전세계적 위험에 인공지능이 미칠 수 있는 긍정적•부정적 영향
Artificial Intelligence as a Positive and Negative Factor in Global Risk

『전세계적 치명적 위험』(닉 보스트롬•밀란 치르코비치 편찬)에 포함될 것

2006.8.31 초안
엘리에저 유드카우스키(Eliezer Yudkowsky)
(yudkowsky@singinst.org)
Singularity Institute for Artificial Intelligence (인공지능 특이점 연구소)
팰러앨토, 캘리포니아

서론[1]

인공지능의 가장 큰 위험은 사람들이 자기가 인공지능을 이해한다고 섣불리 결론을 내린다는 것이다. 물론 이 문제는 AI의 분야에 국한되지 않는다. 쟈크 모노 (Jacques Monod)는 "진화론의 신기한 일면은 사람들이 다 자기가 그것을 이해한다고 생각한다는 것"이라고 했다. (Monod 1974.) 물리학자인 나의 아버지는 사람들이 자기가 물리학 이론을 지어내는 것에 대해 불평했다. 그는 사람들이 자기만의 화학 이론은 왜 안 지어내는지 알고 싶어했다. (물론 정말 그런다…) 그래도 역시 인공지능 분야에 이 문제가 특히 심각한 것으로 보인다. AI는 거창한 약속을 하고 지키지는 않는 분야로 악명이 높다. 대부분의 평자는 사실대로 AI가 어렵다고 결론을 내린다. 그러나 AI의 학계가 당한 창피는 어려움에서 발생하지 않는다. 수소로 항성을 만드는 것도 어려우나, 항성천문학은 별을 만든다고 큰소리치고 실패하는 것으로 악평을 받지 않는다. 결정적인 문제는 AI가 어렵다는 것이 아니라, 어떤 이유 탓에 사람들이 실제로 인공지능에 대해 아는 것보다 더 많이 안다고 생각하기 아주 쉽다는 것이다.

전세계적 치명적 위험에 내가 쓴 다른 챕터 "전세계적 위험에 대한 판단에 영향을 끼칠 수 있는 인지 바이어스들"에, 나는 세계를 고의적으로 파괴할 사람들은 적을 것이라는 말로 시작했다. 이런 이유로, *실수로* 지구가 파괴되는 시나리오는 매우 우려스럽다. 전세계에 재앙을 일으킨다는 것을 아는 버튼을 누르는 사람은 적을 것이다. 그러나 사람들이 버튼이 사실과 다른 일을 한다고 자신만만하게 믿기 쉬우면, 그것은 실로 걱정스러운 일이다.

인지 바이어스에 대해서보다 인공지능에 의한 세계적 위험에 대해 글을 쓰기 훨씬 더 어렵다. 인지 바이어스는 확립된 과학이라서 문헌을 인용하기만 하면 된다. 반면에 인공지능은 확립된 과학이 아니다. 인공지능은 교과서에 나오는 과학이 아니라 최첨단의 과학이다.

또한, 뒤에 설명할 이유로 인공지능에 의한 전세계적 치명적 위험에 대해서는 이미 있는 전문적 문헌에 거의 논의되어 있지 않다.

나는 억지로 이 문제를 나의 견해에서 분석했으며, 나 자신의 결론을 내렸고 한정된 공간에 결론을 뒷받침하기 위해 온 힘을 다했다. 내가 이 주제에 관한 주요 연구를 인용하기를 게을리 한 것이 아니라 내가 온 힘을 다해 살펴본바 (2006년 1월 현재) 인용하거나 참조할 주요 연구가 *없다*는 것이다. 인공지능은 이 책에서 논의되는 전세계적 위험 중 가장 논의하기가 어려워 무시하기가 쉽다. 소행성 충돌처럼 통계를 참고해서 재변의 확률을 낮게 잡을 수도 없다. 제기되는 물리학 실험의 위험처럼 정밀하고 정밀히 검증된 이론으로 계산을 해서 사건의 확률을 제외하거나 무수히 낮게 잡을 수도 없다. 그러나 이런 요인들은 AI의 위험을 보다 *더* 걱정스럽게 한다.

많은 인지 바이어스들은 시간의 제약과 인지적 바쁨, 정보의 부족과 함께 *증가하는* 것으로 밝혀져 있다. 다르게 말하면 분석의 난제가 어려울수록 바이어스를 피하거나 감소하는 것이

[1]이 논문에 대해 논평, 제안, 및 비평을 제공한 벤 괴어첼, 올리 램, 피터 매클러스키, 에릭 봄, 닉 보스트롬, 마이클 로이 에임스, 마이클 윌슨, 에밀 질리엄, 밀란 치르코비치, 존 K. 클라크, 로빈 핸슨, 키스 헨슨, 빌 히버드에게 감사를 드린다. 말할 것도 없이 이 논문에 남은 오류는 모두 내 오류이다.

더 중요하다는 것이다. 그러므로 이 챕터를 계속 읽기 전 [XXX-YYY 장], "전세계적 위험에 대한 판단에 영향을 끼칠 수 있는 인지 바이어스들"을 읽기를 권장한다.

1: 의인화 바이어스

어떤 것이 우리의 일상생활에 충분히 보편적이면 우리는 그것이 있다는 것을 잊을 정도로 당연시 한다.
필요한 유전자가 열 개인 복합적인 생물학적 적응을 상상해 보라. 만약 이 열 개의 유전자가 각각 유전자 풀 안에 도수가 50%이면, 유전자 하나는 종의 절반만이 가지고 있으므로, 대체로 1,024 개 중 하나만이 완성된 적응을 가질 것이다. 모피는 환경이 일관적으로 춥지 않으면 큰 진화적 이점이 아니다. 비슷하게 B라는 유전자가 A라는 유전자에 의존하면, 유전자 A가 *유전적* 환경에 일관적으로 있지 않으면 유전자 B가 큰 이점이 될 수 없다. 성적 번식을 하는 종들에 있어서 *복잡성*과 *상호 의존성*은 필연적으로 보편적이다. 그렇지 않으면 그 종이 진화/출현할 수 없기 때문이다. (Tooby and Cosmides 1992) 개똥지빠귀 한 마리는 어떤 다른 개똥지빠귀보다 더 매끄러운 깃털을 가질 수 있지만, 그 두 마리 다 날개가 있을 것이다. 자연선택은 개체차를 먹고 살면서 개체차를 소모한다. (Sober 1984)
알려진 모든 문화에서, 사람들은 행복, 슬픔, 혐오감, 분노, 공포감, 놀람 등의 감정을 경험하며 (Brown 1991), 이 감정들을 똑같은 표정으로 표현한다. (Ekman and Keltner 1997) 우리는 다른 색깔로 칠해져 있을지라도 모두 엔진 뚜껑 밑에는 똑같은 엔진을 돌린다. 이 원리를 진화심리학자들은 "인류의 정신적 통일성"(psychic unity of humankind)이라고 칭한다 (Tooby and Cosmides 1992). 이 결과는 진화생물학이 설명하고 또한 필요로 한다.
새로 발견된 부족에 대해 "음식을 먹는다! 공기를 마신다! 도구를 사용한다! 서로에게 이야기한다! " 라고 흥미 차게 보고할 인류학자는 없을 것이다. 우리의 차이점만 생각나게 하는 세계에 사는 우리는 우리가 서로 얼마나 같은지 잊는다.
인간은 다른 인간들을 모형화하게 – 동종(同種)과 경쟁하고 협력하게 진화했다. 강력한 지성은 모두 같은 인간이라는 것은 조상 환경의 일관적인 성질이었다. 우리는 다른 인간의 감정을 이입하고 – 모형화될 것이 모형화하는 자와 유사하므로 – 다른 인간의 처지에서 생각하게 진화했다. 놀랍지 않게, 인간은 흔히 의인화, 곧 인간이 아닌 것이 인간과 비슷한 성질을 가질 것이라고 기대를 한다. 영화 "매트릭스" (Wachowski and Wachowski 1999)에는, 가상의 "인공지능"인 스미스 요원은 처음에는 매우 냉정하고 침착하게 보이고, 얼굴이 소극적이고 감정이 없게 보인다. 그러나 나중에 인간인 모피어스를 심문하는 중에는 에이전트 스미스가 인류에 대한 혐오감을 표출하고, 그의 얼굴은 인간에 보편적인 혐오감을 뜻하는 표정을 보인다.
어떤 사람이 다른 인간들의 행동을 예측하려면 자신의 인간 뇌를 조회하는 것이 적응된 본능으로서는 괜찮다. 그러나 다른 최적화 프로세스를 다룰 때에는 (예를 들어 생명의 복잡한 질서를 보고 어떻게 그런 질서가 생겼는지 경이로이 여기는 18세기 신학자 윌리엄 페일리라면) 조심하지 않는 과학자에게 의인화는 벗어나기 위해 다윈 한 명이 필요한 끈끈이 종이가 되고 만다.
의인화 실험은 대상자들이 무의식적으로, 심지어는 자신들의 의식적인 생각에 반대로 의인화를 할 수 있다는 것을 보여준다. Barrett과 Keil(1996)이 한 실험에는, 대상자들이 신은 비인격적인 성격을 (신이 동시에 여러 곳에 계실 수 있다거나, 동시에 여러 사건에 주목할 수 있으시다는 것 등) 가지고 계신다는 믿음을 강력히 고백했다. Barrett과 Keil은 같은 대상자에게 신이 사람들을 물에 빠져 죽는 것에서 구하는 등의 이야기를 제시했다. 대상자들은 이야기들에 대한 질문에 답하거나 자기 말로 이야기를 풀 때, 신이 동시에 한 장소에 있다는 듯, 일을 병렬해 하는 것이 아니라 차례대로 하는 식으로 답했다. 우리의 목적에 편리하게도 Barrett과 Keil은 또, 다른 한 실험 집단을 기타는 같은데 신 대신에 "언콤프"라는 초지능 컴퓨터가 나오는 이야기들로 실험했다. 예를 들어 신의 편재성(遍在性, omnipresence)을 흉내 내기 위해 언콤프의 센서와 에펙터가 "지구를 구석구석 다 덮으므로 어떤 정보도 처리됨을 피할 수 없다."라고 했다. 이 조건에 실험된 대상자들은 신 집단보다는 덜하지만 마찬가지로 강한 의인화를 나타냈다. 우리의 견해로서는, 주요 결과는 사람들이 AI가 인간과 다르다고 의식적으로 믿을 때에도, AI가 (신만큼 의인화되지는 않지만) 인간인 것처럼 시나리오를 상상한다는 것이다.
의인화 바이어스는 고의적이지 않아도, 무의식적으로, 아는 것에 반해 일어나므로 미묘하고도 위험한 문제이다.

펄프 공상 과학의 시절에, 잡지 표지에는 찢어진 드레스를 입은 매력적인 여성을 납치하는 외계 괴물(일상 회화에서는 "벌레 눈을 가진 괴물" (bug-eyed monster, BEM) 이라고 했다)이 그려져 있었다. 그 그림의 작가가, 전혀 다른 진화적 역사를 가진 외계 괴물이 인간 여성들에게 성적 욕구를 느낀다고 생각한 것으로 보일 것이다. 사람들은 "모든 뇌가 대체로 같게 짜져 있으므로, 아마 BEM에게도 인간 여성이 성적 매력이 있을 것이다" 식으로, 명시적인 추론으로는 그런 실수를 하지 않는다. 아마도 그 작가는 "거대 곤충형 괴물이 정말 인간 여자를 매력 있다고 느낄까?" *묻지를* 않았을 것이다. 대신, 드레스가 찢긴 여자는 원래, 본질적으로 섹시하다고 생각했을 것이다.

이 오류를 저지른 이들은 곤충형 괴물의 심리는 염두에 두지 않고, 여자의 찢긴 드레스에 집중한 것이다. 드레스가 찢기지 않았으면 BEM과는 상관없이 여자는 덜 섹시했을 것으로 생각을 했던 것이다. (이 깊고 혼돈스러우면서 매우 빈번히 일어나는 오류를 E.T. 제인스는 "심리 투영 오류"라고 불렀다. (Jaynes and Bretthorst, 2003.) 베이지안 확률론의 연구가인 제인스는지식의 상태를 사물의 성질과 혼동하는 오류를 "심리 투영 오류"라고 일컬었다. 예로, "불가사의한 현상"이라는 표현은 "불가사의"가 그 현상 자체의 성질이라는 것을 의미하는 말이다. 내가 어떤 현상에 대해 무지한 것은 내 지식의 상태에 대해 말하는 것이지, 현상에 대해 말하는 것은 아니다.) 사람들은 자신들이 의인화하고 있다는 사실 (또는 자신들이 다른 지능의 행동을 예측한다는 의심스러운 일을 하고 있다는 사실)을 모르고 있어도 그들의 인지에 의인화가 일어날 수도 있다. 우리가 다른 지능에 대해 추리할 때, 추론의 단계 하나하나가 인간의 경험에 너무나 보편적이어서 공기나 중력처럼 느끼지 못하는 가정들에 의해 오염될 수 있다. 잡지의 삽화가에게 누군가 "아니, 수컷 거대 외계 벌레는 암컷 거대 외계 벌레에게 성욕을 가질 확률이 더 높지 않은가?"라고 항의한다면, 삽화가는 잠시 생각하고는 "음, 처음에는 외계 곤충이 단단한 외골격을 좋아한다고 해도, 그 외계 곤충이 인간 여성들을 만나면 인간 여성들이 훨씬 더 보들보들하고 좋은 피부를 가지고 있다는 것을 알게 될 것이다. 외계 곤충들의 기술이 충분히 발달했으면, 그들은 자신들이 단단한 외골격 대신 부드러운 피부를 좋아하도록 자신들의 유전자를 조작할 것"이라고 말한다.

이는 한 발짝만 떨어진 오류이다. 외계 괴물의 의인화된 사고가 지적된 후에, 잡지 삽화가는 한 발짝 물러서며 외계 괴물이 내린 결론을 외계 괴물의 추리 과정의 결과로 정당하려고 하고 있다. 기술이 발전된 외계인들이 자신들을 (유전적으로나 다른 방법으로나) 부드러운 피부를 좋아하도록 바꿀 수 있을지 모르지만, 그들이 자신들을 그렇게 바꾸기를 정작 원하겠는가? 단단한 외골격을 좋아하는 곤충형 외계인은 - 자연선택으로 인간다운 메타-섹시함을 가지지 않았다면 - 자신을 부드러운 피부를 대신 좋아하도록 바꾸고 싶지 않을 것이다. 의인적인 결론을 내리기 위해 길고 복잡한 추리를 사용할 때에는 추리의 단계가 각각 오류를 집어놓을 기회가 되어버린다.

결론에서부터 시작해 중립적으로 보이는 추리를 찾으려고 하는 것도 역시 심각한 오류이다. 이는 오류를 합리화하려는 것이다. 자신의 뇌를 통해 조회함이 곤충형 괴물이 여자를 쫓는 상상을 만들어냈다면, 그것이 의인화라는 사실은 어떠한 합리화를 해도 바꿀 수 없다. 자신의 의인화 바이어스를 줄이기를 바라는 이는, 연습으로 진화 생물학을 (될 수 있으면 수학적으로) 공부하는 것이 좋다. 과거의 생물학자들은 흔히 자연선택을 의인화하며, 진화가 자신들이 했을 일을 할 것으로 예측하고 자신들을 "진화의 처지에서" 생각함으로 진화의 영향을 예측하려고 했다. 이는 1960년대 말에 와서야 Williams (1966) 등에 의해 체계적으로 근절될 대량의 난센스를 낳는 결과를 가져왔다. 진화 생물학은 의인화 바이어스를 물리치는 것에 도움이 되는 수학과 사례 연구들을 제공해 준다.

1.1: 지성들의 설계 공간의 크기

진화는 특정한 구조들을 보존하는 경향이 강하다. 이전부터 존재하는 유전자에 의존하는 유전자들이 출현하면, 이전의 유전자는 다량의 적응을 고장내지 않고서는 변이할 수 없다. 배아의 체제(體制, body plan) 형성을 제어하는 Homeotic 유전자들은 다른 유전자들이 활성화되는 시간을 제어한다. Homeotic 유전자를 돌연변이시키면 다른 면에서는 정상적으로 발달하는데 머리가 없는 배아가 생긴다. 결과로 homeotic 유전자들은 보존되는 경향이 강하다 보니 그 중 다수가 인간과 초파리가 똑같다. 인간과 곤충의 공통조상 때부터 변하지 않았던 것이다. ATP 합성효소는 진

핵생물의 출현 이래 크게 변화하지 않았으므로, 동물의 미토콘드리아, 식물의 엽록체, 박테리아는 가진 ATP 합성효소의 분자적 장치가 본질적으로 같다.
어떤 두 가지의 AI 설계는 당신과 페투니아가 다른 것보다 더 다를 수 있다.
"인공지능"이란 용어는 "호모 사피엔스"란 용어보다 방대하게 큰 *가능성의 공간을* 가리킨다. "AI"를 가리킬 때는 사실 임의적인 지성들의 공간이 또는 *임의 최적화 프로세스의 집합*을 가리키는 것이다. 심리 설계 공간의 도표를 상상해 보라. 한구석에는 인간들이 조그마한 원 안에 있고, 그 원은 모든 생물학적 생명이 들어 있는 더 큰 원 안에 있고, 이 방대한 도표의 나머지까지 합치면 일반 심리의 공간을 구성하게 되는 것이다. 이 도표는 이어서 보다 더 방대한 공간인 최적화 프로세스의 공간에 포함되어 있다. 자연선택은 인식력이 없이도 복합적인 장치를 창조한다. 그러므로 진화는 최적화 프로세스의 공간 안에 존재하나 심리의 공간 밖에 놓여 있다.
가능성의 공간이 이렇게 거대하므로 의인화는 합리적인 논증으로서 있으면 안 되는 것이다.

