 기초한 경우라면 그 결정을 따르지 않을 수 있는 권리’의 예외가 되는 것이다.

그렇다면 이제 마지막으로 제22조 제4항의 예외 조건 중 제9조 제2항 ‘(a)’와 ‘(g)’를 살펴볼 필요가 있다. 동 조항은 동조 제1항의 적용을 배제하는 예외 규정



으로서 ‘(a)’와 ‘(g)’는 각각 “데이터 주체가 하나 이상의 특정한 목적을 위해 특정 범주의 개인데이터의 프로세싱에 명시적으로 동의한 경우로서, 유럽연합이나 그 회원국 법이 제1항에 언급된 금지가 해제될 수 없다고 정하지 않은 경우”(a)와 “프로세싱이 유럽연합 또는 회원국 법에 근거하여 추구하는 목적에 비례해야 하고, 개인데이터보호권의 본질을 존중해야 하고, 데이터주체의 기본적 권리와 이익을 보호하는 적합하고 특정한 조치를 취해야 하는 중대한 공익을 이유로 필요한 경우”(g)이다.¹¹⁹⁾

‘제9조 제2항 (a)’의 내용은 제9조 제1항에 언급된 특정 범주의 개인데이터에 대해 목적을 정한 프로세싱에 대해 데이터 주체가 동의한 경우라는 점에서 데이터 주체가 명시적으로 동의한 경우를 자동화 프로세싱에만 근거한 결정의 예외로 하는 ‘제22조 제2항 (c)’의 내용과 유사하다. 그러나 유럽연합 또는 회원국 법이 데이터 주체의 동의로 제9조 제1항의 금지를 해제하지 못하게 한 경우 이러한 법이 특수 범주의 개인데이터에 대한 데이터 주체의 동의에 우선한다는 점에서 차이가 있다.

‘제9조 제2항 (g)’는 중대한 공익을 이유로 필요한 경우 특정 범주의 개인데이터에 대한 프로세싱 일반이 허용될 수 있는 근거가 된다. 제22조의 맥락에서 그 프로세싱은 데이터 주체에게 법적 효력을 발생시키거나 이와 유사하게 중대한 영향을 미치는 자동화된 프로세싱이고, 이러한 자동화 프로세싱은 중대한 공익을 이유로 필요한 경우 유럽연합 또는 회원국 법에 근거했을 때 추구하는 목적에 비례하고, 개인 데이터보호권의 본질을 존중하고, 데이터주체의 기본적 권리와 이익을 보호하는 적합하고 특정한 조치가 취해진다면 허용된다는 것이다. 이렇게 제9조 제2항 (a) 또는 (g)의 조건이 충족되고 ‘이와 함께’ 데이터 주체의 권리 및 자유와 정당한 이익을 보호하는 적합한 조치가 마련되어 있는 경우 제9조 제1항의 특정 범주의 개인데이터에 기초한 자동화된 결정이라고 할지라도 그에 ‘따르지 않을 권리’ 또는 ‘중속되지 않을 권리’의 대상이 되지 않는다.

119) Article 9(2) GDPR: “Paragraph 1 shall not apply if one of the following applies: (a) the data subject has given explicit consent to the processing of those personal data for one or more specified purposes, except where Union or Member State law provide that the prohibition referred to in paragraph 1 may not be lifted by the data subject; …(중략)…; (g) processing is necessary for reasons of substantial public interest, on the basis of Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject; (후략)….”



3. 소결

결국 데이터 주체에게 법적 효력을 발생시키거나 이와 유사하게 중대한 영향을 미치는 자동화된 프로세싱에 오로지 근거한 결정이 계약의 체결 및 이행에 필요한 경우, 데이터 주체의 권리 및 자유와 정당한 이익을 보장하는 법이 허용한 경우 또는 데이터 주체가 명시적으로 동의한 경우라고 할지라도 이른바 ‘민감한 개인정보’라고 할 수 있는 특정 범주의 개인데이터에 기초한 경우라면 그 결정을 따르지 않을 수 있는 권리는 인정되지만, 유럽연합 또는 회원국 법이 데이터 주체의 동의로 특정 범주의 개인데이터에 대한 프로세싱의 금지를 해제할 수 있게 허용된 상태에서 데이터 주체가 목적이 정해진 프로세싱에 동의한 경우 또는 유럽연합 또는 회원국 법에 근거하여 넓은 의미의 비례성원칙을 충족하는 중대한 공익을 위해 프로세싱이 필요한 경우에 데이터 주체의 권리 및 자유와 정당한 이익을 보호하는 적합한 조치가 마련되어 있다면 그 결정이 특정 범주의 개인데이터에 근거한 경우라고 할지라도 그 결정을 따르지 않을 수 있는 권리는 인정되지 않는다.

또한 머신러닝 알고리즘 연구자들은 알고리즘 모델을 생성하는 과정에서 민감한 데이터를 이용하는 것이 모델의 정확성을 향상시킬 뿐만 아니라 모델 자체가 비차별적인 것이 되도록 하는 데에도 기여할 수 있다고 주장하기도 한다.¹²⁰⁾ 그렇다면 차별의 관점에서 민감한 데이터의 처리를 제한하는 것은 비차별적인 머신러닝 알고리즘 모델을 생성하는 데에 제약 사항이 될 수도 있다. 다만, 머신러닝 알고리즘의 정확도와 비차별적 모델 생성을 위해 민감한 데이터의 사용이 필요하다고 주장하는 경우에도 실제로 구체적 사안에 모델을 적용하는 단계에서 민감한 데이터를 직접 사용하는 것까지 용인하지는 않는다.

IV. 머신러닝 알고리즘에 대한 설명의무와 설명청구권

1. 자동화된 결정에 구속되지 않을 권리 실현의 전제

데이터 주체가 자기에겐 아무리 불리한 법적 결정이나 중대한 결정이 머신러닝 알고리즘에 의해 자동적으로 생성되었다고 해도 그러한 결정에 도달하기 위해 머신러닝 알고리즘을 사용했다는 사실 자체를 알지 못하거나, 설혹 그 사실을 알았다고

120) Indrė Žliobaitė · Bart Custers, “Using Sensitive Personal Data May Be Necessary for Avoiding Discrimination in Data-Driven Decision Models”, *Artificial Intelligence and Law* 24(2), 2016, pp. 183~201.



하더라도 도대체 머신러닝 알고리즘이 어떤 논리에 따라 그와 같은 결정을 산출해 냈는지 알지 못한다면 정작 그 결정에 대한 이의제기를 하려는 모든 시도는 무력화 된다. 따라서 자동적인 프로세싱이 있었다는 사실과 그 처리 과정이 어떤 규칙에 따랐는지 고지 받고 그 내용을 이해하는 것은 그로부터 생성된 결정의 구속력을 거부하기 위한 권리를 주장에 필수적 요건이 된다. 그리고 과연 이러한 필수적 요건을 GDPR이 법적으로 보장하고 있는지 여부는 차별금지법의 규정 방식을 채택하고 있는 제22조 제4항에 따라 머신러닝 알고리즘의 차별적 결정에 대해 구속되지 않고 이의제기할 수 있는 권리를 데이터 주체가 실제로 행사할 수 있는 것인지 여부를 결정하기도 한다. 비유적으로 말하자면 적어도 오로지 자동화된 결정 시스템에 의거해 내려진 판단에 대해서는 ‘무지의 베일(the veil of ignorance)’¹²¹⁾을 씌우는 것이 공정성을 보장하기 위한 보조적 단순화 장치가 아니라 공정성에 대한 의심을 불러 일으킬 수 있는 복잡성과 비밀성을 보장하는 방해 장치가 되는 셈이다.

2. 모델 중심의 설명과 주체 중심의 설명

머신러닝 시스템에 대해 설명을 요구하는 것 자체는 새로운 대응 방안이 아닐 수 있지만 결정지원시스템을 이용하는 사람을 위한 설명에서 데이터 주체를 위한 설명으로 강조점이 이동했다는 점은 새로운 변화일 수 있다. 에드워즈(L. Edwards)와 벨(M. Veale)은 설명의 중심을 알고리즘 모델에 두는지 아니면 데이터 주체에 두는지에 따라 ‘모델 중심 설명(model-centric explanations)’과 ‘주체 중심 설명(subject-centric explanations)’으로 구분한다.¹²²⁾ 모델 중심 설명은 알고리즘 모델에 관한 폭넓은 정보를 ‘전반적으로’ 제공하는 반면, 주체 중심 설명은 데이터 세트를 둘러싼 제한된 정보를 ‘부분적으로’ 제공한다.

(1) 모델 중심 설명

모델 중심 설명의 대상은 모델의 환경설정 정보, 훈련용 메타데이터, 성능 지표, 추정 전체 논리(estimated global logic), 프로세스 정보 등이다. 환경설정 정보는 모델링 과정의 의도, 신경망, 랜덤 포레스트, 앙상블 조합과 같은 모델군(群), 훈련 전 추가로

121) ‘무지의 베일’에 관해서 John Rawls, *A Theory of Justice*, Original ed., Belknap Press of Harvard University Press, 2005, p. 12 및 pp. 136~142; 본 논문 「제2장 제2절 II. 4. 모델 정립을 위한 단순화와 인간의 사고실험」 참조.

122) Lilian Edwards · Michael Veale, “Slave to the Algorithm? Why a ‘Right to an Explanation’ Is Probably Not the Remedy You Are Looking For”, *Duke Law & Technology Review* 16(1), 2017/2018, pp. 18~84: pp. 55~59.



지정하는 데 사용되는 매개 변수이다. 훈련용 메타데이터는 모델을 훈련시키는 데 사용된 입력 데이터의 요약 통계 및 질적 설명, 그 데이터의 출처, 이 모델에서 예상되는 출력 데이터 또는 분류이다. 성능 지표는 데이터의 눈에 띄는 특정 하위 카테고리의 성공과 같은 고장을 포함하여 보이지 않는 데이터에 대한 모델의 예측 기술에 대한 정보를 말한다. 추정 전체 논리는 인간이 이해할 수 있도록 단순화하고 평균화한 형태의 투입이 산출물로 바뀌는 것으로 정의가 완전하지 않거나 그렇지 않으면 복잡한 모형 대신에 그것을 사용하여 동일한 결과를 얻을 수 있는 것을 말한다. 여기에는 변수 가중치 점수, 규칙 추출 결과 또는 민감도 분석이 포함된다. 프로세스 정보는 원하지 않는 특성을 위해 시험되고, 훈련되고, 심사된 방법을 말한다.

(2) 주체 중심 설명

데이터의 변수에 따라 달라지는 차원의 조합 속에서 복잡성의 양상이 달라진다. ‘차원의 저주’¹²³⁾라고도 불리는 이런 상황에서 변수의 개수가 늘어나면서 차원의 개수도 늘어나고 그에 따라 조합되는 잠재적인 값은 기하급수적으로 증가하게 된다. 그렇기 때문에 주체 중심 설명은 변수가 적을 때 효과적이다. 반면에 변수가 많아질 경우 설명은 인간이 해석 가능한 영역을 쉽게 벗어나 버리게 된다. 그래서 주체 중심 설명에서 인간의 해석가능성 영역을 벗어나지 않도록 설명이 이루어져야 한다는 또 다른 조건이 추가된다. 이 조건이 목표하는 바는 인간이 해석할 수 있도록 설명 시스템을 최적화하는 것이다. 주체 중심 설명은 기초로 삼는 요소에 따라 민감성 기반의(sensitivity-based) 설명, 사례 기반의(case-based) 설명, 인구학 기반의(demographic-based) 설명, 성능 기반의(performance-based) 설명으로 구분하기도 한다. 민감성을 기초로 한 설명은 입력 데이터의 변화에 주목하며, 사례를 기초로 한 설명은 모델의 훈련에 사용된 데이터 기록에 주목하고, 인구학을 기초로 한 설명은 자신과 동일한 대우를 받은 개인들의 특징에 주목하며, 성능에 기초를 둔 설명은 자신과 근접한 위치에 분류된 개인들이 잘못 분류된 것은 아닌지 그 결과에 대한 확실성에 주목한다.

3. 사전 설명과 사후 설명

또한 설명이 언제 이루어질 것을 요구하는지 그 시점이 문제된다. 데이터 처리 과정에 대한 설명은 관리자가 데이터를 처리하기 이전에 이루어질 수도 있지만,

123) 이에 관해서 Bellman, Richard E., *Adaptive Control Processes*, Princeton University Press, 1961 및 본 논문 「제4장 제4절 I. 3. 복잡한 머신러닝 알고리즘의 작동방식과 불투명성」 참조.



데이터를 처리하여 그 결과가 나온 이후에 그러니까 구체적인 의사결정을 한 이후에도 이루어질 수 있다. 그런데 데이터를 처리하여 구체적인 의사결정을 한 경우 그 의사결정에 대한 설명은 시기적으로 데이터 처리 과정이 종료된 이후에 이루어질 수밖에 없다. 물론 데이터를 처리하기 이전에 그 결과를 예상할 수 있으므로 예상되는 의사결정, 즉 미래에 지향된 의사결정을 사전에 설명할 수는 있을 것이다. 이렇게 예상된 의사결정과 구체적인 의사결정을 구별하는 것은 머신러닝 알고리즘 시스템을 본질적으로 확률론에 의거한 것으로 취급할 때에 특별한 의미를 갖게 된다. 예를 들어 바흐터(S. Wachter) 및 그 동료들은 “복잡한 확률론의 분석을 사용 (the use of complex probabilistic analytics)”¹²⁴⁾하는 것이 구체적 결정에 대한 설명을 저해하는 요인이 된다고 본다. 그렇기 때문에 구체적 결정을 사전에 설명하는 것은 논리적으로 불가능하고, 사후에 설명하는 것은 대단히 복잡한 문제라 설명에 관한 권리의 대상이 되기 어려운 것으로 취급한다. 그러나 대부분의 시스템에서는 시스템 수준에서만 완벽한 설명이 이루어져도 구체적 사례에 대한 모든 것을 알 수 있다고 보는 경우¹²⁵⁾ 구체적 결정에 대한 사후 설명의 필요성은 감경된다. 예를 들어 켈프스트(A. Selbst)와 파울스(J. Powles)는 현재 머신러닝 시스템이 생성하는 모델이 결정론에 의거한다고 보면서 시스템에 대해 설명하는 것만으로 충분하고 구체적 결정을 별도로 구분하는 것이 반드시 필요하지는 않다고 본다.¹²⁶⁾

4. ‘로직(logic)에 관한 의미 있는 정보’ 제공의무와 접근권

“프로파일링을 포함해서 제22조 제1항 및 제4항에 언급된 자동화된 결정의 존재, 적어도 이 경우에 포함된 로직에 관한 의미 있는 정보와 데이터 주체에 대한 프로세싱의 유의성 및 예상되는 결과”¹²⁷⁾라는 표현은 GDPR 전체에 걸쳐서 제13조(2)(f),

124) Sandra Wachter · Brent Mittelstadt · Luciano Floridi, “Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation”, *International Data Privacy Law* 7(2), 2017, pp. 76~99: p. 79.

125) Eric Horvitz, “On the Meaningful Understanding of the Logic of Automated Decision Making”, BCLT Privacy Law Forum, 24 March 2017, https://www.law.berkeley.edu/wp-content/uploads/2017/03/BCLT_Eric_Horvitz_March_2017.pdf.

126) Andrew D. Selbst · Julia Powles, “Meaningful Information and the Right to Explanation”, *International Data Privacy Law* 7(4), 2017, pp. 233~242: pp. 239~241.

127) Article 13(2)(f), 14(2)(g) and 15(1)(h) GDPR: “the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.”



