9-2018

Morality as Social Software

Jongjin Kim
The Graduate Center, City University of New York

MORALITY AS SOCIAL SOFTWARE

by

JONGJIN KIM

A dissertation submitted to the Graduate Faculty in philosophy in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2018

Morality as Social Software

by

Jongjin Kim

This manuscript has been read and accepted for the Graduate Faculty in philosophy in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Date	Jesse Prinz
	Chair of Examining Committee
Date	Nickolas Pappas
	Executive Officer

Supervisory Committee:

Rohit Parikh

Melvin Fitting

Jesse Prinz

THE CITY UNIVERSITY OF NEW YORK

Abstract

Morality as Social Software

by

Jongjin Kim

Advisor: Rohit Parikh

The dissertation research is a project to understand morality better through the concept of 'Social Software.' The dissertation is, consequently, to argue that the morality in a human society functions as a form of social software in the society. The three aspects of morality as social software are discussed in detail: the evolutionary, anti-entropic, and epistemic game-theoretic aspect.

We humans 'usually' think that, for example, (a) killing other humans without any necessary reason is morally wrong, and (b) helping other humans in need is morally right. We want to know, in this dissertation research project, why we think in such ways. Myriads of answers to this question have already been offered. We will pursue an answer that has more explanatory power and enlightening lucidity.

The term, 'Social Software' was coined by Rohit Parikh to connote, broadly, social "procedures that structure social reality" (van Eijck and Parikh 2009, p. 2). The term can be understood, "more or less equivalently," (Parikh 2002-1, note 2) as 'social procedure,' 'social algorithm,' or 'social game.'

(1) The first aspect of 'morality as social software,' to be discussed is the evolutionary: human morality has emerged and developed further through the process of evolution; (2) the second aspect

is the anti-entropic: human morality is human resistance against the universal law of entropy that tends to annihilate everything from order to disorder; (3) the third aspect is the epistemic game-theoretic: human morality is understood better by epistemic game theory, which is a combination of 'classical game theory' and relatively new 'epistemic logic.'

As more specific case studies for the epistemic game-theoretic aspect, the concepts of backward induction and "the less we know, the more rational and moral," are discussed. Finally, a thorough discussion on the naturalistic fallacy instills more philosophical rigor into the dissertation.

Dedication

To my mother and in memory of my father

who let me learn

what is morally right and what is morally wrong, so that

how I ought to live with other beings.

Acknowledgements

It is my delightful duty to acknowledge the help that I have received in the writing of this dissertation.

I appreciate the four elements (earth, water, fire, and air) and the six kinds of quarks (up, down, charm, strange, top, and bottom) that have allowed me to write this dissertation. I appreciate that the food grown on earth came to me with water and air to make the fire within me.

I would like to thank my dissertation supervisory committee comprised of Professors Rohit Parikh, Melvin Fitting, and Jesse Prinz.

I am deeply grateful to my dissertation adviser Rohit Parikh for his judicious advice and firm encouragement (though sometimes "tough but fair"). Literally and in any counterfactual way, without Professor Parikh as a scholar, teacher, mentor, and friend of mine, this dissertation would not have been shaped into the present form. Though, as a science-trained philosophy disciple, I do not endorse any supernatural entity and concept, I have sometimes wondered that Professor Parikh and I might have had another teacher-student relationship in another trans-world. Especially, for the last few years, due to a strange University Rule between Colleges, he has not received the advising credit from advising me; still we have met his office, apartment, and cafés, once, twice, or three times a week whenever any issue has arisen. (Rohit 'bhai,'[1] I will remember this as long as my personal identity and memory persist.)

To Professor Melvin Fitting, I am deeply indebted for his teaching, encouragement, and trust. Since I first took Professor Fitting's *First-Order Modal Logic* course, his lucidly written logic textbooks[2] and articulately spoken logic classes have been the major sources of my 'humble'

[1] "(Indian) Added to proper names to form an affectionate form of address to an older person," *Oxford Living Dictionaries (English)*, https://en.oxforddictionaries.com/definition/bhai.

[2] Such as, among others, Melvin Fitting (1990/1996 2nd ed.), *First-order Logic and Automated Theorem Proving*; Melvin Fitting and Richard L. Mendelsohn (1998), *First-order Modal Logic.*

understanding of logic; since he became my first dissertation adviser and then, before he retired (I had not expected that fast retirement), his encouragement to me to pursue what I really wanted to research has come to fruition in this dissertation as an interface between ethics and logic; and since he has joined the prospectus (proposal) committee, and then later the supervisory committee, (I was fortunate to have him in those committees even after his 'half'- retirement) his trust in my work has made me expedite the completion of this dissertation. (Mel, I will show you what I promised in the defense.)

To Professor Jesse Prinz, I owe a great deal of the content and the form of this dissertation. I should confess that, while reading Professor Prinz's works on moral philosophy, I have woken up to shatter my dogmas and to construct my view on morality (if I may, as Kant admitted that Hume interrupted Kant's "dogmatic slumber"). I thank Professor Prinz for his professional services as the chair of both my prospectus exam and dissertation defense, especially when the unexpected happened in the prospectus exam. I appreciate Professor Prinz's policies, "Students first!," and further, "Crying students first!." (Thanks, Jesse, my teacher.)

I would like to thank Professors Richard Mendelsohn and Samir Chopra, who were the members of my prospectus committee, for their helpful suggestions and critical comments. I hope Professor Mendelsohn is now savoring his time after retirement; I will remember the long discussion with Professor Chopra at the first-floor cafeteria of the Graduate Center, and I believe we will have more time to share together.

I am grateful to the committee of the comprehensive exam in ethics: Professors Stefan Bernard Baumrin, Douglas Lackey, and Steven Ross. As examiners, their standards were really high and hard to pass; while striving to meet their standards, I was able to come up with many of the ethical ideas discussed in this dissertation.

I am also grateful to the former and the current chairpersons (Executive Officers) of the philosophy program of the Graduate Center, CUNY, for their inexhaustible support: Professors John Greenwood, Iakovos Vasiliou, and Nickolas Pappas. During the countless times that my Ph. D. education was in danger of being derailed, especially, Professor Greenwood's considerate and encouraging involvement resolved the (administrative and financial) issues. (Thanks, John, I will remember you and Singapore.)

In addition to those professors, I owe lots of ideas discussed here to the conversations in classrooms, cafés, and pubs with fellow students and colleagues. They are: Kathleen R., Todd, Cosim, Tudor, Len Mitchell, Hidenori, Hirohiko, Naoko, Yoko, Ren-June, Junhua, Yunqui (QiQi), Can Başkent, Çagil, Yoonhee, Sung Oo; and my friends whom I met at the University at Buffalo (SUNY): Konishi, Ping, S. Choi.