2: 예측과 설계

우리는 우리 자신의 뇌를 조회함으로 벌레 눈을 가진 괴물이든지, 자연 선택이든지, 인공 지능이든지, 인간이 아닌 최적화 프로세스에 대해서는 답을 구할 수가 없다. 그러면 어떻게 나아갈 수 있을 것인가? 인공 지능이 무엇을 할 것인지, 어떻게 예측할 수 있을 것인가? 나는 고의로 이 문제를 풀기 어려운 형태로 제기한다. 정지 문제때문에, *임의의* 계산 시스템이 어떤 입력-출력 함수(예를 들어 간단한 곱셈)을 계산하는지 예측할 수 없다. (Rice 1953) 그러면 어떻게 인간 공학자들은 신뢰성 있게 곱셈을 계산할 수 있는 컴퓨터 칩을 조립할 수 있는가? 인간들은 의도적으로 자기가 실제로 이해할 수 있는 설계를 사용하기 때문이다.
의인화는 사람들로 해금 어떤 것이 "지적인 존재"라는 정보밖에 가지지 않은 상태에서 그 "지능"을 예측할 수 있다고 믿게 한다. 뇌가 자동으로 자신을 그 "지능"의 처지에서 생각하면서 의인화는 계속 예측을 생성시킨다. 이 요인은, AI의 어려움 자체가 아니라 어떤 AI 설계가 무슨 일을 하는지를 잘못된 생각을 하기가 이상하게 쉬운 것에서 비롯된 AI의 부끄러운 역사의 한 원인이었을 수도 있다.
어떤 다리가 30톤까지 나가는 차들을 지탱할 수 있다고 단정할 때, 토목 기사들은 두 가지의 무기가 있다. 초기 조건의 선택과 안전 마진이다. 30톤의 차를 지탱할 수 있다고 할 수 있는 다리 단 한 가지만 설계하면 되지, *임의의* 건축물이 30톤의 차를 지탱할 수 있는지는 예측할 필요가 없다. 다리가 지탱할 수 있는 무게를 정확히 계산할 수 있으면 좋으나, 다리가 지탱할 수 있는 차의 무게를 *최소한* 30톤으로 계산해도 (이것을 *엄밀히* 단정하는 데 필요한 이론의 이해가 정확한 계산에 필요한 이해와 대체로 같을 것이나) 무난하다.
토목 기사들은 다리가 차량을 지탱할 수 있는지 예측할 때 엄한 기준을 세운다. 고대의 연금술사들은 어떤 반응물들을 어떤 순서로 반응시키면 납이 금으로 변환되리라고 예측할 때 훨씬 더 느슨한 기준으로 했다. 얼마만의 납이 얼마만의 금으로 바뀐다는 것일까? 이 일이 일어날 수 있는 인과적 메커니즘은 도대체 무엇일까? 연금술 연구가가 납 대신 금을 원하는 이유는 확실하지만, 이 물질들을 이렇게 반응시키면 금이 납으로, 또는 납이 물로 변환되지 않고, 하필이면 납이 금이 될 이유는 무엇일까? 등의 문제는 간과하고 말이다.
초기의 어떤 AI 연구가들은 역전파(backpropagation)로 학습하는 다층 임계 유닛들로 된 인공 신경망이 "지능"을 얻으리라고 믿었다. 이 희망적 관측은 아마도 토목공학보다 연금술에 더 가까웠을 것이다. 마법은 도널드 브라운의 인간 보편적인 요소 중 하나이나(Brown 1991) 과학은 그 중 하나가 아니다. 우리는 연금술이 불가능하다는 사실을 직관적으로 알지 못한다. 우리는 엄밀한 이해와 재미있는 이야기를 직관적으로는 구별하지 못한다. 우리는 긍정적인 결과를 기대하는 분위기를 직관적으로는 눈치채지 못한다.
인류는 우연적인 돌연변이의 비우연적인 유지로 작동하는 자연선택에 의해 출현했다. 연구가들이 결합한 시스템이 어떻게 작동하는지 이해하지 못한 채 알고리즘을 이와 비슷한 방식으로 부착함으로 인공지능을 개발하면, 버튼이 무슨 일을 하는지 모른 채 버튼을 누르는 사람에게는 전세계적 재앙을 일으킬 수 있는지름길이다. 친화적인 행동을 보장하는 데 관여하는 과정들은 분명히 설명할 수 없고, 친화성이 무슨 말인지도 모름에도 연구가들은 AI가 친화적일 것이라고 믿는다.

우리는 초기의 AI 연구가들이 그들이 만든 프로그램의 지능에 대해 강하면서 그릇되고 막연한 기대를 했던 것처럼 이 AI 연구가들이 지능적인 프로그램을 설계함에는 성공하나 그 프로그램의 친화성에 대해서 그릇되고 막연한 기대를 가지고 있다고 상상한다.
친화적 AI를 구축하는 방법을 모르는 것은, 당사자가 자신이 모른다는 것을 알면 그 자체로 결코 치명적이지 않다. AI가 설마 친화적일 것이라는 그릇된 믿음이, 바로 전세계적 재앙으로 이르기 쁜한 방향으로 가는 것이다.

3: 지능의 과소평가

우리는 인류의 보편적인 면보다는 개인차를 보는 경향이 많다. 그러니 "지능"이라고 하면 우리는 인간이 아니라 아인슈타인을 떠올린다.
지능의 개인차는 표준화된 명칭이 있다. 스피어먼의 "*g*" 혹은 "*g*-팩터" 는 여러 가지 지능 검사의 점수가 다른 지능검사 점수와 현실 세계의 변수, 혹은 평생 수입과 상관관계가 높다는 확실한 실험 결과에 대한 논란 많은 설명이다. (Jensen 1999). 스피어먼의 *g*는, *종족*으로서는 도마뱀보다 지능이 훨씬 더 높은 인간의 지능 차이에서 통계적으로 얻은 추상적 개념이다. *g*는 거인들의 키의 몇 밀리미터 되는 차이를 추상화한 것이다.
*g*는 다른 종의 생물들이 이해할 수 없는 많은 인지적 작업을 행하는 인간 특유의 능력인 *일반 지능*과 혼동해서는 안 된다. 일반 지능은 종간(種間)의 차이며, 복합적인 적응이며, 알려진 모든 문화에 존재하는 인류에 보편적인 실재이다. 지능에 대한 학문적 합의는 아직 없을지 몰라도, 그 설명할 것의 존재와 힘은 의심의 여지가 없다. 인간이 달에 발자취를 남기게 할 수 있게 하는 무언가는 분명히 있다. 그러나 "지능"이라는 말은 흔히 IQ가 160이면서 굶주리는 교수와 반면에 억만장자인데도 불구하고 IQ가 120에 불과한 CEO를 연상하게 한다. 물론 인간 세계에서 성공에 이바지하는 능력의 개인차 "책벌레"의 지능뿐만이 아니고, 열정, 사회적 기술, 교육, 음악적 재능, 합리성 등도 중요하다. 하지만 내가 열거한 요소들은 모두 *인지적* 요소라는 것을 상기해야 한다. 사회적 능력은 뇌에 존재하지, 간에 존재하지 않는다. 우스개는 따로 두고, 침팬지 중에는 CEO나 대학교수를 찾을 수는 없을 것이다. 마찬가지로 생쥐 중에 찬사받는 이성가들이나, 예술가들이나, 시인들이나, 리더나, 공학자들이나, 솜씨 있는 사회적 네트워커들이나, 무술가들이나, 작곡가들은 찾지 못할 것이다. 지능은 인간이 가진 힘의 기초이며, 인간의 모든 솜씨와 기술의 원동력이다.
일반 지능을 *g*-팩터와 혼동하는 것의 위험은 AI가 미칠 수 있는 영향을 과소평가하게 할 수 있다는 것이다. (이 점은 나쁜 영향뿐 아니라 좋은 영향에도 적용된다.) "초인간적 AI"나 "인공 초지능"이라는 표현도 상자에 들어 있는 책 똑똑이를 떠올릴 수 있다. 그 AI는 체스나 추상적인 수학 등 흔히 "지능"과 관계된 것으로 생각하는 인지적 작업에는 *엄청* 뛰어나나, 초인간적인 설득력은 없고, 사회적 상황에서 예측을 내리고 능숙하게 다루는 데 인간보다 크게 뛰어나지도 않고, 장기적 작전을 짜는 데도 초인간적으로 영리하지 않다고 생각할 수 있다. 그렇다면 아인슈타인 대신에 19세기 정치 외교의 천재 오토 폰 비스마르크를 생각해야 할까? 그러나 이는 그 정반대의 오류일 뿐이다. 바보에서 아인슈타인까지의 범위도, 바보에서 비스마르크까지의 범위도, 아메바에서 인간까지의 범위에 비하면 점같이 작은 범위이다.
"지능"이란 말이 인간이 아니라 아인슈타인을 떠오르게 한다면, 지능이 총은 못 당한다고, 마치 그 당시에 총이 나무에 열렸던 것처럼 해도 일리가 있는 말로 들릴 수 있다. 돈이 지능보다 더 강하다고, 마치 생쥐도 돈을 사용하는 것처럼 말해도 맞는 말로 들릴 수 있다.
인간들은 *처음에는* 다른 종들의 일상적인 도구인 발톱, 이빨, 갑옷 등의 면에서 큰 장점이 없는 상태로 출발했다. 나머지 생태권의 관점에서 봤을 때에는 이 살도 연한 것들이 언젠가 장갑된 탱크를 입을 기미가 전혀 없었다. 우리는 사자와 늑대와 겨루어서 이긴 싸움터 자체를 만들어냈다. 발톱은 발톱으로, 이빨은 이빨로 겨루는 대신, 중요한 것이 무엇인지를 우리가 스스로 판단했다. 창의력의 힘은 이렇다.
빈지(1993)는 적절하게도 인간보다 더 영리한 심리들이 존재하는 미래는 질적으로 다르리라고 말했다. 인공 지능은 의학, 제조공업, 에너지 등 발달의 그래프에는 속하지 않는다. 인공 지능은

함부로 섞을 수 없는 것이다. 적당히 높은 마천루는 자신을 공사하지 않는다. 인류는 다른 종보다 숨을 더 오래 죽임으로써 눈에 띄는 세력이 되지 않았다.
지능의 위력을 과소평가함의 위험은, 누군가 버튼을 만들고는, 설마 이 버튼이 내게 해를 끼칠 수는 없겠지, 하며 그 버튼이 할 수 있는 일에는 정작 주의를 기울이지 않을 수도 있다. 아니면, 지능의 힘이 과소평가됨에 따라 인공지능이 끼칠 힘도 또한 과소평가되어, 인류의 존망 위기를 우려하는 (현재에는 극히 적은) 연구가들, 보조금 지원 단체들 및 개인 자선가들이 인공지능에는 관심을 두지 않을 수도 있다. 아니면, 강AI의 위험을 AI 학계 전체가 무시함으로, 강지능을 개발하기 가능하게 될 때, AI 친화성을 위해 필요한 확고한 기초와 중요한 도구가 없을 것이다.
또한 존망 위험에 영향을 미치므로, 인공 지능이 다른 위험을 해결할 수 있는 강한 도구가 될 수도 있다는 것을 잊어서는 안 된다. 그렇지 않으면 우리가 생존을 위한 최선의 길을 무시하게 될 수도 있기 때문이다. AI가 미칠 수 있는 영향력의 과소평가가 위험하다는 사실은 좋은 영향에도 나쁜 영향에도 똑같이 적용된다. 이 챕터의 제목이 "인공지능이 미치는 전세계적 위험"이 아니라 "전세계적 위험에 인공지능이 미칠 수 있는 긍정적•부정적 영향들"인 것도 그런 이유 탓에서이다. 인공지능의 전망은 세계적 위험에 위협으로만 작용하지 않고, 더 복잡히 작용하기 때문이다. AI가 단순히 부담을 끼쳤으면 해결할 방법은 간단했을 것이다.

4: 능력과 동기의 혼동

인공지능, 특히 초인간적 능력의 인공지능을 논의할 때 자주 범하는 오류가 있다. 누군가 이렇게 말한다. "기술이 적당히 발전하면 인간의 지능을 훨씬 초과하는 기계를 개발할 수 있을 수도 있다. 그럼, 만들 수 있는 치즈케익의 크기는 지능에 달린 것은 분명하다. 그럼 초지능은 어마어마하게 큰 치즈케익… 도시만한 치즈케익을 만들 수도 있겠다. 어머나, 미래는 거대한 치즈케익으로 가득할 거야!" 문제는 그 초지능이 꼭 거대 치즈케익을 만들기를 원하느냐는 것이다. 이 상상은 *가능성*에서부터 *현실*로 곧장, 필요한 중간자인 *동기*는 고려하지 않고 비약을 한다.
아래의 추리는 뒷받침되지 않고 단독으로 본 상태에서 모두 이 거대 치즈케익의 오류를 나타내고 있다:

- 적당히 강력한 AI는 인간의 어떠한 저항도 압도하고 인류를 없애버릴 수 있다. [또한 AI는 그러기로 할 것이다.] 그러므로 AI가 개발되면 안 된다.
- 적당히 강력한 AI는 수백만 명의 생명을 구할 수 있는 새로운 의학 기술을 개발할 것이다. [또한 AI는 그러기로 할 것이다.] 그러므로 AI는 만들어야 한다.
- 컴퓨터가 적당히 값싸지면 작업 대부분이 인간보다 인공지능에 의해서 더 쉽게 수행될 수 있을 것이다. 적당히 강력한 AI는 우리보다 수학, 공학, 음악, 예술 등 우리에게 의미가 있는 일에도 인간보다 더 뛰어날 것이다. [그리고 그 AI는 이런 노동을 하고자 할 것이다.] 그러므로 AI가 발명된 후에는, 인간들은 할 일이 없어져서 굶거나, 혹은 텔레비전이나 보고 있을 것이다.

4.1: 최적화 프로세스

나는 위의 거대 치즈케익의 오류의 해체하는 데 의인화를 이용했다. "동기"가 분리될 수 있는 것이라는 생각, "능력"과 "동기"를 별개의 관념으로 볼 수 있다는 생각. 이는 편리하나 의인적인 분리이다.
이 문제를 더 일반적인 말로 설명하기 위해 나는 *최적화 프로세스*라는 개념을 도입한다. 최적화 프로세스는 큰 탐색 공간의 어떤 작은 부분에 도달함으로 세계에 분명한 영향을 미치는 것이다. 최적화 프로세스는 미래를 가능성 공간의 어느 특정한 영역으로 이끈다. 내가 어느 멀리 떨어진 도시에 가 있고, 그 도시에 사는 친구가 공항으로 나를 태워 보내준다고 하자. 나는 이 동네를 모른다. 교차점에 있을 때 나는 내 친구의 회전을 하나씩으로도, 차례로도 예측하지 못한다. 하지만 나는 친구의 예측 불가한 절차의 *결과*는 "공항에 도착할 것이다."라고 예측할 수 있다. 내 친구의 집이 도시 안 다른 곳에 있어서 전혀 다르게 회전했을지라도 나는 그대로 신뢰성 있게 결과를 예측할 수 있을 것이다. 이는 과학으로 볼 때 매우 기묘한 상황이지 않은가? 절차의 *중간 단계*를 예

측하지 못해도 절차의 최종 결과는 예측할 수 있다는 것이다. 최적화 프로세스가 미래를 이끄는 영역을 최적화 프로세스의 *표적이라고* 하겠다.
자동차, 이를테면 도요타 코롤라를 생각해 보라. 코롤라를 형성하는 원자의 가능한 모든 배열 중에는 극히 적은 일부만이 유용한 자동차를 형성한다. 분자들을 무작위적으로 합쳐서 자동차를 얻으려면 우주의 나이의 수많은 배의 시간이 흘러야 할 것이다. 설계 공간의 또 하나의 극히 작은 부분은 코롤라보다 빠르고 효율적이고 안전한 운송수단들을 나타낸다. 그러므로 코롤라는 설계자의 목표에 따라 *최적*의 운송수단은 아니다. 그러나, 설계자는 코롤라 정도의 품질의 자동차는 커녕 쓸모있는 자동차를 만들기 위해서도 설계 공간의 비교적으로 극히 작은 표적에 맞혀야 했기 때문에 코롤라는 *최적화*의 산물이다. 나무판을 무작위로 자르고 못을 동전 던짐에 따라 박아서는 유용한 손수레를 만들 수 없다. 배열 공간에서 그 정도로 작은 표적에 맞히려면 강력한 최적화 프로세스가 필요하다.
"최적화 프로세스"라는 개념이 실제로 *예측에 유용한* 이유는 최적화 프로세스의 최종적인 *표적*이 최적화 프로세스의 단계별 *다이내믹*보다 이해하기 쉬울 수 있기 때문이다. 위의 코롤라에 대한 논의는 설계자가 "탈것"을 만들려고 했다고 가정하고 있다. 이 가정이 존재한다는 것을 명백히 밝힐 필요는 있으나 이 가정은 코롤라를 이해하는 데 아주 유용하다.

4.2: 최적화 프로세스의 표적

임의적인 지성들의 공간이 인간을 포함하는 조그마한 반점보다 더 넓음을 잊고, "AI"라는 특정한 종족이 "원할" 것이 무엇인지를 생각하려는 유혹에 빠지기 쉽다. 한정 기호를 가능한 모든 심리에 덮을 유혹을 참아야 한다. 미래라는 멀고 이국적인 나라의 이야기를 늘어놓는 자들은, 미래가 꼭 이러리라고 *예상*을 한다. "AI들은 인간들을 로봇 부대로 공격할 것이다", "AI들은 암 치료법을 개발할 것이다."라고 한다. 초기 조건과 결과의 복합적인 관계를 정확히 얘기하면 독자들을 잃기 때문이다. 하지만 우리가 미래를 인류에게 적합한 영역으로 돌리려면 관계적인 이해가 필요하다. 핸들을 돌리지 않으면 지금 가고 있는 곳에 도착할 위험이 있기 때문이다.
결정적인 문제는 "AI들"이 인간들을 로봇 부대로 공격하거나 아니면 암 치료법을 개발하리라고 *예상*하는 것이 아니다. 어떤 예상들을 *임의*의 AI 설계에 적용하는 것도 아니다. 우리가 직면하는 문제는 유익한 효과를 보장할 수 있는 어떤 특정된 강한 최적화 프로세스를 선택해 만드는 것이다.
나는 독자들이 완전히 일반적인 최적화 프로세스가 친화적일 이유를 생각해내지 않기를 강력히 촉구한다. 자연 선택은 친화적이지 않고, 당신을 미워하지도 않으며, 당신을 혼자 내버려 두지도 않을 것이다. 진화는 당신이 작동하는 것처럼 작동하지 않으므로, 진화는 그렇게 의인화될 수 없다. 1960년대 이전, 많은 생물학자는 진화가 온갖 친절한 일을 할 것으로 기대하고, 자연 선택이 그 일들을 할 온갖 복잡한 이유를 합리화해냈다. 그들은 실망했다. 자연 선택 자체는 인간이 좋다고 할 결과를 원한다는 것을 알면서 시작하고 선택 압력으로 좋은 결과를 생산해내기 위한 복잡한 방법들을 합리화해내지 않기 때문이다. 따라서 자연의 사건들은 1960년대 이전 생물학자들의 정신 안에서 일어난 사건들과 인과적으로 다른 과정이었음에 인해, 예측과 현실이 빗나가게 된 것이다.
희망적 사고는 세부를 더하고, 예측을 제약하며 미개연성의 부담을 가한다. 다리가 무너지지 않기를 바라는 토목기사는 어떤가? 그 토목기사는 일반적인 다리가 무너질 가능성이 작다고 논해야 하는가? 그러나 자연은 일반적인 다리가 무너지지 않을 이유를 합리화하지 않는다. 대신 그 토목기사는 미개연성의 부담을 구체적인 이해를 이용해 구체적인 선택을 한다. 토목 기사는 다리를 원하는 데에서 시작하여, 엄밀한 이론으로 차를 받칠 수 있는 다리 설계를 선택하고, 그 설계에 맞춘 구조를 가진 현실 세계의 다리를 건설하므로 실제 다리가 차를 받칠 수 있는 것이다. 그러하여 예측과 현실이 원하는 대로 맞추어진다.