제14조(2)(g) 및 제15조(1)(h) 이 세 조항에서 동일하게 사용된다. 그런데 맥락은 조금씩 차이가 있다. 먼저 제13조는 제1항에서 개인정보가 데이터 주체로부터 수집되는 경우에 컨트롤러가 데이터 주체에게 제공해야 하는 정보를 규정하면서¹²⁸⁾ 제2항에서 공정하고 투명한 프로세싱을 보장하기 위해 개인정보가 획득된 때에 데이터 주체에게 제공해야 하는 추가 정보를 규정하고 있다.¹²⁹⁾ 제14조는 제1항에서 개인정보가 데이터 주체로부터 획득되지 않은 경우에 컨트롤러가 데이터 주체에게 제공해야 하는 정보를 규정하면서¹³⁰⁾ 제2항에서 공정하고 투명한 프로세싱을 보장하기 위해 데이터 주체에게 제공해야 하는 추가 정보를 제시하고 있다.¹³¹⁾ 제13조 제2항과 제14조 제2항은 개인정보를 획득한 컨트롤러가 데이터 주체에게 일정한 정보를 제공해야 할 의무를 부담한다는 점을 규정한다는 점에서 차이가 없고, 다만 개인정보의 획득이 데이터 주체로부터 수집된 것인지 여부에서 차이가 있을 뿐이다. 제15조 제1항은 데이터 주체가 자신에 관한 개인정보가 처리되고 있는지 여부에 관해 컨트롤러로부터 확인 받을 권리를 규정하면서 이 경우에 데이터 주체가 갖는 접근권의 대상으로서 개인정보와 몇 가지 추가적인 정보를 실시하고 있다.¹³²⁾ “프로파일링을 포함해서 제22조 제1항 및 제4항에 언급된 자동화된 결정의 존재, 적어도 이 경우에 포함된 로직에 관한 의미 있는 정보와 데이터 주체에 대한 프로세싱의 유의성 및 예상되는 결과”는 제13조 제2항, 제14조 제2항 및 제15조 제1항에서 추가 정보로서 공통적으로 규정된 것이다. 다만, 제13조 제2항 및 제14조 제2항에서는 컨트롤러가 부담하는 정보 제공의무의 대상이 되는 정보인 반면, 제15조 제1항에서는 데이터 주체가 갖는 정보 접근권의 대상이 되는 정보라는 점에서 차이가 있다.

128) Article 13(1) GDPR: “Where personal data relating to a data subject are collected from the data subject, the controller shall, at the time when personal data are obtained, provide the data subject with all of the following information: (후략)…”

129) Article 13(2) GDPR: “In addition to the information referred to in paragraph 1, the controller shall, at the time when personal data are obtained, provide the data subject with the following further information necessary to ensure fair and transparent processing: (후략)…”

130) Article 14(1) GDPR: “Where personal data have not been obtained from the data subject, the controller shall provide the data subject with the following information: (후략)…”

131) Article 14(2) GDPR: “In addition to the information referred to in paragraph 1, the controller shall provide the data subject with the following information necessary to ensure fair and transparent processing in respect of the data subject: (후략)…”

132) Article 15(1) GDPR: “The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the following information: (후략)…”



5. 설명에 관한 권리 중심 접근의 한계

‘설명에 관한 권리’가 어떤 내용을 갖는지는 설명의 대상이 되는 머신러닝 알고리즘에 관한 정보를 어떤 시점에 어느 정도의 범위에서 컨트롤러로부터 확인할 수 있는지에 따라 결정된다. GDPR에서 구체적으로 ‘설명에 관한 권리’를 어떻게 규정하는지는 ‘프로세싱의 예상되는 결과’ 및 ‘로직에 관한 의미 있는 정보’에 대한 해석에 따라 달라질 수 있다. 먼저 ‘프로세싱의 예상되는 결과’는 자동화된 처리 과정을 거칠 경우 ‘예상되는’ 결과이지 데이터 주체에게 직접 효과를 미치는 구체적 결정은 아니라는 점에서 실제로 구체적 결정이 산출되기 이전에 제공되는 것으로 해석될 수 있다. 즉, 사전 설명으로 보는 것이다. 문제는 ‘의미 있는 정보’의 범위가 불확정적이어서 그 구체적 범위를 설정하는 방식에 따라 GDPR에서 규정한 ‘설명에 관한 권리’의 내용이 달라질 수 있다는 것이다. 프로세싱에 사용된 알고리즘 모델에 관한 일반사항만으로도 ‘의미 있는’ 정보가 될 수 있지만, 세부사항까지 모두 제공되어야 ‘의미 있는’ 정보가 될 수 있기 때문이다.

또한 로직에 관한 의미 있는 정보에 접근할 권리가 영업비밀이나 지적재산권 또는 소프트웨어에 관한 저작권 등과 충돌하는 경우를 상정하여 “프로파일링과 관련한 로직을 실행하는 알고리즘의 세부사항은 정보주체에게 공개되지 않을 것”¹³³⁾이라는 전망도 제시된다. 그러나 로직에 관한 정보를 제공하도록 특별히 규정한 것이 자동화된 결정에 인간이 구속되지 않을 권리 및 차별로부터 보호 받을 권리와 연결되어 있어서 제공되는 알고리즘에 관한 정보는 이러한 권리를 주장하여 이의를 제기할 수 있는 수준에 이르러야 한다는 점이 고려되면 알고리즘에 대한 접근권에 상대적으로 높은 비중이 부여될 수 있기 때문에 ‘알고리즘의 세부사항이 정보주체에게 공개되지 않을 것’이라고 단정할 수만은 없다.

여기에 덧붙여 설명의 수준이 문제될 수 있다. 설명의 상대가 데이터 주체라는 점에서 해당 알고리즘에 대한 이해도가 비교적 낮은 일반인을 대상으로 할 것인지,¹³⁴⁾ 실제 알고리즘에 대한 감사(audit)를 시행하는 주체는 데이터 주체의 위임을 받거나 공공기관의 위탁을 받은 전문가라는 점을 고려해 해당 알고리즘에 대한 이해도가 비교적 높은 전문가를 대상으로 할 것인지 결정해야 한다. 그런데 복잡한 알고리즘을 간단하고

133) 박노형·정명현, “EU GDPR상 프로파일링 규정의 법적 분석”, 안암법학 56, 2018, 283~315쪽: 303쪽.

134) 기술적 문맹(technological illiteracy)은 알고리즘을 불투명하게 하는 주요 원인 중 하나이다.

본 논문 「제4장 제4절 I. 2. 기술적 문맹 상태와 불투명성」 참조.



쉽게 설명하는 것만큼 어려운 일이 없으며, 머신러닝 알고리즘의 작동방식이 인간의 이해를 넘어서는 경우 전문가로 이해할 수 없는 경우도 있을 수 있다는 점 즉, 설명해도 이해할 수 없는 내용이 있을 수 있다는 점을 고려하면 설명이 데이터 주체를 보호하는 궁극적 해결책이 될 수 없다는 한계에 봉착하게 된다.

마지막으로 굿먼과 플렉스먼이 ‘설명에 관한 권리’의 근거로 GDPR 제15조 제1항을 제시한 것은 제15조 제1항이 데이터 주체의 권리를 규정했기 때문이다. 논의를 권리에만 한정하지 않는다면 제13조 제2항과 제14조 제2항을 근거로 ‘설명 의무 (obligation to explain)’도 함께 논의될 수 있을 것이다. 자동화된 프로세싱의 컨트롤러 같은 특정 책임자에게 설명의무를 부담시킬 경우 설명청구권을 보유한 권리주체가 권리를 행사하지 않더라도 의무부담자는 설명의무를 이행해야 한다는 점에서 설명에 관한 의무 중심 구성은 규범적 측면에서 권리 중심 구성과 차이가 있다. 특히 차별 금지의무의 한 유형으로서 적극적 조치를 권리에 상응하지 않는 의무로 구성할 경우 자동화된 결정이 차별 관련적일 때 그에 대한 설명의무는 적극적 조치의무로 구성될 수도 있다.¹³⁵⁾ 또한 권리와 결합된 의무 이외에 권리와 결합되지 않은 공공 의무로서 설명의무가 폭넓게 인정될 경우 설명을 제공 받는 기회가 확대될 수 있기 때문에 이를 바탕으로 차별적 결정에 대해 구속 받지 않고 이의제기할 수 있는 권리를 보다 실질적으로 보장하는 데에 기여할 수 있다.

135) 차별금지의무로서 적극적 조치에 관해서 본 논문 「제5장 제1절 III. 차별금지의무의 성격과 내용」 참조.



제6장

결론

정부와 국회, 기업, 그리고 언론의 관심 속에서 이른바 ‘4차 산업혁명론’은 유독 대한민국에서 단시간에 광범위하게 확산됐지만 정작 그 실체는 첨단 과학기술 용어들의 화려함 뒤에 불분명하게 가려져 있다. 인간의 육체적 노동을 대체하는 기계가 등장했을 때부터 이를 받아들이는 인간의 태도는 각각의 타당한 근거를 가진 대등한 관점들로 팽팽한 대립 관계를 형성해 왔다. 이러한 대립 관계의 한편에는 인류의 진보에 대한 열망이 담겨 있고, 다른 한편에는 인류의 멸망에 대한 공포가 담겨 있다. 그리고 과학기술을 대하는 기본적 관점에 따라 그 열망과 공포는 과학기술의 발전에 따르는 위험을 관리하는 국가의 헌법상 책무를 이해하는 입장에도 암암리에 반영되어 있다.

인공지능이나 로봇, 지능적 에이전트를 조종하고 제어하는 중심에는 알고리즘이 있다. 기술적이고 전문적인 기능을 수행하는 것처럼 보이는 알고리즘은 인간이 문제를 해결해 온 방법의 다른 이름이기도 하다. 머신러닝 알고리즘은 인간이 설계하지 않고도 알고리즘을 생성할 수 있다. 그리고 머신러닝 알고리즘이 생성하는 모델은 인간의 사고과정에서 흔히 사용되는 단순화를 가정한다. 인간의 인식과 판단 작용에서도 그 과정을 세세하게 드러내지 않는 이상 그 내부의 복잡성을 알 수 없는 것과 마찬가지로 단순화는 의사결정을 할 때 개입될 수 있는 선입견이나 편견을 ‘모델의 맹점’¹⁾ 속으로 들어가게 함으로써 지능적인 기계의 판단이 인간보다 객관적이고 합리적이며 공정한 것이라는 기대 또는 신뢰를 갖게 할 수 있다.

그런데 머신러닝 알고리즘에 의해 만들어진 결정 모델이 오늘날의 상식에 반하거나 규범적 직관에 맞지 않게 편향적으로 보이는 결정을 내릴 수 있다는 점은 실증적인 사례들로 나타나고 있다. 머신러닝 알고리즘 개발의 구심점이 되는 컴퓨터 과학 분야 연구자들이 그러한 차별을 증명하려는 연구를 시도하면서 머신러닝 알고리즘의 차별에 관한 논의가 본격화되었지만, 해당 분야의 비전문가라고 할지라도 일상적으로 노출

1) 이에 관해서 본 논문 「제2장 제2절 II. 3. 데이터 마이닝과 모델의 단순화」 참조.



되어 있는 검색엔진에 몇 가지 간단한 질문을 입력해 보는 것만으로도 머신러닝 알고리즘의 결정에 대해 어렵지 않게 차별의 의심을 던질 수 있다.

머신러닝의 기본적인 방식으로 지도학습이 어떻게 알고리즘을 생성하는지 그 과정을 살펴보는 것은 단순히 머신러닝 알고리즘의 작동에서 오류나 편향이 발생할 때 그에 대해 차별 개념이 적용될 것인지 대입해보는 차원에 그치지 않고, 머신러닝 알고리즘이 사회에서 무수히 생산해 낸 데이터를 통해 재현해 낸 사회의 구조가 어떤 방식으로 차별을 양산하는지 반사적 또는 반성적으로 살펴본다는 의미를 갖는다. 인간의 주관적 레이블링이 개입하지 않는 비지도학습이라고 할지라도 훈련용 데이터의 원천이 사회라면 인간이 사회의 집단을 구별하는 방식과 다른 관점에서 머신러닝 알고리즘이 집단을 구별하는 방식을 통해 인간이 이해하기 어렵거나 이해할 수 없는 차별 사유와 영역이 발생할 수 있다는 점은 머신러닝 알고리즘에 의한 결정에 잠재되어 있는 위험이기도 하다.

법학에서 차별은 헌법과 차별금지법 분야에서 집중적으로 다룬다. 대한민국에서 ‘대한민국헌법’이라는 명칭을 가진 헌법은 제정되어 있으나, ‘차별금지법’이라는 명칭을 가진 법률은 아직 제정되어 있지 않다. 하지만 차별금지 사유나 차별금지 영역과 관련된 개별 법률이나 법규정은 법체계 내부에 편입되어 있다. 그리고 그 이론적·실천적 출발을 이끄는 것은 ‘누구든지 성별·종교 또는 사회적 신분에 의하여 정치적·경제적·사회적·문화적 생활의 모든 영역에 있어서 차별을 받지 아니한다.’는 법명제이다. 차별을 중심으로 한 법체계를 차별법체계라고 부르고 그에 관한 법이론을 차별법이론이라고 한다면 이와 관련된 법체계와 법이론을 구성할 때 차별 개념은 핵심적인 지위를 획득하게 된다.

차별은 경험의 문제이면서 규범의 문제로 다루어질 수 있다. 특히 서술적 의미에서 세계와 사물을 구별하는 차별 개념은 머신러닝 알고리즘에 기반을 둔 에이전트가 환경을 지각하고 또 환경에 작용하기 위해 사물을 분류하고 예측하며 군집을 형성하는 과정에서 공통적으로 적용될 수 있다. 그러나 넓게 이해된 서술적 의미의 차별 개념을 법적 개념으로 그대로 수용할 것인지는 별개의 문제이다. 서술적 의미의 차별을 법과 불법으로 이진화한 법체계로 끌어들이기 때 사용되는 규범적 의미의 차별 개념은 모든 구별을 규범적으로 부당한 차별로 보게 될 경우 야기되는 문제들의 복잡성을 경감시킬 수 있기 때문이다.

일반화 또는 보편화 가능성을 염두에 둔 차별법체계에 관한 논의는 일관된 규범적 기초를 탐색하는 데에 지향되곤 한다. 하지만 복잡한 층위를 가지고 있는



차별 개념, 차별 사유와 영역을 중심으로 적용 범위가 형성되는 차별법체계 및 그와 연결된 헌법적 가치에 대한 다양한 규범적 해석 가능성은 이를 곤란하게 한다. 차별에 관한 법적 이해는 주로 평등과의 관계 설정에 바탕을 둔다. 그리고 평등 관념에 수반되는 비교는 차별을 구체적으로 인식하는 방편이기도 하다. 그런데 차별과 평등의 관계를 관념적인 반대 관계로 설정하는 것을 제쳐두고 역사적으로 축적된 차별 형식을 전통적인 평등의 관념과 대조해 보면 평등과 논리적 반대 관계에 있는 불평등과 차별이 형식적 측면이나 내용적 측면에서 부분적인 일치와 부분적인 불일치 속에서 관계를 형성하고 있음을 확인할 수 있다. 또한 평등뿐만 아니라 자유, 존엄성, 사회통합의 관점에서 차별을 이해하려는 시도들은 차별법체계의 통일적 기초를 마련하는 것이 과연 가능한 것인지에 대한 의문을 남긴다.

시대와 지역적 상황에 맞추어 차별에 대응하여 형성된 차별금지법의 특수성을 고려할 때 차별법체계의 차별 개념을 하나의 기준에 입각해 통일되게 정의하는 것은 어려운 문제이다. 다만, 편견이나 고정관념, 간접차별, 부작위에 의한 차별 등 차별의 여러 형식 중에서 중요한 기능을 담당하는 것이 프락시(proxy) 즉, 대용물이라는 점은 확인할 수 있다. 대용물은 어떤 특성이나 지표를 다른 특성이나 지표로 지시하거나 대체한다. 그동안 지식의 원천이었던 인류의 진화나 경험, 문화에 더해 현대 사회에서 컴퓨터는 새로운 정보나 지식을 산출하는 기반이 되고 있다. 빅데이터 환경에서 지식을 산출하는 것이 더욱 용이해진 머신러닝 알고리즘에서도 대용물은 데이터세트의 특성을 추론하는 규칙을 생성할 때 중요한 기능을 수행한다. 머신러닝 알고리즘이 산출하는 새로운 정보와 지식에는 인간이 쉽게 파악할 수 없는 특성이 있을 수 있고, 이러한 특성은 사회적으로 의미 있는 집단을 범주화하여 구별하는 기준이 될 수 있다는 점에서 규범적인 의미의 차별 개념으로 포착될 수 있는 긴장 관계에 놓일 수 있다.

머신러닝 알고리즘에 의한 결정이 차별과 관련하여 야기하는 긴장 관계는 두 가지 차원의 함의를 갖는다. 한편으로 머신러닝 알고리즘이 산출한 정보와 지식이 차별에 관한 사회의 의미체계를 형성하는 차별법체계와 상충할 수 있다는 점에서 머신러닝 알고리즘에 의한 결정이나 판단은 규제 대상이 될 수 있지만, 다른 한편으로는 그 정보와 지식은 비가시적인 사회의 심층적인 차별의 구조를 드러낼 수 있다는 점에서 차별법체계에 오히려 부합하는 측면이 있다. 이러한 대조적인 함의는



데이터 분석 단계에서 머신러닝 알고리즘을 활용하는 것과 분석을 통해 머신러닝 알고리즘이 생성한 모델을 활용하는 것을 단계적으로 구분해서 살펴보도록 하는 이유가 되기도 한다.