My language is often "shaky," both written and spoken, and both in my mother tongue and English. To my editor, Jen, for her thorough and creative edit, I am deeply grateful. Or, more than that: It has been an "ontological turn" that her editing work has made me feel free to write whatever I want to write.

Instead of resorting to circumlocution, let me be direct: Without money, this work would not have been possible. The following universities' students and chairpersons have provided me with financial support (as well as academic) by hiring me as an adjunct instructor: Professors Mark Halfon of Nassau Community College, Enrique Chávez-Arvizo of John Jay College of Criminal Justice, Julie Maybee of Lehman College, Arlig Andrew of Brooklyn College, Howard Ruttenberg, George White, and Timothy Kirk of York College, and Tiger Roholt of Montclair State University. I appreciate the shining eyes of those students.

However, “adjuncting” in New York City would not be enough to make ends meet, and so my “budget deficit” has had to be filled by the financial support of numerous friends, in addition to brothers, sisters, and their families. I would like to express and “engrave” my deep gratitude to them: Kyu-Tae Jeong, Kyu-Dong, Jae-Myeong Kim, Sang-Gyun, Hee-Baek, Chang-Ryung, Ho-Seok, and brothers and sisters. So, not rhetorically, but literally, they all have some shares of my life and what I will achieve.

Finally, all the help appreciated here is my fortune; still, all errors and mistakes remaining in this dissertation are my sole responsibility. I will fix them in my next version. Whereas the entire universe is trembling to blossom a flower, it may not be necessary so to blossom a Ph.D. dissertation. Still, through the writing of this dissertation, I have learned and felt that we do and should, live together, and, thanks to all the help, this dissertation project has been possible.

May 31, 2018

Jongjin Kim

Table of Contents

List of Tables, Figures, and Diagrams

Chapter 1. Introduction: Philosophical Motivation and Background

§1-1. The Spirit: From the Ultimate Questions to Morality as Social Software

"How ought I to live?" This is one of the ultimate questions that make inquisitive children or adults become interested in philosophy. This question is investigated mainly, among others, in the field of ethics (moral philosophy). When an additional phrase is added to the ethical question, we have the next ultimate question, "How ought I to live 'together with other human beings'?" This question is investigated mainly in socio-political philosophy. If, once more, an additional phrase is added to the socio-political question, we have a third ultimate question, "How ought I to live together with other human beings and 'all other sentient beings'[1]?" Now this question is 'metamorphosed' from axiological into ontological, and so it becomes closer to another ultimate question, "Who am I? (or What am I?)."

I believe that discussions about ethical problems cannot be separated from discussions about socio-political and ontological problems, because a human being is, figuratively speaking, not an individual atom (or island), but rather a multifaceted jewel tied to a knot of the 'Indra's Net.'[2] All the jewels in Indra's Net reflect, and are reflected in, all the others repeatedly and infinitely. Following this metaphor, we cannot discuss completely the morality of an action of a person without discussing his relationships with others, his knowledge of the morality of others, others' knowledge of his morality, and so on, ad infinitum (like 'common knowledge' in epistemic logic).

Until I entered a college, though I was a strict rule-following teenager, my curiosity about the foundation of moral rules was trivial. I was more inclined to ask questions such as "What is the beginning (or end) of the universe?" and "What is the meaning of life?." So, perhaps, I was more

[1] By the term, "sentient beings," I mean, in this dissertation, a vague definition, "things that can feel."

[2] A recent philosophical discussion on Indra's net is found in Priest (2014), *One: Being an Investigation into the Unity of Reality and of its Parts, including the Singular Object which is Nothingness*, Ch. 11. Absence of Self, and the Net of Indra.

of a metaphysical kid than a practical one. When I started college, then everything changed. In my college years, I was a feeble sentient being who suffered, most seriously among others, from some fellow young men's tragic instances of self-immolation for social causes. They screamed: for example, "Justice for equal labor!," "Free general vote in presidential elections!," and "Restore democracy!" Those young men did not harm any other persons, but themselves; they did not follow any others, but their own autonomy and conscience. I did not participate in the student protests actively; I was just trying to resolve differential equations in natural science, while feeling a guilty conscience sharply. Ever since I watched or heard those instances, I have been eager to know the answers to the questions, such as "What makes an action right or wrong?," "What is the foundation of morality?," and, hence "How ought I to live?." The research project of this dissertation is my quest for answers to such questions. I argue, throughout this project, that better answers, if not the best, can be centered on the concept of 'Morality as Social Software.'

§1-2. A Thought on the Foundation of Morality

An agent's morality is composed of the standards that the agent holds regarding what is right or wrong. The agent's moral standards are ideal goals that the agent tries to follow when choosing what is right or wrong. A moral standard is like a ruler. Just as an agent may use a ruler when she measures the length of a thing, so she depends on her moral standards when she performs some actions. These words describing an individual agent's morality should equally apply to a group of agents, too. For example, the morality of a society is made up of the moral standards that the members of the society hold, though, of course, it is altogether different and difficult matter to determine how to measure the 'average' morality, if ever, of the society.

The concept of morality need not be so mysterious. Though people may not have an explicit sense of their morality, they usually have the sense that certain ways of living are more worthwhile than others. And, that sense becomes more explicit, when the issue is involved in bringing up their children. Bernard Williams (1985/2006, p. 48) writes: "we have much reason for, and little reason against, bringing up children within the ethical world we inhabit." The moral standards which parents try to raise their children to share are likely to be the morality of the parents. Perhaps, not many parents intentionally raise their children with some vices. Imagine the long list of hot moral issues in the contemporary world. We might not be sure whether we are praising, endorsing, or blaming those controversial moral claims. But, if we apply those claims to our children (e.g., if a father imagines that his son supports such-and-such a controversial moral claim), we become much more certain of our positions for or against those claims. Of course, it is another matter whether those children obey their parents.

The foundation of morality is the origin of moral standards. When we ask about the foundation of morality, we ask who or what has given us the moral standards, and how they have been made. The current standard of, for example, length and time have been stipulated by scientists. The meter is, 'operationally,' defined as the "distance traveled by light in vacuum during a time interval of 1/299,792,458 of a second."[3] Likewise, there are some moral standards. Then, there must be someone, something or some entity that has made the moral standards. The way in which morality has been constructed, I argue, is not very different form the ways in which the units of the meter and the second have been stipulated by scientists, that is, human beings.

[3] https://physics.nist.gov/cuu/Units/meter.html. In addition, the "second is defined as the duration of 9,192,631,770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium-133 atom." https://physics.nist.gov/cuu/Units/second.html.