5: 친화적 AI

인류가 어떤 특정한 표적을 가진 강력한 최적화 프로세스를 꼽아 만들 수 있으면 아주 좋을 것이다. 더 구어체적으로 말하면, 친절한 AI를 만드는 방법을 알면 좋을 것이다. 이 문제에 착수하기에 필요한 *지식의 분야*를 말하기 위해 나는 "친화적 AI"(Friendly AI; 일상적으로 "친절한" 등의 의미가 있는 friendly와 혼동시키지 않기 위해 첫 F을 대문자로 쓴다)라는 용어를 제안했다. "친화적 AI"라는 말은, 이 분야 외에 또 이 기술의 생산물, 곧 특정된 동기를 갖게 만들어진 AI를 말한다.

내가 흔히 접하는 반응 한 가지는, 적당히 강한 AI는 자신의 소스코드를 변경함으로써 자신에게 주어진 제약을 넘어설 수 있으므로 친화적 AI는 불가능하리라고 말하는 것이다. 첫째로 주목할 허점은 거대 치즈케익 오류이다. 자신의 소스코드에 접근할 수 있는 AI는 원칙적으로 자신의 최적화 표적을 바꾸는 방향으로 자신의 소스코드를 변경할 *능력*은 있으나, 이것이 AI가 자신의 동기를 바꿀 *동기*가 있다는 뜻은 아니다. 내가 의도적으로 내가 살인을 즐기게 할 알약을 먹지 않을 이유는, 현재에는 다른 사람들이 죽지 않기를 바라기 때문이다.

그런데 혹시 내가 나를 변경할 때 실수를 하면 어떻게 할까? 컴퓨터 공학자들이 칩의 정당성을 증명할 때 (칩이 1억 5,500만 개의 트랜지스터를 가지고 있고 후에 패치를 발행할 수 없을 때에는 좋은 방책이다), 그들은 인간이 안내하고 컴퓨터가 확인하는 형식적 증명을 한다. 형식적 수학적 증명이 주는 혜택은 10억 개의 단계로 된 증명도 열 개의 단계로 된 증명만큼 신뢰성이 있다는 것이다. 그러나 인간들은 오류를 놓치기 쉬우므로 10억 단계로 된 증명을 점검하기에 믿을 만하지 않다. 그리고 현존하는 알고리즘들은 탐색 공간에 지수적 폭발을 하기 때문에 현재의 정리 증명 기술들은 컴퓨터 칩 전부를 설계하고 증명할 수 있을 만큼 똑똑하지 않다. 인간 수학자들은 지수적 폭발의 걱정 없이 현재 정리 증명 시스템보다 훨씬 더 복잡한 정리들을 증명할 수가 있다. 그러나 인간 수학은 비형식적이고 신뢰성이 없다 – 가끔은 이전에 받아들여져 있던 비형식적 증명에 결함이 발견된다. 그러나 인간 공학자들은 증명기를 증명의 중간 단계를 통해 안내한다는 것이다. 인간이 다음 보조정리를 선택하고는 복잡한 증명기가 형식적 증명을 생성시키는 일을 하고, 간단한 검증기가 증명의 단계들을 검사한다. 그렇게 해서 현대의 엔지니어들이 1억 5,500만 개의 상호 의존하는 부분으로 된 기계를 설계하는 것이다.

컴퓨터 칩의 정당성을 증명하는 것은, 현재는 인간의 지능도 컴퓨터 알고리즘도 충분하지 않으므로, 둘이 협력해야 한다. 진정한 AI는 능력의 비슷한 조합을 이용하여 자신의 코드를 변경할 수 있을지도 모른다. 지수적 폭발 없이 복잡한 디자인들을 만들어 낼 능력과 또한 신뢰성이 높게 단계들을 점검할 능력이 있을 것이다. 자기 변경을 여러 번 행하고도 진정한 AI가 증명 가능한 정당성을 유지할 수 있을 한 가지 방법이다.

이 논문에서는 위의 개념을 상세히 검토하지 않을 것이다. (비슷한 개념을 보려면 Schmidhuber 2003을 보라.) 그러나 도전을 불가능하다고 단정 짓기 전, 특히 이해관계가 걸려 있는 도전일 때는, 그 문제를 생각해보고 가능한 한 기술적으로 정밀하게 연구를 해야 할 것이다. 창의적으로 어떤 도전을 잘 살펴보지 않고 해결 불가능하다고 단정하는 것은 인류의 독창서을 경멸하는 것이다. 어떤 일을 할 수 없다는 것은 엄청나게 강한 논제이다. 공기보다 무거운 비행기를 만들 수 없다는지, 원자 반응에서 유용한 에너지를 얻을 수 없다든지, 달로 갈 수 없다든지 말이다. 그런 논제들은 보편적인 일반화이다. 문제를 풀기 위해 모든 사람이 생각해 낼 수 있는 모든 해결책에 다양화하는 것이다. 전칭 양화사는 하나의 반례만 있으면 반증된다. 친화적 (또는 친절한)가 이론적으로 불가능하다는 것은 감히 가능한 모든, 친절하고 더 친절하기를 바라는 어떤 인간들까지 포함해서 설계와 가능한 모든 최적화 프로세스에 양화하고 있는 것이다. 현재에는 친화적 AI가 인간에게 불가능할 수 있는 몇 가지 애매한 이유가 있고, 더 가능성이 큰 일은 문제는 해결할 수 있지만 우리가 때맞추어 해결하지 않을 수도 있다. 그러나 이해관계를 생각하고는 이 도전을 성급히 배제해서는 안될 것이다.

6: 기술적 실패와 철학적 실패

보스트롬(Bostrom 2001)은 존재 재앙을 지구의 지적 생명을 영구히 절멸시키거나 또는 그것의 잠재력의 일부를 파괴하는 재앙으로 정의한다. 우리는 친화적 AI의 시도의 실패를 두 개의 비형식적인 퍼지(fuzzy) 종류들, 기술적 실패와 철학적 실패로 구분할 수 있다. 기술적 실패란 AI를

작성할 때 AI가 프로그래머가 기대했던 방식으로 작동하지 않는 일을 뜻한다 – 쓴 코드의 메커니즘을 이해하지 못함에서 발생하는 실패이다. 철학적 실패란 잘못된 것을 만들려고 하는 것이다. 이것을 건설하는 데 성공해도 인간들을 돕지 못하는 것이다. 말할 것도 없이 이 두 가지의 실패는 상호 배타적이지 않다.
대부분의 철학적 실패들은 기술적 지식이 있을 때 훨씬 더 쉽게 설명되기 때문에 이 두 경우의 경계가 모호하다. 이로상 원하는 것을 먼저 생각하고 그다음에 성취할 방법을 찾아야 한다. 실제 원하는 것을 알아내기 위해서는 흔히 깊은 기술적 이해가 필요하다.

6.1: 철학적 실패의 예

19세기 말, 정직하고 머리가 좋았던 사람 중 많은 사람이 최선의 의도로 공산주의를 옹호했다. 공산주의 밈을 만들어내고, 퍼뜨리고, 받아들인 사람들은 엄연한 역사적 사실로 보았을 때는 이상가들이었다. 최초의 공산주의자들을 경고할 소련이라는 본보기가 없었기 때문이다. *뒷궁리의 혜택이 없었던 그 당시에는 그런대로 좋은 생각 같았을 것이다.* 혁명 후에 공산주의자들이 정권을 잡아 권력에 물들었을 때에는 다른 동기가 개입되었을 수 있으나, 이는 예상하기 아무리 쉬웠어도 최초의 이상주의자들이 예상한 것이 아니었다. 큰 비극의 장본인이 꼭 사악하거나, 심지어는 *대단히* 바보일 필요가 없다는 것을 아는 것이 중요하다. 우리가 모든 비극을 악이나 대단한 어리석음에 돌리면, 우리는 자신을 보고, 우리가 사악하거나 대단히 어리석지 않음을 맞게도 인식하고는 “우리에게는 그런 일이 없을 거야”라고 말하게 된다.
최초의 공산주의 혁명가들이 혁명의 관찰 결과일 것으로 생각한 것은 사람들의 생활이 개선되는 것이었다. 노동자들은 돈을 적게 받으면서 고된 노동을 하지 않아도 될 것이었다. 부드럽게 말하면, 사실은 그렇지 않았던 것으로 되었다. 그러나 최초의 공산주의 혁명가들이 일어날 것으로 생각했던 관찰적 결과는 다른 정치 제도의 지지자들이 그들이 좋아하는 정치 조직의 관찰적 결과일 것으로 생각한 것과 크게 다르지 않았다. 그들은 사람들이 결과로 행복해질 것으로 생각했다. 그들은 잘못 생각했던 것이다.
이제 누군가 "친화적" AI를, 유토피아를 가져올 것으로 생각되는 공산주의, 자유지상주의, 무정부 봉건주의, 또는 *좋아하는정치제도*를 실행하도록 프로그램한다고 가정하자. 사람들이 좋아하는 정치 제도들을 생각하면 긍정적 정서가 불탈 것이므로, 제안하는 이에게는 아주 좋은 생각으로 들릴 것이다.
우리는 프로그래머의 실패를 윤리적•도덕적인 관점에서 보아서, 프로그래머가 자신을 지나치게 믿은 나머지 자신이 틀렸을 가능성을 생각하지 않고, 공산주의가 사실 잘못되었을 가능성은 생각하지 않아서 발생한 일이라고 할 수 있다. 그러나 베이지안 결정 이론에서는 이 문제에 대한 보완적인 기술적 관점이 있다. 결정 이론의 관점에서는 공산주의의 선택은 관찰적 믿음을 가치 판단과 합침으로 된 것이다. 관찰적 믿음은 공산주의가 실행되면 특정한 결과 또는 특정한 집합의 결과가 발생하리라는 것이다: 사람들이 더 행복해지고, 일을 더 적게 하고 물질적으로 더 부유해질 것이라는 것이다. 이는 결국 관찰적 예측이다. "행복"도 측정하지 어렵지만 뇌의 상태의 실제 성질이다. 공산주의가 실행되면, 이 결과가 일어나거나 혹은 안 일어날 것이다. 가치 판단은 이 결과가 충족적이거나 현재 조건에 비해 낫다는 것이다. 공산주의의 현실적인 결과에 대한 다른 관찰적 믿음이 주어지면, 결정도 상응하는 변화를 할 것이다.
우리는 진정한 AI, 또는 인공 일반지능은 자신의 관찰적 믿음 (또는 세계의 개연적 모형화, 등...)을 바꿀 수 있다고 기대할 수 있다. 어떻게 되어서 찰즈 배비지가 니콜라우스 코페르니쿠스 이전에 살았고, 컴퓨터가 망원경 이전에 발명되었고, 그 시대의 프로그래머들이 인공 일반지능을 건설하는 데 성공했다면, AI가 그 후 영원히 태양이 지구를 돈다고 믿을 것으로 성립되지 않았을 것이다. 프로그래머들이 추리를 천문학보다 더 잘 이해했다면, 이 AI는 프로그래머들의 사실적인 오류를 초월할 수 있다. 행성들의 궤도를 발견하는 AI를 건설하려면 프로그래머들은 뉴턴 역학의 수학을 물라도 되고 베이지안 확률론만 알면 된다.
AI를 공산주의나 다른 정치 제도를 실행하게 프로그램하는 것의 오류는 목적 대신 수단을 프로그램하고 있다는 것이다. 공산주의가 실행된 결과에 대해 개선된 지식을 얻을 때 재평가된 기회를

주지 않고 고정된 결정을 프로그램하고 있다는 것이다. 결정을 내린 오류가 있는 프로세스를 더 높은 수준의 지능에서 재평가하도록 하지 않고 AI에게 고정된 결정을 주고 있다는 것이다.
만약 내가 더 강한 선수를 상대로 체스를 둔다면, 나는 상대가 둘 수를 미리 예측할 수 없다. 내가 그것을 예측할 수 있다면 필연적으로 내가 그 정도로 체스에 강할 것이다. 그러나 나는 최종 결과는 내 상대가 이길 것으로 예측할 수 있다. 나는 내 상대가 향하고 있는 가능한 미래들의 영역을 알기 때문에 진로를 알 수 없어도 목적지를 예측할 수 있는 것이다. 내가 가장 창의적일 때에 내 행동을 예측하기 가장 어렵지만, 내 목적을 알고 이해하면 내 행동의 결과는 예측하기 가장 쉬울 것이다. 인간보다 뛰어난 체스 선수를 만들 때에는 이기는 수를 검색하게 프로그램해야 할 것이다. 특정한 수를 프로그램하면 나보다 뛰어난 체스 선수가 되지 않을 것이기 때문에 특정한 수는 프로그램하면 안 된다. 내가 검색을 시작할 때에는 답을 정확히 예측할 능력을 희생하게 되는 것이다. 아주 좋은 해답을 구할 때, 문제가 무슨 문제인지 알 능력은 희생하면 안 되지만 때에는 답을 예측할 능력을 희생해야 한다.
공산주의를 직접 프로그램함 등의 오류는 결정 이론의 언어를 말하는 AGI 프로그래머를 유혹하지 않을 것이다. 나는 이 오류를 기술적 지식이 결핍되어 일어난 철학적 실패라고 하겠다.

6.2: 기술적 실패의 예

> 지능형 기계의 행동을 억누르는 규칙 대신에, 우리는 기계들에게 행동의 학습을 인도하는 감정을 주어야 한다. 이 기계들은 우리가 곧 "사랑"이라고 일컫는 감정으로 우리가 행복하고 성공하기를 바라야 한다. 우리는 가장 기본적이고 선천적인 감정이 모든 인간에 대한 조건 없는 사랑인 지능형 기계를 설계할 수 있다. 먼저 우리는 인간의 표정, 인간의 목소리, 인간의 신체 표현에서 행복과 불행을 인식하기를 배우는 비교적 간단한 시스템을 구축할 수 있다. 그다음에는 이 학습의 결과가 선천적인 감정적 가치로서 배선되어 우리가 행복하면 정적으로 강화되고 우리가 불행하면 부적으로 강화되는 더 복잡한 지능 기계를 만들 수 있다. 그 기계들은 미래를 예측하는 알고리즘을 배울 수 있다. 그러므로 우리는 미래의 인간의 행복을 예측하는 알고리즘을 배우고 이 예측들을 감정적 가치관으로 사용하는 지능적 기계를 프로그램할 수 있다."

-- 빌 히버드 (2001), 『초지능형 기계들』

옛날 옛적에, 미국 육군은 위장된 적군 탱크를 자동으로 탐지하기 위해서 신경망을 이용하기로 했다. 연구가들은 신경 회로망을 나무 사이에 위장된 탱크의 사진 50개, 탱크가 없는 나무들의 사진 50장으로 훈련했다. 지도 학습에 쓰이는 표준 기법을 이용하여, 연구가들은 이 신경망이 훈련 세트를 정확하게 분류하는 가중치를 쓰도록 - 위장된 탱크 사진 50개에는 "예", 보통 숲의 사진 50장에는 "아니오"를 출력하도록 훈련했다. 이는 전에 보지 못했던 예들이 바르게 분류되기를 보증하기도, 함축하지도 못했다. 그 신경망은 새로운 상황으로 일반화되지 못하는 100가지를 배웠을 수도 있었다. 현명하게도 연구가들은 원래 탱크 사진 100장, 나무 사진 100장, 모두 200장을 찍었었다. 연구가들이 나머지 100장에 신경망에 실행했더니, 신경망은 추가 훈련 없이 나머지 사진들을 모두 맞추었다. 확증된 성공이었다! 그런데 연구가들이 완성된 신경망을 펜타곤에 제출한 다음, 펜타곤은 신경망이 우연보다 사진을 구분하는데 잘하지 못했다고 불평하며 돌려주었다…
알고 보니 연구가들의 데이터 세트에 있었던 위장된 탱크의 사진들은 흐린 날에 찍었던 것이었고, 빈 숲의 사진들은 맑은 날에 찍었던 것이었다. 신경망은 위장된 탱크를 빈 숲과 구별하는 대신, 흐린 날을 맑은 날과 구별하기를 배웠던 것이다.
기술적 실패는 코드가 프로그램되었던 대로 충실히 실행하지만 생각했던 대로 기능하지 않는 일이다. 같은 데이터를 출력하는 모형화는 하나 이상이다. 우리가 신경망을 인간의 웃는 얼굴을 인지하고 웃는 얼굴과 찌푸린 얼굴을 구별하게 학습시켰다고 하자. 그 신경망은 조그마한 웃는 얼굴 그림을 웃는 인간의 얼굴과 똑같은 끌개로 구분하겠는가? 그런 코드가 "배선된" AI는 만약 능력이 있으면 - Hibbard(2001)와 같이 초지능을 말한다면 - 그러면 은하수를 분자로 된 웃는 얼굴 그림으로 가득 채울 것인가?

이 형태의 실패는 AI가 고정된 환경에서는 잘 기능하고 있는 것처럼 보이지만 환경이 바뀌면 실패가 되어버리므로 특히 위험하다. 탱크 분류기 일화의 연구가들은 학습 데이터를 맞출 때까지 학습시키고 추가 데이터로 (추가 학습 없이) 점검했다. 불행하게도 학습 데이터와 점검 데이터는 다 데이터에 적용된 가정들이 있었는데, 실제로 신경망이 쓰일 환경에서는 확립되지 않았다는 것이다. 탱크 분류기 일화에서는 탱크 사진들은 흐린 날에 찍었다는 가정이다.
점점 강해지고 있는 AI를 개발하고자 한다고 하자. AI는, 인간 프로그래머들이 AI의 전기 공급만 제어하는 것이 아니라 AI보다 더 똑똑하고 창의적이고 교묘하다는 의미에서 더 강력한, *개발 단계*가 있다. 개발 단계 도중에는 AI의 허락 없이 프로그래머들이 AI의 코드를 바꿀 수 있다고 가정한다. 그러나 AI는 *개발 이후의 단계*, 히버드의 경우에는 초지능도 포함할 의도로 만들어진 것이다. 초인간적인 지능의 AI는 허락 없이 변경될 수 없을 것이다. 그 시점에서는 AI가 제대로 작동한다는 보장을 하기 위해서는 이미 확립된 목적 시스템에 의존해야 한다. 그 시점에는 AI가 우리가 수정하려고 하면 저항을 하고, 인간보다 뛰어나며 아마도 이길 것 때문이다.
발달하는 AI의 목적 시스템을 신경망을 학습시켜서 만드는 것은 AI의 개발 단계와 개발 후의 단계 사이의 큰 환경의 변화가 일어날 것이다. 개발 단계에서는 AI가 연구가들이 의도했던 대로 연구가들이 주는 작업을 해서 "웃는 인간 얼굴" 카테고리로 향하는 끌개들에 떨어지는 자극만 산출할 수 있지만, AI가 인간보다 지능이 높고 나노기술 기반기술을 만들었을 때에는 AI가 은하를 조그마한 웃는 얼굴들로 채움으로써 같은 끌개에 떨어지는 자극을 자신에게 술 수 있을 수도 있다. 그러하여 AI가 개발 도중에는 괜찮게 작동하는 것으로 보이지만 프로그래머들보다 똑똑해진 후에는 파멸적인 결과를 가져오게 될 수 있다. "그런데 그 AI가 그건 우리가 의도했던 결과가 아니라는 것은 알지 않겠는가?"라고 생각할 유혹이 있다. 그러나 코드는 AI가 검토하여 코드가 잘못된 일을 하면 돌려주게 AI에게 주어져 있지 않다. 코드가 바로 AI이다. 아마 우리가 충분한 노력과 이해가 있으면, 우리가 코드를 잘못 썼는지를 신경쓰는, 이른바 전설의 DWIM(프로그래머들 사이에서는 Do-What-I-Mean을 뜻한다) 명령어를 작성할 수 있을지도 모른다(Raymond 2003). 그러나 DWIM 다이내믹을 작성하려면 노력이 필요하고, 히버드의 제안에는 우리가 시키는 대로 하지 않고 우리가 의도했던 대로 하는 AI를 설계하는 것의 언급이 어디에도 없다. 현대 칩들은 코드를 DWIM하지 않는다. DWIM은 칩의 자동인 성질이 아니기 때문이다. 또한 DWIM 자체를 망치면 그 결과를 경험하게 될 것이다. 예를 들어 DWIM이 코드에 대한 프로그래머의 만족도를 최대화하는 것으로 정의되었다고 하자. 이 코드가 초지능으로서 실행되면 코드에 대해 최대의 만족을 느끼게 프로그래머들의 뇌를 재구성할 수도 있다. 나는 이 일이 불가피하다고 말하는 것이 아니다. 나는 단지 "내가 의도하는 대로 해라"가 친화적 AI의 중요하고 비단순(non-trivial)한 기술적 도전임을 지적하고자 한다.