특히 사람에 관한 데이터는 이미 법체계에서 작동하고 있는 개인정보 보호법 체계의 규율 대상이기 때문에 개인정보자기결정권으로 개별화된 데이터 주체의 권리는 데이터처리에 대한 통제의 헌법적 근거가 된다. 머신러닝 알고리즘은 프로파일링을 통해 개인정보를 추론할 수 있고 그 중에는 차별법체계에서 판단의 사유로 이용하지 않도록 정한 특성도 포함될 수 있다. 특히 민감한 정보로 분류되는 개인 정보는 차별금지법체계에서 분류, 배제, 우대 등의 금지 사유로 삼는 집단의 특성과 상당부분 일치하는 측면이 있다. 머신러닝 알고리즘은 직접 데이터주체로부터 개인정보를 수집하지 않고도 개인정보를 추론하거나 집단적 정보를 다룸으로써 개인정보 보호법 체계의 근간이 되는 개인정보처리에 관한 정보주체의 동의를 쉽게 우회할 수 있다.

머신러닝 알고리즘이 집단적 정보를 처리하는 것에 대해 차별금지법체제로 규율할 때에도 의도나 동기를 차별에 대한 책임의 근간으로 삼을 경우 머신러닝 알고리즘 에이전트 자체가 어떤 의도나 동기를 가지고 있다고 보기 어려울 뿐만 아니라 있다고 보더라도 그 의도나 동기를 확인하는 것이 가능할지도 불확실하다. 특히 의도나 동기를 확인하는 것의 어려움은 인간 에이전트의 차별을 다룰 때에도 마찬가지이다. 따라서 결과를 중심에 두거나 결과를 고려하는 차별법이론 구성이 그 흠결을 보충할 수 있을 것이라고 기대해 볼 수 있다. 그러나 머신러닝 알고리즘이 데이터를 기반으로 알고리즘을 생성한다는 점은 합리성의 레이블이 달린 통계적 근거를 기반으로 하고 있다는 것을 의미하기 때문에 최종적인 차별 판단에서 쉽게 정당화될 수 있다는 점은 대안적 이론구성의 의미를 퇴색시킨다.

머신러닝 알고리즘이 가장 효율적인 알고리즘을 찾는 과정은 비용을 최소화 하면서 성능을 최대화하는 것이기도 하다. 이때 비용은 계산의 복잡도와 소요 시간으로 나타난다. 사회의 문제를 해결하는 방법으로서 법 역시 추상적 수준에서 그 복잡성이 높아질 경우 이를 사례에 구체화하는 과정에서 복잡성은 더 증가하게 된다. 복잡성의 증가와 시간의 지연을 막기 위해서는 문제 해결을 위한 단순화된 모델을 적용하거나 어느 지점에서는 세부적 논증을 중단해야 한다. 가치나 이익을 사이에 둔 복잡한 형량의 과정은 이와 같은 단순화나 논증의 중단을 통해 비로소 결론에 도달할 수 있게 된다.



단순화나 논증중단 절차에서 숫자화를 통해 가치나 이익의 비중을 분류하는 것은 숫자를 비중의 대용물로 사용하고, 비중을 가치나 이익의 대용물로 사용하며, 가치나 이익을 그 향유 집단의 대표적 특성에 대한 대용물로 사용하는 먼 경로를 통해 잠재적 차별 가능성과 연결된다. 머신러닝 알고리즘의 작동방식은 인간 에이전트가 차별을 정당화할 때에 결정자로서 판단을 위해 직업적으로 학습한 법률이나 판례가 이를 형성해 온 정치인이나 법률가의 선입견이나 무의식적 편향에서 비롯될 수 있고, 그 토대는 수많은 시민의 의지나 쉽게 포착되거나 합의되기 어려운 사회의 위계적 구조로부터 생성된 무수한 양의 합리적인 정보와 지식일 수 있다는 점을 시사한다. 이와 같은 맥락에서 차별은 과연 어떤 에이전트의 판단과 결정이 합리적이고 정당한 것인지에 대해 자기반성적 고려를 지속하게 하는 계기를 마련해 준다. 나아가 법체계에 수용된 차별 개념은 법체계 스스로의 합리성과 정당성을 성찰적으로 갱신할 수 있는 기회를 제공한다.

평등원칙을 논증부담 또는 입증부담의 원칙으로 재해석하는 절차주의적 이해에 따르면 불평등으로서 차별을 정당화하는 측이 논증책임 또는 입증책임을 부담하게 된다. 이때 책임의 주체는 차별의 가해자를 국가와 사인으로 구분하여 각각 책임의 여부나 정도를 분담시킬 수도 있지만 머신러닝 알고리즘은 그 범용성으로 인해 이용자의 측면에서 국가와 사인을 구별하지 않는다. 오히려 사이버-물리적 시스템 환경에서 그 중심이 되는 알고리즘을 둘러싼 권력 관계를 고려하여 알고리즘의 설계자, 심사자, 감독자로 책임의 주체를 재구성하고, 알고리즘의 제어를 받는 시스템의 운영자와 차별적 결정을 직접 실행한 알고리즘 에이전트에 대한 책임 부담도 시도해 볼 수 있다.

알고리즘 에이전트를 법적 주체로 인정할 것인지 여부는 인공지능의 중흥기를 맞으면서 자연인을 중심으로 법인에게도 법인격을 부여하는 전통적 법체계나 법이론에 대해 제기되는 근본적인 문제로서 차별법체계나 차별법이론에만 국한되지 않는다. 차별의 행위 중심 구성이나 결과 중심 구성 역시 차별에 관한 개념이나 법체계 구성의 특수성에도 불구하고 그 기저에는 법 개념이나 법체계의 목적이나 기능에 대한 법철학적 또는 도덕철학적 입장의 목시적 차이가 있다. 이러한 입장 차이는 인간이 아닌 머신러닝 알고리즘 에이전트의 규범적 지위를 구성하는 데에도 영향을 미쳐 그 차별적 결정에 대한 책임 분담의 내용을 달라지게 할 수 있다.



알고리즘을 설계하는 것은 직관에 따른 문제 해결 방법을 구체적이고 명시적으로 표현한다는 측면에서 객관성의 확보와 밀접한 관련을 맺는다. 그런데 머신러닝 알고리즘은 법적인 이유나 기술적인 이유로 비밀에 부쳐질 수 있고, 머신러닝 알고리즘의 작동 결과를 이해하는 인간 능력의 한계까지 더해지면 머신러닝 알고리즘에 의한 판단이나 결정이 도대체 어떤 과정과 절차를 거쳐 산출된 것인지 알기 어려운 불투명한 상태에 놓이게 된다. 이는 마치 자신에게 적용될 법률의 제정 이유나 과정, 자신에게 집행될 행정처분이 부과된 이유나 이행절차 또는 자신에게 효력을 미치는 법적 결정이 내려진 이유나 근거를 알 수 없는 경우와 유사하다.

머신러닝 알고리즘의 투명성 문제에서 차별은 그 서술적 의미를 공유하는 구별을 실행하는 머신러닝 알고리즘을 이용해 내려진 판단이나 결정에 관한 이유나 과정을 설명해 줄 것을 요청하는 하나의 논거가 될 수 있다. 머신러닝 알고리즘이 어떻게 차별적 결정을 산출하게 됐는지 알 수 없다면 그에 대한 책임 소재를 밝히기도 어렵기 때문이다. 최근 시행된 유럽연합의 일반데이터보호법(GDPR, 일반개인정보 보호법)과 관련해서 해당 법이 자동화된 결정에 대해 설명을 요구하는 권리를 규정하고 있는지 그리고 그 대상으로 알고리즘도 포함되는 것인지 논의할 때 차별에 관한 법이론을 적용해 보는 것은 온라이프 사회를 염두에 둔 관련 법률의 제·개정과 해석에서 참고할 만한 예시가 될 수 있을 것이다.





참고문헌

단행본

- 계희열, 헌법학(중), 박영사, 2007.
- 김도균·최병조·최종고, 법치주의의 기초: 역사와 이념, 서울대학교출판부, 2006.
- 김소영·김우재·김태호·남궁석·홍기빈, 4차 산업혁명이라는 유령: 우리는 왜 4차 산업혁명에 열광하는가, Humanist, 2017.
- 김의중, (알고리즘으로 배우는) 인공지능, 머신러닝, 딥러닝 입문, 위키북스, 2016.
- 김지연·박주영·박정호·김재인·김태환, 머신러닝 기술의 이해: 기술사회학과 공학적 측면을 중심으로, 드림미디어, 2018.
- 박노형·고환경·구태언·김경환·박영우·이상직·이창범·정명현·최주선, EU 개인정보보호법-GDPR을 중심으로-, 박영사, 2017.
- 박호성, 평등론, 창작과비평사, 1995.
- 서울대 법과경제연구센터, 데이터 이코노미, 한스미디어, 2017.
- 서울대학교 기술과법센터, 과학기술과 법, 박영사, 2007.
- 선우현, 평등, 책세상, 2012, 20쪽.
- 송주영·송태민, 빅데이터를 활용한 범죄 예측, 황소걸음아카데미, 2018.
- 양천수, 빅데이터와 인권: 빅데이터와 인권의 실제적 조화를 위한 법정책적 방안, 영남대학교출판부, 2016.
- 유네스코한국위원회(편), 과학기술과 인권, 당대, 2001.
- 이광근, 컴퓨터 과학이 여는 세계: 세상을 바꾼 컴퓨터, 소프트웨어의 원천 아이디어 그리고 미래, 인사이트, 2015.
- 이광석, 데이터 사회 비판, 책읽는수요일, 2017.
- 이대열, 지능의 탄생: RNA에서 인공지능까지, 바다출판사, 2017.
- 이준일, 차별금지법, 고려대학교출판부, 2007.
- _____, 차별없는 세상과 법, 홍문사, 2012.
- _____, 감시와 법, 고려대학교출판부, 2014.
- _____, 헌법학강의, 제6판, 홍문사, 2015.
- _____, 인권법, 제7판, 홍문사, 2017.



- 이준원, 법 논리학, 동방문화사, 2017.
- 임흥빈, 기술문명과 철학, 문예출판사, 1995.
- 장애인법연구회, 장애인 차별 금지법 해설서, 나남, 2017.
- 정인섭(편), 사회적 차별과 법의 지배, 박영사, 2004.
- 조순경 · 한승희 · 정형옥 · 정경아 · 김선옥, 간접차별의 이론과 여성노동의 현실, 푸른사상, 2007.
- 책갈피 편집부(편), 계급, 소외, 차별, 책갈피, 2017.
- 한국분석철학회(편), 합리성의 철학적 이해, 철학과현실사, 1998.
- 한국포스트휴먼학회(편저), 포스트휴먼 시대의 휴먼, 아카넷, 2016.
- 홍성준, 고객을 유혹하고 기업을 성장으로 이끄는 차별화의 법칙, 21세기북스, 2013.
- 東浩紀, 一般意志 2.0 ルソー, フロイト, グーグル, 講談社, 2011; 안천(역), 일반의지 2.0: 루소, 프로이트, 구글, 현실문화, 2012; Azuma Hiroki, *General Will 2.0: Rousseau, Freud, Google*, John Person · Naoki Matsuyama(Trans.), Vertical, 2014.
- 莊子, 조현숙(역), 장자(莊子), 책세상, 2016.
- Alexy, Robert, *Theorie der Juristischen Argumentation: Die Theorie des Rationalen Diskurses als Theorie der Juristischen Begründung*, 3. Aufl., Suhrkamp, 1996[1., 1978]; Ruth Adler · Neil MacCormick(Trans.), *A Theory of Legal Argumentation*, Clarendon Press, 1989; 변종필 · 최희수 · 박달현(역), 법적 논증 이론: 법적 근거제시 이론으로서의 합리적 논증대화 이론, 고려대학교출판부, 2007.
- _____, *Theorie der Grundrechte*, 1. Aufl., Suhrkamp, 1994[1., 1985]; Julian Rivers(Trans.), *A Theory of Constitutional Rights*, Oxford University Press, 2002; 이준일(역), 기본권이론, 한길사, 2007.
- _____, *Begriff und Geltung des Rechts*, 3. Aufl., Verlag Karl Alber, 2002[1., 1992]; Stanley L. Paulson · Bonnie Litschewski Paulson(Trans.), *The Argument from Injustice: A Reply to Legal Positivism*, Clarendon Press; Oxford University Press, 2002; 이준일(역), 법의 개념과 효력, 고려대학교출판부, 2007[지산, 2000].
- Anscombe, Gertrude Elizabeth Margaret, *Intention*, 2nd ed., Harvard University Press, 2000[1st, Basil Blackwell, 1957].
- Aristoteles, 정치학[*Politika*], 천병희(역), 제2판, 숲, 2013.
- Asimov, Isaac, *I, Robot*, Bantam Books, 2004[1st, 1950]; 김옥수(역), 아이 로봇, 우리교육, 2008.
- Ball, Kirstie · Kevin D. Haggerty · David Lyon(Eds.), *Routledge Handbook of Surveillance Studies*, Routledge, 2012.
- Bamforth, Nicholas · Peter Leyland(Eds.), *Accountability in the Contemporary Constitution*, Oxford University Press, 2014.
- Barak, Aharon, *Proportionality: Constitutional Rights and Their Limitations*, Cambridge University Press, 2012.
- Barrat, James, *Our Final Invention: Artificial Intelligence and the End of the Human Era*, Thomas Dunne Books, 2013; 정지훈(역), 파이널 인벤션: 인공지능 인류 최후의 발명, 동아시아, 2016.
- Bayertz, Kurt(Hrsg.), *Verantwortung: Prinzip oder Problem?*, WBG(Wissenschaftliche Buchgesellschaft), 1995.



- Beck, Ulrich, *World at Risk* [Weltrisikogesellschaft, 2007], Ciaran Cronin(Trans.), Polity Press, 2009; 박미애 · 이진우(역), 글로벌 위험사회, 길, 2010.
- Bellman, Richard E., *Adaptive Control Processes*, Princeton University Press, 1961.
- Bengio, Yoshua, *Learning Deep Architectures for AI*, Foundations and Trends in Machine Learning 2(1), Now, 2009.
- Bhuta, Nehal · Susanne Beck · Robin Geiss · Claus Kress · Hin Yan Liu(Eds.), *Autonomous Weapons Systems: Law, Ethics, Policy*, Cambridge University Press, 2016.
- Bingham, Tom, *The Rule of Law*, Penguin, 2011[1st, Allen Lane, 2010]; 김기창(역), 법의 지배, 이음, 2013.
- Borking, John J. · B. M. A. van Eck · P. Siepel, *Intelligent Software Agents: Turning a Privacy Threat into a Privacy Protector*, Registratiekamer, 1999.
- Borowski, Martin(Ed.), *On the Nature of Legal Principles*, Stuttgart: Franz Steiner and Nomos, 2010.
- Bostrom, Nick, *Superintelligence: Paths, Dangers, Strategies*, 2014; 조성진(역), 슈퍼인텔리전스: 경로, 위험, 전략, 까치, 2017.
- Böckenforde, Ernst-Wolfgang, 헌법과 민주주의, 김효진 · 정태호(편역), 법문사, 2003.
- Braidotti, Rosi, *The Posthuman*, Polity Press, 2013; 이경란(역), 포스트휴먼, 아카넷, 2015.
- Brandom, Robert, *Making It Explicit: Reasoning, Representing and Discursive Commitment*, Harvard University Press, 1994.
- Brugger, Winfried · Ulfrid Neumann · Stephan Kirste, *Rechtsphilosophie im 21. Jahrhundert*, Suhrkamp, 2008.
- Brynjolfsson, Erik · Andrew McAfee, *The Second Machine Age: Work, Progress and Prosperity in a Time of Brilliant Technologies*, New York London: W.W. Norton & Company, 2016.
- Buchanan, James M. · Gordon Tullock, *The Calculus of Consent: Logical Foundations of Constitutional Democracy*, Liberty Fund, 1999.
- Burazin, Luka · Kenneth Einar Himma · Corrado Rovorsi(Eds.), *Law as an Artifact*, Oxford University Press, 2018.
- Calo, Ryan · Michael Froomkin · Ian Kerr(Eds.), *Robot Law*, Edward Elgar Publishing, 2016.
- Čapek, Karel, 로봇: 로숨의 유니버설 로봇[R. U. R.(Rossum's Universal Robots), 1920], 김희숙(역), 모비딕, 2015[1st, 길, 2002].
- Caro, Robert A., *The Power Broker: Robert Moses and the Fall of New York*, Vintage Books, 1975[1st, 1974].
- Chopra, Samir · Laurence F. White, *A Legal Theory for Autonomous Artificial Agents*, University of Michigan Press, 2011.
- Cohen, Julie E., *Configuring the Networked Self: Law, Code, and the Play of Everyday Practice*, Yale University Press, 2012.
- Cohen, Robert S. · Marx W. Wartofsky(Eds.), *Epistemology, Methodology and the Social Sciences*, Springer Netherlands, 1983.