According to common theories of the foundation of morality in ethics (for example, in (Rachels 1986) and (Pojman 1990)), it can be argued that the moral standards are given by ***God*** (Divine Command Theory), ***Nature*** (Natural Law Ethics), ***human reason*** (Kantian deontology), ***the consequences of actions*** (of an individual, Ethical Egoism; of the group involved, utilitarianism), ***Social agreement*** (Social Contract Theory); or moral standards depend on our ***characters*** (Aristotelian virtue ethics): Many different foundations of moralities have been claimed. Of course, each claim has problems. For example, there are many gods worshiped in many religions, and the moral standards of one religion often conflict with those of others. We do not know which god is the absolute one; therefore we do not know which moral standards are the absolute ones. Regarding Nature, there are many human interpretations of Nature, though there may be only one Nature. People have different views on how to live according to Nature. If we understand the situation wherein many different kinds of foundations of morality are claimed to be correct, and many different kinds of problems are also raised, then we may accept that ethical relativism is unavoidable. And also, we may pursue a less problematic and more explanatory foundation of morality.

§1-3. Morality by Social Contract

It is social contract theory that has a deep evolutionary connection with the main thesis of this dissertation, morality as social software. According to social contract theory, broadly, by Hobbes, Locke, and Rousseau, the common morality (and its realization through laws of nations) has been agreed upon by people (members of society) via a shared social contract. On this view, morality is a human invention and construct, not given to us by other entities such as Gods or Nature. It is often pointed out, for example, by Hume, that a typical problem of social contract theory is that

real historical persons have neither signed the contract nor witnessed, say, the signing ceremonies: the contract is simply a fiction fabricated by imaginative theorists. One of my theses in this dissertation is that we can explain the existence of this contract based on an evolutionary epistemic game-theoretic process which could change this fictional fabrication into a real historical story.

In Western history, there is a strong intellectual tradition claiming that morality is a human invention, not given by Gods or Nature. Callicles, in Plato's *Gorgias*, may be the first prominent figure who explicitly argues for the idea that morality is a human invention. Then, the idea was followed by Thrasymachus in Plato's *Republic*, Machiavelli, Hobbes, Locke, Rousseau, Kant and Marx. Perhaps, Nietzsche was the last prominent philosopher before the 20th century who argued for this idea.[4] Contemporary theorists of social contact theory include Rawls, Scanlon and Gauthier, among others.[5]

Callicles (483b)[6] claims that "the people who institute our laws [that is, justice or morality] are the weak and the many. They do this, and they assign praise and blame with themselves and their own advantage in mind."[7] The most distinctive points of this claim are that laws are made by us, humans, and that these humans are weak and many (like the humans in democracy).

Unlike Callicles, Thrasymachus argues that "justice [that is, law or morality] is the advantage of the stronger" (338c), not the weaker. Nevertheless, like Callicles, Thrasymachus' argument is based on the idea that morality is a human invention. A succinct summary of his argument is as follows: Rulers make laws for their own advantage; they are the stronger; therefore, justice is nothing other than the advantage of the stronger. In Book 1 of *Republic*, Socrates' refutation of

[4] A recent discussion on Nietzsche's moral philosophy is found in Prinz (2007), *The Emotional Construction of Morals*, Chapter 6, The Genealogy of Morals.

[5] The more specific thoughts and books of these philosophers, from Callicles to Gauthier, will be discussed through the dissertation, when appropriate.

[6] When quoting from Plato's works, we use the traditional "Stephanus Pagination."

[7] On a discussion about the claims of Callicles and Thrasymachus, refer to Barney (2004/2017).

Thrasymachus is humble and weak. It looks as if the author of the book, i.e., Plato, was somewhat sympathetic with Thrasymachus' argument. I think that Thrasymachus' claim is more subversive and revelatory than Callicles', in the sense that Thrasymachus' claim highlights the stronger people, not the weaker, so that it foreshadows Machiavelli, Marx, and Nietzsche, among others.

The idea in the tale of the "Ring of Gyges" in *Republic* is also similar to the idea that morality is a human invention (359a-360d). In the tale's lesson, humans are no longer moral when they become invisible because they do not fear punishment for their wrong actions. I do not think it is a coincidence that the tale appears in Book [chapter] 2 of *Republic*, which is near Thrasymachus' fervent speech in Book [chapter] 1: the two ideas may be purposely arranged that way in the books [chapters] by Plato, to deal with similar ideas. Since the tale will be investigated as a typical example of a context in which epistemic logic plays an important role, I think we may develop a further thought experiment, in which, not just one person has such a ring, but two, many, or all persons have their own rings of Gyges (another epistemic issue!).

Social contract theories can be divided into two categories: the egoistic (self-interested) Hobbesian line and the altruistic Kantian line (cf. Cudd 2000/2017, "Contractarianism."). The egoistic Hobbesian line holds that humans are self-interested, so that humans consent to the social contract in order to maximize their self-interest. By contrast, the Kantian line holds that human rationality requires us to respect other persons, and therefore, humans consent to the social contract. On the one hand, the Hobbesian line is often connected to the term, 'contractarianism,' while the Kantian line is connected to 'contractualism' (Ibid.). On the other hand, Ashford and Mulgan (2012) distinguish contractualism from Kantian moral philosophy, and instead, they connect it to Rousseau. ("Contractualism," Section 2. How does contractualism differ from other social contract theories?) In this dissertation, I avoid using the terms 'contractarianism' and

'contractualism'; rather, I prefer using the terms, 'egoistic Hobbesian line' and 'altruistic Kantian line,' since I believe these terms are simpler than others, intuitively, at least for the current purpose. Cudd (Ibid.) puts Gauthier (1986 *Morals by Agreement*) in the Hobbesian line; while Rawls (1971/1998 *A Theory of Justice*) and Scanlon (1998 *What We Owe to Each Other*) in the Kantian line. I think Rawls can also be regarded as the self-interested Hobbesian line, too, since humans behind the veil of ignorance in Rawls's theory try to maximize their self-interest. (See chapter 6 below for further discussion.) [8]

§1-4. The Three Aspects of Morality

The topic of morality raises many questions. Everyone has his or her own views on morality, regardless of her or his circumstances such as religion, political views, and education. I think that discussing morality can be compared to the famous ancient parable of 'the blind men and an elephant.' If a person's sight is disabled, he may recognize an elephant as a snake, a fan, a pillar, a wall, or a rope, depending on which part of the elephant he is touching with his hands. So, if the topic of morality is like an elephant, then I am like a man whose sight is disabled. Given this limiting condition, my discussion in this dissertation will focus on only three aspects of morality: evolutionary, anti-entropic, and epistemic game-theoretic aspects.

[8] It will be continued and repeated later in this dissertation to discuss the theme of comparison between the egoistic Hobbesian line and the altruistic Kantian line, together with some variants of the theme, such as: the debate between Nature and Nurture; between Darwin's moral theory and Huxley's moral theory; and, in ancient Chinese philosophy, the debate between Mencius (circa 372-289 BCE) and Xunzi (circa 310-220 BCE) on whether human nature is good or evil: briefly, Mencius argues for the innate goodness of humans, basing on four sprouts of human nature; while Xunzi argues that human nature is evil, so humans need to reform it radically, to live together. It may not be exact, but still it can be contrasted that the Mencius thought is related to the altruistic line above, and the Xunzi thought is related to the egoistic (self-interested) line.