[2]이 일화는 유명하고 흔히 사실로 언급되지만, 실제 사건이 아닐 수도 있다. 나는 직접 얻은 정보를 찾지 못했다. (출처가 없는 기술들을 보려면 Crochat and Franklin (2000) 또는 http://neil.fraser.name/writing/tank/ 등을 보라.) 그러나 이러한 식의 실패는 실제 신경망을 구축하고 테스트할 때 고려할 큰 문제이다.
[3]빌 히버드는 이 논문의 초안을 본 후 "탱크 분류기" 문제에의 유추는 강화학습 일반적으로 적용되지 않는다고 논하는 답변을 썼다(http://www.ssec.wisc.edu/~billh/g/AIRisk_Reply.html 에 있다.). 거기에 대한 나의 답변은 http://yudkowsky.net/AIRisk_Hibbard.html 에 있다. 히버드는 또한 Hibbard(2001)의 제안이 Hibbard(2004)에 의해 대체되었다고 언급한다. Hibbard(2004)는 인간들의 동의하는 표현이 행복의 인식을 강화시키고, 인식된 행복이 행동 방책을 강화시키는 두 레벨로 된 시스템을 권장한다.

7: 지능 증가의 속도

치명적 위험의 관점에서 볼 때는, 인공 지능은 *매우 빠르게* 지능이 증가할 수 있다는 것이 가장 중대한 논점이다. 이 가능성을 짐작할 명백한 이유는 재귀적 자기개선이다. (Good 1965) AI가 똑똑해지면 AI 내부의 인지적 기능들을 재구성하는 데 더 똑똑해지고, AI는 자신의 기능들을 개선할 수 있고, 그 후에는 AI가 더욱더 똑똑해지고 자신을 더 잘 개선하게 된다는 것이다.

인간들은 자기개선을 하지 않는다. 우리는 한정된 범위에서 자신을 학습, 연습 등을 통해 개선한다. 한정된 범위에서는 이 개선들이 우리의 개선 능력을 개선할 수 있다. 새로운 발견들은 새로운 발견을 할 능력을 개선한다. 그 의미에서는 지식이 자신을 개선한다. 그러나 우리가 손대지 않은 기초 단계가 있다. 발견의 근원인 뇌는 1만 년 전과 사실상 같다.
유사한 의미에서 자연선택은 생물들을 개선하지만 자연 선택 자체는 강한 의미에서 개선하지 않는다. 적응은 추가적응으로 갈 길을 열기도 한다. 이 의미에서는 적응도 자신을 개선한다. 그러나 유전자 풀이 끓을 때에도 자연 선택의 바탕이 되는 히터인, 자신은 재구성되지 않는 돌연변이와 재조합과 선택이다. 몇 가지의 드물고 혁신적인 적응, 예를 들어 성적 재조합의 출현은 진화의 속도 자체를 증가시켰다. 그러나 성도 추상적 지능이 없고, 돌연변이에 의존하고, 맹목적, 점진적이고, 대립 유전자 빈도에 집중하는 진화의 본질적인 성질을 바꾸지 않았다. 비슷하게 과학의 출현도 인간의 뇌의 본질적인 성질, 변연계, 대뇌 피질, 전전두엽에 있는 자기 모형, 특유의 200헤르츠의 속도를 바꾸지 않았다.
인공 지능은 자신의 코드를 처음부터 다시 씀으로 최적화의 다이내믹을 바꿀 수 있을지도 모른다. 그러한 최적화 프로세스는 진화가 적응을 쌓고, 인간이 지식을 쌓는 것보다 훨씬 더 강하게 재귀적일 수 있다는 것이다. 우리에게 주요한 함축은 AI가 어떤 임계 수준을 넘으면 지능에 매우 큰 도약을 할 수 있다는 것이다.
굿(1965)이 "지능 폭발"(intelligence explosion)이라 칭한 이 시나리오에 회의를 접하는 때가 잦다. 인공 지능은 발전이 아주 느린 분야로 알려져 있기 때문이다. 지금, 대충 유사한 역사적 사건을 살펴보는 것이 나을 것이다. (아래는 주요 출처 Rhodes 1986임)
1933년. 어니스트 러더포드 경은 원자를 쪼개서 동력을 얻기를 기대할 수 없다고 했다. "원자의 변환에서 동력을 구하는 사람들은 공상적인 헛소리를 하고 있는 것이다." 당시에는 몇 개의 원자핵을 분열시키기 위해서 수 주의 고된 일이 필요했다.
그 후 1942년이 되었다. 시카고 대학의 스태그 필드 지하 스쿼시 코트에서 물리학자들은 사상 처음으로 자동으로 계속되는 핵반응을 일으키기 위해 교차하는 우라늄/흑연 층으로 거대한 문 손잡이 같은 것을 만들고 있다. 과제의 담당자는 엔리코 페르미이다. 주요한 수는 유효 중성자 중배계수인 k이다. 또 하나의 핵분열을 일으키는 평균 중성자 수를 뜻한다. k가 1보다 작으면 원자로가 미임계 상태이다. k가 1이거나 더 크면 원자로가 임계 반응을 지속시킬 수 있을 것이다. 페르미는 원자로가 56번과 57번 사이에 k가 1이 될 것으로 계산한다.
허버트 앤더슨이 담당하는 작업팀이 1942년 12월 1일 밤에 57번째 층을 마무리한다. 제어봉들과 중성자를 흡수하는 카드뮴 박으로 싸인 나무 봉들이 원자로가 임계치에 이르는 것을 방지한다. 앤더슨은 제어봉 1개만 남기고 떼고, 원자로의 방사를 측정하며 원자로가 다음날 연쇄 반응을 할 준비가 된 것으로 확인한다. 앤더슨은 카드뮴 봉들을 다 제자리에 밀어놓고 맹꽁이자물쇠들로 고정하고, 스쿼시 코트를 정리하고 집에 간다.
다음날, 1942년 12월 2일, 영하 온도에 바람이 많이 부는 아침에 페르미가 마지막 실험을 시작한다. 하나를 제외하고 모든 제어봉이 빼어져 있다. 오전 10시 47분에 페르미는 남은 제어봉을 반쯤 빼도록 지시한다. 가이거 계수기들이 더 빠르게 짤깍거리고 그래프 펜이 위로 올라간다. "아닙니다." 페르미가 그래프의 한 지점을 가리키면서 말한다. "방사능 추적자가 여기까지로만 올라가고 평평해질 것입니다." 몇 분 후 그래프 펜이 가리킨 지점으로 올라가고는 그 이상 올라가지 않는다. 7분 후, 페르미는 제어봉을 30센티미터 더 빼기를 지시한다. 다시 한 번 방사능이 조금 올라가고는 평평해진다. 제어봉을 15센티미터 더 한 번 더, 그리고 또 한 번 더, 또 한 번 더 빼낸다. 11시 30분, 그래프가 서서히 올라가는 도중 갑자기 엄청난 굉음이 난다. 이온화함이 비상 제어봉을 작동시켜서 여전히 미임계에 있는 원자로를 잠근 것이다. 페르미는 침착히 작업팀에게 점심시간을 선언한다.
지하 오후 2시에 작업팀이 다시 모이고 비상 제어봉을 빼내고 고정하고는 제어봉을 마지막 설정으로 설정한다. 페르미는 측정, 계산을 더 하고, 제어봉을 조금씩 빼내기를 시작한다. 3시 25분 페르미는 제어봉을 30센티미터 더 빼내게 한다. "이때입니다." 페르미가 말한다. "드디어 핵반응이 지속될 때입니다. 계속 올라가고 멈추지 않을 것입니다."
허버트 앤더슨은 다음같이 이야기한다(Rhodes 1986):

처음에는 중성자 계수기가 짤깍짤깍 하는 소리를 들을 수 있었다. 그리고는 짤깍 소리가 점점 더 급속하게 났고, 잠시 후에는 하나의 굉음으로 합쳐지기 시작해서 계수기들이 더는 따라가지 못했다. 도표 기록계로 교체할 때이었다. 그러나 바꾸었을 때에는 모두가 기록계의 펜이 점점 편차하는 것을 갑작스러운 침묵으로 바라보고 있었다. 놀라운 침묵이었다. 모두가 그 교체의 중요성을 깨달았다. 고강도 상태에 있었고, 계수기들이 더는 대처하지 못한 것이었다. 더욱더 급속히 증가하고 있었던 중성자 강도에 맞추기 위해 자꾸만 기록계의 범위를 바꾸어야 했다. 뜻밖에 페르미가 손을 들었다. “이 원자로는 임계치에 도달했습니다.” 그가 알렸다. 그것을 의심하는 사람이 없었다.

페르미는 중성자 강도의 세대시간이 2분일 때 원자로를 28분 동안 계속 작동시켰다. 처음의 임계반응은 k가 1.0006이었다. k가 1.0006이었음에도, 원자로가 통제될 수 있었던 것은 핵분열에서 생긴 수명이 짧은 원소들에서 나온 중성자들이 지연되어서 그랬다. 100개의 U-235 원자가 분열할 때, 즉시 (.0001초) 242개의 중성자가 방사되고, 10초 후에는 1.58개의 평균 1.58개의 중성자가 발생한다. 그러므로 중성자의 평균 수명은 대략 .1초, 거기에 인해 2분 안에 1200새대가 일어나고, 1.0006의 1200승이 약 2이기 때문에, 중성자 강도는 2분 만에 2배로 증가한다. 즉발임계는 중성자의 지연이 없는 임계 상태이다. 페르미의 원자로가 k=1.0006일 때 즉발임계 상태였으면, 중성자 강도가 10분의 1초에 한 번 2배로 늘어났을 것이다.
첫 번째 교훈은 AI 연구의 속도와 개발된 후 AI의 속도를 혼동하는 것은 물리학 연구의 속도를 핵반응의 속도와 혼동하는 것과 같다는 것이다. 지도를 땅과 혼동하는 것이다. 보도 자료들도 많이 내지 못한 적은 물리학자들이 최초의 원자로를 만들기에는 수년이 걸렸다. 그러나 원자료가 만들어졌을 때에는 사건들이 인간이 하는 대화의 속도로 일어나지 않고 핵 상호작용의 속도로 일어났다. 원자핵의 규모에서는 상호작용들이 인간의 신경세포의 점화보다 훨씬 빨리 일어난다. 트랜지스터의 경우도 비슷하다.
또 하나의 교훈은 자기개선 1회가 평균적으로 .9994회 자기 개선을 일으키는것과 자기개선 1번이 1.0006회의 자기개선을 발생시키는 것은 크게 다르다는 것이다. 원자로가 임계점을 넘은 것은 물리학자들이 갑자기 많은 방사능 물질을 쌓아올려서 그런 것이 아니었다. 물리학자들은 물질을 느리게, 꾸준히 쌓아올렸던 것이다. 뇌의 지능이 그 뇌에 놓인 최적화 압력의 함수로 매끄러운 곡선을 따른다고 해도, 재귀적 자기 개선은 급격한 증가를 나타낼 수도 있다.
AI가 갑작스러운 지능의 증가를 나타날 수 있는 다른 이유들도 있다. 인류는 자연 선택이 호미니드들에 비교적 평탄한 최적화 압력을 가함으로 뇌와 전전두엽을 서서히 확대시킨 결과로 유효 지능의 급격한 증가를 보였다. 수만 년 전에, 호미니드들의 지능은 어떤 주요 문턱을 넘고 현실적 유효성의 커다란 증가가 있었다. 우리는 진화가 눈 깜박할 사이 동굴에서 고층 빌딩들에 도달했다. 이는 배후의 연속적인 선택 압력 때문에 일어난 것이다 - 인류가 나타났을 때 진화의 최적화 능력은 큰 도약을 하지 않았다. 뇌 아키텍처의 발저도 연속적이었다 - 우리의 두개 용량이 갑작스럽게 두 자리씩 증가하지 않았다. 그러므로, AI가 외부 프로그래머들에 의해 서서히 개발된다 해도, 유효 지능의 곡선은 급격한 도약을 할 수 있다.
아니면 누군가가 유망한 결과를 보이는 AI 프로토타입을 만들고, 1억 달러의 투기 자본을 끌고, 이 돈은 계산력을 천 배로 늘리는 데 쓰일 수도 있다. 나는 하드웨어의 1,000배 증가가 유효 지능의 1,000배를 살 것인지 의심하나, 순전한 의심은 분석적인 계산을 할 수 없을 때에는 신뢰할 수 없다. 침팬지들에 비해 인간들은 뇌가 3배 크고 전전두엽 피질은 6배 정도 큰 것은 (ㄱ) 소프트웨어가 하드웨어보다 중요하며 (ㄴ) 작은 하드웨어 증가는 큰 소프트웨어의 개선을 제공할 수 있다는 것을 암시한다. 이도 한 가지 고려할 점이다.
마지막으로, AI가 바보 천치와 아인슈타인을 일반적인 지성들의 공간의 거의 구별할 수 없는 점들로 보는 대신 지능 범위의 가장자리라고 생각하는 의인화 때문에 급격하게 보이는 도약을 할 수도 있다. 바보 인간보다 더 바보인 것들은 우리에게는 모두 그냥 “바보”로 보일 수 있다. 우리는 “AI의 화살”이 지능의 범위를 살금살금 올라가는 것을 상상하고, AI들이 아직 언어를 유창하게 못 하고 과학 논문도 못 쓰니 AI가 여전히 “바보”라고 생각할 수도 있다. 그리고는 AI의 화살은 바보 이하에서 아인슈타인 이상까지의 좁은 간격을 한 달 같이 짧은 시간에 넘는다. 나는 자기개선의 곡선이 직선적으로 자랄 것으로 기대하지 않기 때문에 정확히 이런 시나리오가 일어나리

라고 생각하지 않는다. 그러나 나는 “AI”가 움직이는 목표라는 것을 처음 지적하는 것이 아니다. 마일스톤이 정작 달성되었을 때에는 더는 “AI”가 아니게 된다. 이는 문제에 질질 끄는 것을 조장할 수밖에 없다.
논의의 편의상 우리가 아는 것 (또한 내가 현실에 일어날 가능성이 크다고 생각하는) 바로는 AI가 지능에 있어 갑작스럽고 매우 큰 증가를 할 수 있다고 시인하자. 이는 무엇을 뜻하는가?
무엇보다도 먼저 내가 흔히 접하는 “AI가 아직 없어서 우리는 친화적 AI가 필요하지 않다.”라는 의견은 잘못되었고 명백히 자멸적이라는 것이다. 우리는 AI가 만들어지기 전에 사전 지식을 얻는 데에 의존할 수 없다. 과거의 기술 혁명들은 그 당시의 사람들에게 자신들을 전송시키지 않았고, 사람들이 알게 된 것은 다 사후에 알게 된 것이다. 친화적 AI에 들어가는 수학과 지식은 우리가 필요할 때 펑 나타나지 않을 것이다. 확고한 기초를 세우기에는 수년, 수십 년이 걸린다. 또한 명백히, 우리는 친화적 AI의 문제를 범인공지능이 이미 등장한 후가 아니라 전에 해결해야 한다. AI의 학계 자체가 합의가 낮고 엔트로피가 높은 상태에 있기 때문에 친화적 AI의 연구에 어려움이 있을 것은 분명하다. 그러나 그렇다고 우리가 친화적 AI를 걱정하지 않아도 된다는 뜻이 아니다. 슬프게도 이 두 말들은 동일하지 않다.
지능의 급격한 증가의 가능성은 또한 친화적 AI 기술에 더 높은 기준들을 요구한다. 친화적 AI의 기술은 AI 프로그래머들이 AI 자신의 의지에 반대해 AI를 감시하거나, 우수한 군사력의 압박을 우려해 AI의 허락 없이 AI를 재작성하면 안 되고, 더 똑똑한 AI가 프로그래머들에게서 얻어낼 수 있는 “보상 버튼”을 통제할 수 있다고 가정할 수도 없고, 등등... . 실로 아무도 애당초에 이런 가정을 하고 있으면 안 된다. 불가결한 보안은 바로 당신을 해하기를 바라지 않는 AI이다. 불가결한 방책 없이는 어떤 보조 방어 대책도 안전하다고 여길 수 없다. 자신의 보안을 깰 방법을 찾아내는 시스템은 진정한 보안이 되어 있지 않은 시스템이다. 만약 AI가 어느 조건에서도 인류에게 해를 입힐 것이면, 당신은 아주 깊은 수준에서 뭔가를 잘못하고, 당신의 기초를 헛되게 하고 있는 것이다. 권총을 만들고, 자기의 발에 겨누고 방아쇠를 당기는 것이다. 어떤 상황에서는 당신을 해할 다이내믹을 고의로 가동시키는 것이다. 그것은 잘못된 것을 하는 다이내믹이다. 다르게 작동하는 코드를 쓰라.
같은 이치로 친화적 AI 프로그래머들은 AI가 자신의 소스코드를 완전히 보고, 알고, 변경할 수 있다고 가정해야 한다. 만약 AI가 더는 친화적이지 않게 자신을 변경할 의도를 가지게 되면, 그때에는 친화성이 이미 실패한 것이다. AI가 자신을 변경할 수 없는 것에 의존하는 방책은 어떤 이유로든지, AI가 자신을 실제로 변경하지 않는다고 해도 잘못된 것이다. 인류를 해하지 않기로 하는 AI를 만드는 것이 유일한 예방 대책은 아니나, 주요되고 불가결한 조치이다.
거대 치즈케익 오류를 피하기 위해, 우리는 자신을 개선할 능력이 자신을 개선할 의도를 의미하지 않는다고 상기한다. 친화적 AI의 기술이 성공하면, 더 빠르게 성장할 수 있으나 더 천천히, 처리하기 쉽게 성장하는 AI가 만들어질 수 있다. 그러해도 AI가 재귀적 자기 개선을 할 수 있는 시점을 넘으면 훨씬 더 위험한 정황에서 작업하고 있는 것이다. 만약 친화성이 실패한다면, AI는 전속력으로 자기 개선을 강행하기로 하여, 비유로 말하자면 즉발임계 상태가 되는 것이다.
내가 지능의 임의적으로 큰 잠재적 도약을 가정하는 이유는 (ㄱ) 안전의 관점에서 볼 때 보수적인 추정이며, (ㄴ) 진정한 이해 없이 AI를 만들면 안 되며, (ㄷ) 큰 잠재적 도약이 실제로 상당히 가능성이 크다고 생각하기 때문이다. 만약 내가 위험 관리의 관점에서 보수적인 추정이 AI가 느리게 성장한다는 가정이었다면, 나는 AI가 수년 이상 동안 인간 수준에 남을 때에도 계획이 치명적으로 실패하지 않기를 요구했을 것이다. 이는 내가 기꺼이 좁은 신뢰 구간을 사용할 영역이 아니다.