- Collins, Hugh · Tarunabh Khaitan(Eds.), *Foundations of Indirect Discrimination Law*, Hart Publishing, an imprint of Bloomsbury Publishing Plc, 2018.
- Cormen, Thomas H. · Charles E. Leiserson · Ronald Rivest · Clifford Stein, *Introduction to Algorithms*, 3rd ed., MIT Press, 2009.
- Custers, Bart · Toon Calders · Bart Schermer · Tal Z. Zarsky(Eds.), *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, Springer, 2013.
- Dicey, Albert Venn, *Introduction to the Study of the Law of the Constitution*, E. C. S. Wade(Ed.), 10th ed., 1959[1st, 1885]; 안경환 · 김중철(역), 헌법학입문, 경세원, 1999.
- Delacampagne, Christian, 인종차별의 역사[*Une Histoire du Racisme*, 2000], 하정희(역), 예지, 2013.
- Domingos, Pedro, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, New York: Basic Books, a member of the Perseus Books Group, 2015; 강형진(역), 마스터 알고리즘: 머신러닝은 우리의 미래를 어떻게 바꾸는가, 비즈니스북스, 2016.
- Durante, Massimo, *Ethics, Law and the Politics of Information: A Guide to the Philosophy of Luciano Floridi*, Springer Berlin Heidelberg, 2017.
- Dworkin, Ronald, *Taking Rights Seriously*, Cambridge, Mass: Harvard University Press, 1977; 염수균(역), 법과 권리, 한길사, 2010.
- _____, *Sovereign Virtue: The Theory and Practice of Equality*, Cambridge, Mass: Harvard University Press, 2000; 염수균(역), 자유주의적 평등, 한길사, 2005.
- Dyzenhaus, David · Malcolm Thorburn(Eds.), *Philosophical Foundations of Constitutional Law*, Oxford University Press, 2016.
- Ellis, Evelyn · Philippa Watson, *EU Anti-Discrimination Law*, 2nd ed., Oxford University Press, 2012.
- Ely, John Hart, *Democracy and Distrust: A Theory of Judicial Review*, Harvard University Press, 1980; 전원열(역), 민주주의와 법원의 위헌심사, 나남, 2006.
- Eubanks, Virginia, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, St. Martin's Press, 2018.
- Fischer, John, Martin · Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge University Press, 2000.
- Flach, Peter A., *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*, Cambridge University Press, 2012; 최재영(역), 머신 러닝: 데이터를 이해하는 알고리즘의 예술과 과학, 비제이퍼블릭, 2016.
- Floridi, Luciano(Ed.), *The Onlife Manifesto: Being Human in a Hyperconnected Era*, Springer, 2015.
- Fredman, Sandra, *Discrimination Law*, 2nd ed., Oxford University Press, 2011.
- Frege, Gottlob, *Begriffsschrift und Andere Aufsätze*, Ignacio Angelelli(Hrsg.), 6. Nachdr. der 2. Aufl.(1964), Olms, 1993 [1., 1879]; 전응주(역), 개념표기: 수리학의 공식 언어를 본뜬 순수 사유의 공식 언어, 이제이북스, 2015.
- Friedman, Lawrence M., *Impact: How Law Affects Behavior*, Harvard University Press, 2016.



- Goldstein, Brett · Lauren Dyson(Eds.), *Open Data and the Future of Civic Innovation*, Code for America Press, 2013.
- Greenwald, Glenn, *No Place to Hide: Edward Snowden, the NSA and the U. S. Surveillance State*, Metropolitan Books/Henry Holt, 2014; 박수민 · 박산호(역), 더 이상 숨을 곳이 없다: 스노든, NSA, 그리고 감시국가, 모던타임스, 2014.
- Gutmann, Amy(Ed.), *Multiculturalism: Examining the Politics of Recognition*, Princeton University Press, 1994.
- Habermas, Jürgen, *Faktizität und Geltung: Beiträge zur Diskurstheorie des Rechts und des Demokratischen Rechtsstaats*, 4. Aufl., Suhrkamp, 1994[1., 1992]; Reh, William(Trans.), *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*, Polity Press, 1996; 한상진 · 박영도(역), 사실성과 타당성: 담론적 법이론과 민주적 법치국가 이론, 나남, 2007.
- _____, 인간이라는 자연의 미래[*Die Zukunft der menschlichen Natur: auf dem Weg zu einer liberalen Eugenik?*, 2001], 장운주(역), 나남, 2003.
- Hart, Herbert L. A., *The Concept of Law*, 3rd ed., Oxford University Press, 2012[1st, 1961]; 오병선(역), 법의 개념, 아카넷, 2001.
- Hassemer, Winfried · Ulfried Neumann · Frank Saliger(Hrsg.), *Einführung in Rechtsphilosophie und Rechtstheorie der Gegenwart*, 9. Aufl., C. F. Müller, 2016[1., 1976].
- Hastie, Trevor · Robert Tibshirani · J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed., Springer, 2009.
- Hegel, Georg Wilhelm Friedrich, *Phänomenologie des Geistes*[1., 1807], Eva Moldenhauer · Karl Markus Michel(Hrsg.), 2nd ed., Suhrkamp, 1989[1., 1986]; 임석진(역), 정신현상학 1, 2, 한길사, 2004.
- Heidegger, Martin, *Der Technik und die Kehre*, 8. Aufl., Neske, 1991[1., 1962]; 이기상(역), 기술과 전향, 서광사, 1993.
- _____, *Vorträge und Aufsätze*, Verlag Günther Neske, 2000; 이기상 · 신상희 · 박찬국(역), 강연과 논문, 이학사, 2008.
- Hellman, Deborah, *When Is Discrimination Wrong?*, Cambridge, Mass: Harvard University Press, 2008; 김대근(역), 차별이란 무엇인가: 차별은 언제 나쁘고 언제 그렇지 않은가, 서해문집, 2016.
- Hellman, Deborah · Sophia Reibetanz Moreau(Eds.), *Philosophical Foundations of Discrimination Law*, Oxford University Press, 2013.
- Hesse, Konrad, 헌법의 기초이론, 계획열(역), 박영사, 2001.
- Hildebrandt, Mireille · Katja de Vries(Eds.), *Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology*, Routledge, 2013.
- Hobbes, Thomas, *Leviathan*, J. C. A. Gaskin(Ed.), Oxford University Press, 1998[1st, 1651]; 진석용(역), 리바이어던 1, 2, 나남, 2008.
- Hoerster, Nobert, *Was ist Recht?: Grundfragen der Rechtsphilosophie*, C. H. Beck, 2006; 윤재왕(역), 법이란 무엇인가?: 어느 법실증주의자가 쓴 법철학 입문, 세창출판사, 2009.
- Jackson, Vicki C. · Mark V. Tushnet(Eds.), *Proportionality: New Frontiers, New Challenges*, Cambridge University Press, 2017.



- Jellinek, Georg, *Allgemeine Staatslehre*, 3. Aufl., Verlag von Julius Springer, 1929[1., 1900]; 김효진(역), 일반 국가학, 법문사, 2005.
- Jonas, Hans, *Das Prinzip Verantwortung: Versuch Einer Ethik Für Die Technologische Zivilisation*, Insel-Verlag, 1979; 이진우(역), 책임의 원칙: 기술 시대의 생태학적 윤리, 서광사, 1994.
- _____, *Technik, Medizin und Ethik: zur Praxis des Prinzips Verantwortung*, Suhrkamp, 1987; 이유태(역), 기술 의학 윤리: 책임 원칙의 실천, 숲, 2005.
- Jones, Steven E, *Against Technology: From the Luddites to Neo-Luddism*, Taylor and Francis, 2013.
- Kant, Immanuel, *Critique of Pure Reason*[*Kritik der reinen Vernunft*, 1st, 1781], Paul Guyer · Allen W. Wood(Eds. and Trans.), Cambridge University Press, 1998; 백종현(역), 순수이성비판 1, 아카넷, 2006.
- _____, *Critique of Practical Reason*[*Kritik der praktischen Vernunft*, 1st, 1788], Mary J. Gregor(Ed. and Trans.), Revised ed., Cambridge University Press, 2015[1st, 1997]; 백종현(역), 실천이성비판, 아카넷, 2009.
- _____, *Die Metaphysik der Sitten*, 15. Aufl., Suhrkamp, 2009[1., 1797]; 백종현(역), 윤리형이상학, 아카넷, 2012.
- Kaplan, David M.(Ed.), *Readings in The Philosophy of Technology*, 2nd ed., Rowman & Littlefield Publishers, 2009.
- Kelsen, Hans, *Allgemeine Staatslehre*, Österreichische Staatsdruckerei, 1925; 민준기(역), 일반 국가학, 민음사, 1990.
- _____, *Allgemeine Theorie der Normen*, Kurt Ringhofer · Robert Walter(Hrsg.), Manz Verlags- und Universitätsbuch handlung, 1979; 김성룡(역), 규범의 일반이론 1, 2, 아카넷, 2016.
- _____, *Verteidigung der Demokratie*, Matthias Jestaedt · Oliver Lepsius(Hrsg.), Mohr Siebeck, 2006.
- Khaitan, Tarunabh, *A Theory of Discrimination Law*, Oxford University Press, 2016.
- Klatt, Matthias(Ed.), *Institutionalized Reason: The Jurisprudence of Robert Alexy*, Oxford University Press, 2012.
- Kurki, Visa A. J. · Tomasz Pietrzykowski(Eds.), *Legal Personhood: Animals, Artificial Intelligence and the Unborn*, Springer, 2017.
- Kurzweil, Ray, *The Singularity Is Near: When Humans Transcend Biology*, Viking, 2005; 김명남 · 장시형(역), 특이점이 온다: 기술이 인간을 초월하는 순간, 2007.
- Kymlicka, Will, *Contemporary Political Philosophy: An Introduction*, 2nd ed., Oxford University Press, 2002; 장동진 · 장휘 · 우정열 · 백성욱(역), 현대 정치철학의 이해, 개정판, 동명사, 2018[초판, 2006].
- Latour, Bruno, *Science in Action: How to Follow Scientists and Engineers through Society*, 11th print, Harvard University Press, 2003[1st, 1987].
- Lessig, Lawrence, *Code: And Other Laws of Cyberspace, Version 2.0*, 2nd Revised ed., New York: Basic Books, 2006; 김정오(역), 코드 2.0, 나남, 2009.
- Leavitt, David, 너무 많이 알았던 사람: 앨런 튜링과 컴퓨터의 발명[*The Man Who Knew Too Much*, 2007], 고종숙(역), 승산, 2008.
- Lippert-Rasmussen, Kasper(Ed.), *The Routledge Handbook of the Ethics of Discrimination*, Routledge, 2018.



- Liu, Huan · Hiroshi Motoda(Eds.), *Feature Extraction, Construction and Selection: A Data Mining Perspective*, Springer US, 1998.
- Lodder, Arno R. · Anja Oskamp(Eds.), *Information Technology and Lawyers: Advanced Technology in the Legal Domain, from Challenges to Daily Routine*, Springer, 2006.
- Luhmann, Niklas, *Recht und Automation in der öffentlichen Verwaltung: eine verwaltungswissenschaftliche Untersuchung*, 2. Aufl., Duncker & Humblot, 1997.
- _____, *Das Recht der Gesellschaft*, 8. Aufl., Frankfurt am Main: Suhrkamp, 2013[1995]; 윤재왕(역), *사회의 법*, 새물결, 2014.
- _____, *Einführung in die Systemtheorie*, 4. Aufl., Carl-Auer, 2008[1., 2002]; 윤재왕(역), *체계이론 입문*, 새물결, 2014.
- Lyon, David · Elia Zureik(Eds.), *Computers, Surveillance, and Privacy*, University of Minnesota Press, 1996.
- Lyon, David, *Surveillance after September 11*, Malden, Mass: Polity Press in association with Blackwell Pub. Inc, 2003; 이혁규(역), 9월 11일 이후의 감시, 울력, 2011.
- _____(Ed.), *Surveillance as Social Sorting: Privacy, Risk and Digital Discrimination*, Routledge, 2003.
- MacCormick, Neil · Ota Weinberger, *An Institutional Theory of Law: New Approaches to Legal Positivism*, Kluwer Academic Publishers, 1986.
- Marshall, Thomas H., *Citizenship and Social Class*, Cambridge University Press, 1950; 조성운(역), *시민권*, 나눔의집, 2014.
- Merat-Bruns, *Discrimination at Work: Comparing European, French and American Law*, Marie, Elaine Holt(Trans.), University of California Press, 2016.
- Mindell, David A., *Our Robots, Ourselves: Robotics and the Myths of Autonomy*, Viking, 2015.
- Mitchell, Tom, *Machine Learning*, McGraw-Hill, 1997.
- Möllers, Christoph, *The Three Branches: A Comparative Model of Separation of Powers*, Oxford University Press, 2013.
- Moravec, Hans, *Mind Children: The Future of Robot and Human Intelligence*, Harvard University Press, 1988; 박우석(역), *마음의 아이들: 로봇과 인공지능의 미래*, 김영사, 2011.
- Mulhall, Stephen · Adam Swift, *Liberals and Communitarians*, 2nd ed., Blackwell Publishing, 1996[1st, 1992]; 김해성 · 조영달(역), *자유주의와 공동체주의*, 한울 아카데미, 2001.
- Neumann, Ulfrid, *Recht als Struktur und Argumentation: Beiträge zur Theorie des Rechts und zur Wissenschaftstheorie der Rechtswissenschaft*, Nomos-Verl.-Ges, 2008; 윤재왕(역), *구조와 논증으로서의 법*, 세창출판사, 2013.
- Neumann, Ulfrid · Lorenz Schultz(Hrsg.), *Verantwortung in Recht und Moral*, Steiner, 2000.
- Newton-Smith, W. H. · K. V. Wilkes(Eds.), *Modelling the Mind*, Oxford University Press, 1990.
- Nilsson, Nils J., *Introduction to Machine Learning*, Stanford University, 1998.
- Noble, Safiya Umoja, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York University Press, 2018.