§1-5. A Longer Abstract: Nine Points

In addition to the shorter abstract above, the following nine points are argued in more specific detail.

(1) **Social software is a complex concept**: The initial coinage of the term 'social software' need not restrict its enlarging and enriching use. Once artwork such as novels, poems, paintings, and music are released to the audience, the interpretations are open to them, the audience. Analogously, the initial concepts of 'social algorithm' or 'social procedure' can be expanded to contain morality. I will focus on this point and argue that if we view morality as social software, we can better understand morality.

(2) **Societies with better social software have advantages in evolution**: The first of my three aspects of morality as social software is the evolutionary aspect. I endorse the evolutionary view of morality, that is, the view that morality has been developed through the process of evolution in human history. For instance, we may view that, 'Homo Sapiens' or 'we' have flourished while Neanderthal people went extinct, because Homo Sapiens have developed more advantageous morality as social software than Neanderthals.[9]

(3) **Humans have evolved to be cooperative**: In the evolutionary process of morality, cooperation plays a crucial role. I discuss further the similarities and differences among these concepts: 'cooperation,' 'altruism,' 'mutualism,' 'collaboration,' and 'selfishness.' Michael

[9] Whether this view is true or not is not a topic of the dissertation. I am just introducing an exemplary discourse. Still I believe that the claims in anthropology that Homo Sapiens and Neanderthal met, mate, and (sometimes) ate each other, so that 3 to 5 percent of genes of Homo Sapiens stem from our Neanderthal ancestors.

Tomasello (2008) gives dramatic examples of how helpful even toddlers tend to be. This cooperative nature allows humans to carry out large and complex enterprises.

(4) **Complex software requires coordination and cooperation**: While cooperation is crucial in morality, how humans coordinate cooperation is a significant question. Here, the second of the three aspects of morality is discussed: I argue that some common concepts in epistemic game theory such as backward induction, common knowledge, and logical powers are often used in human social coordination.

(5) **Individuals making plans within the context of social software use backward induction**: In this dissertation, I discuss backward induction in more detail. (Common knowledge will not be discussed, only mentioned.) When an individual is making plans such as to go somewhere, she asks, "Where do I want to go? What are the intermediate steps? What resources will I need?" and then after performing the backward induction, working backward from her goal, she then relies on the existence of social software like a bus system or the currency system to pay her fare. And honesty plays an important role.

Adam Smith says,

> *"It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest."*

But Smith is mistaken because self-interest is only part of the story. The morality to which the butcher and the baker are committed also plays a large role. For example, when a toddler helps his mother vacuum her apartment he is not thinking of what is "in it for him." He wants to help! Some butchers are dishonest, but if most butchers were dishonest, the system would break down.[10]

(6) **Societies with less entropy are able to have better algorithms**: The third of the three aspects of morality as social software is that morality is against entropic disorder. The universe has an intrinsic inclination to make things disorderly, following the law of entropy. By contrast, living organisms, from amoebas to humans, are the exceptions to this inclination. In order for an organism (a life form) to live its life, it must resist this inclination, and its resistance tends to be a struggle. I suggest that humans build societies to win this struggle, and the morality of a society functions as an algorithm to reduce the entropic disorder.

(7) **These three aspects can only work if most people are moral**: This dissertation project is based on the relativistic view of morality. There is no such thing as absolute morality given by any supernatural entity; rather, we humans construct our own morality. There is no absolute moral or immoral person. To say that most people in a society are moral means that those people recognize these three aspects of morality as social software, regardless of whether it is conscious or not.

(8) ***Our* sense of morality is not vulnerable to Hume's Law and the naturalist fallacy**: As a concluding philosophical remark on my work so far, I discuss Hume's Law (no 'ought' can be

[10] See Fukuyama (1995) for the role which trust plays in the running of society.

derived from an 'is') and its sibling, the naturalistic fallacy (drawing non-natural moral values from natural values), to argue that these criticisms are not as powerful as they might seem at first glance. While I do not completely rely on emotion in order to refute the naturalistic fallacy, I find that emotion – and an emotional feeling of the desire to be moral can also have its own function in moral discourses to a degree.[11]

(9) **Humans need knowledge to do their moral jobs**: In order for humans to do moral works, we need some, or any, form of knowing. A person taking part in a social algorithm needs to know when to act and how to act. The bus driver needs to know when he *should* show up at the bus stop and the passenger needs to know when he *will* show up.

§1-6. A Background: Teaching Artificial Intelligence Morality as Social Software

While focusing on only the three aspects above and below, I should mention a motivational background to this entire dissertation project in a larger landscape. We humans have been watching the advent of artificial intelligence (AI): rumors, omens, warnings, and curses, as well as the good news of a utopian society. In 2016, AlphaGo of Google Deep Mind, an AI program that plays the board game Go, amazed the world by defeating the reigning human world Go champion. This victory of AI is in addition to AI's history of defeating humans in chess by IBM's Deep Blue in 1997 and in the TV quiz show by IBM's Watson in 2011. If the advancement of AI continues at this speed, then it is claimed that, 'soon' (within 30 years, 50 years or some uncertain time in the future,) AI will reach so-called 'Singularity' (Kurzweil 2006, Chalmers 2010, 2012; and as a critique of Singularity, Prinz 2012) or 'Superintelligence' (Bostrom 2014) where AI exceeds

[11] See Prinz (2007)

humans in (almost) all aspects of humanity (more human than humans): e.g. intentionality. Since I watched the 'AlphaGo Shock,' I have thought that this kind of discourse is not frivolous science fiction. As I watched AlphaGo's 'mysterious' moves in the Go games (which, at first sight, were regarded as blatant mistakes, but then later, confirmed as great moves), I felt as if AlphaGo were whispering in my ear, "Hey Humans, shut up!" I believe, with other philosophers, computer scientists, and policy makers, that teaching AI morality is a vital issue that we as the human race must deal with right now to maintain the continuity of humanity. Just now, in February of 2018, a group of scholars published a warning report: *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation* (Miles Brundage, Avin Shahar et al.). "Are humans going to be irrelevant?" So, the questions arise, "How can artificial agents act morally?" and "What moral actions can we humans perform in order to counter the threat to humans from cheap AI agents?" A guaranteed basic income for all citizens, as a palliative, has been suggested by some, to maintain society when AI programs take most of the jobs.[12] Here, the concept of morality as social software is well-timed. I argue that, in order to 'teach AI morality,' viewing morality as a form of software is a legitimate idea. Again, this topic of 'teaching AI morality' will not be discussed any further but will be functioning as 'a background and a motivation' of the entire dissertation project.[13]

[12] Van Parijs, Philippe. "Basic Income: A simple and powerful idea for the 21st century." Redesigning Distribution (2001): 4.