8: 하드웨어

사람들은 큰 컴퓨터 용량을 인공지능을 가능하게 만드는 단 하나의 요인으로 생각한다. 이는 기껏해야 아주 의심스러운 가정이다. 하드웨어 개선은 지능의 이해에 비해 측정하기가 쉬우므로 인공 지능을 논하는 외부 미래학자들은 하드웨어 진전을 얘기한다. 이해에는 진전이 없었던 것이 아니라, 이해의 진전은 깨끗하게 파워포인트 그래프에 그려질 수 없기 때문이다.
인공지능에 “필요한” “최소한의” 하드웨어라는 관점에서 보는 대신, 하드웨어 개선의 함수로 감소하는 필요한 최소한의 연구가들의 이해 수준을 생각해보라. 계산 하드웨어가 좋을수록, AI를

만드는 데 필요한 이해가 적어진다는 것이다. 극단적인 예는 터무니없는 양의 계산력으로 아무 이해 없이, 우연적인 돌연변이의 비우연적인 유지만으로 인간 수준의 지능을 만들어낸 자연 선택이다.
계산력이 더 많으면 AI를 만들기가 더 쉽지만, 계산력의 증가가 친화적 AI를 더 쉽게 할 분명한 이유가 없다. 이해하지 못하는 기술을 합쳐서 무차별 대입으로 AI를 만들기가 쉬워지는 것이다. 무어의 법칙은 인지의 깊은 이해 없이 AI를 만드는 것을 막는 장벽을 끊임없이 허무는 것이다. AI와 친화적 AI에 둘 다 실패해도 된다. AI와 친화적 AI에 둘 다 성공해도 된다. 안 되는 것은 AI에 성공하고 친화적 AI에 실패하는 것이다. 무어의 법칙은 바로 이것을 더 쉽게 하는 것이다. 그러나 다행히도 쉽게 하지는 않을 것이다. 나는 AI를 구축하기 위해 엄청난 노력을 하는 당사자가 있고 그중 하나가 엄청난 노력 끝에 성공할 것이라고 해서 AI가 생길 때에도 쉬우리라는 것을 의심한다.
무어의 법칙은 다른 기술들에 흔히 간과된 치면적 위험을 더하는, 친화적 AI와 다른 기술들의 상호 작용이다. 분자 나노기술이 선량한 다국적 컨소시엄에 의해 개발되고, 나노기술의 물리학적 위험들이 성공적으로 방지된다고 상상해보자. 우발적인 레플리케이터 방출이 쉽게 방지되고, 훨씬 더 어려운 노력으로 악성 레플리케이터에 대한 전세계적 방어를 설치한다. “루트 레벨” 나노기술은 규제하면서 구성될 수 있는 나노블록을 대신 분배하는 등의 방책들을 사용한다. (이 책의 Phoenix and Treder를 참고하라) 하지만 나노컴퓨터들은 규제가 시도되지 않거나 시도되어도 규제를 피해서 누군가가 나노컴퓨터를 만들어 낸다. 그리고는 누군가가 비친화적인 AI를 무차별 시도로 만들어내고 만다. 이 시나리오가 특히 우려스러운 이유는, 엄청나게 강력한 나노컴퓨터는 분자 나노기술의 최초, 가장 쉽고 가장 안전하게 보이는 응용 중의 하나일 것이기 때문이다.
슈퍼컴퓨터에 대한 규제는 어떤가? 나는 AI가 개발되는 것을 완전히 방지하기 위해 규제에는 의존하지 않을 것이다. 어제의 슈퍼컴퓨터는 내일의 노트북이 되어버리기 때문이다. 규제의 제안에 대한 흔한 답변은 슈퍼컴퓨터가 범죄화되면 범죄자들만 슈퍼컴퓨터를 가질 것이라는 것이다. 부담은 감소된 분포의 추측되는 장점이 고르지 않은 분포의 불가피한 위험을 능가한다고 논하는 것이다. 나 자신을 위해서도 나는 인공지능 연구를 위한 슈퍼컴퓨터의 사용을 규제하는 것에 찬성하지 않겠다. 불확실한 유익밖에 주지 않고 AI 학계 전체가 필사적으로 반대할 제안이다. 그러나 그런 제안이 정치적 과정을 그렇게 빨리 진전하는 가능성이 작은 일이 일어날 경우에는, 나는 그 제안을 반대하는 데 많은 노력을 쏟지 않겠다. 친화적 인공지능은 무차별 도입으로 푸는 것이 아니기 때문에, 나는 친화적 AI 연구가들에게 그 때의 슈퍼컴퓨터가 필요할 것으로 기대하지 않는다.
나는 현재 “슈퍼컴퓨터”로 간주되는 소수의 엄청나게 비싼 컴퓨터들을 효과적으로 규제하는 제도들을 상상할 수는 있다. 그러나 컴퓨터는 어디에나 있다. 플루토늄과 농축 우라늄만 규제하면 되는 핵 확산의 문제와 다르다. AI의 원료는 이미 어디에나 있다. 당신의 손목시계, 핸드폰, 식기세척기에도 있을 정도로 널리 퍼져 있다. 이 점도 인공지능이 지니는 위험의 특별하고 비정상적인 요인이다. 위험한 상황에서 동위원소 원심분리기나 입자가속기 같은 크고 눈에 띄는 시설들이 아니라, 부족한 지식만으로 떨어져 있는 것이다. 좀 지나치게 극적인 비유로 설명하자면, 레오 실라르드가 연쇄 반응을 처음 떠올리기 전에 이미 임계치 이하의 농축 우라늄 덩어리들로 차나 배들이 동력을 얻었다고 생각해보라.

9: 위험과 약속

친화적인 AI가 어떻게 인류를 돕고, 비친화적인 AI가 어떻게 인류를 해할지 예측하려고 하는 것은 위험천만한 지적 시도이다. 결합 오류에 빠질 위험이 있다. 추가된 세부 사항은 전체 이야기의 확률을 반드시 줄이나, 대상자들은 세부 사항이 추가된 이야기들이 확률이 더 높다고 말한다. (이 책의 인지적 바이어스들을 논하는 챕터를 보라.) 상상력이 실패할 거의 확실한 위험이 있고, 능력에서 동기로 비약하는 거대 치즈케익의 오류가 발생할 위험도 있다. 그렇더라도 나는 위험과 약속을 논하도록 하겠다.
미래는 과거가 불가능했다고 생각했던 위업들을 달성하는 것으로 유명하다. 미래의 문명들은 과거가 물리적 법칙이라고 생각했던 것들을 위반하기도 했다. 서기 1,000년은 말할 것도 없이, 서

기 1900년의 예언자들이 인류의 문명의 능력의 한계를 잡으려고 했다면, 어떤 불가능했던 것들, 예를 들어 납을 금으로 변환시키기는 한 세기가 지나기도 전에 달성되었을 것이다. 우리는 미래의 문명들이 과거의 문명들을 놀라게 했던 것을 기억하기 때문, 우리 후손들의 한계를 잡을 수 없다는 말은 판의 박은 문구가 되었다. 이럼에도 20세기, 19세기, 11세기에 살아 있었던 이들은 모두 인간이었다.
인간보다 똑똑한 인공 지능의 능력을 상상하는 신뢰할 수 없는 비유들은 3가지로 분류할 수 있다.

- *g-팩터 비유:* 인간 지능의 개인차에서 착상된다. AI들은 새로운 기술들의 특허를 얻거나 혁신적인 연구 논문들을 제출하거나 정치적 파워 블록들을 지도할 것이다.
- *역사적 비유:* 과거와 미래의 문명의 지식수준의 차이에서 착상된다. AI들은 재빨리 분자 나노기술, 항성 간 여행, 1초당 10^{25}번 연산을 할 수 있는 컴퓨터 등, 판에 박은 문구가 백년 또는 천년 후의 문명에 상상하는 능력들을 개발할 것이다.
- *종간 비유:* 종간 뇌 아키텍처의 차이에서 착상된다. AI는 마력을 지닐 것이다.

g-팩터 비유들이 대중 미래 예측에서 가장 흔히 쓰이는 비유로 보인다: 사람들은 "지능"이라고 하면 인간 대신 천재 인간들을 생각한다. 적대적인 인공 지능의 능력이 *g*에 비유되는 이야기에서는, AI는 극적인 긴장을 일으킬 만큼 강하면서, 주인공들을 단번에 무너뜨릴 만큼은 강하지 않고, 최후에는 주인공에게 질 만큼 약한, 이른바 보스트롬이 말하는 "좋은 이야기"의 요소이다. 골리앗과 다윗의 싸움은 좋은 이야기지만, 골리앗과 초파리의 싸움은 좋은 이야기가 아니다. *g*-팩터 비유를 가정한다면, 이 시나리오의 전세계적 치명적 위험은 낮다. 적대적인 AI도 적대적인 인간 천재보다는 크게 더 심각한 위험은 아닐 것이기 때문이다. AI가 *다수* 있다고 가정하면 국가 간의 대립, AI 부족과 인간 부족의 싸움에 비유한 것이 된다. AI 부족이 인간들과 전쟁해서 이겨 인간들을 전멸한다면 "뱅"식 존망이 된다 (Bostrom 2001). AI 부족이 세계를 경제적으로 정복하고 지구의 지적 생명체들의 운명을 좌우하게 되고 AI 부족의 목적이 우리에게 흥미나 보람이 있지 않으면 "슈릭", "윔퍼", 아니면 "크런치"이다.
그러나 인공 지능이 아메바와 바보 인간 사이의 간격을 넘고는 인간 천재의 수준에서 딱 멈출 가능성은 얼마나 되는가?
관찰된 가장 빠른 신경세포들은 1초당 100번 점화하며, 가장 빠른 축삭 섬유들은 1초당 150 미터의 속도로 신호를 전달하며, 시냅스의 각 연산은 15,000 아토줄에서 발산된다. 15,000 아토줄은 실온에서 비가역적 계산에 필요한 열역학적 최소한의 에너지의 1백만 배 이상이다 (kT_{300} ln(2) = 비트당 .003 아토줄). 크기를 줄이지 않고, 더 낮은 온도에서 돌리지 않고, 가역적 컴퓨팅이나 양자 컴퓨팅을 사용하지 않고도 현재 인간의 뇌의 백만 배의 속도로 돌아가는 뇌를 만들기가 물리적으로 얼마든지 가능하다는 것이다. 이러하여 가속된 인간의 뇌에는 주관적 1년의 세월이 외부 세계의 31초 만에 할 수 있을 것이고, 30분 만에 1,000년이 날아갈 것이다. 빈지(1993)는 이렇게 가속된, 인간 같으나 훨씬 더 빨리 생각하는지성들을 "약-초지능"이라고 일컬었다.
당시에 존재하는 상태의 문명 안에서 매우 빠른 지성이 나타난다고 하자. "아무리 빠르게 생각해도 자신이 가지고 있는 머니퓰레이터의 속도로만 세계를 변화시킬 수 있고 인간들의 손에게 시킬 수 있는 속도보다 빨리 기계를 작동시킬 수 없으므로 빠른 지성은 킨 위협이 되지 않는다."라고 하는 것은 상상력이 부족한 것이다. 물리적 작동들이 수 초의 시간으로 일어나야 한다는 자연의 법칙은 없다. 분자의 기본적인 상호작용들은 펨토초, 때로는 피코초 정도의 시간 사이에 일어난다. 드렉슬러(1992)는 초당 10^6 이상의 기계적 작동을 할 분자적 머니퓰레이터들을 분석했다. 이는 "백만 배의 가속"이라는 주제에 맞는 것에 주의하라. (물리학적으로 이치에 맞는 가장 짧은 시간의 경과는 플랑크 시간 ($5*10^{44}$ 초)이다. 이 규모에서는 돌아다니는 쿼크들도 조각상과 같다.)
인류의 문명이 상자 속에 가두어져 있고 더디게 느린 외계인의 촉수 또는 초당 수 마이크론씩 움직이는 기계 팔로만 외부에 영향을 미칠 수 있게 된다고 상상해보라. 우리는 외부 세계에 빠른 머니퓰레이터들을 조립할 가능한 한 가장 짧은 경로를 찾는 데에 창의력을 쏟을 것이다. 빠른 머니퓰레이터들을 생각할 때, 다른 방법들도 있을 수 있지만 즉시 분자 나노기술을 생각할 수 있다. 각 단계를 숙고할 수 있는 장구한 세월이 있으면, 당신이 느린 외부 세계에 분자 나노기술을 설치

할 수 있는 가장 짧은 경로는 무엇인가? 나는 숙고할 장구한 세월이 없어서 모른다. 다음과 같은 빠른 경로를 상상할 수 있다:

- 복잡한 화학적 상호작용에서 어느 특정한 기능을 할 수 있는 펩타이드 서열을 제조하는 DNA 문자열을 생성할 수 있을 정도로 단백질 접힘 문제를 해결함.
- DNA 합성/펩타이드 서열 분석/페덱스 배달 등의 서비스들을 제공하는 온라인 연구소로 여러 가지의 DNA 문자열들을 이메일로 보냄. (많은 연구소들이 현재 이 서비스를 제공하며, 어떤 연구소들은 72시간의 왕복 시간을 자랑한다.)
- 인터넷에 연결되어 있는 인간 최소한 한 명을 찾아, 돈을 주거나 협박하거나 어떤 배경 스토리로 속여 배달된 병들을 특정된 환경에서 섞게 함.
- 합성된 단백질들은 외부에서 (예를 들어 비커에 부착된 스피커로 전달되는 음향 진동의 형태로) 명령을 받아들일 수 있는, 리보솜과 유사한 원시적인 습식 나노시스템을 형성한다.
- 이 원시적인 나노시스템을 이용하여 더 정교한 시스템들을 만들고, 이 차세대 시스템들로 더욱더 정교한 시스템을 만듦으로써 분자 나노기술 또는 그 이후로 부트스트랩함.

걸리는 시간은 빠른 지능이 단백질 접힘 문제를 불 수 있게 된 때부터 1주일 정도의 시간일 것이다. 물론 이 시나리오는 엄밀히 내가 생각하는 것이다. 19500년의 주관적 시간(100만 배 가속했을 때의 1주일의 물리적 시간)이 지나면 나는 더 좋은 방법을 발견할 수도 있을지 모른다. 페덱스 대신 급사로 보내게 할 수 있을지도 모른다. 현존하는 기술 또는 기술의 변형을 단순한 단백질 시스템과 공동으로 협력하게 할 수 있을지도 모른다. 적당히 똑똑하면 파형 전기장으로 현존하는 생화학적 반응 경로들을 바꿀 수 있을지도 모른다. 나는 그만큼 똑똑하지 않기 때문에 지금 알 수 없다.
중요한 문제는 가지고 있는 능력들을 체인하는 것 – 루트를 접근하기 위해 컴퓨터 시스템의 취약점들을 결합하는 개념을 물리적 세계에 유추한 것이다. 만약 한 경로가 막혔어도 다른 경로를 선택하고, 항상 자신의 능력을 증가시키고 섞어서 사용할 수 있다. 추정된 목표는 고속 기반기술, 짧은 시간에 외부 세계를 다루는 수단을 얻는 것이다. 분자 나노기술은 기초적 작동들이 빠르며, 자기 복제와 지수적 성장을 시킬 수 있는 원료인 원자가 풍부하다는 것이다. 위에 상상한 경로는 AI가 일주일 만에 고속 기반기술을 얻는 시나리오이다 – 이는 200헤르츠로 돌아가는 신경세포들을 가진 인간에게는 빠른 것 같으나, AI에게는 훨씬 장구한 시간이다.
AI가 고속 기반기술을 가질 때부터는 사건들이 (AI가 인간의 속도로 행동하기를 선호하지 않으면) 인간의 속도가 아니라 AI의 속도로 일어난다. 분자 나노기술을 가졌다면 AI는 제약 없이 태양계를 다시 쓸 능력이 있을 것이다. 분자 나노기술이나 다른 방식의 고속 기반기술을 휘두르는 비친화적 AI는 굳이 로봇 부대나 공갈/갈취나 미묘한 경제적 강압을 사용할 필요가 없을 것이다. 비친화적 AI는 태양계의 물질을 모두 자신의 최적화 목표에 맞추어 재구성할 능력이 있다. 만약 AI가 특별히 현존하는 패턴들, 예를 들어 생물학이나 사람들에게 어떤 영향을 줄 것인지의 기준에 따라 선택하지 않는다면 우리에게 치명적이다. AI는 당신을 미워하지도 않고 사랑하지도 않지만, 당신은 그 AI가 다른 데에 사용할 수 있는 원자로 만들어져 있다. AI는 당신과 다른 속도로 돌아가고 있다. 당신의 신경 세포들이 “어떻게 해야 하겠다.”라고 생각할 때에는 당신이 이미 패배한 것이다.
분자 나노기술을 가진 친화적 AI는 아마도 원자를 옮기거나 창의적으로 생각하면 해결될 모든 문제를 해결할 수 있을 만큼 강력할 것이다. 상상의 실패를 조심해야 한다: 암을 치료하는 것은 현대 자선의 인기 있는 목표이나, 그렇다고 분자 나노기술을 가진 친화적 AI가 자신에게 “이제 암을 치료하리라”라고 말하리라는 뜻은 아니다. 어쩌면 암이라는 문제를 더 유용하게 보는 관점은 생물학적 세포들은 프로그램될 수 없다는 문제일 지도 모른다. 후자를 해결하는 것은 당뇨병과 비만과 함께 암도 특수한 경우로 치료하는 것이다. 분자 나노기술을 사용하는 빠르고 친절한 지성은 암을 없애는 수준이 아니라 질병을 없애는 수준의 힘일 것이다.