- Nozick, Robert, *Anarchy, State and Utopia*, Reprint ed., Blackwell, 2013[1st, 1974]; 남경희(역), *아나키에서 유토피아로: 자유주의 국가의 철학적 기초*, 문학과지성사, 1997.
- O'Neil, Cathy, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Allen Lane, Penguin Books, 2016; 김정혜(역), *대량살상 수학무기: 어떻게 빅데이터는 불평등을 확산하고 민주주의를 위협하는가*, 흐름출판, 2017.
- Ottmann, Thomas · Peter Widmayer, *Algorithmen und Datenstrukturen*, 6th ed., Springer Vieweg, 2017[1st, 1986].
- Pasquale, Frank, *The Black Box Society: The Secret Algorithms That Control Money and Information*, Harvard University Press, 2015; 이시은(역), *블랙박스 사회: 당신의 모든 것이 수집되고 있다, 돈과 빅데이터를 통제하는 정보 제국주의의 비밀*, 안티고네, 2016.
- Peters, William, *A Class Divided: Then and Now*, Expanded ed., Yale University Press, 1987; 김희경(역), *푸른 눈 갈색 눈: 세상을 놀라게 한 차별 수업 이야기*, 한겨레출판사, 2012.
- Platon, *국가[Politeia]*, 천병희(역), 숲, 2013.
- Posjman, Louis P. · Robert Westmoreland(Eds.), *Equality: Selected Readings*, Oxford University Press, 1997.
- Rahwan, Iyad · Guillermo R. Simari(Eds.), *Argumentation in Artificial Intelligence*, Springer, 2009.
- Rawls, John, *A Theory of Justice*, Original ed., Belknap Press of Harvard University Press, 2005[1st, 1971]; 황경식(역), *정의론*, 이학사, 2003.
- Ray, Larry J. · R. Andrew Sayer(Eds.), *Culture and Economy after the Cultural Turn*, SAGE, 1999.
- Russell, Stuart J. · Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed., Prentice Hall, 2010; 류광(역), *인공지능: 현대적 접근방식 1, 2*, 제이펍, 2016.
- Saliger, Frank, *라트브루흐 공식과 법치국가[Radbruchsche Formel und Rechtsstaat]*, C. F. Müller, 1995; 윤재왕(역), 제2판, 세창출판사, 2011.
- Saussure, Ferdinand de, *Course in General Linguistics[Cours de Linguistique Générale]*, 1916, Perry Meisel · Haun Saussy(Eds.), Wade Baskin(Trans.), Columbia University Press, 2011; 최승연(역), *일반언어학 강의*, 민음사, 2006.
- Schauer, Frederick, *Profiles, Probabilities, and Stereotypes*, Belknap Press of Harvard University Press, 2003.
- Schwartz, Steve, *A Brief History of Analytic Philosophy: From Russell to Rawls*, Wiley-Blackwell, 2012; 한상기(역), *분석철학의 역사: 러셀에서 롤스까지*, 서광사, 2017.
- Searle, John R., *Expression and Meaning: Studies in the Theory of Speech Acts*, Cambridge University Press, 1979.
- Sedgewick, Robert · Philippe Flajolet, *An Introduction to the Analysis of Algorithms*, 2nd ed., Addison-Wesley, 2013.
- Selden, Larry · Geoffrey Colvin, *회사를 먹여 살리는 착한고객[Angel Customers & Demon Customers: Discover Which Is Which and Turbo-Charge Your Stock]*, 2003, 황숙혜(역), 위즈덤하우스, 2010.
- Shannon, Claude E. · Warren Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, 1963[1st, 1949]; 백영민(역), *수학적 커뮤니케이션 이론*, 커뮤니케이션북스, 2016.
- Sheplyakova, Tatjana(Hrsg.), *Prozeduralisierung des Rechts*, Mohr Siebeck, 2018.



- Silver, Nate, *The Signal and the Noise: Why So Many Predictions Fail – But Some Don't*, Penguin Books, 2012;
이경식(역), 신호와 소음: 미래는 어떻게 당신 손에 잡히는가, 더퀘스트, 2014.
- Simon, Herbert A., *Models of Bounded Rationality*, Cambridge, Mass: MIT Press, 1982.
- Singer, Peter, *Practical Ethics*, 3rd ed., Cambridge University Press, 2011[1st, 1980]; 황경식 · 김성동(역), 실천윤리학, 연암서가, 2013.
- Solanke, Iyiola, *Discrimination as Stigma: A Theory of Anti-Discrimination Law*, Oxford; Portland, Oregon: Hart Publishing, 2017.
- Solove, Daniel J., *The Digital Person: Technology and Privacy in the Information Age*, New York University Press, 2004.
- Somek, Alexander, *Rationalität und Diskriminierung: zur Bindung der Gesetzgebung an das Gleichheitsrecht*, Springer, 2001.
- _____, *The Legal Relation: Legal Theory after Legal Positivism*, Cambridge University Press, 2017.
- Spencer-Brown, George, *Laws of Form*, US ed., Julian Press, 1972[UK ed., George Allen and Unwin, 1969].
- Stiller, Sebastian, *Planet der Algorithmen: ein Reiseführer*, München: Knaus, 2015; 김세나(역), 알고리즘 행성 여행자들을 위한 안내서-쇼핑부터 인공지능까지, 우리 삶을 움직이는 알고리즘에 관한 모든 것, 와이즈베리, 2017.
- Turner, Bryan, *Equality*, Ellis Horwood Limited and Tavistock Publications, 1986.
- Voigt, Paul · Axel von dem Bussche, *The EU General Data Protection Regulation (GDPR): A Practical Guide*, Springer, 2017.
- Wachsmuth, Ipke, *Menschen, Tiere und Max: Natürliche Kommunikation und Künstliche Intelligenz*, Springer Spektrum, 2013; 장병탁 · 최윤영(역), 커뮤니케이션: 인간, 동물, 인공지능, 서울대학교출판문화원, 2014.
- Waldron, Jeremy, *Dignity, Rank and Rights*, Meir Dan-Cohen(Ed.), Oxford University Press, 2012.
- Webster, Frank, *Theories of the Information Society*, 4th ed., Routledge, 2014; 조동기(역), 현대 정보사회이론, 나남, 2016.
- Whitaker, Reginald, *The End of Privacy: How Total Surveillance Is Becoming a Reality*, New Press: Distributed by W. W. Norton, 1999; 이명균 · 노명현(역), 개인의 죽음, 개정판, 생각의나무, 2007.
- Wiener, Norbert, *Cybernetics: or Control and Communication in the Animal and the Machine*, 2nd ed., MIT Press, 1961[1st, 1948].
- Winner, Langdon, *Autonomous Technology*, MIT Press, 1977; 강정인(역), 자율적 테크놀로지와 정치철학, 아카넷, 2000.
- _____, *The Whale and the Reactor: A Search for Limits in an Age of High Technology*, University of Chicago Press, 1986; 손화철(역), 길을 묻는 테크놀로지: 첨단 기술 시대의 한계를 찾아서, 씨아이알, 2010.
- Witten, Ian H. · Eibe Frank · Mark A. Hall · Christopher J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed., Elsevier, 2017; 이승현(역), 데이터 마이닝 Data mining: 데이터 속 숨은 의미를 찾는 기계 학습의 이론과 응용(제3판), 에이콘출판, 2013.
- Young, Iris Marion, *Justice and the Politics of Difference*, Princeton University Press, 1990; 김도균 · 조국(역), 차이의 정치와 정의, 모티브북, 2017.



논문

- 고인석, “아시모프의 로봇 3법칙 다시 보기”, 철학연구 93, 2011, 97~120쪽.
- 고학수, “인공지능 알고리즘과 시장”, 서울대 법과경제연구센터, 데이터 이코노미, 한스미디어, 2017, 11~38쪽.
- 권건보 · 이한주 · 김일환, “EU GDPR 제정 과정 및 그 이후 입법동향에 관한 연구”, 미국헌법연구 29(1), 2018, 1~38쪽.
- 김건우, “포스트휴먼의 개념적, 규범학적 의의”, 한국포스트휴먼학회(편자), 포스트휴먼 시대의 휴먼, 아카넷, 2016, 29~66쪽.
- _____, “로봇윤리 vs. 로봇법학: 따로 또 같이”, 법철학연구 20(2), 2017, 7~44쪽.
- 김광수, “인공지능 규제법 서설”, 토지공법연구 81, 2018, 279~310쪽.
- 김대환, “유럽연합법원(EuGH) 판결에 나타난 비례성원칙의 내용과 심사강도”, 세계헌법연구 17(3), 2011, 1~27쪽.
- 김민호 · 이규정 · 김현경, “지능정보사회의 규범설정 기본원칙에 대한 고찰”, 성균관법학 28(3), 2016, 293~320쪽.
- 김선희, “인공지능과 이해의 개념”, 인지과학 8(1), 1997, 37~56쪽.
- 김소영, “4차 산업혁명, 실체는 무엇인가?”, 김소영 · 김우재 · 김태호 · 남궁석 · 홍기빈, 4차 산업혁명이라는 유럽: 우리는 왜 4차 산업혁명에 열광하는가, Humanist, 2017, 11~26쪽.
- 김연식, “과학기술의 발달에 따른 탈인간적 법이론의 기초 놓기”, 법과 사회 53, 2016, 71~107쪽.
- 김하열, “민주주의 정치이론과 헌법원리”, 공법연구 39(1), 2010, 161~192쪽.
- 김희숙, “역자 후기: 로봇, 현대SF의 탄생”, Karel Čapek, 김희숙(역), 로봇 R. U. R, 모비딕, 2015, 191~214쪽.
- 김희연, “미국 국가과학기술위원회(NSTC)의 인공지능(AI) 기반 준비를 위한 권고안”, 정보통신방송정책 28(19), 2016, 12~19쪽.
- 남중권, “헌법의 몇 가지 법치주의 모델-개념과 구조-”, 법학연구 59(3), 2018, 1~34쪽.
- 박노형 · 정명현, “EU GDPR상 프로파일링 규정의 법적 분석”, 안암법학 56, 2018, 283~315쪽.
- 박상돈, “헌법상 자동의사결정 알고리즘 설명요구권에 관한 개괄적 고찰”, 헌법학연구 23(3), 2017, 185~218쪽.
- 손제연, “위상적 개념으로서의 인간존엄”, 법철학연구 21(1), 2018, 295~338쪽.
- 송석윤, “차별의 개념과 법의 지배”, 정인섭(편), 사회적 차별과 법의 지배, 박영사, 2004, 3~24쪽.
- 신기현, “한국의 전통 사상과 평등 인식”, 한국정치학회보 29(2), 1995, 407~430쪽.
- 심우민, “인공지능의 발전과 알고리즘의 규제적 속성”, 법과 사회 53, 2016, 41~70쪽.
- _____, “인공지능과 법패러다임 변화 가능성: 입법 실무 거버넌스에 대한 영향과 대응 과제를 중심으로”, 법과 사회 56, 2017, 351~385쪽.



- 안형준, “알고리즘 안에 내재된 사회적 차별: 빅데이터에 대한 미국 정부의 우려”, 과학기술정책 (214), 2016, 4~7쪽.
- 양종모, “인공지능 알고리즘의 편향성, 불투명성이 법적 의사결정에 미치는 영향 및 규율 방안”, 법조 66(3), 2017, 60~105쪽.
- _____, “인공지능에 의한 판사의 대체 가능성 고찰”, 홍익법학 19(1), 2018, 1~29쪽.
- 유승익, “인공지능의 추론방식을 활용한 법적 논증 - 폐기가능성, 비단조논리, 형량 - ”, 원광법학 32(2), 2016, 299~323쪽.
- 윤영미, “평등규범의 사인에 대한 적용”, 헌법학연구 19(3), 2013, 39~78쪽.
- 윤재왕, “개인주의적 절대주의: 토마스 홉스의 국가철학과 법철학에 관하여”, 원광법학 28(2), 2012, 7~35쪽.
- 이준일, “법학에서 최적화”, 법철학연구 3(1), 2000, 101~130쪽.
- _____, “소수자(Minority)와 평등원칙”, 헌법학연구 8(4), 2002, 219~243쪽.
- _____, “헌법상 비례성원칙”, 공법연구 37(4), 2009, 25~44쪽.
- _____, “차별, 소수자, 국가인권위원회”, 헌법학연구 18(2), 2012, 177~222쪽.
- _____, “헌법재판소의 평등심사기준과 국가인권위원회의 차별판단기준”, 세계헌법연구 18(2), 2012, 333~356쪽.
- 이 철, “스펜서브라운의 ‘재진입’과 그 과학철학적 의의 - $X+1=0$ 에 숨겨진 시간과 상상의 세계”, 사회사상과 문화 18(2), 2015, 111~137쪽.
- 정영기, “비단조 논리적 합리성”, 합리성의 철학적 이해, 한국분석철학회(편), 철학과현실사, 1998, 261~288쪽.
- 정준현 · 김민호, “지능정보사회와 헌법상 국가의 책무”, 법조 66(3), 2017, 106~145쪽.
- 정재연, “법페러다임 변화의 관점에서 인공지능과 법담론 : 법에서 탈근대성의 수용과 발전”, 법과 사회 53, 2016, 109~136쪽.
- 조병훈 · 박성빈, “컴퓨터 과학 교육: 웹 접근성 분석”, 한국컴퓨터교육학회 학술발표대회논문집 13(2), 2009, 243~246쪽.
- 조한상 · 이주희, “인공지능과 법, 그리고 논증”, 법과 정책연구 16(2), 2016, 295~320쪽.
- 차진아, “독일의 차별금지법 체계와 「일반적 평등대우법」의 역할”, 공법연구 40(1), 2011, 327~356쪽.
- 한수웅, “헌법 제37조 제2항의 과잉금지원칙의 의미와 적용범위”, 저스티스 (95), 2006, 5~28쪽.
- 허유선, “인공지능에 의한 차별과 그 책임 논의를 위한 예비적 고찰- 알고리즘의 편향성 학습과 인간 행위자를 중심으로 -”, 한국여성철학 29, 2018, 165~210쪽.
- 현윤경, “니콜라스 루만의 체계이론에서 ‘형식’ 개념의 수용과 응용”, 사회와이론 31, 2017, 211~251쪽.
- 홍성수, “인간이 없는 인권이론? - 루만의 체계이론과 인권 -”, 법철학연구 13(3), 2010, 251~280쪽.
- 홍성욱, “과학과 기술의 상호작용: 지식으로서의 기술과 실천으로서의 과학”, 창작과 비평 22(4), 1994, 329~350쪽.
- _____, “왜 ‘4차 산업혁명’이 문제인가?”, 김소영 · 김우재 · 김태호 · 남궁석 · 홍기빈, 4차 산업혁명이라는 유령: 우리는 왜 4차 산업혁명에 열광하는가, Humanist, 2017, 29~52쪽.



- Abbott, Russ, "Meaning, Autonomy, Symbolic Causality, and Free Will", *Review of General Psychology* 22(1), 2018, pp. 85~94.
- Alarie, Benjamin, "The Path of the Law: Towards Legal Singularity", *University of Toronto Law Journal* 66(4), 2016, pp. 443~455.
- Alarie, Benjamin · Anthony Niblett · Albert H. Yoon, "How Artificial Intelligence Will Affect the Practice of Law", *University of Toronto Law Journal* 68(1), 2018, pp. 106~124.
- Alexander, Larry, "What Makes Wrongful Discrimination Wrong? Biases, Preferences, Stereotypes and Proxies", *University of Pennsylvania Law Review* 141(1), 1992, pp. 149~219.
- Alexander, Larry · Kevin Cole, "Discrimination by Proxy", *Constitutional Commentary* 14(3), 1997, pp. 453~463.
- Alexy, Robert, "On Balancing and Subsumption", *Ratio Juris* 16(4), 2003, pp. 433~449.
- _____, "Two or Three?", in *On the Nature of Legal Principles*, Martin Borowski(ed.), Stuttgart: Franz Steiner and Nomos, 2010, pp. 9~18.
- _____, "Constitutional Rights and Proportionality", *Revus: Journal for Constitutional Theory and Philosophy of Law* 22, 2014, pp. 51~65.
- _____, "Proportionality and Rationality", in *Proportionality: New Frontiers, New Challenges*, Vicki C. Jackson · Mark V. Tushnet(Eds.), New York: Cambridge University Press, 2017, pp. 13~29.
- Allan, Trevor R. S., "Accountability to Law", in *Accountability in the Contemporary Constitution*, Nicholas Bamforth · Peter Leyland(Eds.), Oxford University Press, 2014, pp. 77~104.
- Arneson, Richard J., "What Is Wrongful Discrimination?", *San Diego Law Review* 43, 2006, pp. 775~808.
- _____, "Discrimination, Disparate Impact and Theories of Justice", in *Philosophical Foundations of Discrimination Law*, Hellman, Deborah · Sophia Reibetanz Moreau(Eds.), Oxford University Press, 2013, pp. 87~111.
- Appiah, K. Anthony, "Stereotypes and the Shaping of Identity", *California Law Review* 88(1), 2000, pp. 41~53.
- Ashley, Kevin D., "An AI Model of Case-Based Legal Argument from a Jurisprudential Viewpoint", *Artificial Intelligence and Law* 10(1~3), 2002, pp. 163~218.
- Ashley, Kevin D., "Case-Based Reasoning", in *Information Technology and Lawyers: Advanced Technology in the Legal Domain, from Challenges to Daily Routine*, Arno R. Lodder · Anja Oskamp(Eds.), Springer, 2006, pp. 223~260.
- Bagenstos, Samuel R., "'Rational Discrimination,' Accommodation and The Politics of (Disability) Civil Rights", *Virginia Law Review* 89(5), 2003, pp. 825~923.
- _____, "Implicit Bias, 'Science', and Antidiscrimination Law", *Harvard Law & Policy Review* 1(2), 2007, pp. 477~493.
- Bakan, Abbie, "마크스주의 차별론", 계급, 소외, 차별, 차승일(역), 책갈피 편집부(편), 책갈피, 2017, 85~121쪽.
- Baker, Aaron · Gavin Phillipson, "Policing, Profiling and Discrimination Law: US and European Approaches Compared", *Journal of Global Ethics* 7(1), 2011, pp. 105~124.