[13] For further discussion, see *Moral Machines: Teaching Robots Right from Wrong* (Wendell Wallach and Colin Allen 2009)

§1-7. The Structure of the Dissertation with a Diagram

The following diagram is to explain the structure of the dissertation more visually:

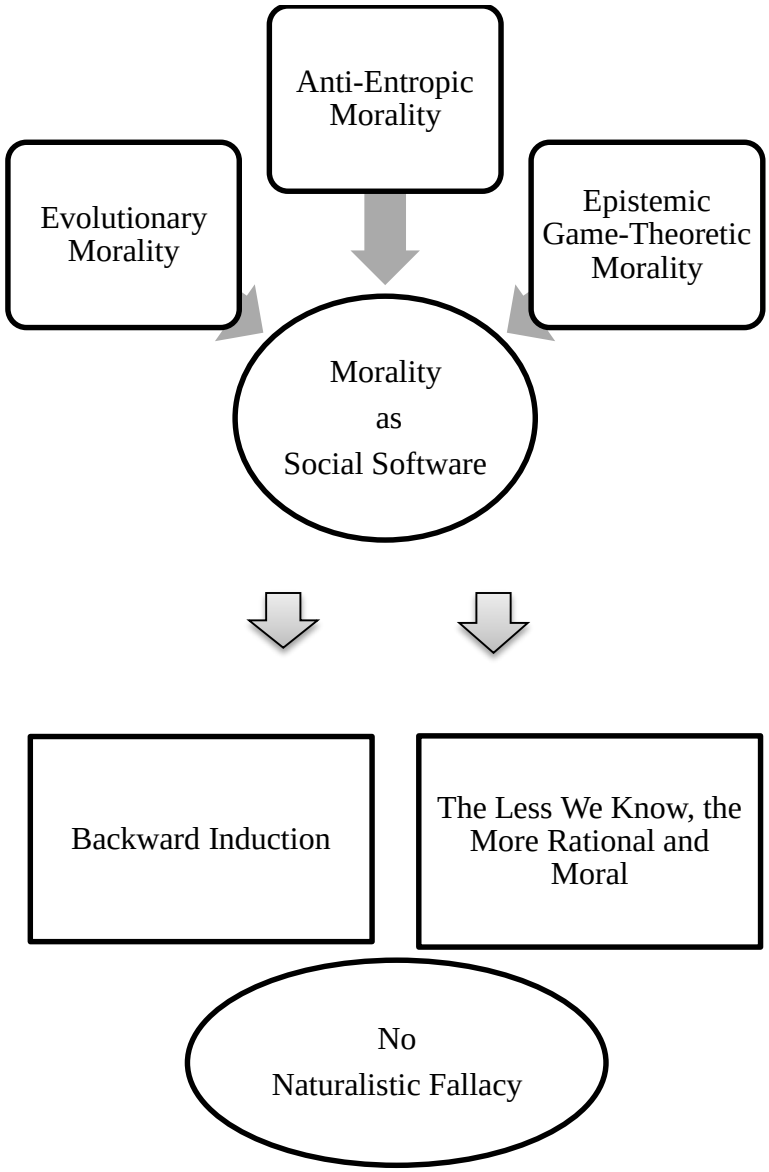


Diagram 1: The Structure of the Dissertation

§1-8. Plan of the Chapters[14]

Chapter 2: The purpose of chapter 2 is to incorporate the relatively new term, 'Social Software,' into my discourse on morality. I first discuss the essential characteristics of the concept, 'social software.' Then, I introduce the reader to three prototypical examples that demonstrate the concept, 'morality as social software': Those are 1) King Solomon's dilemma, 2) cake cutting procedure, and 3) freedom and knowledge-based obligation. Some brief and less technical discussions on these topics are summarized, for the purpose of providing the reader with a foretaste of later discussions on 'morality as social software.' Finally, I introduce and discuss an analogy between evolutionary morality and Monte Carlo tree search in artificial intelligence.

Chapter3: The main subject to be discussed in this chapter is my interpretation of the thesis, "Morality is evolutionary." I will make it clear what I mean by that, among a discussion of a variety of interpretations of scholars and thinkers. More specifically, I first discuss the concept, 'Natural Selection' and its units (or levels). Second, I introduce three cases that demonstrate the concept of natural selection through morality. I then discuss altruism and cooperation, which have been central to the debates about evolutionary morality. Finally, I summarize my interpretation, while discussing Jesse Prinz's 'sentimentalist theory of morality.' The whole purpose of this chapter is to set a 'playground,' where the two concepts, social software and evolutionary morality can 'play' with each other. While I touch on subtle and controversial matters, I will mainly try to summarize the standard views of the major topics in evolutionary moral philosophy.

[14] N.B. This summary of chapters is copied from the beginning paragraphs of each chapter, and is pasted here, "verbatim," in order to help the reader grasp the main theses of the dissertation at one place, more conveniently. Still, the reader's discretion is recommended, while reading later the same parts in each chapter.

Chapter 4: The single main thesis that I argue in this chapter is as follows:

(1) Main Thesis: Morality as social software is anti-entropic, that is, against entropy.

This main thesis has three connections here: (a) foundation (or background); (b) argument; and (c) corollary. The (a) foundation part is, 'roughly put,' as follows:

(2) The second law of thermodynamics holds that the total entropy of a system tends to increase over time.

(3) The concept of entropy often translates, 'mistakenly,' as 'disorder' in ordinary language.

(4) So, the second law may be interpreted, 'mistakenly,' as stating that the disorder of a system tends to increase over time.

(5) The inexorable tendency of entropy governs everything in the universe.

(6) Or, for some moment of time, there are exceptions, which are living organisms.

(7) From this ontology of nature, a normative rule can be deduced: organisms such as amoebae and my life do, and should do, resist the law of entropy in order to live lives.

The (b) argument for the main thesis (that morality as social software is anti-entropic) is as follows:

(8) Premise 1: Entropy and information are sibling concepts.

(9) Premise 2: Entropy and equilibrium are also sibling concepts.

(10) Conclusion IE (also, Premise 3) ('I' for 'information'; 'E' for 'equilibrium'): Therefore, information and equilibrium are also sibling concepts.

(11) Premise 4: Information and software are sibling concepts.

(12) Premise 5: Equilibrium and morality are also sibling concepts.

(13) Conclusion MS (also, Premise 6) ('M' for 'morality'; 'S' for 'software'): Therefore, morality and (social) software are sibling concepts, that is, morality as social software.

(14 = 1) Main Thesis: Therefore, morality as social software is anti-entropic, that is, against entropy.

Lastly, two corollaries, which result from this main thesis, will be discussed.