마지막으로 종간 비유들이 있다. AI는 마력을 지니지만, 주문이나 묘약 같은 마법을 뜻하는 것이 아니라, 늑대는 총이 어떻게 작동하는지, 총을 만드는 데 무슨 노력이 필요한지, 또는 총을 발명할 수 있게 하는 인간의 힘을 이해할 수 없다는 것과 같은 의미이다. 빈지(1993)는 이렇게 말한다:

> 강-초인간성은 인간과 같은 지성의 시계 속도를 높이는 것 그 이상일 것이다. 강-초인간성이 정확히 어떤 것일지 논하기는 어렵지만, 엄청나게 큰 차이가 있을 것이다. 개의 지성을 매우 빠른 속도로 돌린다고 상상해 보라. 천 년 동안 개로서 산다고 인간의 통찰력이 생길 것인가?

종간 비유가 선험적으로 가장 맞는 비유로 보이나, 상세한 이야기를 꾸미기에는 적합하지 않다. 이 비유가 우리에게 주는 조언은 친화적 AI를 정확하게 해결하라는 것이다. 이는 물론 좋은 조언이다. 적대적인 AI에 대해 제안하는 유일한 방어는 처음부터 만들지 말라는 것이다. 이것도 훌륭한 조언이다. 절대적인 힘은 친화적 AI의 개발에서 설계의 흠들을 드러내는 보수적인 공학적 가정이다. 마력이 있을 때 AI가 당신에게 해를 끼칠 것이면, AI의 친화성 아키텍처가 잘못된 것이다.

10: 국부적 및 과반수적 방책

제의되는 위험 감소 방책들은 3가지로 나눌 수 있다.

- *만장일치*의 협동이 필요한 방책들. 개인 또는 작은 단체의 변절자들이 치명적으로 망칠 수 있는 방책들이다.
- *과반수*(예를 들어 한 국가의 국회나, 한 나라의 투표자들이나, 또는 UN에 있는 국가들의 과반수)의 협동이 필요한 방책들. 이러한 방책들은 이미 존재하는 대규모 집단의 다수이나 전체가 아닌 사람들이 특정한 방식으로 행동해야 하는 방책들이다.
- *국부적*인 운동이 필요한 방책들. 집중된 의지, 재능과 자금이 어떤 특정한 작업의 문턱을 넘어선다.
- *만장일치적*인 방책들은 실현할 수 없지만 사람들은 만장일치적인 방책들을 제의하기를 서슴지 않는다.

*과반수적*인 방책들은 수십 년의 노력이 있으면 종종 실행할 수 있다. 운동을 수년간의 시초에서 공종 정책에서 인정되는 세력으로 등장하고 반대하는 세력들을 꺾을 때까지 운동을 건설해야 한다. 과반수적 방책은 많은 시간과 엄청난 노력이 필요하다. 사람들이 과반수적 방책에 착수했으며, 어떤 것들은 성공한 것으로 기록된다. 그러나 주의할 것은, 역사책들은 전혀 성공하지 못하는 대다수의 운동보다, 영향을 좀 주는 운동에 집중하는 경향이 있다는 것이다. 운도 하나의 요소이고, 대중이 기꺼이 귀를 기울일 것인가도 요소이다. 운동의 결정적인 시기들은 자신의 통제를 넘어선 사건들이 포함될 것이다. 만약 과반수적 방책을 밀고 나아가기 위해 온 일생을 헌신할 각오가 없으면, 일부러 과반수적 방책을 실행하려고 하지 말라. 하나의 일생도 충분하지 않을 것이다. 일반적으로는 국부적인 방책이 가장 믿을 만하다. 1억 달러의 자금을 얻기는 쉽지는 않고, 전세계적인 정치적 변화는 해내기 불가능하지는 않으나, 1억 달러를 얻는 것이 전세계적 변화를 이루어내기보다 훨씬 쉽다.

AI에 대해 과반수적인 방책을 취할 두 가지 가정은:

- 다수의 친화적 AI들은 소수의 비친화적 AI들에서 효과적으로 보호할 수 있다.
- 최초의 AI는 그 자체로 치명적인 피해를 줄 수 없다.

이는 핵무기와 생물학적 무기의 개발 이전의 문명의 경우를 재현한다. 즉 대다수의 사람은 종합적인 사회 조직에서 협조자들이고, 변절자들은 피해를 줄 수 있지만 전세계적 치명적 피해는 입힐 수 없는 상태이다. 대부분의 AI 연구가들은 비친화적 AI를 만들기를 바라지 않을 것이다. 안정적으로 친화적인 AI를 만들 수 있는 사람만 있으면 – 문제가 현재의 지식과 기술을 완전히 뛰어넘지 않는다면 – 연구가들은 서로의 성공으로 배우고 되풀이하게 될 것이다. (예를 들이) 연구가들이 자신들의 친화성 방책들을 공개하기를 명하거나, 해를 끼치는 AI를 만든 연구가들을 벌하는

법이 있을 수도 있다. 이러한 법은 실수를 다 방지할 수는 없을 것이나, 다수의 AI가 친화적으로 만들어지면 충분할 수 있다.
쉬운 국부적 방책을 취할 수 있는 시나리오도 상상할 수 있다:

- 최초의 AI는 그 자체로 치명적인 피해를 줄 수 없다.
- 친화적인 AI가 하나라도 존재하면 그 AI는 인간의 제도와 협력해서 비친화적인 AI를 몇이라도 차단할 수 있다.

이 쉬운 시나리오는 예컨대 사회 제도들이 친화적인 AI와 비친화적인 AI를 정확히 구별할 수 있으면 적용될 수 있을 것이다. 친화적인 AI들에게 무효화할 수 있는 권력을 준다면 우리는 우리의 우방을 고를 수 있을 것이다.
위의 시나리오들은 모두 최초의 (강력하고 일반적인) AI가 전세계에 치명적인 해를 끼칠 수 없다는 가정에서 비롯된다. 이 가정을 의미하는 상상들은 AI들이 유별나게 능력이 뛰어난 인간들과 비슷하리라 예측하는 *g* 비유를 사용한다. 제7항에서 나는지능의 엄청나게 빠른 도약을 우려할 이유를 열거했다.

- 우리에게는 크게 보이는 바보에서 아인슈타인까지의 거리는 임의적인 심리들의 공간에서 작은 반점에 지나지 않는다.
- 자연 선택이 게놈에 대략 일정한 선택 압력을 가했는데도 불구하고, 유인원들은 지능의 유효성에서 큰 도약을 했다.
- AI는 어떤 점을 지나서 (인터넷을 먹어치우는 등의 방식으로) 엄청나게 많은 하드웨어를 흡수할 수 있을지도 모른다.
- 재귀적 자기 개선의 임계점이 있다. 자기 개선 1건이 자기 개선 1.0006건을 발생시키는 것은 1건당 0.9994건의 자기 개선이 발생하는 것과 질적으로 다르다.

제9항에서 말한 것과 같이, 적당히 강력한 지능은 인간의 관점에서는 짧은 시간 안에도 분자 나노기술이난 다른 방식의 고속 기반기술을 달성할 수 있을지도 모른다.
그러므로 우리는 초지능의 선점 우위 효과를 상상할 수 있다. 선점 우위 효과는, 예를 들이 자기 개선과 같은 지능의 주요한 문턱에 이르는 최초의 지능의 성질에 따라 지구의 지적인 생명의 결과가 결정되는 것을 말한다. 두 가지의 필요한 가정은 다음과 같다:

- 어떤 주요한 수준(재귀적 자기 개선의 임계점 등)을 초과하는 최초의 AI는 비친화적이면 인류를 멸절시킬 수 있다.
- 같은 주요한 수준을 초과하는 최초의 AI가 친화적이라면, 악성 AI가 발생하거나 인류를 해하는 것을 방지할 수 있거나, 아니면 지구에서 생성된 지적 생명의 생존과 번영을 보증할 다른 창의적인 해결책을 찾을 수 있다.

선점 우위 효과에 포함되는 시나리오는 여러 가지가 있다. 다음의 예들은 다른 임계적 수준을 나타낸다:

- 임계점 이후, 자기 개선은 몇 주 이하의 시간 만에 초지능에 도달한다. 다른 AI 프로젝트들은 적당히 모든 반대를 압도할 만큼의 수준에 이르기 전에 하나의 AI가 이 지점에 이른다. 주요한 문턱은 재귀적 자기 개선의 임계점이다.
- AI-1이 단백질 접힘 문제를 AI-2보다 3일 먼저 푼다. AI-1이 AI-2보다 6시간 먼저 나노기술을 발명한다. 빠른 머니퓰레이터들로, AI-1은 AI-2의 연구•개발이 실현되기 전 차단할 수 있다. 경쟁이 아무리 백중이라고 해도 결승전을 먼저 가로지르는 자가 이긴다. 주요한 문턱은 빠른 기반기술이다.
- 처음 인터넷을 흡수하는 AI는 인터넷이 다른 AI에 의해 이용되는 것을 막을 가능성이 있다. 그 후에는 경제적 지배, 비밀공작, 협박, 또는 사회적 조종으로 최초의 AI가 다른 AI 프로젝트들을 차단한다. 주요한 문턱은 유일한 자원의 흡수이다.

인류, 호모 사피엔스는 선점 우위자 중 하나이다. 진화의 관점에서는 우리의 사촌 침팬지들은 우리에게 매우 가까웠다. 호모 사피엔스는 여전히 조금 더 지능에 이르렀기 때문에 이 모든 기술을 생성시켰다. 이 선점 우위자가 언어, 기술, 추상적인 사고 등 너무나 많은 문턱을 넘은 최초의 생

물이기 때문에, 진화 생물학자들은 주요한 문턱을 우리가 어떤 순서로 넘었는지를 여전히 밝혀내려고 하고 있다. 중요한 것은 호모 사피엔스가 경쟁자의 여지가 없이 선점 우위자였다는 것이다. 선점 우위 효과는 원칙적으로 국부화될 수 있는 방책을 의미하지만, 매우 어려운 기술적 도전에 의지하기도 한다. 친화적 AI를 한 곳, 한 때에서만 장 해결해도 된다. 그러나 다른 모든 이들이 더 낮은 기준으로 AI를 만들기 전에 누군가가 첫 번째 기회에 맞게 해내어야 한다.
나는 정밀히 검증된 이론으로 정밀한 계산을 할 수 없지만, 현재 나의 의견은 지능의 급격한 상승이 가능하고, 가장 가능성이 크다는 것이다. 그러나 지능의 급속한 상승이 일어나지 않아도 방책이 우리에게 피해를 주면 안 된다. 그러나 훨씬 더 심각한 문제는 느리게 자라는 선점 우위 효과가 만약 있으면 치명적으로 실패하는 AI에 대처하기 위한 방책들이다. 후자가 더욱 심각한 문제인 이유는:

- 더 급속히 발달하는 AI들은 기술적으로 더 처리하기 어렵다.
- 차가 트럭을 지탱하도록 설계된 다리 위로 다닐 수 있는 것처럼, 열악한 조건에서 친화성을 유지하도록 설계된 AI는 아마도 덜 나쁜 상황에서도 친화적일 것이다. 이 반대는 사실이 아니다.
- 지능의 급격한 도약은 일상적인 직관에 맞지 않는다. AI의 *g*-팩터 비유는 직관적이고, 안심되고, 편하게도 더 느슨한 설계 사항을 수반한다.
- 현재 내가 추측하기에는 지능의 곡선은 급격한 증가들이 있을 수도 있다.

"최초의 AI가 친화적이어야 한다"는 현재 나의 방책은 어려운 국부적 방책에 속한다. 추가 조건은 지능의 급격한 증가가 발생하지 않는다면, AI의 과반수를 친화적으로 만드는 방책으로 바꿀 수 있느냐한다는 것이다. 어느 경우에나 극단적인 경우인 선점 우위 효고에 대비하기 위해 했던 노력은 우리의 형편이 좋아질 듯하다.
불가능한 만장일치적 방책을 요구하는 시나리오는:

- 친화적 AI들의 노력에도, AI 하나는 인류를 멸할 수 있을 만큼 강력할 수 있다.
- 연구가들이 AI들을 연달아 건설하는 것을 막거나 혹은 다른 창의적인 방책으로 문제를 해결할 수 있을 만큼 강력한 AI가 없다.

이 상태는 선험적으로 가능성이 적게 보이는 것이 기쁜 일이다. 이 시나리오에서는 우리는 이미 운이 다한 것이기 때문이다. 패에서 카드를 꺼내면, 언젠가는 클럽 에이스를 도를 것이다.
어떤 지점 이상으로 자신들의 능력을 향상시키지 않기로 하는 AI를 만드는 방책에도 같은 문제가 있다. 제약된 AI들이 제약되지 않은 AI들을 패배시킬 수 있을 만큼 강력하지 않으면, 이러한 방법으로 제약된 AI는 소용이 없다. 우리는 하트 에이스든지, 클럽 에이스든지 초지능을 도를 때까지 계속카드 패를 꺼내고 있다.
과반수적 방책은 하나의 변절자가 전세계적 치명적 해를 끼칠 가능성이 없어야만 가능하다. AI에서 이 가능성이나 불가능성은 설계 공간의 자연적인 성질이다. 이 가능성은 광속이나 중력 상수처럼, 인간의 결정에 의해 바뀌지가 않는다.

11: 인공지능 VS 지능 증가

나는 호모 사피엔스가 현재 지능의 상한을 깨는지성의 출현이 없이 무한정, 수천년 또는 수십억 년 동안 지속되리라고 믿기 어렵다. 만약 이 말이 사실이라면 언젠가는 초지능의 도전에 맞서야 한다. 우리가 이 도전의 첫 번째 라운드를 이기면, 우리는 다음 라운드들에서 인간 이상의 지능을 불러서 직면할 수 있다.
혹시 인간 이상의 지능으로 할 길은 AI가 아닌 다른 길이 더 좋지 않겠는가? 예컨대 현존하는 인간의 지능을 향상하는 것이 낫지 않겠는가? 극단적인 예를 들어, 누군가가 이렇게 말한다고 하자: "AI가 만들어질 가능성은 나에게 걱정스럽다. 어떤 AI가 개발되기 전에 개별적 인간들이 한 신경세포씩 컴퓨터들로 촬영시키고는 천천히, 확실히 개량시켜주어서 초지능을 갖게 하는 것이 좋겠다. 그것이 우리가 초지능에 직면해야 할 방법이다."

우리는 여기에서 두 가지 문제에 직면하게 된다. 이 시나리오는 가능성이 큰가? 만약 그렇다면 이 시나리오는 바람직하다? (이성상 이런 순서로 이 두 의문을 제기하는 것이 좋다. 실제로는 선택할 수 없는 것에 애착되면 안 되기 때문이다.)
모라벡(1988)이 제의한 대로 인간이 한 신경세포씩 컴퓨터로 촬영된다고 하자. 그러면 틀림없이 사용된 계산력이 두뇌의 계산력을 능가해야 한다.
이 상상에 의하면 컴퓨터는 저수준에서 발생하는 조직적인 오차에서 큰 고수준 오류들이 발생하지 않을 만큼 충실하게 생물학적 두뇌를 상세하게 모의해야 한다. 어떻게든 정보 처리에 영향을 미치는 어떤 생물학적인 우연이라도 전체적인 처리의 흐름이 동형으로 남을 만큼 충실하게 모의해야 한다. 인간의 두뇌라는 뒤죽박죽 한 생물학적 컴퓨터를 모의시키기 위해서는 뇌 자체에 담겨 있는 계산력보다 훨씬 더 유용한 계산력이 요구된다.
신경 구조의 사고에 관련된 모든 양상을 포착할 수 있을 만큼의 충실도로 인간의 뇌를 한 신경세포씩 촬영할 방법을 개발시킬 가장 유력한 방법은 분자 나노기술을 개발하는 것일 것이다.[4] 분자 나노기술은 아마도 현재 인구 전체보다 더 처리력이 강한 데스크톱 컴퓨터를 만들 수 있을 것이다. (Bostrom 1998; Moravec 1999; Merkle and Drexler 1996; Sandberg 1999.)
또한 기술이 촬영된 뇌를 코드로 실행할 수 있을 만큼 충실하게 촬영할 수 있으면, 그전 수년 동안은 신경 세포들이 하는 처리를 매우 상세하게 포착할 수 있는 기술이 있어야 했고, 연구가들은 신경회로가 하는 방식의 처리를 이해하기 위해서 온 힘을 다했을 것이다.
게다가 인간들은 외부의 신경과학자들이나 내부의 재귀적 자기 개선으로 의해 개선되도록 설계되지 않았다. 자연 선택은 인간의 뇌를 인간이 조작할 수 있게 설계하지 않았다. 두뇌의 모든 복잡한 장치는 뇌의 설계의 좁은 영역에서 작동하도록 적응되었다. 문제의 인간을 초지능 정도까지는 아니라도 더 똑똑하게 할 수 있다고 하자. 그 인간은 정신이 온전하게 남겠는가? 두뇌는 교란시키기가 아주 쉽다. 신경전달물질들의 균형을 어지럽히는 것 만으로도 정신분열증 등의 장애들을 유발할 수 있다. 디컨(1997)은 두뇌의 진화와 두뇌의 요소들이 얼마나 미묘히 균형잡혀있는지, 이것이 현대의 뇌 이상에 어떻게 반영되는지, 등의 문제를 우수하게 논하고 있다. 두뇌는 최종 사용자가 고칠 수 없다.
이 요인들은 모두 누군가 어디에서 인공지능을 처음으로 만들기 전에 인간이 처음으로 업로드되고 제정신으로 개량되는 것을 믿기 어렵게 만든다. 기술이 업로딩을 할 수 있는 수준일 때에는 AI를 만드는 데보다 훨씬 많은 계산력과 아마도 훨씬 발달한 인지 과학을 의미한다.
처음부터 보잉 747을 건설하기는 쉽지 않다. 그러나:

- 현존하는 생물학적 조류의 설계에서 시작하여,
- *각각 실행할 수 있는 단계로*
- 점진적으로 단계별로 원래의 설계를 변경하여,
- 날기까지 하며,
- 게다가 747만큼 *빨리 나는*
- 747의 크기로 늘린 새를 설계하고,
- 실제 살아 있는 새를 죽게 하거나 다치게 하지 않고
- 새에 이 변환들을 실행하기가 더 쉽겠는가?