- Balkin, Jack M., "The Constitution in the National Surveillance State", *Minnesota Law Review* 93(1), 2008, pp 1~25.
- _____, "The Path of Robotics Law", *California Law Review Circuit* 6, 2015, pp. 45~60.
- _____, "Information Fiduciaries and the First Amendment", *U. C. Davis Law Review* 49(4), 2016, pp. 1183~1234.
- _____, "The Three Laws of Robotics in the Age of Big Data", *Ohio State Law Journal* 78(5), 2017, pp. 1217~1241.
- Bamberger, Kenneth A., "Technologies of Compliance: Risk and Regulation in a Digital Age", *Texas Law Review* 88, 2010, pp. 669~740.
- Barocas, Solon · Andrew D. Selbst, "Big Data's Disparate Impact", *California Law Review* 104, 2016, pp. 671~732.
- Bench-Capon, Trevor · Henry Prakken · Giovanni Sartor, "Argumentation in Legal Reasoning", in *Argumentation in Artificial Intelligence*, Iyad Rahwan · Guillermo R. Simari(Eds.), Springer, 2009, pp. 363~382.
- Bench-Capon, Trevor et al., "A History of AI and Law in 50 Papers: 25 Years of the International Conference on AI and Law", *Artificial Intelligence and Law* 20(3), 2012, pp. 215~319.
- Bennett, Colin J., "The Public Surveillance of Personal Data: A Cross-national Analysis", in *Computers, Surveillance, and Privacy*, David Lyon · Elia Zureik(Eds.), University of Minnesota Press, 1996, pp. 237~259.
- Bringsjord, Selmer · Paul Bello · David Ferrucci, "Creativity, the Turing Test and the (Better) Lovelace Test", *Minds and Machines* 11(1), 2001, pp. 3~27.
- Brożek, Bartosz, "Legal Rules and Principles: A Theory Revisited", *i-Lex* 7(17), 2012, pp. 205~226.
- _____, "The Troublesome 'Person'", in *Legal Personhood: Animals, Artificial Intelligence and the Unborn*, Kurki, Visa A. J. · Tomasz Pietrzykowski(Eds.), Springer, 2017, pp. 3~13.
- Brożek, Bartosz · Marek Jakubiec, "On the Legal Responsibility of Autonomous Machines", *Artificial Intelligence and Law* 25(3), 2017, pp. 293~304.
- Bryson, Joanna J. · Mihailis E. Diamantis · Thomas D. Grant, "Of, for, and by the People: The Legal Lacuna of Synthetic Persons", *Artificial Intelligence and Law* 25(3), 2017, pp. 273~291.
- Burrell, Jenna, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms", *Big Data & Society* 3(1), 2016, pp. 1~12.
- Calders, Toon · Indrė Žliobaitė, "Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures", in *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, Bart Custers · Toon Calders · Bart Schermer · Tal Z. Zarsky(Eds.), Springer, 2013, pp. 43~57.
- Calo, Ryan, "Robots as Legal Metaphors", *Harvard Journal of Law & Technology* 30(1), 2016, pp. 209~237.
- _____, "Artificial Intelligence Policy: A Primer and Roadmap", *U. C. Davis Law Review* 51(2), 2017, pp. 399~436.
- Campbell, Colin · Dale Smith, "Deliberative Freedoms and the Asymmetric Features of Anti-Discrimination Law", *University of Toronto Law Journal* 67(3), 2017, pp. 247~287.
- Carusi, Annamaria, "Data as Representation: Beyond Anonymity in E-Research Ethics", *International Journal of Internet Research Ethics* 1, 2008, pp. 37~65.



- Chapman, Nathan S. · Michael W. McConnell, “Due Process as Separation of Powers”, *Yale Law Journal* 121(7), 2012, pp. 1672~1807.
- Chopra, Samir, “Rights for Autonomous Artificial Agents?”, *Communications of the ACM* 53(8), 2010, pp. 38~40.
- Chopra, Samir · Laurence F. White, “Artificial Agents and Agency”, *A Legal Theory for Autonomous Artificial Agents*, University of Michigan Press, 2011, Chap. 1, pp. 5~28.
- Citron, Danielle Keats, “Technological Due Process”, *Washington University Law Review* 85, 2007, pp. 1249~1313.
- Citron, Danielle Keats · Frank A. Pasquale, “The Scored Society: Due Process for Automated Predictions”, *Washington Law Review* 89(1), 2014, pp. 1~33.
- Clarke, Roger A., “Information Technology and Dataveillance”, *Communications of the ACM* 31(5), 1988, pp. 498~512.
- Colb, Sherry F., “Innocence, Privacy, and Targeting in Fourth Amendment Jurisprudence”, *Columbia Law Review* 96, 1996, pp. 1456~1525.
- Collins, Hugh, “Discrimination, Equality and Social Inclusion”, *Modern Law Review* 66(1), 2003, pp. 16~43.
- Collins, Hugh · Tarunabh Khaitan, “Indirect Discrimination Law: Controversies and Critical Questions”, in *Foundations of Indirect Discrimination Law*, Hugh Collins · Tarunabh Khaitan(Eds.), Hart Publishing, an imprint of Bloomsbury Publishing Plc, 2018, pp. 1~30.
- Crawford, Kate · Jason Schultz, “Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms”, *Boston College Law Review* 55, 2014, pp. 93~128.
- Davidson, Donald, “Turing’s Test”, in *Modelling the Mind*, W. H. Newton-Smith · K. V. Wilkes(Eds.), Oxford University Press, 1990, pp. 1~11.
- Dietrich, Bryce J. · Ryan D. Enos · Maya Sen, “Emotional Arousal Predicts Voting on the U. S. Supreme Court”, 2017, https://scholar.harvard.edu/files/msen/files/scotus_audio.pdf, pp. 1~9.
- Doneda, Danilo · Virgilio A. F. Almeida, “What Is Algorithm Governance?”, *IEEE Internet Computing* 20(4), 2016, pp. 60~63.
- Domingos, Pedro, “A Few Useful Things to Know about Machine Learning”, *Communications of the ACM* 55(10), 2012, pp. 78~87.
- Dreier, Horst, “Verantwortung im Demokratischen Verfassungsstaat”, in: *Verantwortung in Recht und Moral*, Ulfrid Neumann · Lorenz Schultz(Hrsg.), Steiner, 2000, S. 9~38.
- Dürig, Günter, “Der Grundrechtssatz von der Menschenwürde: Entwurf eines Praktikablen Wertsystems der Grundrechte aus Art. 1 Abs. I in Verbindung mit Art. 19 Abs. II des Grundgesetzes”, *AöR(Archiv des öffentlichen Rechts)* 81(2), 1956, S. 117~157.
- Dwork, Cynthia · Moritz Hardt · Toniann Pitassi · Omer Reingold · Rich Zemel, “Fairness Through Awareness”, 2011, <https://arxiv.org/pdf/1104.3913v2.pdf>, pp. 1~23.
- Edwards, Lilian · Michael Veale, “Slave to the Algorithm? Why a ‘Right to an Explanation’ Is Probably Not the Remedy You Are Looking For”, *Duke Law & Technology Review* 16(1), 2017/2018, pp. 18~84.



- _____. "Enslaving the Algorithm: From a 'Right to an Explanation' to a 'Right To Better Decisions'?", *IEEE Security & Privacy* 16(3), 2018, pp. 46~54.
- Fayyad, Usama, "The Digital Physics of Data Mining", *Communications of the ACM* 44(3), 2001, pp. 62~65.
- Fenster, Mark, "The Opacity of Transparency", *IOWA Law Review* 91, 2006, pp. 885~949.
- Fix, Evelyn · J. L. Hodges, Jr., "Discriminatory Analysis – Nonparametric Discrimination: Consistency Properties", *International Statistical Review / Revue Internationale de Statistique* 57(3), 1989, pp. 238~247
- Floridi, Luciano, "A Look into the Future Impact of ICT on Our Lives", *The Information Society* 23(1), 2007, pp. 59~64.
- _____, "Introduction", in *The Onlife Manifesto: Being Human in a Hyperconnected Era*, Luciano Floridi(Ed.), Springer, 2015, pp. 1~3.
- Fraser, Nancy, "Social Justice in the Age of Identity Politics: Redistribution, Recognition and Participation", in *Culture and Economy after the Cultural Turn*, Larry J. Ray · R. Andrew Sayer(Eds.), SAGE, 1999, pp. 25~52.
- Fredman, Sandra, "Substantive Equality Revisited", *International Journal of Constitutional Law* 14(3), 2016, pp. 712~738.
- Gal, Michal S. · Niva Elkin-Koren, "Algorithmic Consumers", *Harvard Journal of Law & Technology* 30(2), 2017, pp. 309~353.
- Gandy, Oscar H., "Engaging Rational Discrimination: Exploring Reasons for Placing Regulatory Constraints on Decision Support Systems", *Ethics and Information Technology* 12(1), 2010, pp. 29~42.
- Gellert, Raphaël · Katja de Vries · Paul de Hert · Serge Gutwirth, "A Comparative Analysis of Anti-Discrimination and Data Protection Legislations", in *Discrimination and Privacy in the Information Society Data Mining and Profiling in Large Databases*, Bart Custers · Toon Calders · Bart Schermer · Tal Z. Zarsky(Eds.), Springer, 2013, pp. 61~88.
- Gellman, Robert, "Fair Information Practices: A Basic History", Ver. 2.18, 2017, <https://bobgellman.com/rg-docs/rg-FIPshistory.pdf>, pp. 1~46.
- Gilman, Michele · Rebecca Green, "The Surveillance Gap: The Harms of Extreme Privacy and Data Marginalization", *New York University Review of Law & Social Change* 42(3), 2018, pp. 253~307.
- Goldberg, Suzanne B., "Discrimination by Comparison", *The Yale Law Journal* 120(4), 2011, pp. 728~812.
- Goodman, Bryce · Seth Flaxman, "EU Regulations on Algorithmic Decision-Making and a 'right to explanation'", ver. 1, ICML Workshop on Human Interpretability in Machine Learning, June 2016, pp. 26~30.
- Guerra-Pujol, Enrique, "The Turing Test and the Legal Process", *Information & Communications Technology Law* 21(2), 2012, pp. 113~126.
- Hage, Jaap, "A Theory of Legal Reasoning and a Logic to Match", *Artificial Intelligence and Law* 4(3-4), 1996, pp. 199~273.
- _____, "Theoretical Foundations for the Responsibility of Autonomous Agents", *Artificial Intelligence and Law* 25(3), 2017, pp. 255~271.
- Heesen, Constantijn · Vincent Homburg · Margriet Offereins, "An Agent View on Law", *Artificial Intelligence and Law* 5(4), 1997, pp. 323~340.



- Heidegger, Martin, "Die Frage nach der Technik", in: ders., *Die Technik und die Kehre*, 8. Aufl., Neske, 1991 [1. 1962], S. 5~36; Lovitt, William(Trans.), "The Question Concerning Technology", in *The Question Concerning Technology and Other Essays*, Harper & Row, 1977, pp. 3~35.
- Hellman, Deborah, "Two Concepts of Discrimination", *Virginia Law Review* 102(4), 2016, pp. 895~952.
- Hildebrandt, Mireille, "Radbruch's Rechtsstaat and Schmitt's Legal Order: Legalism, Legality and the Institution of Law", *Critical Analysis of Law* 2(1), 2015, pp. 42~63.
- _____, "Law as Information in the Era of Data-Driven Agency", *The Modern Law Review* 79(1), 2016, pp. 1~30.
- _____, "Law as Computation in the Era of Artificial Legal Intelligence: Speaking Law to the Power of Statistics", *University of Toronto Law Journal* 68(suppl. 1), 2018, pp. 12~35.
- Hohfeld, Wesley Newcomb, "Some Fundamental Legal Conceptions as Applied in Judicial Reasoning", *The Yale Law Journal* 23(1), 1913, pp. 16~59.
- Holmes, Elisa, "Anti-Discrimination Rights without Equality", *The Modern Law Review* 68(2), 2005, pp. 175~194.
- Jackson, Vicki C., "Constitutional Law in an Age of Proportionality", *Yale Law Journal* 124(8), 2015, pp. 3094~3196.
- _____, "Proportionality and Equality", in *Proportionality: New Frontiers, New Challenges*, Vicki C. Jackson · Mark V. Tushnet(Eds.), Cambridge University Press, 2017, pp. 171~196.
- Jolls, Christine, "Antidiscrimination and Accommodation", *Harvard Law Review* 115(2), 2001, pp. 642~699.
- Just, Natascha · Michael Latzer, "Governance by Algorithms: Reality Construction by Algorithmic Selection on the Internet", *Media, Culture & Society* 39(2), 2017, pp. 238~258.
- Kamiran, Faisal · Toon Calders · Mykola Pechenizkiy, "Techniques for Discrimination-Free Predictive Models", in *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, Bart Custers · Toon Calders · Bart Schermer · Tal Z. Zarsky(Eds.), Springer, 2013, pp. 223~239.
- Kerr, Ian R., "Prediction, Pre-emption, Presumption: The Path of Law after the Computational Turn", in *Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology*, Mireille Hildebrandt · Katja de Vries(Eds.), Routledge, 2013, pp. 91~120.
- Khaitan, Tarunabh, "Dignity as an Expressive Norm: Neither Vacuous Nor a Panacea", *Oxford Journal of Legal Studies* 32(1), 2012, pp. 1~19.
- Klatt, Matthias · Moritz Meister, "Verhältnismäßigkeit Als Universelles Verfassungsprinzip", in: *Prinzipientheorie und Theorie der Abwägung*, Matthias Klatt(Hrsg.), Mohr Siebeck, 2013, S. 62~104.
- Krawietz, Werner, Theorie der Verantwortung – neu oder alt?: Zur normativen Verantwortungsattribution min Mitteln des Rechts, in: *Verantwortung: Prinzip oder Problem?*, Kurt Bayertz(Hrsg.), WBG, 1995, S. 184~216.
- Kroll, Joshua A. · Joanna Huey · Joel R. Reidenberg · David G. Robinson · Harlan Yu, "Accountable Algorithms", *University of Pennsylvania Law Review* 165(3), 2017, pp. 633~705.



- Kuner, Christopher, "The European Commission's Proposed Data Protection Regulation: A Copernican Revolution in European Data Protection Law", *Bloomberg BNA Privacy and Security Report*, 2012, pp. 1~15.
- Ladeur, Karl-Heinz, "A Critique of Balancing and the Principle of Proportionality in Constitutional Law – A Case for 'Impersonal Rights'?", *Transnational Legal Theory* 7(2), 2016, pp. 228~256.
- Lee, Irene · Fred Martin · Jill Denner · Bob Coulter · Walter Allan · Jeri Erickson · Joyce Malyn-Smith · Linda Werner, "Computational Thinking for Youth in Practice", *ACM Inroads* 2(1), 2011, pp. 32~37.
- Lerman, Jonas, "Big Data and Its Exclusions", *Stanford Law Review Online* 66, 2013, pp. 55~63.
- Lippert-Rasmussen, Kasper, "Indirect Discrimination, Affirmative Action and Relational Egalitarianism", in *Foundations of Indirect Discrimination Law*, Hugh Collins · Tarunabh Khaitan(Eds.), Hart Publishing, an imprint of Bloomsbury Publishing Plc, 2018, pp. 173~196.
- Luhmann, Niklas, "Verfassung als Evolutionäre Errungenschaft", *Rechtshistorisches Journal* 9, 1990, pp. 176~220.
- Lyon, David, "Surveillance as Social Sorting: Computer Codes and Mobile Bodies", in *Surveillance as Social Sorting: Privacy, Risk and Digital Discrimination*, David Lyon(Ed.), Routledge, 2003, pp. 13~30.
- MacCarthy, Mark, "Standards of Fairness for Disparate Impact Assessment of Big Data Algorithms", *Cumberland Law Review* 48(1), 2017/2018, pp. 67~147.
- MacCormick, Neil, "Law as Institutional Fact", in Neil MacCormick · Ota Weinberger, *An Institutional Theory of Law: New Approaches to Legal Positivism*, Kluwer Academic Publishers, 1986, pp. 49~76.
- Markov, Andrei Andreyevich, "An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains", *Science in Context* 19(4), 2006, pp. 591~600.
- Martin, Andrew D. · Kevin M. Quinn · Theodore W. Ruger · Puline T. Kim, "Competing Approaches to Predicting Supreme Court Decision Making", *Perspectives on Politics* 2(4), 2004, pp. 761~767.
- McColgan, Aileen, "Cracking the Comparator Problem: Discrimination, 'Equal' Treatment and the Role of Comparisons", *European Human Rights Law Review* 6, 2006, pp. 650~677.
- McCrudden, Christopher, "Human Dignity and Judicial Interpretation of Human Rights", *European Journal of International Law* 19(4), 2008, pp. 655~724.
- Medina, Eden, "Rethinking Algorithmic Regulation", *Kybernetes* 44(6/7), 2015, pp. 1005~1019.
- Mercat-Bruns, "From Disparate Impact to Systemic Discrimination", in *Discrimination at Work: Comparing European, French, and American Law*, Marie, Elaine Holt(Trans.), University of California Press, 2016, Chapter 4: pp. 82~144.
- Mitchell, Tom M., "The Discipline of Machine Learning", CMU-ML-06-108, Carnegie Mellon University, July 2006, <http://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf>, pp. 1~7.
- Mochales, Raquel · Marie-Francine Moens, "Argumentation Mining", *Artificial Intelligence and Law* 19(1), 2011, pp. 1~22.
- Moon, Gay · Robin Allen, "Dignity Discourse in Discrimination Law: A Better Route to Equality?", *European Human Rights Law Review* 6, 2006, pp. 610~649.