Chapter 5: The main subject to be discussed in this chapter is the inference of backward induction. I argue that backward induction is one of the most common and vital kind of inferences related to morality as social software. I first introduce backward induction intuitively by discussing the centipede game and a dialogue between Alexander the Great and a philosopher. I then argue that Popperian creatures in Dennett's theory are, basically, animals who exercise backward induction, and I introduce some exemplary animals from an Indian tale. I finally discuss the relationship between morality and backward induction. At the end of the chapter, a very brief précis of epistemic logic and epistemic game theory is attached, as an appendix.

Chapter 6: The main subject to be discussed in this chapter is a combination of a series of ideas: that is, (1) if we know less, then it is sometimes possible that we become more rational; (2) being rational is being moral; 'therefore,' (3) if we know less, it is sometimes possible that we become more moral. The first idea of (1) above is discussed in Parikh (2017), and I develop it further by adding the ideas of (2) and (3). I first discuss two simple cases, cake cutting and a tiger as a

motivational introduction of this chapter. I then discuss John Rawls's 'veil of ignorance' in the 'original position.' I finally discuss a verse of T.S. Eliot's poem "April," comparing to Norbert Wiener's thought of information.

Chapter 7: The main thesis to be argued in this chapter is that the naturalistic fallacy and its 'living ancestor,' Hume's Law, can be resolved from the viewpoint of morality as social software that we have developed so far. The naturalistic fallacy is committed, on G. E. Moore's accusation, when one draws X's goodness from any of its natural properties; and, an 'ought' (value), on Hume's contention, cannot be deduced from 'is' (fact). Since this dissertation project is, essentially, based on the evolutionary concept of morality, the project may not be completely immune to the criticisms from the naturalistic fallacy and Hume's Law. That is, it may be critiqued that the construction of evolutionary morality as social software here is not enough to explain why we ought to follow 'that' morality just constructed. It seems inevitable, therefore, that I should provide sound replies to these kinds of criticisms before concluding the dissertation. I hope that the discussion in this chapter instills a bit more philosophical rigor into the theses of the dissertation.

Chapter 8: I conclude the dissertation by mentioning the topic of ethical relativism, which is seemingly necessary but has been neglected so far; and some future research topics, which have been recognized, but not developed in my dissertation research, and still look promising. And finally, some last words.

Chapter 2. Social Software

The purpose of this chapter is to incorporate the relatively new term, 'Social Software,' into my discourse on morality. First, I discuss the essential characteristics of the concept, 'social software.' Then, I introduce the reader to three prototypical examples that demonstrate the concept, 'morality as social software': 1) King Solomon's dilemma, 2) cake cutting procedure, and 3) freedom and knowledge-based obligation. Some brief and less technical discussions on these topics are summarized, to provide the reader with a 'foretaste' of later discussions on 'morality as social software.' Finally, I introduce and discuss an analogy between evolutionary morality and the Monte Carlo tree search in artificial intelligence.

§2-1. Various Definitions of Social Software

The term, 'Social Software,' was coined by Rohit Parikh in his paper "Language as Social Software" (1995), and developed further in his later literature (especially 2001, 2002-1, 2002-2, 2014) and in collaborations (especially Pauly 2001; Pacuit 2005; van Eijck and Verbrugge 2009, 2012; van Eijck 2009, 2014).

The concept, 'social software,' can be understood interchangeably as: 1) 'social procedure,' if the reader is not familiar with computer science, 2) 'social algorithm,' if familiar, and 3) 'social game' if familiar with game theory. Parikh uses these terms "more or less equivalently" in his paper "Social Software" (Parikh 2002-1, note 2), which was monumental to the research program on social software. He uses 'social software' broadly as "procedures that structure social reality," without fixing the term to a precise definition (van Eijck and Parikh 2009, p. 2). This broad usage seems to me to have let the term 'evolve' more autonomously to include some other broader concepts.

Eric Pacuit later (2005, p. 3) says:

> *Social software is an interdisciplinary **research program** [emphasis is mine] that combines mathematical tools and techniques from game theory and computer science in order to analyze and design social procedures.*

I emphasize the expression, "research program." The term 'social software' in Pacuit's definition denotes, not only specific social procedures for social realities, but also the "emerging interdisciplinary field" (Pacuit 2005, p. iv) itself. This kind of broader connotation is similar to that of, for example, 'artificial intelligence,' in that the term 'artificial intelligence' now denotes, not only computers or computer programs that can perform human-like intelligence, but also the 'research program' itself. Van Eijck and Verbrugge give us a more recent definition (2014) of 'social software' as denoting: "the emerging interdisciplinary enterprise that is concerned with the design and analysis of algorithms that regulate social processes," which is similar to Pacuit's.

The analogies between 'social' software/hardware and 'computer' software/hardware are useful to explicate our current topic, though there are still differences between them.

Firstly, on the one hand, it may be hard to draw a clear-cut line of demarcation between computer "hardware (the machine) and software (the programs running on the machine)," since many of the built-in system jobs are done by hardware (van Eijck and Parikh 2009, p. 2). Nevertheless, the demarcation is not always impossible. "[W]hat can be changed without changing the machine itself is called software" (Ibid.).[15]

[15] Still, there may be the metaphysical issues of identity over time such as, famously, the "ship of Theseus."

Analogously, there also seems a demarcation problem between social hardware and software, but nevertheless demarcating is not always impossible. Van Eijck and Parikh (2009) say that:

> *[Social] hardware consists of institutions such as schools, churches, law courts, parliaments, banks, newspapers, supermarkets and prisons, while social software consists of the more specific procedures followed in these institutions.*

We may simply regard, as social hardware, optical fibers that connect bank buildings; and, as social software, the banking systems. But, what about "banks as institutions," in the definition above? Are "banks as institutions" social hardware or software? The demarcation issue here is not simple, so we need further discussion.[16]

Secondly,

> *[Computer] software is roughly divided into system software, namely, the software that is needed to make other software run, and application software, the software that turns the computer into a tool for a specific task* (van Eijck and Parikh 2009, p. 2).

Analogously, I argue in this dissertation that the morality in a society corresponds to social system software; by contrast, a specific ethical code for a particular behavior corresponds to application software.

Thirdly, although most computer software is designed directly by programmers, some other software systems result from evolutionary processes.

[16] Perhaps, there are more than two levels. We may consider "harder software" and "softer hardware" as the third and fourth level, based on how easily it changes or how long it lasts.

> *[Large] software systems such as the Linux operating system, --- can certainly be viewed as products of evolution of a certain kind, --- and by a process of genesis and natural selection* (van Eijck and Parikh 2009, pp. 2-3).

Note those familiar terms in evolutionary theories, such as 'genesis' and 'natural selection.' Analogously, again,

> *There is a large class of social practices that have evolved in the course of development of a civilization. --- Other social practices were designed and redesigned over a long period of time, e.g., the principles of common law. --- It is obvious that the foundations and principles of legislation are part and parcel of the broad field of social software* (van Eijck and Parikh 2009, pp. 3-4).