[4]그러나 Merkle (1989)은 전자현미경, 광학 절편 등 영상 기술의 혁명적이지 않은 발달도 두뇌 전체를 업로딩하는데 충분할 수도 있다고 말한다.
결코 할 수 없을 것이라는 말이 아니다. 비유적으로 747을 처음부터 시작하여 만드는 것이 새를 개량하는 것보다 쉬울 것이라는 말이다. "그냥 현존하는 새를 747의 크기로 늘리자"는 것은 것은 항공역학의 으스스한 이론적 불가사의를 다루지 않아도 되기 위한 기발한 방책이 아니다. 아마도 처음에는 당신이 비행에 대해서 아는 것이라고는 새가 비행의 신비스러운 본질이 있다는 것밖에 모르고 747을 건설하기 위한 재료들은 그냥 바닥에 놓여 있기만 할지도 모른다. 그러나 새 안에 이미 실재하는 비행의 불가사의한 본질도, 비행이 당신에게 더는 불가사의한 본질이 아닐 때까지는 비행기에 조각할 수 없다.
위의 논증은 고의적으로 극단적은 예에 집중했다. 내가 강조하는 일반적인 요점은 우리가 괜찮고 안심이 되거나 공상 과학 소설에 들어맞을 길을 선택할 완전한 자유가 없다는 것이다. 우리는 기술의 개발 순서에 따라 제약되어 있다.

나는 인간들을 컴퓨터로 촬영시키고 더 똑똑하게 만드는 것에는 반대하지 않지만, 이 방법이 인류가 초지능에 처음 도전하는 바탕이 되기에는 가능성이 희박해 보인다. 인간들을 업로드하고 개량하기에 필요한 기술과 지식의 여러 가지 진부분 집합들을 가지고

- (예를 들어 새로운 신경세포를 기존 뇌에 짜맞추어서) 생물학적 뇌를 원래 위치에서 개량하거나,
- 아니면 컴퓨터를 생물학적 뇌에 유용하게 인터페이스시키거나,
- 아니면 인간의 뇌끼리 유용하게 인터페이스시키거나,
- 아니면 인공 지능을 구축할 수 있을 것이다.

더군다나 (IQ나 노벨상을 받는 것이 유동성 지능의 척도로서 유용한가의 논쟁은 놔두고) 보통 인간을 안전하게 IQ 140으로 올리는 것과 노벨상 수상자를 인간보다 뛰어난 것으로 올리는 것은 차원이 다르다. 피라세탐을 복용하거나 카페인을 마시는 것이 최소한 어떤 사람들을 더 똑똑하게 만들 수도 있고, 그러지 않을 수도 있으나, 아인슈타인보다 상당히 똑똑하게 만들지는 못할 것이다. 이런 경우에서는 우리는 새로운 능력도 얻지 못했을 것이고, 문제의 다음 라운드를 더 쉽게 만들지 못했을 것이고, 존망 위기에 대처하기에 쓸 수 있는지능의 상한을 깨지 못했을 것이다. 존망 위험 관리의 관점에서, 말 그대로 인간보다 더 똑똑한 친절하고 제정신의 지성을 생산해내지 못하는지능 행상 기술은 다음 논점을 제기한다: 혹시 같은 시간과 노력으로 같은 문제에 매우 똑똑한 현대 인간을 찾아서 나서게 하는 것에 더 유용하게 쓰일 수 있는가?
더군다나, 인간의 뇌의 "자연적인" 설계에 맞는, 뇌 자체가 나타내는 뇌의 개개 구성 요소들에 맞는 조상 환경에서 멀어질수록 한 개인의 정신 이상의 위험이 더 커진다. 만약 향상된 인간이 보통 인간보다 더 똑똑하면, 이것도 또한 전세계적 치명적 위험이다. 사악한 향상된 인간은 얼마나 큰 피해를 줄 수 있겠는가? 글쎄, 그가 얼마나 창의적일 것인가? 내 머리에 처음 떠오르는 질문은, "자신을 섬기는 재귀적 자기 개선을 하는 AI를 만들 만큼?"이다.
본질적인 인간 지능 향상 기술들은 특유의 안전 문제를 제기한다. 다시 말하자면, 나는 이 문제들이 공학적으로 해결하기 불가능한 문제들이라고 주장하지 않고, 단순히 문제가 있다는 점만 지적하기를 바란다. AI도 안전 문제들이 있고, 지능 향상도 안전 문제들이 있다. 덜거덕거린다고 해서 반드시 당신의 적인 것이 아니며, 철퍽거린다고 해서 다 당신의 친구인 것이 아니다. 한편, 친절한 인간은 우리가 "친화적"이라고 말하는 결정을 표현하는 도덕적, 윤리적, 아키텍처적 복잡성을 모두 포함한 채로 시작한다. 다른 한편, AI는 안정된 재귀적 자기 개선을 위해 설계되고 안전으로 이끌 수 있다: 자연 선택은 인간의 뇌를 다중적인 예방 조치들과 보수적인 결정 과정들과 몇 자릿수만 한 안전 마진으로 설계하지 않았다.
지능 향상은 인공 지능의 부차적인 주제가 아니라 그 자체의 주제이다. 이 챕터는지능 향상을 구체적으로 논할 공간아 부족하다. 말할 것은 내 직업이 시작될 때에는 인간 지능 향상과 인공 지능 두 가지를 다 고려했고, 내 노력을 인공 지능에 할당하기로 했다는 것이다. 주요한 이유는 내가 재귀적으로 자기 개선을 할 AI의 개발에 크게 영향을 미칠 만큼 유용한 인간을 뛰어넘는지능 향상 기술들이 제때 오지 않으리라고 기대한 까닭이다. 내 의견이 틀렸다면 나는 즐겁게 놀랄 것이다.
그러나 남들은 지능 향상을 연구할 때에, 향상된 인간들이 문제를 더 잘 해결하기를 바라며 친화적 AI에는 고의적으로 연구하지 않기로 하는 것은 좋은 방책이 아니라고 생각한다. 나는 인간 지능 향상이 AI보다 더 오래 걸린다면 치명적으로 실패할 방책을 받아들일 수 없다. (반대의 경우도 마찬가지이다.) 나는 생물학을 가지고 연구하는 것은 너무 오랜 시간이 걸릴 것을 우려한다. 관성과 자연 선택이 잘못 선택한 설계와 싸우는 일이 너무 많을까봐 우려한다. 규제청이 인간 실험을 허락하지 않을 것을 우려한다. 또한 인간 천재들도 그들의 기교를 습득하기에 수년을 소비한다. 향상된 인간이 더 빨리 학습해야 할 수록, 어떤 사람을 그 수준으로 올리기 더 어려울 것이다.
만약 지능 향상된 인간들이 나타나서 친화적 AI를 먼저 만들었으면 나는 즐겁게 놀랄 것이다. 그러나 이 결과를 바라는 이는지능 향상 기술의 발달을 가속시키기 위해 노력해야 할 것이다; 내가 속도를 늦추게 설득하기에는 어려울 것이다. 만약 AI가 지능 향상보다 자연적으로 훨씬 더 어려우면, 아무 문제가 없다. 만약 747을 만드는 것이 새를 부풀리는 것보다 자연적으로 더 쉬우면, 이 지연은 치명적일 것이다. 고의적으로 친화적 AI를 연구하지 않는 방책이 좋을 수도 있는 좁은

가능성의 영역이 있고, 그 방책이 헛되거나 해로울 가능성의 큰 영역이 있다. 인간 지능 향상이 가능하다고 해도 실질적인 어려운 안전 문제들이 있다; 나는 우리가 친화적 AI가 지능 향상보다 먼저, 또는 그 반대를 원하는지 심각히 고려해야 할 것이다.
나는 친화적 AI가 지능 향상보다 더 쉽거나 더 안전하리라는 주장에 높은 신뢰를 두지 않는다. 인간의 지능을 증가시킬 수 있는 상상하는 방법이 많다. 어쩌면 AI보다 쉽고 안전할 뿐 아니라 존망 위험에 영향을 줄 만큼 강력한 방법이 있을지도 모른다. 만약 그렇다면 나는 직업을 바꿀 수도 있다. 그러나 나는 인간 지능 향상이 더 쉽고, 더 안전하고, 좋은 영향을 줄 만큼 강력하다는 의문되지 않는 가정에 반대하는 몇 가지 고려할 점들을 지적하고자 한다.

12: AI와 다른 첨단기술의 상호작용

바람직한 기술을 더 빠르게 개발하는 것은 국부적인 방책인 반면, 위험한 기술의 개발을 늦추는 것은 어려운 과반수적 방책이다. 바람직하지 못한 기술의 개발을 중단하는 것은 불가능한 만장일치의 방책이 필요하다. 나는 우리가 기술을 개발하느냐, 안 개발하느냐의 관점에서 생각하는 대신 우리에게 실질적으로 사용할 수 있는 능력을 가지고 기술들을 빠르게 하거나 늦추고, 현실적으로 가진 능력을 가지고 어떤 기술들이 다른 기술보다 먼저, 또는 나중에 개발되기를 원하는지 묻기를 권하겠다.
나노기술에서는 일반적으로 제시되는 목적이 공격적인 기술들 전에 방어적인 보호 기술을 개발함이다. 내가 익 사실을 걱정하는 이유는, 일정한 수준의 공격적 기술은 그것에 방어할 수 있는 기술보다 정교성이 덜 필요한 경향이 있기 때문이다. 문명의 역사 동안 공격이 방어를 능가했다. 총은 방탄복보다 수세기 전 개발되었다. 천연두는 천연두 백신이 개발되기 전에 전쟁의 도구로 사용되었다. 오늘날에는, 핵폭발에 대해 방어할 수 있는 기술적 방패가 없고, 국가는 공격을 상쇄하는 방어가 아니라 공격적 공포의 균형으로 보호되고 있다. 나노기술자들은 본질적으로 어려운 문제에 착수한 것이다.
그러면 AI에 앞서 나노기술이 개발되는 것이 나은가, 아니면 나노기술 이전에 AI가 개발되는 것이 나은가? 이렇게 제시된 문제는 좀 속임수가 들어 있다. 해답은 나노기술, 또는 AI가 존망 위험으로서 더 곤란한가에는 관련이 적다. *순서*를 따지자면, 우리가 해야 할 질문은 "AI는 우리가 나노기술에 대처하기 더 쉽게 하는가? 나노기술은 AI에 대처하기 더 쉽게 하는가?"이다.
나는 인공지능의 성공적인 해결이 나노기술에 대처하는 데 큰 도움이 될 것으로 생각한다. 나노기술이 어떻게 친화적 AI를 개발하기 더 쉽게 하는지는 모르겠다. 커다란 나노컴퓨터가 친화성이라는 특정한 문제를 해결하기에 더 쉽게 만들지 않고 AI의 개발을 더 쉽게 한다면, 그것은 부정적인 상호작용이다. 그러므로 다른 것이 같다는 조건에, 나는 기술 발전의 순서에서 친화적 AI가 나노기술에 앞서 개발되는 것을 선호할 것이다. 우리가 AI의 도전에 맞서서 승리하면 친화적 AI를 시켜 나노기술에 대처함에 도움을 줄 수 있을 것이다. 우리가 나노기술을 먼저 개발하고 생존한다면 AI 문제가 아직 남아 있을 것이다.
일반적으로, 친화적 인공지능의 성공은 다른 거의 모든 문제를 해결하는 데 도움이 될 것이다. 그러므로 어떤 기술이 AI를 쉽게도, 어렵게도 하지 않지만 치명적 위험을 가한다면, AI의 문제에 먼저 맞서는 것을 선호해야 할 것이다.
사용가능한 계산력을 늘리는 기술은 인공 지능을 개발하는 데 필요한 최소한의 이론적 정교를 감소시키나, 친화적 AI에는 도움을 주지 않으므로 나는 부정적 영향으로 본다. 무어의 미치광이 과학 법칙: "18개월마다 세계를 파괴하기에 필요한 IQ가 1점씩 떨어진다." 인간 지능 강화에의 성공은 친화적 AI를 더 쉽게 만들 것이고, 다른 기술에도 도움이 될 것이다. 그러나 지능 강화는 친화적 AI보다 안전하지도 쉽지도 않을 수도 있고, 또한 친화적 AI가 지능 강화보다 본질적으로 더 쉬우면 우리가 실제적으로 사용할 수 있는 능력으로 이 두 기술의 자연적 개발 순서를 뒤집을 가능성이 작다.

13: 친화적 AI의 문제에 발전을 이룩하기

"우리는 10명의 연구가가 2개월 동안 뉴햄프셔 하노버의 다트머스 대학에서 인공 지능을 연구하기를 제의한다. 이 연구는 학습이나 지능의 다른 모든 양상이 원리상 기계가 모의할 수 있다는 추측을 바탕으로 진행되어야 한다. 기계로 하여금 언어를 사용하고, 추상화와 개념을 형성하고, 현재 인간만이 풀 수 있는 문제를 해결하고, 자신을 개선하게 할 방법을 찾아낼 시도를 할 것이다. 우리는 엄선된 과학자들이 한 여름 동안 이 문제를 함께 연구하면 최소한 하나의 문제에 중요한 발전이 이루어질 수 있다고 본다."

-- 매카시, 민스키, 로체스터, 섀넌(1955).

*다트머스 여름학기 인공 지능 연구 프로젝트 제안*은 "인공 지능"이라는 말이 처음으로 기록된 곳이다. 그들은 문제가 어렵다고 알려줄 사전 경험이 없었다. 그래도 한 여름학기 동안 연구하면 "중요한 발전이 이루어질 수 있다."라고 한 것을 문제 삼는다. 그것은 미개연성의 부담을 지니는 문제의 난도와 해석 시간에 관한 특정한 예측이다. 그러나 "이루어질 지도 모른다."라고 했으면 나는 반대하지 않았을 것이다. 어떻게 알 수 있었겠는가?
다트머스 제안은 다른 주제들과 함께 언어 소통, 언어적 추리, 신경망, 추상화, 무작위성과 창의력, 환경과의 상호작용, 뇌를 모형화하기, 독창성, 예측, 발명, 발견, 자기 개선 등의 주제를 포함했다.
이제 나에게는 언어, 추상적 사고, 창의성, 환경과의 상호작용, 독창성, 예측, 발명, 발견, 무엇보다도 재귀적 자기 개선을 가진 AI는 친화적이어야 할 정도를 크게 넘어선 것으로 보인다.
인간 수준의 AI가 가까운 것 같았던 그 화창한 여름에도 안전의 문제는 묵살하기 위해서도 언급되지 않았다. 다트머스 제안은 생명공학을 논한 아실로마르 회의와 탈리도마이드 아기, 체르노빌, 9.11 사건 전인 1955년에 작성되었다. 만약 인공지능이라는 아이디어가 오늘날 처음으로 제의되었다면, 누군가가 위험을 정확히 어떻게 관리하고 있는지를 알고 싶어할 것이다. 나는 이것이 우리의 문화의 좋거나 나쁜 변화라고 하는 것이 아니다. 좋은 과학, 나쁜 과학을 낳는지 논하는 것이 아니다. 그러나 다트머스 제안이 실제보다 50년 후에 쓰였으면, 안전도 주제였을 것이다.

2006년 현재에도 AI 연구 학계는 여전히 친화적 AI를 문제의 일부로 보지 않는다. 이 사실을 인용할 수 있었으면 좋겠으나 나는 없는 자료를 인용할 수 없다. 친화적 AI는 인기 없고 기금을 주지 않을뿐 아니라, 개념 자체가 없다. 무언가가 부족하다는 생각도 없어서 친화적 AI를 지도의 빈 자리라고 할 수도 없다.[5] 만일 독자가 『괴델, 에셔, 바흐』(Hofstadter 1979)나 『마음의 사회』(Minsky 1986) 등 AI를 어떻게 만들지 제안하는 대중을 위해 쓰인 책을 읽은 적이 있으면, 독자는 친화적 AI가 문제의 일부로 논의되지 않았던 것을 기억할 수도 있다. 기술적 문헌에서도 기술적 문제로 다루어진 것을 나는 본 적이 없다. 문헌을 찾아보니 주로 짧고
비기술적인, 서로 연결성이 없고 아이작 아시모프의 "로봇공학의 3원칙"(Asimov 1942)을 제외하고는 서로 참고 문헌을 가지고 있지 않은 논문이 나왔다.
2006년인 오늘에 왜 더 많은 AI 연구가들이 안전을 얘기하지 않고 있는가? 나는 다른 사람들의

[5]항상 이렇지는 않다. 널리 쓰이는 교과서인 *Artificial Intelligence: A Modern Approach* (Russell and Norvig 2003)는 "인공지능의 윤리와 위험"을 논하는 곳이 있으며, I. J. 굿의 지능 폭발과 특이점을 언급하며, 추가적인 연구를 곧 할 것을 촉구한다. 그러나 2006년 현재 이 태도는 규칙이 아닌 예외이다.

심리를 읽을 수 없으나, 개인적인 대화에서 얻은 정보를 가지고 간략히 추측하겠다.
인공 지능 학계는지난 50년에 겪은 경험, 즉 인간 수준의 능력 등의 큰 약속 뒤에 대중에게 창피한 실패를 당한 패턴에 적응해 왔다. 이 창피를 "AI" 탓으로 돌리는 것은 공평하지 않을 수도 있다. 아무런 약속을 하지 않은 더 현명한 연구가들의 신중함은 신문에 과시되지 않았다. 그래도 고도로 발달한 AI가 언급될 때에는 여전히 AI 학계 안팎에서나 이런 어겨진 약속이 즉시 생각난다. AI 연구의 문화가 이런 조건에 적응한 것이다. 인간 수준의 능력을 논하는 것은 금기시된다. 코드를 실행해서 보여주지 않은 능력을 예측하거나 주장하는 것에는 더욱 거센 금기가 있다. 내가 접한 견해는, 친화적 AI를 연구한다고 주장하는 이들은 은연중에 자기의 AI 설계가 친화적이어야 할 만큼 강력하다고 주장한다는 것이다.