- Moreau, Sophia, “What Is Discrimination?”, *Philosophy & Public Affairs* 38(2), 2010, pp. 143~179.
- _____, “Equality Rights and Stereotype”, in *Philosophical Foundations of Constitutional Law*, David Dyzenhaus · Malcolm Thorburn(Eds.), Oxford University Press, 2016, pp. 283~303.
- Nelkin, Dorothy · Lori Andrews, “Surveillance Creep in the Genetic Age”, in *Surveillance as Social Sorting: Privacy, Risk and Digital Discrimination*, David Lyon(Ed.), Routledge, 2003, pp. 94~110.
- Neumann, Ulfrid, “Theorie der Juristischen Argumentation”, in: *Rechtsphilosophie im 21. Jahrhundert*, Winfried Brugger · Ulfrid Neumann · Stephan Kirste(Hrsg.), Suhrkamp, 2008, S. 233~260; 윤재왕(역), “법적 논증이론”, 법과 논증이론, 세창출판사, 2009, 177~212쪽.
- _____, “Zur Veränderung von Verantwortungsstrukturen unter den Bedingungen des wissenschaftlich-technischen Fortschritts”, in: ders., *Recht als Struktur und Argumentation: Beiträge zur Theorie des Rechts und zur Wissenschaftstheorie der Rechtswissenschaft*, Nomos-Verl.-Ges, 2008, S. 188~202.
- Nickerson, Raymond S., “Confirmation Bias: A Ubiquitous Phenomenon in Many Guises”, *Review of General Psychology* 2(2), 1998, pp. 175~220.
- Norton, Helen, “The Supreme Court’s Post-Racial Turn towards a Zero-Sum Understanding of Equality”, *William and Mary Law Review* 52(1), 2010, pp. 197~259.
- O’Reilly, Tim, “Open Data and Algorithmic Regulation”, in *Beyond Transparency: Open Data and the Future of Civic Innovation*, Brett Goldstein · Lauren Dyson(Eds.), Code for America Press, 2013, pp. 289~300.
- Pasquale, Frank A., “Internet Nondiscrimination Principles: Commercial Ethics For Carriers and Search Engines”, *University of Chicago Legal Forum* 2008, pp. 263~299.
- _____, “Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility and Explainability in an Algorithmic Society”, *Ohio State Law Journal* 78(5), 2017, pp. 1243~1255.
- Pasquale, Frank · Glyn Cashwell, “Prediction, Persuasion and the Jurisprudence of Behaviourism”, *University of Toronto Law Journal* 68(suppl. 1), 2018, pp. 63~81.
- Pedreschi, Dino · Salvatore Ruggieri · Franco Turini, “Measuring Discrimination in Socially-Sensitive Decision Records”, *Proceedings of the SIAM International Conference on Data Mining*, SIAM, 2009, pp. 581~592.
- _____. _____. _____. “The Discovery of Discrimination”, in *Discrimination and Privacy in the Information Society: Data Mining And Profiling In Large Databases*, Bart Custers · Toon Calders · Bart Schermer · Tal Z. Zarsky(Eds.), Springer, 2013, pp. 91~108.
- Pierik, Roland · Wibren Van der Burg, “What Is Neutrality?”, *Ratio Juris* 27(4), 2014, pp. 496~515.
- Poscher, Ralf, “The Principles Theory: How Many Theories and What is Their Merit?”, in *Institutionalized Reason: The Jurisprudence of Robert Alexy*, Matthias Klatt(Ed.), Oxford University Press, 2012, pp. 218~247.
- Pojman, Louis P., “Introduction: The Nature and Value of Equality”, in *Equality: Selected Readings*, Louis P. Posjman · Robert Westmoreland(Eds.), Oxford University Press, 1997, pp. 1~14.



- Primus, Richard A., "Equal Protection and Disparate Impact: Round Three", *Harvard Law Review* 117(2), 2003, pp. 493-587.
- Réaume, Denise G., "Discrimination and Dignity", *Louisiana Law Review* 63(3), 2003, pp. 645-695.
- _____, "Dignity, Equality and Comparison", in *Philosophical Foundations of Discrimination Law*, Deborah Hellman · Sophia Reibetanz Moreau(Eds.), Oxford University Press, 2013, pp. 7-27.
- Richard, Neil M., "The Dangers of Surveillance", *Harvard Law Review* 126, 2013, pp. 1934-1965.
- Richards, Neil M. · William D. Smart, "How Should the Law Think about Robots?", in *Robot Law*, Ryan Calo · Michael Froomkin · Ian Kerr(Eds.), Edward Elgar Publishing, 2016, pp. 3-22.
- Romei, Andrea · Salvatore Ruggieri, "A Multidisciplinary Survey on Discrimination Analysis", *Knowledge Engineering Review* 29(5), 2014, pp. 582-638.
- Rumelhart, David E. · Geoffrey E. Hinton · Ronald J. Williams, "Learning Representations by Back-Propagating Errors", *Nature* 323(6088), 1986, pp. 533-536.
- Saliger, Frank, "Prozeduralisierung im (Straf-)Recht", in: *Einführung in Rechtsphilosophie und Rechtstheorie der Gegenwart*, Winfried Hassemer · Ulfrid Neumann · Frank Saliger(Hrsg.), 9. Aufl., C. F. Müller, 2016 [1. 1976], S. 434-452.
- Samaha, Adam M., "Government Secrecy, Constitutional Law, and Platforms for Judicial Intervention", *UCLA Law Review* 53(4), 2006, pp. 909-976.
- Sartor, Giovanni, "Doing Justice to Rights and Values: Teleological Reasoning and Proportionality", *Artificial Intelligence and Law* 18(2), 2010, pp. 175-215.
- Sartor, Giovanni · Andrea Omicini, "The Autonomy of Technological Systems and Responsibilities for their Use", in *Autonomous Weapons Systems: Law, Ethics, Policy*, Nehal Bhuta · Susanne Beck · Robin Geiss · Claus Kress · Hin Yan Liu(Eds.), Cambridge University Press, 2016, pp. 39-74.
- Saurwein, Florian · Natascha Just · Michael Latzer, "Governance of Algorithms: Options and Limitations", *Info* 17(6), 2015, pp. 35-49.
- Schauer, Frederick, "Transparency in Three Dimensions", *University of Illinois Law Review* 2011(4), 2011, pp. 1339-1357.
- Schlink, Bernhard, "Freiheit durch Eingriffsabwehr – Rekonstruktion der klassischen Grundrechtsfunktion", *Europäische GRUNDRECHTE-Zeitschrift* 11, 1984, S. 457-468.
- Schwartz, Paul M., "Property, Privacy, and Personal Data", *Harvard Law Review* 117(7), 2004, pp. 2056-2128.
- Searle, John R., "Minds, Brains and Programs", *Behavioral and Brain Sciences* 3(3), 1980, pp. 417-424.
- Segall, Shlomi, "What's So Bad about Discrimination?", *Utilitas* 24(1), 2012, pp. 82-100.
- Segev, Re'em, "Making Sense of Discrimination", *Ratio Juris* 27(1), 2014, pp. 47-78.
- Selbst, Andrew D. · Julia Powle, "Meaningful Information and the Right to Explanation", *International Data Privacy Law* 7(4), 2017, pp. 233-242.
- Selmi, Michael, "Indirect Discrimination and the Anti-Discrimination Mandate", in *Philosophical Foundations of Discrimination Law*, Deborah Hellman · Sophia Reibetanz Moreau(Eds.), Oxford University Press, 2013, pp. 250-268.



- Shannon, Claude E, “A Mathematical Theory of Communication”, *The Bell System Technical Journal* 27, 1948, pp. 379-423(July) and pp. 623-656(October).
- Shin, Patrick S., “Is There a Unitary Concept of Discrimination?”, in *Philosophical Foundations of Discrimination Law*, Deborah Hellman · Sophia Reibetanz Moreau(Eds.), Oxford University Press, 2013, pp. 163-181.
- Silverman, Bernard Walter · M. C. Jones, “E. Fix And J. L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951)”, *International Statistical Review / Revue Internationale de Statistique* 57(3), 1989, pp. 233-238.
- Solaiman, S. M., “Legal Personality of Robots, Corporations, Idols and Chimpanzees: A Quest for Legitimacy”, *Artificial Intelligence and Law* 25(2), 2017, pp. 155-179.
- Solum, Lawrence, “Legal Personhood for Artificial Intelligences”, *North Carolina Law Review* 70, 1992, pp. 1231-1287.
- Somek, Alexander, “Equality, Freedom and Dignity”, *The Legal Relation: Legal Theory after Legal Positivism*, Cambridge University Press, 2017, pp. 133-155.
- Stich, Stephen · Ian Ravenscroft, “What Is Folk Psychology?”, *Cognition* 50(1), 1994, pp. 447-468.
- Strahilevitz, Lior Jacob, “Privacy versus Antidiscrimination”, *University of Chicago Law Review* 75, 2008, pp. 363-381.
- Sturm, Susan, “Second Generation Employment Discrimination: A Structural Approach”, *Columbia Law Review* 101(3), 2001, pp. 458-568.
- Surden, Harry, “Machine Learning and Law”, *Washington Law Review* 89(1), 2014, pp. 87-115.
- Sweeney, Latanya, “Discrimination in Online Ad Delivery”, *Communications of the ACM* 56(5), 2013, pp. 44-54.
- Taylor, Charles, “The Politics of Recognition”, in *Multiculturalism: Examining the Politics of Recognition*, Amy Gutmann(Ed.), Princeton University Press, 1994, pp. 25-73.
- Teubner, Gunther, “Rights of Non-Humans? Electronic Agents and Animals as New Actors in Politics and Law”, *Journal of Law and Society* 33, 2006, pp. 497-521.
- Thornhill, Chris, “Legal Proceduralization and the Fictions of the Political”, in: *Prozeduralisierung des Rechts*, Tatjana Sheplyakova(Hrsg.), Mohr Siebeck, 2018, S. 161-189.
- Turing, Alan M., “On Computable Numbers, with an Application to the Entscheidungsproblem”, *Proceedings of the London Mathematical Society* s2-42(1), 1937, pp. 230-265.
- _____, “Computing Machinery and Intelligence”, *Mind* 59(236), 1950, pp. 433-460.
- Tutt, Andrew, “An FDA for Algorithms”, *Administrative Law Review* 69(1), 2017, pp. 83-123.
- Wachter, Sandra · Brent Mittelstadt · Luciano Floridi, “Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation”, *International Data Privacy Law* 7(2), 2017, pp. 76-99.
- Wachter, Sandra · Brent Mittelstadt · Chris Russell, “Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR”, *Harvard Journal of Law & Technology* 31(2), 2018, pp. 842-887.



- Waldron, Jeremy, "The Concept and the Rule of Law", *Georgia Law Review* 43(1), 2008, pp. 1~61.
- Wang, Ke · Suman Sundaresh, "Selecting Features by Vertical Compactness of Data", in *Feature Extraction, Construction and Selection: A Data Mining Perspective*, Huan Liu · Hiroshi Motoda(Eds.), Springer US, 1998, pp. 71~84.
- Westen, Peter, "The Empty Idea of Equality", *Harvard Law Review* 95(3), 1982, pp. 537~596.
- Wing, Jeannette M., "Computational Thinking", *Communications of the ACM* 49(3), 2006, pp. 33~35.
- Winner, Langdon, "Technology as Forms of Life", in *Epistemology, Methodology and the Social Sciences*, Robert S. Cohen · Marx W. Wartofsky(Eds.), Springer Netherlands, 1983, pp. 249~263.
- _____, "Do Artifacts Have Politics?", in *The Whale and the Reactor: A Search for Limits in an Age of High Technology*, University of Chicago Press, 1986, pp. 19~39.
- _____, "Techne and Politeia"[1st, 1983], in *The Whale and the Reactor: A Search for Limits in an Age of High Technology*, University of Chicago Press, 1986, pp. 40~58.
- Winter, Jenifer Sunrise, "Introduction to the Special Issue: Digital Inequalities and Discrimination in the Big Data Era", *Journal of Information Policy* 8, 2018, pp. 1~4.
- Yoshino, Kenji, "Assimilationist Bias in Equal Protection: The Visibility Presumption and the Case of "Don't Ask, Don't Tell"", *The Yale Law Journal* 108(3), 1998, pp. 485~571.
- _____, "Covering", *The Yale Law Journal* 111(4), 2002, pp. 769~939.
- _____, "The New Equal Protection", *Harvard Law Review* 124(3), 2011, pp. 747~803.
- Zambonelli, Franco · Flora Salim · Seng W. Loke · Wolfgang De Meuter · Salil Kanhere, "Algorithmic Governance in Smart Cities: The Conundrum and the Potential of Pervasive Computing Solutions", *IEEE Technology and Society Magazine* 37(2), 2018, pp. 80~87.
- Zarsky, Tal Z., "Governmental Data Mining and its Alternatives", *Pennsylvania State Law Review* 116, 2012, pp. 285~330.
- _____, "Transparent Predictions", *University of Illinois Law Review* 2013(4), 2013, pp. 1503~1570.
- _____, "Understanding Discrimination in the Scored Society", *Washington Law Review* 89(4), 2014, pp. 1375~1412.
- _____, "An Analytic Challenge: Discrimination Theory in the Age of Predictive Analytics", *I/S: A Journal of Law and Policy for the Information Society* 14(1), 2018, pp. 11~35.
- Zatz, Noah D., "Disparate Impact and the Unity of Equality Law", *Boston University Law Review* 97, 2017, pp. 1357~1425.
- Zhao, Jieyu · Tianlu Wang · Mark Yatskar · Vicente Ordonez · Kai-Wei Chang, "Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints", *Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2979~2989.
- Žliobaitė, Indrė · Bart Custers, "Using Sensitive Personal Data May Be Necessary for Avoiding Discrimination in Data-Driven Decision Models", *Artificial Intelligence and Law* 24(2), 2016, pp. 183~201.



판례 및 법문서

경제개발협력기구(OECD)

“OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data”, 1980, <http://www.oecd.org/internet/ieconomy/oecdguidelinesontheprivacyandtransborderflowsofpersonaldata.htm#part1>.

“The OECD Privacy Framework”, 2013, http://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf.

뉴질랜드

Te Awa Tupua (Whanganui River Claims Settlement) Act 2017.

대한민국

국가인권위원회 2011. 9. 27.자 10진정0480200 결정, 결정례집(차별시정분야) 4, 262.

헌재 1989. 5. 24. 89헌가37 등, 판례집 1, 48.

헌재 1995. 12. 27. 95헌마224 등, 판례집 7-2, 760.

헌재 1999. 1. 28. 97헌마253 등, 판례집 11-1, 54.

헌재 1999. 12. 23. 98헌마363, 판례집 11-2, 770.

헌재 2001. 9. 27. 2000헌마159, 판례집 13-2, 353.

헌재 2001. 10. 25. 2000헌마92 등, 판례집 13-2, 502.

헌재 2004. 5. 14. 2004헌나1, 판례집 16-1, 609.

헌재 2007. 6. 28. 2004헌마644 등, 판례집 19-1, 859.

헌재 2010. 11. 25. 2006헌마328, 판례집 22-2하, 446.

헌재 2014. 10. 30. 2012헌마192 등, 판례집 26-2상, 668.