Let's remember that one of the main theses of this dissertation is that morality has evolved through the evolutionary, epistemic, game-theoretic process. The morality in a society is certainly one of "a large class of social practices" above. And also, let's remember the common saying that "Obeying the law is the minimum level of ethical conduct enforced in society." If the law in a society is social (system) software, so is the morality.

Fourthly, Parikh, as another approach, divides social software into two categories: individual level and social (societal) level (Parikh 2014). Parikh's own example demonstrates the individual level. If one goes from his apartment to a hotel in Chicago, his schedule will consist of several steps such as:

1. *Taking a cab to the airport,*
2. *Boarding the plane,*
3. *In Chicago, taking a cab to the hotel,*

and these steps constitute a social procedure, which is social software in an *individual* level (Parikh 2014, p. 2). By contrast, on the *societal* level, more (or multi-) agents and social structures are involved, and the interactions among the multi-agents are important factors of that social software. I think that, roughly, the concept of social system software matches social software in the societal level, and the concept of social application software matches social software in the individual level (though further investigation will be needed). As Parikh observes, one role of our society is "to serve as an operating system within which we can write our individual program" (Parikh 2014, p. 3). Common operating systems are those old-fashioned UNIX and DOS, and more recent Apple OS, Microsoft Windows, and Google Android. If we regard a function of our society as an operating system like those, the morality of a society serves as built-in social system software.

The most recent discussion on social software that I can find in the literature is Başkent's (2017) "A Non-classical Logical Approach to Social Software," in the Book, *Rohit Parikh on Logic, Language and Society* (eds. Başkent, Moss, and Ramanujam 2017). Başkent argues that "non-classical logical approaches can enrich, broaden and support the agenda of social software" (Başkent 2017, p. 91). By "non-classical logics" (and, interchangeably in the paper, by logical pluralism), Başkent means paraconsistent logics, where the "rule of explosion" fails, that is, "non-trivial inconsistent theories" are possible; in short, paraconsistent logics allow "inconsistent-tolerant models," and, Başkent believes that "this is a key notion in understanding social software."

It may seem no wonder that "enriching, broadening and supporting" the discourses of social software can be done by adding the non-classical perspective, once we recall that logic developed throughout history when various non-classical logics were added into the classical Aristotelian logic. What is novel in Başkent's discussion, I think, is that he applies this paraconsistent addition to the broader fields of ethics and economics, which are outside logic. He discusses, in ethics, the incident of 'Kitty Genovese'[17] and Priest's (1987/2006 2nd (Extended) edition) example of the situation where someone is obliged to do x and not to do x, and in economics, the real-world economics movement. In all these examples, it is possible that inconsistent normativity (that is, inconsistent action-guiding rules) can arise. I think that Başkent's discussion is a natural form of development of the term, 'social software' as part of the process of its evolution, though he seems to focus his attention on "using tools in logic and computer science" (Başkent 2017, p. 91).

I do not impose such restrictions to the definition as "tools in logic and computer science" in my theses of morality as social software in this dissertation. My argument will be that the concept, "social software" does not have to be restricted within concepts such as 'procedure,' 'algorithm,' or 'tool,' to which we have been introduced so far. My argument will be that morality itself (not just some rules or procedures that may constitute the higher-level morality) can be understood as social software. This kind of enlarging of a concept is not unusual, if we look around at things in the world and in the universe. Artworks such as novels, poems, paintings, songs, and movies are open to various interpretations: they are, on Sir Karl Popper's terminology, the products in 'World 3'[18]; organisms such as amoebas, roses, and human beings are open to the process of evolution.

[17] This tragic incident is discussed in many ethics textbook, as a provoking moral question. An original newspaper report: Martin Gansberg, "37 Who Saw Murder Didn't Call the Police," *New York Times*, March 27, 1964.

[18] As well known, Popper, while developing his theory of evolutionary epistemology, distinguishes the world (and, actually, the universe) into three categories: World 1, World 2, and World 3. See (Kim 1998, "Cyberspace and Karl Popper's World 3,")

Our very concept, social software, has been evolved through the discussions of various philosophers and logicians since its birth, and will continue to evolve. I am attempting, in this dissertation, to add one more pebble into the evolutionary process.

§2-2. Fair Division 1: King Solomon's Dilemma

Fair division of goods (and services) has been a vital practice for human societies in order to survive and flourish, since, probably, the human species started thinking about the concepts of property and its ownership. If we set aside the prehistory of the world for a moment, human history shows that a society that fails to divide its goods (and services) fairly to its members tends to collapse in the end. I am now considering, for example, the Roman Empire and the current 21st century global village of human society. The decline and fall of the Roman Empire resulted partially from its failure to divide goods fairly; and the 21st century global village might collapse unless the current convergence of economic wealth toward a relatively small number of people, is fixed.

'Distributive justice' has been a crucial element of the just society, since the time of, at least, the hunter-gatherer societies needed to discuss important issues like who eats the most delicious part of their hunted prey (e.g. the sirloin part of a hunted buffalo), and how to share what is gathered, especially since their prey had to be hunted by groups of people who all contributed to the effort. And later, 'distributive justice' became significant due to, for example, the work of Adam Smith and Karl Marx, which has continued with John Rawls in the 20th century, and to Joseph Stiglitz, Paul Krugman, and Thomas Piketty in the 21st century. The common English expression that represents fair division between two people is: 'I cut, you choose.' It is the procedure in which one person divides, and then the other person chooses one of the two pieces. This procedure is believed

to make the division fair. We can apply this procedure to large groups of people, not just two, and this is further discussed in the examples below.

2-2-1. The First Dispute

The Old Testament (1 Kings 3:16-28) tells a story of two women claiming the same baby, and going to see King Solomon for his adjudication. Solomon settles the dispute by threatening to cut the baby in half. The real mother refuses this option preferring to give up her baby in order to save its life. The fake mother agrees to share half of a dead baby rather than give up a living infant. King Solomon knows which woman is the real mother by their responses and gives the child to the woman who agreed to give it away. So, the real mother gained her portion by agreeing to give away her entire portion.

2-2-2. A Second Dispute

Van Eijck and Parikh (2009, pp. 4-6) discuss a possible second dispute about a baby between two women. Suppose that another pair of women, who have heard about Solomon's first judgment, come to Solomon claiming one baby as their own. Now, Solomon's threatening 'software' to cut the baby in half as he did in the first dispute may not work, since both of the women may say that they want the other to have the baby. "Almost all social procedures are susceptible to strategic behavior," so, for example, the fake mother acts strategically by pretending she is willing to give up the baby (van Eijck and Parikh 2009, pp. 4-6). What procedure can Solomon use when he resolves the matter of the two women with one baby so that the justice of fair division is fulfilled?