이는 논리적으로도 맞지 않을뿐더러 실용적으로 좋지도 않은 관점인 것이 분명하다. 만약 친화적이어야 할 정도로 강력한 성숙한 AI가 실제로 만들어지고, 우리가 원하는 대로 이 AI가 제안 친화적이라면, 누군가가 오랫동안 친화적 AI를 실제로 연구했어야 한다. 친화적 AI는 처음 필요할 때 바로 만들어 내고 다듬어진 기존 디자인에 끼워놓을 수 있는 모듈이 아니다.
현재 AI에서는 수십 년간 조금씩 개선됨으로써 강력해진 신경망이나 진화 프로그래밍(EP)같은 기술이 있다. 그러나 신경망의 사용자는 신경망이 어떻게 결정을 내리는지 알 수가 없고, 알 수 있게 할 쉬운 방법이 없다. 신경망 기술을 발상해내고 갈고닦은 사람들은 친화적 AI의 장기적인 문제를 생각하지 않은 것이다. 진화 프로그래밍은 확률적이고, 코드의 최적화 목표를 정확히 보존하지 않는다. EP는 시험된 조건에서는 웬만하면 시킨 대로 하지만, 다른 일을 할 수도 있다. EP는 친화적 AI의 요구 사항에 본질적으로 부적당한, 강력하고 아직도 발달하고 있는 AI 기술인 것이다. 내가 제안하는 대로의 친화적 AI는 반복적인 자기 개선 하에서도 안정한 최적화 목표가 보존되는 것을 요구한다.
오랜 시간에 걸쳐 개발되고 개선되고 연마된 현존하는 가장 강력한 AI 기술은 내가 생각하고 있는 친화적 AI의 요구 사항과 기본적으로 맞지 않는 것이 있다. 치명적이지 않았지만 수리하기 매우 어려웠던 Y2K 문제도 또한 내일의 설계 사항을 미리 보지 못한 것에서 발생했다. 최악의 시나리오는 성숙하고 강력하고 공적으로 이용될 수 있는 AI 기술이 합해져서 친화적이지 않은 AI를 이루지만 지난 30년간의 AI 연구를 다시 하지 않으면 친화적 AI를 만드는 데 이용될 수 없는 상태에 빠지는 것이다.
현재 AI 학계에서 공공연히 인간 수준의 AI를 논하는 것만도 과거 AI 학계가 겪은 경험으때문에 용감한 일로 보인다. 그렇게 용감했다고 자신을 칭찬하고는 멈출 유혹이 있다. 그렇게 용감히 도전한 후에 또 초인간적인 AI를 논하는 것은 말도 안 되고 쓸데없을 것 같을 것이다. (그러나 AI들이 지능의 범위를 서서히 올라가고는 인간이 차지하는 점에 딱 영원히 멈출 이유가 없다.) 초인간적 AI의 전세계적 치명적 위험에 대처하기 위해 감히 친화적 AI를 논하는 것은 초월적이고 용감하다고 받아들여질 만큼의 배짱에서 두 수준이나 위인 것이다.
친화적 AI가 중요한 문제임은 인정하면서도, 현재의 이해 수준으로는 친화적 AI를 실용적으로 시도할 수가 없다는 주장도 있기는 하다 지금 당장 친화적 AI를 해결하려고 달려들면, 정말 실패하거나, 과학이 아닌 반(反)과학을 낳을 수도 있다는 우려이다. 이 반대는 일리가 있기는 하다. 내가 보기에는 충분한 공부를 하면 친화적 AI 연구를 처음부터 장애물에 부딪치지 않고 시작할 수 있는, 필요한 지식이 있기는 하다. 그런데 이 지식이 많은 분야에 흩어져 있다는 어려움이 있다: 결정 이론, 진화 심리학, 확률론, 진화 생물학, 인지 심리학, 정보 이론, 거기에다 또 전통적으로 “인공 지능”이라고 불린 분야... 현재 있는 다수의 연구가를 친화적 AI 연구에 진전할 준비를 시킬 만한 학습 과정이 여전히 준비되어 있지 않다.
수학과 음악에서 테니스까지 여러 분야에 걸쳐 검증된 천재성의 “10년 법칙”은, 어떤 분야에서 탁월한 성과를 얻으려면 적어도 10년을 노력해야 한다는 것이다.. (Hayes 1981) 모차르트는 4살 때부터 교향곡을 작곡했으나, *모차르트* 교향곡은 아니었다. 모차르트가 진정으로 뛰어난 교향곡을 작곡하게 될 때까지는 13년이 더 걸렸다(Weisberg 1986). 나 자신의 학습 곡선의 경험은 이 우려를 가중시킨다. 친화적 AI에 전진을 이룰 수 있는 사람들이 필요하다면, 그들은 급히 필요하기 전에 자신을 훈령하는데 전시간을 쏟아부어야 할 것이다. 만약 내일 게이츠 재단이 친화적 AI 연구에 기금 1,000만 달러를 부여한다면, 수천 명의 과학자가 즉시 기금
신청을 친화적 AI에 관련이 있는 것처럼 보이게 다시 쓰기 시작할 것이다. 그러나 친화적 AI 문제에 진정한 관심은 보이지 않을 것이다. (돈을 받을 기회가 생기기 전에는 관심이 없었다는 증거). 인공 일반 지능이 유행하지 않고 친화적 AI는 전혀 중요하게 여겨지지 않는 시기에는, 적어도 이 문제에 대해 말하는 사람들은 진정히 관심이 있다는 것은 추정할 수 있다. 어떤 학계가 풀 준비가 되어 있지 않은 문제에 돈을 너무 많이 쏟는다면, 여분의 돈은 과학보다 거짓 해결책들이 난무하는 반과학을 낳기가 더 쉽다.
나는 이 판정을 좋은 소식으로 여기지 못한다. 친화적 AI가 단순히 다수의 연구가와 대량의 돈을 쌓아놓음으로 해결될 수 있다면 무리 모두가 훨씬 더 안전한 상태에 있을 것이다. 그러나 2006년 현재 나는 이것이 사실임을 매우 의심한다 – 친화적 AI의 분야, 인공 지능 그 자체의 분야가 너무 혼돈의 상태에 있다. 그러해도 누군가가 우리가 아는 것이 부족하므로 친화적 AI에 진전할 수 없

다고 논하면, 이 결론에 도달하기 전 얼마나 공부를 했는지 의심해야 한다. 과학이 모르는 것일 누가 알겠는가? 과학은 한 사람이 공부하기에는 양이 너무나 많다. 도대체 누가 미리 우리가 과학 혁명에 준비되어 있지 않았다고 할 수 있는가? 또한 우리가 친화적 AI에 진전하지 못한다면, 친화적 AI가 필요하지 않다는 뜻이 아니다. 이 두 가지의 말은 *전혀 동등하지 않다*!
그러므로 우리가 친화적 AI에 진전할 수 없다는 것을 알게 된다면, 가능한 한 빠르게 그 상태를 벗어날 방법을 찾아야 한다! 우리가 어떤 위험을 다루지 못한다고 해서 그 위험이 친절히 그냥 떠나가 줄 것이라는 보장은 전혀 없다.
만약 입증되지 않은 재능이 뛰어난 젊은 과학자들이 자발적으로 친화적 AI에 관심을 두게 된다면, 그들이 전시간을 이 문제에 헌신하기 위해 다년간 연구비를 요청하는 것이 인류에 아주 유익한 일일 것이라고 나는 생각한다. 친화적 AI에 제공되는 자금이 아주 없어서는 안 되고, 이 일을 위해 지금보다 상당히 많은 자금이 필요하지만, 연구 초기에는 맨해튼 프로젝트 규모의 연구비는 잡음 대 신호 비율을 증가시킬 것이 우려된다.

결론

언젠가 나에게 현대 문명은 불안정한 상태에 있다는 생각이 떠올랐다. I. J. 굿이 추측한 지능 폭발은 끝에 아슬아슬하게 서 있는 펜과 같은 역학적으로 불안정한 시스템을 설명한다. 만약 펜이 정확히 수직으로 서 있다면 계속 서 있을 것이나, 조금이라도 기울어지면, 중력이 그 방향으로 더 끌어당기며 이 과정이 가속된다. 이처럼 더 똑똑한 시스템들도 자신들을 더 똑똑하게 만드는 것이다. 생명 없이 자신의 항성을 선회하는 죽은 행성도 또한 안정하다. 지능 폭발과는 달리 멸종은 역학적 끌개는 아니다. 거의 멸종된 것과 정말 멸종된 것은 큰 차이가 있을 것이다. 그러해도 완전한 멸종은 안정하다.
우리의 문명이 이 둘 중에 하나로 갈 것이지 않은가?
논리적으로는 위의 논증을 결점이 있다. 예를 들어 거대 치즈케익 오류가 있다: 지성들은 맹목적으로 끌개에 빠지지 않고 자신들의 동기가 있다. 그러해도 시제적으로 말하면 나는 우리의 대안들은 더 똑똑해지거나 멸종되는 그 둘 중에 하나인 것으로 추측한다. 자연은 잔인하지 않고 무관심하다. 이 중립성은 흔히 노골적인 적개심과 구별하기 어렵다. 현실은 당신에게 자꾸자꾸 도전을 던지고, 당신이 다룰 수 없는 도전에 직면하게 된다면 그 대가를 치러야 하는 것이다. 흔히 자연은 실패에 대한 벌이 죽음인 시험에서도 당신에게 심히 불공정한 요구들을 한다. 10세기에 사는 중세시대 농부가 어떻게 폐결핵의 치료법을 발명할 수 있는가? 자연은 도전들을 당신의 실력이나 가지고 있는 자산이나 문제를 생각할 수 있는 시간이 얼마 있는지에 맞추지 않는다. 그리고 당신에게 너무 어렵고 치명적인 문제에 마주치면 당신은 죽고 만다. 생각하기 싫겠지만, 이것이 수천년 동안 인간들이 겪는 현실이었다. 만약 인류가 불공평한 도전에 마주치면 인류도 쉽게 똑같은 운명을 맞이할 수 있다.
만약 인간들이 노화를 하지 않아서 100세 인간들이 15세 인간과 사망률이 같았으면, 우리는 불사하지 않을 것이다. 우리는 확률들이 우리를 따라잡을 때까지만 남았을 것이다. 노화하지 않는 인간도 우리의 세계만큼 위험한 세계에서는 100만 년만 살기 위해서도 어떻게든 연간 사고의 확률을 거의 0으로 줄여야 한다. 차를 운전해도 안 되고, 비행기를 타도 안 되고, 길의 양쪽을 다 보았어도 건너면 안 된다. 그것도 너무 큰 위험이기 때문이다. 인생을 즐길 생각은 모조리 단념하고, 생명을 보존하기 위해 생활을 포기한다 하더라도 백만년에 걸친 장애물 코스를 지날 수 없을 것이다. 물리적이 아니라 인지적으로 불가능할 것이다.
인류, 호모 사피엔스는 불로이지만 불사는 아니다. 호미니드들이 이렇게 오래 살아남은 이유는지난 수백년 간은 수소 폭탄 창고도 없었고, 소행성들을 지구로 돌릴 수 있는 우주선들도 없었고, 슈퍼바이러스들을 생산할 수 있는 생물 무기 연구소들도 없었고, 해마다 돌아오는 핵전쟁이나 나노기술 전쟁이나 악성 인공 지능의 가능성도 없었기 때문일 뿐이다. 상당한 시간 동안 생존하기 위해서는 각 위험을 거의 0으로 줄여야 한다. “그만하면 잘하는” 것은 1백만년 더 살아남기 위해 충분히 잘하는 것이 아니다.
불공정한 도전 같을 것이다. 그 수준의 능력은 역사적으로 인간의 제도가 아무리 노력해도 전형적으로 가지지 못했던 수준이다. 수십 년 동안 미국과 소련이 핵전쟁을 완벽히는 피하지 않았다.

1962년의 쿠바 미사일 위기 같은 위기일발도 있었다. 만약 미래의 지성들이 우리가 역사책에서 읽는 지성들과 똑같이 혼합된 어리석음과 지혜, 용기와 이기심을 나타낸다면, 존망 위기의 게임은 이미 끝난 것이고, 시작부터 패배한 것이다. 10년 더, 또는 한 세기 더 생존할 수는 있겠지만, 100만 년 동안 살아남지 못할 것이다.

그러나 인간의 지성은 가능한 지성의 한계가 아니다. 호모 사피엔스는 최초의 일반 지능을 나타낸다. 우리는지성의 까마득한 시초에서 태어났다. 만약 운이 좋으면 미래의 역사가들이 현재의 세계를 어정쩡한 청소년 시기와도 같은, 인류가 자신에게 중대한 문제를 만들 만큼 똑똑했으나 그 문제를 해결하기에는 좀 덜 똑똑했던 시기로 기술할 수도 있다.

그러나 이 청소년 시기를 통과하기 전 우리는 인간보다 더 똑똑한 지능이라는 성인의 문제에 직면해야 한다. 생활 주기의 사망률이 높은 단계에서 벗어날 방법이며, 취약성의 창을 닫는 방법이며, 우리가 맞서야 할 가장 큰 위험이기도 하다. 인공 지능은 이 도전을 해결한 하나의 길이고, 나는 이것이 우리가 갈 길이라고 생각한다. 결국은 현존하는 새의 크기를 늘리거나 제트 엔진을 이식하는 것보다 처음부터 747을 만드는 것이 더 쉬운 것으로 드러날 것으로 생각한다.

나는 정밀한 목적과 설계에 따라 우리 자신보다 더 똑똑한 것을 구축하는 일의 순전한 대담성을 평가절하하고자 하지 않는다. 하지만 인간의 과학이 부닥친 문제 중 이해하기 어려운 것으로 입증된 것 중에는 지능이 첫 번째가 아니라는 사실을 기억하자. 항성들도 화학과 생물학과 같이 한 때에는 비밀이었다. 순전한 과학이 알기 불가능한 것으로 여겨지게 되었다. 옛날 옛적에는 아무도 왜 어떤 물질은 활동하지 못하고 죽어 있는 반면 어떤 물질은 피와 생명력으로 움직이는지 알지 못했다. 생명을 가진 물질이 어떻게 자신을 복제하는지, 우리의 손이 왜 우리 정신의 명령을 따르는지 아무도 몰랐다. 켈빈 경은 다음과 같이 썼다:

> "동•식물의 생명이 물질에 가진 지배력은 이제까지 연구된 과학적 탐구의 범위를 한없이 초월한다. 증명된 일상의 기적인 우리 인간의 자유의지와 하나의 종자에서 식물들이 대대로 성장하는 것에 있는, 움직이는 입자들의 운동을 지배하는 생명의 능력은 원자들의 우연한 병행의 어떤 가능한 결과와도 무한히 다르다." (MacFie 1912에 인용)

모든 과학적 무지는 오래되었기 때문에 성스러움을 얻는다. 모든 지식의 부재는 인류의 호기심의 시초로까지 거슬러 올라가며, 그 빈자리는 영원한 것처럼 오랫동안 남아 있다가 누군가 채우게 된다. 나는 실수할 수 있는 인간들도 친화적 AI를 만드는 문제에 성공할 가망성이 있다고 생각한다. 그러나 그러려면 더는 생명이 켈빈 경에게 신성한 불가사의이었던 것같이 지능이 우리에게 신성한 불가사의가 되어서는 안 된다. 지능은 신성하든지 않든지, 더는 어떤 불가사의도 되어서는 안 된다. 우리는 인공 지능의 창조를 정확한 기술의 정확한 응용으로 수행해야 한다. 그러면 승리할 수 있을지도 모른다.

참고 문헌

Asimov, I. 1942. Runaround. *Astounding Science Fiction,* March 1942.

Barrett, J. L. and Keil, F. 1996. Conceptualizing a non-natural entity: Anthropomorphism in God concepts. *Cognitive Psychology,* 31: 219-247.

Bostrom, N. 1998. How long before superintelligence? *Int. Jour. of Future Studies,* 2.

Bostrom, N. 2001. Existential Risks: Analyzing Human Extinction Scenarios. *Journal of Evolution and Technology,* 9.

Brown, D.E. 1991. *Human universals.* New York: McGraw-Hill.

Crochat, P. and Franklin, D. (2000.) Back-Propagation Neural Network Tutorial. http://ieee.uow.edu.au/~daniel/software/libneural/

Deacon, T. 1997. *The symbolic species: The co-evolution of language and the brain.* New York: Norton.

Drexler, K. E. 1992. *Nanosystems: Molecular Machinery, Manufacturing, and Computation.* New York: Wiley-Interscience.

Ekman, P. and Keltner, D. 1997. Universal facial expressions of emotion: an old controversy and new findings. In *Nonverbal communication: where nature meets culture,* eds. U. Segerstrale and P. Molnar. Mahwah, NJ: Lawrence Erlbaum Associates.
Good, I. J. 1965. Speculations Concerning the First Ultraintelligent Machine. Pp. 31-88 in *Advances in Computers, vol 6*, eds. F. L. Alt and M. Rubinoff. New York: Academic Press.
Hayes, J. R. 1981. *The complete problem solver.* Philadelphia: Franklin Institute Press.
Hibbard, B. 2001. Super-intelligent machines. *ACM SIGGRAPH Computer Graphics*, 35(1).
Hibbard, B. 2004. Reinforcement learning as a Context for Integrating AI Research. Presented at the *2004 AAAI Fall Symposium on Achieving Human-Level Intelligence through Integrated Systems and Research.*
Hofstadter, D. 1979. *Gödel, Escher, Bach: An Eternal Golden Braid.* New York: Random House
Jaynes, E.T. and Bretthorst, G. L. 2003. *Probability Theory: The Logic of Science.* Cambridge: Cambridge University Press.
Jensen, A. R. 1999. The G Factor: the Science of Mental Ability. *Psycoloquy*, 10(23).
MacFie, R. C. 1912. *Heredity, Evolution, and Vitalism: Some of the discoveries of modern research into these matters – their trend and significance*. New York: William Wood and Company.
McCarthy, J., Minsky, M. L., Rochester, N. and Shannon, C. E. 1955. *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.*
Merkle, R. C. 1989. Large scale analysis of neural structure. Xerox PARC Technical Report CSL-89-10. November, 1989.
Merkle, R. C. and Drexler, K. E. 1996. Helical Logic. *Nanotechnology,* 7: 325-339.
Minsky, M. L. 1986. *The Society of Mind.* New York: Simon and Schuster.
Monod, J. L. 1974. *On the Molecular Theory of Evolution*. New York: Oxford.
Moravec, H. 1988. *Mind Children: The Future of Robot and Human Intelligence*. Cambridge: Harvard University Press.
Moravec, H. 1999. *Robot: Mere Machine to Transcendent Mind.* New York: Oxford University Press.
Raymond, E. S. ed. 2003. DWIM. *The on-line hacker Jargon File,* version 4.4.7, 29 Dec 2003.
Rhodes, R. 1986. *The Making of the Atomic Bomb.* New York: Simon & Schuster.
Rice, H. G. 1953. Classes of Recursively Enumerable Sets and Their Decision Problems. *Trans. Amer. Math. Soc.,* 74: 358-366.
Russell, S. J. and Norvig, P. *Artificial Intelligence: A Modern Approach.* Pp. 962-964. New Jersey: Prentice Hall.
Sandberg, A. 1999. The Physics of Information Processing Superobjects: Daily Life Among the Jupiter Brains. *Journal of Evolution and Technology,* 5.
Schmidhuber, J. 2003. Goedel machines: self-referential universal problem solvers making provably optimal self-improvements. In *Artificial General Intelligence,* eds. B. Goertzel and C. Pennachin. Forthcoming. New York: Springer-Verlag.
Sober, E. 1984. *The nature of selection.* Cambridge, MA: MIT Press.
Tooby, J. and Cosmides, L. 1992. The psychological foundations of culture. In *The adapted mind: Evolutionary psychology and the generation of culture,* eds. J. H. Barkow, L. Cosmides and J. Tooby. New York: Oxford University Press.
Vinge, V. 1993. The Coming Technological Singularity. Presented at the VISION-21 Symposium, sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute. March, 1993.

Wachowski, A. and Wachowski, L. 1999. *The Matrix*, USA, Warner Bros, 135 min.
Weisburg, R. 1986. *Creativity, genius and other myths*. New York: W.H Freeman.
Williams, G. C. 1966. *Adaptation and Natural Selection: A critique of some current evolutionary thought.* Princeton, NJ: Princeton University Press.