헌재 2017. 3. 10. 2016헌나1, 판례집 29-1, 1.

독일

BVerfG, Beschluss des Ersten Senats vom 04. April 2006 – 1 BvR 518/02 – Rn. [1-184], http://www.bverfg.de/e/rs20060404_1bvr051802.html.

Oberverwaltungsgericht NRW(Nordrhein-Westfalen), 17 A 805/03, 24. 6. 2009.

미국(US)

City of Cleburne v. Cleburne Living Ctr., 473 U. S. 432 (1985).



Frontiero v. Richardson. 411 U. S. 677 (1973).
 Griggs v. Duke Power Co. 401 U. S. 424 (1971).
 Hopwood v. Texas 78 F 3d 932 (1996).
 Parents Involved in Community Sch. v. Seattle School Dist. No. 1, 551 U. S. 701 (2007).
 Regents of University of California v. Bakke, 438 U. S. 265 (1978).
 Ricci v. DeStefano, 557 U. S. 557 (2009).
 State v. Loomis, 2016 WI 68, 881 N. W. 2d 749 (2016).
 Sweatt v. Painter, 339 U.S. 629. (1950).
 Watkins v. United States Army 875 F.2d 699 (1989).

영국(UK)

James v. Eastleigh Borough Council (1990) 2 AC 751.

유럽연합(EU)

“Council Directive 2004/113/EC of 13 December 2004 Implementing the Principle of Equal Treatment between Men and Women in the Access to and Supply of Goods and Services”, *Official Journal of the European Union*, 21 December 2004, L 373, pp. 37~43.

“Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data”, *Official Journal of the European Union*, 23 November 1995, L 281, pp. 31~50.

“Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)”, *Official Journal of the European Union*, 4 May 2016, L 119, pp. 1~88.

Association Belge des Consommateurs Test-Achats ASBL and Others v Conseil des ministres, C-236/09, Judgment of the Court of 1 March 2011, EU:C:2011:100.

CHEZ Razpredelenie Bulgaria AD v. Komisia za Zashtita ot Diskriminatsia, C-83/14, Judgment of the Court of 16 July 2015, EU:C:2015:480.

Heinz Huber v Bundesrepublik Deutschland, C-524/06, Judgment of the Court of 16 December 2008, EU:C:2008:724.

캐나다

British Columbia (Public Service Employee Relations Commission) v. British Columbia Government Service Employees' Union, Case No. 26374, [1999] 3 S. C. R. 3.

R. v. Oakes(Sa Majesté La Reine c. David Edwin Oakes), Case No. 17550, [1986] 1 S. C. R. 103.



보고서

- 권건보, 영국의 포스트 휴먼 기술법제에 관한 비교법적 연구-드론과 자율주행차를 중심으로-, 한국법제연구원, 2016.
- 나채준, 일본의 포스트 휴먼 기술법제에 관한 비교법적 연구-드론과 자율주행차를 중심으로-, 한국법제연구원, 2016.
- 손승우 · 김윤명, 인공지능 기술 관련 국제적 논의와 법제 대응방안 연구, 한국법제연구원, 2016.
- 윤성현, 캐나다의 포스트 휴먼 기술법제에 관한 비교법적 연구-드론과 자율주행차를 중심으로-, 한국법제연구원, 2016.
- 윤인숙, 미국의 포스트 휴먼 기술법제에 관한 비교법적 연구-드론과 자율주행차를 중심으로-, 한국법제연구원, 2016.
- 이원태, 인공지능의 규범이슈와 정책적 시사점, KISDI Premium Report, 정보통신정책연구원, 2015.
- 장원규, 독일의 포스트 휴먼 기술법제에 관한 비교법적 연구-드론과 자율주행차를 중심으로-, 한국법제연구원, 2016.
- 정관선, 프랑스의 포스트 휴먼 기술법제에 관한 비교법적 연구-드론과 자율주행차를 중심으로-, 한국법제연구원, 2016.
- “Treasury Responds to Suggestion that Robots Pay Income Tax”, Tax Notes 25, 30 September 1984.
- Crawford, Kate · Meredith Whittaker · Madeleine Elish · Solon Barocas · Aaron Plasek · Kadija Ferryman, “The AI Now Report: The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term”, A Summary of the AI Now Public Symposium(6 July 2016), Hosted by the White House and New York University’s Information Law Institute, 2016, https://ainowinstitute.org/AI_Now_2016_Report.pdf.
- Crawford, Kate · Meredith Whittaker · Alex Campolo · Madelyn Sanfilippo, “AI Now 2017 Report”, The AI Now Institute at New York University, 2017, https://ainowinstitute.org/AI_Now_2017_Report.pdf.
- Executive Office of the President, “Big Data: A Report on Algorithmic Systems, Opportunity and Civil Rights”, May 2016, https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.
- _____, “Artificial Intelligence, Automation and the Economy”, December 2016, <https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.PDF>.
- McCrudden, Christopher · Sacha Prechal, “The Concepts of Equality and Non-Discrimination in Europe: A Practical Approach”, European Commission Directorate-General for Employment, Social Affairs and Equal Opportunities Unit G. 2, <http://ec.europa.eu/social/BlobServlet?docId=4553&langId=en>.
- National Science and Technology Council, Executive Office of the President, “Preparing for the Future of Artificial Intelligence”, October 2016, https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.
- Porat, Marc Uri, “The Information Economy: Definition and Measurement”, Office of Telecommunications (U. S. Department of Commerce), 1 May 1977.
- Reisman, Dillon · Jason Schultz · Kate Crawford · Meredith Whittaker, “Algorithmic Impact Assessments”, The AI Now Institute at New York University, 2018, <https://ainowinstitute.org/aiareport2018.pdf>.



Stanford University, “Artificial Intelligence and Life in 2030”, One Hundred Year Study on Artificial Intelligence, Report of the 2015 Study Panel, September 2016.

Whittaker, Meredith · Kate Crawford · Roel Dobbe · Genevieve Fried · Elizabeth Kaziunas · Varoon Mathur · Sarah Myers West · Rashida Richardson · Jason Schultz · Oscar Schwartz, “AI Now Report 2018”, The AI Now Institute at New York University, December 2018, https://ainowinstitute.org/AI_Now_2018_Report.pdf.

웹 페이지

“뜯구름 4차 산업혁명” 공약...창조경제 전철 우렁, 연합뉴스 TV, 2017. 4. 14., <http://www.yonhapnewstv.co.kr/MYH20170414000900038/>.

네이버 이미지 검색, “쇼팽”, https://search.naver.com/search.naver?sm=tab_hy.top&where=image&query=%EC%87%BC%ED%95%91&oquery=%EC%87%BC%ED%95%91&tqi=TzVQespySD8ssvNQk1ZssssssDV-289909.

네이버 이미지 검색, “장보기”, https://search.naver.com/search.naver?sm=tab_hy.top&where=image&query=%EC%9E%A5%EB%B3%B4%EA%B8%B0&oquery=%EC%9E%A5%EB%B3%B4%EA%B8%B0&tqi=TzVj1dpySE4ssaRnKedsssssteR-336156.

이설영, “4차 산업혁명 핵심 ‘데이터 경제’ 활성화 위한 TF 발족”, 2018. 9. 6., <http://www.fnnews.com/news/201809061409461558>.

정원영, “국회 4차산업혁명 특위, 4차산업혁명 국가로드맵 초안 발표”, 로봇신문사, 2018. 4. 24., <http://www.irobotnews.com/news/articleView.html?idxno=13764>.

“A Day of Terror; Bush's Remarks to the Nation on the Terrorist Attacks”, The New York Times, 12 September 2001, [NYTimes.com](http://www.nytimes.com/2001/09/12/us/a-day-of-terror-bush-s-remarks-to-the-nation-on-the-terrorist-attacks.html), <http://www.nytimes.com/2001/09/12/us/a-day-of-terror-bush-s-remarks-to-the-nation-on-the-terrorist-attacks.html>.

“Algorithm Appointed Board Director”, BBC News, 16 May, 2014, <https://www.bbc.com/news/technology-27426942>.

Andreas von der Heydt, “First Time Ever: Artificial Intelligence Nominated as a Board Member”, 20 May 2014, <https://www.linkedin.com/pulse/20140520045550/-175081329-first-time-ever-artificial-intelligence-nominated-as-a-board-member>.

Angwin, Julia · Jeff Larson, “The Tiger Mom Tax: Asians Are Nearly Twice as Likely To...”, ProPublica, 1 September 2015, <https://www.propublica.org/article/asians-nearly-twice-as-likely-to-get-higher-price-from-princeton-review>.

_____, “Machine Bias”, ProPublica, 23 May 2016, <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

Blanchard, Dave, “Musk's Warning Sparks Call For Regulating Artificial Intelligence”, NPR.org, 19 July 2017, <https://www.npr.org/sections/alltechconsidered/2017/07/19/537961841/musks-warning-sparks-call-for-regulating-artificial-intelligence>.

Cheng, Selina, “An Algorithm Rejected an Asian Man's Passport Photo for Having ‘closed Eyes’”, Quartz, 7 December 2016, <https://qz.com/857122/an-algorithm-rejected-an-asian-mans-passport-photo-for-having-closed-eyes/>.



- Dastin, Jeffrey, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women", Reuters, 10 October 2018, <https://uk.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUKKCN1MK08G>.
- Guyon, Jessica, "Google Photos Labeled Black People 'Gorillas'", USA TODAY, 1 July 2015, <https://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465/>.
- Hill, Kashmir, "How Target Figured Out a Teen Girl Was Pregnant Before Her Father Did", Forbes, 16 February 2012, <https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/>.
- Horcher, Gary, "Woman Says Her Amazon Device Recorded Private Conversation, Sent It out to Random Contact", KIRO7, 25 May 2018, <https://www.kiro7.com/news/local/woman-says-her-amazon-device-recorded-private-conversation-sent-it-out-to-random-contact/755507974>.
- Kohli, Sonali, "Bill Gates Joins Elon Musk and Stephen Hawking in Saying Artificial Intelligence Is Scary", Quartz, 29 January 2015, <https://qz.com/335768/bill-gates-joins-elon-musk-and-stephen-hawking-in-saying-artificial-intelligence-is-scary/>.
- Kuczmarski, James, "Reducing Gender Bias in Google Translate", Google, 6 December 2018, <https://www.blog.google/products/translate/reducing-gender-bias-google-translate/>.
- Lessig, Lawrence, "Against Transparency", New Republic, 9 October 2009, <https://newrepublic.com/article/70097/against-transparency>.
- Loizos, Connie, "This Famous Robotist Doesn't Think Elon Musk Understands AI", TechCrunch, 19 July 2017, <http://social.techcrunch.com/2017/07/19/this-famous-robotist-doesnt-think-elon-musk-understands-ai/>.
- Pichai, Sundar, "AI at Google: Our Principles", Google, 7 June 2018, <https://www.blog.google/topics/ai/ai-principles/>.
- Prodhan, Georgena, "European Parliament Calls for Robot Law, Rejects Robot Tax", Reuters, 16 February 2017, <https://www.reuters.com/article/us-europe-robots-lawmaking/european-parliament-calls-for-robot-law-rejects-robot-tax-idUSKBN15V2KM>.
- Rose, Adam, "Are Face-Detection Cameras Racist?", Time, 22 January 2010, <http://content.time.com/time/business/article/0,8599,1954643,00.html>.
- Simonite, Tom, "When It Comes to Gorillas, Google Photos Remains Blind", WIRED, 18 January 2018, <https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/>.
- Smith, Carl, "AI That Can Teach? It's Already Happening", ABC News, 16 June 2018, <http://www.abc.net.au/news/science/2018-06-16/artificial-intelligence-that-can-teach-is-already-happening/9863574>.
- Sofge, Erik, "Why Artificial Intelligence Will Not Obliterate Humanity", Popular Science, 20 March 2015, <https://www.popsci.com/why-artificial-intelligence-will-not-obliterate-humanity>.
- Zolfaghari, Ellie, "ROBOT Becomes the World's First Company Director", Mail Online, 19 May 2014, <http://www.dailymail.co.uk/sciencetech/article-2632920/Would-orders-ROBOT-Artificial-intelligence-world-s-company-director-Japan.html>.



기타

“Fuel of the Future; The Data Economy”, *The Economist*; London, 423(9039), 6 May 2017, pp. 17~20.

Čapek, Karel, “The Meaning of ‘R. U. R.’”, *The Saturday Review of Politics, Literature, Science and Art* 136(3534), 21 July 1923, p. 79.

Čapek, Karel, 李光洙(역), “人造人”, 東明 31, 1923. 4.

Čapek, Karel, 朴英熙(역), “人造勞働者”, 開闢 56~59, 1925. 2~5.

Horvitz, Eric, “On the Meaningful Understanding of the Logic of Automated Decision Making”, BCLT Privacy Law Forum, 24 March 2017, https://www.law.berkeley.edu/wp-content/uploads/2017/03/BCLT_Eric_Horvitz_March_2017.pdf.

Northpointe, Inc., “Practitioner's Guide to COMPAS Core”, 19 March. 2015, http://www.northpointeinc.com/files/technical_documents/Practitioners-Guide-COMPAS-Core-_031915.pdf.



Abstract

Machine Learning Algorithms and Discrimination

Nam, Joong-Kweon[✉]

Department of Law
The Graduate School of Korea University

Machine learning algorithms used for autonomous agents that receive percepts from the environment and perform actions are the basis for the algorithmic society constituted of cyber-physical systems. The benefits of machine learning algorithms are accompanied by the risks of adverse effect. It is usually assumed that algorithmic decision-making is objective, rational, and fair. But machine learning algorithms writing down an algorithms on their own by making inferences from data are able not only to learn from biased dataset reflecting the differential and hierarchical structure of society, but also to result in discriminatory decisions unintentionally.

This dissertation addresses several legal problems and challenges focusing on algorithmic discrimination, and thus considers both mechanisms of machine learning and theories of anti-discrimination law. It investigates why a decision-making using machine learning algorithms can be suspected of discrimination. For this purpose various cases and studies are presented. It examines which concepts of discrimination can apply to the suspicious situations. And it also analyzes on the mechanism of machine learning, the overlapping meanings of discrimination and the relation of discrimination to constitutional virtues and interests. This analysis implies that even if the concept of discrimination is not unitary so that theories of anti-discrimination law are also complicated, a legal model of discrimination may simplify the process of awareness and judgment of discrimination so that it may not grasp the suspicious situations.

One of the main impacts of machine learning algorithms used for artificial intelligence agents is substitution effect. Legal rules may be substituted with algorithms inferred

[✉] namc@korea.ac.kr



by machine learning algorithms suitable to predict and optimize rules. This means that argumentation may be replaced by computation. Machine learning algorithms conduct classification, prediction, clustering etc., and inferring algorithms using unknown or unaccountable features as proxy can not only indirectly circumvent reasons prohibited from using as basis for decision-making by law, but also produce new groups, classes or categories which are not able to be grasped in the enumerated reason-based legal model of anti-discrimination. Furthermore, in justification of discrimination based on the proxies and new features, discrimination may be classified as rational, since the proxies and new features are outcome from data-driven algorithms based on the statistical dataset.

It is difficult to deal with the problem of discrimination by machine learning algorithms, for they are opaque. Although the request for transparency is the default setting of democratic state, legal secrecy, technical complexity or cognitive obfuscation can make machine learning algorithms opaque. The responsibility for discrimination by machine learning algorithms is dealt with in the context of changes in the structure of responsibility correlations due to the development of science and technology. The way to attribute responsibility for discrimination to so-called ‘discriminator’ divided into public and private can be reconstructed into the processes of designing, operating and monitoring for accountable machine learning algorithms. Moreover, it is examined through the confrontation of restrictivism and permissivism whether algorithmic agents as such are the subjects of responsibility for discrimination.

The fact that automated algorithms process personal data facilitates data privacy protection-related approach to the algorithmic discrimination. Personal information as data may represent the identity or identifiability of individual, but may also be the ground to build a group. This dissertation divides data processing procedure using machine learning algorithms into three steps: collection, analysis and use of data. Then it applies data privacy and anti-discrimination approach to each step, and articulates their limitations. In addition, it demonstrates there is room for discussion on the ‘right to explanation’ of automated decision-making algorithms to be reinterpreted focusing on discrimination.

Keywords Algorithmic Discrimination · Artificial Intelligence and Law(AI & Law) · Theories of Anti-Discrimination Law · Machine Bias · Disparate Impact · Substitution Effect · Governance of Algorithms · Data Privacy Protection · Automated Decision Making · Transparency and Accountability of Algorithms