In a nutshell, Solomon can settle this second dispute by proposing to sell the baby to the highest bidder, while he assumes the highest must be the real mother. The following are several

possibilities that van Eijck and Parikh (2009, pp. 4-5) discuss, which I believe show some characteristics of morality as social software very well:

> *- Solomon offers the two women a generous loan from the Temple funds, to be paid back in monthly installments plus interest.*
>
> *- The rules are publicly announced: bids in closed papyri, the highest bidder gets the baby at the offered price, and the loser pays a fee into the Temple funds to cover court expenses.*
>
> *-This might not work if the fake mother has more money than the real mother, so it is better to ask them how many times their annual income they are willing to bid for the baby.*
>
> *-If, in closed papyri, the first mother offers A times her annual income and the second B times, with A > B, then the child should go to the first mother for '**B times her**' annual income.*
>
> *-N.B. What the first mother has to pay is neither 'A times' her income (which she initially offered) nor B times 'the second' mother's income, since 'B times the first mother's income' is the amount that she would have paid in an open auction, when the second mother would drop out at B times her income.*[19]

[19] Some other variations of this solution by Moore (1992) and Pauly (2005) are also introduced in van Eijck and Parikh (2009, pp. 5-6):

-Suppose the baby is worth A times the real mother's annual income, and B times the fake mother's, with A > B.

-After their bids in sealed papyri, Solomon announces the procedure in which he will toss a coin to decide who gets the baby, and will rule that the woman who gets the baby pays M times, with A > M > B, and the other pays a small fine.

-Then, since the fake mother is not likely to run the risk of having to pay more than the baby is worth to her, the baby goes to the real mother.

§2-3. Fair Division 2: Cake Cutting Procedure

In Solomon's dilemma, the baby is not divisible, and three persons (two women and Solomon) are involved. We discuss now the cake cutting procedure, which is about divisible goods that can be shared among many persons. Due to this divisibility and the multi-agents involved, the cake cutting procedure is a better example of "morality as social software" than Solomon's.

The cake cutting procedure can work as "a metaphor for a *division of a single heterogeneous good*" (van Eijck and Parikh 2009, p. 7); (also, Brams and Taylor 1996; Brams 2005). For example, dividing an inheritance among *n* heirs is similar to dividing a cake among *n* participants, when both the inheritance and the cake are heterogeneous. That is, the inheritance (say, a plot of land) can be composed of parts that are worth different values like a cake can be composed of parts that have different toppings and cream.

A basic procedure to cut the cake (or land) fairly is as follows:

> *The first person among n participants cuts out a piece. He is satisfied with that piece if he can take it, but first he offers it to the n-1 other participants. If someone else wants it, that person takes it, and other n-1 participants, including the first person, continue dividing; if no one else wants it, the first person takes it and let the other n-1 participants, now excluding the first person, continue* (see van Eijck and Parikh 2009, p. 7).

Here, the order is important: who cut first, and next; who chooses first, and so on. Nevertheless, this basic cake cutting procedure is not *envy-free*. "A cake division is called *envy-free* if each person feels that nobody else received a larger piece" (van Eijck and Parikh 2009, p. 7). For example, when a second person takes the piece that the first person cuts and if a third person is

envious of it, the division procedure is not envy-free. (A division between two persons is envy-free by nature, since there is no third person.) As a fairer division, 'Banach and Knaster cake cutting algorithm'[20] is *envy-freer* than the basic one above:

> *The first person among n participants cuts out a piece for him and claims that it is a fair share. All other n-1 participants examine it in turn. If nobody objects, the first person takes it. If a second person raises an objection, he has a right to cut off a slice and put that back with the rest of the cake. Then, he asks if he can take the reduced piece. This procedure will continue until someone gets the trimmed piece* (see van Eijck and Parikh 2009; Parikh 2002-1).

§2-4. Freedom and Knowledge-based Moral Obligation

As a third example of 'morality as social software,' I now discuss the thesis that the moral obligation of a moral agent depends on the agent's knowledge about the circumstances (Pacuit, Parikh and Cogan 2004; Pacuit and Parikh 2006). And I mention, considering the common discussions on freedom and obligation in ethics, that this dependence on knowledge is essentially related to the agent's freedom. In a nutshell, 'no knowledge, no freedom, no obligation.'

Some typical cases are as follows:

a) *A female physician Uma has no obligation to treat a patient Sam who is a neighbor of hers, if Uma does not know Sam is ill*.

[20] Which was attributed to Banach and Knaster by Steinhaus (1949).

b) *Sam's daughter comes to Uma's house, and tells her. Now Uma does have an obligation to treat Sam* (Pacuit and Parikh 2006, p. 18).

These cases remind us of the famous 'Problem of Evil' in the (Judeo-Christian-Islamic) philosophy of religion. According to the argument based on the manifestation (or existence) of evils, God is responsible for his 'non-existence' (simply put, God does not exist). For, if he is omniscient, omnipresent, omnipotent, and omnibenevolent, he must stop the occurrence of evils. But evils occur, so he does not exist. Relating to our current discussion, God's all-knowing (omniscient) attribute results in his obligation.

c) *Mary having a heart attack is a patient in a hospital. Now the hospital has an obligation to be aware of Mary's condition at all times and to provide emergency treatment as appropriate. There is not only a knowledge-based obligation, but also the obligation to* ***have*** *the knowledge* (Pacuit and Parikh 2006, p. 18).

People who argue against the existence of God may regard God as the hospital above that is neglecting its obligation to look after its patients.

d) *Uma is about to inject a patient with drug d, without knowing that the patient is allergic to drug d and there is a good alternative drug d'. Nurse Rebecca helping Uma is aware of these two facts. It is then Rebecca's obligation to inform Uma* (Ibid.).

In these cases (a)-(d), the obligations arise depending on the circumstances, mainly the moral agent's knowledge.[21]

It is a common discussion in moral philosophy (and legal philosophy, too) whether a moral agent (or a criminal) is responsible for his action (or crime) when it does not seem that he had the freedom not to do that action (or crime.)[22] Our current discussion on 'knowledge-based obligation' sheds some light on the relationship between freedom and responsibility, by showing that, if an agent has no knowledge, then he has no freedom, so no obligation.

It is also interesting to note that one of the strategies in theodicy relating to the problem of evil above is to give human beings freedom, in order to argue for the existence of God. God exists despite evils. For evils are a kind of "collateral," or the price of enjoying the freedom which God allows us. When we don't exercise our freedom correctly due to our ignorance (that is, no knowledge), some evils may occur.

§2-5. Evolutionary Morality and Monte Carlo Tree Search in Artificial Intelligence

One of the main theses with which this dissertation project is 'hardwired' is that "Morality is evolutionary." While I will discuss in more detail what I mean by 'evolutionary' in the next chapter, here in the last section of the chapter on social software I introduce an interesting analogy between the concept of evolutionary morality and the Monte Carlo tree search (MCTS) in artificial

[21] Formalization of knowledge-based obligation is another matter. Pacuit et al. (2004) develop a logic that captures the dependence of obligation on knowledge. The semantics of the logic extends the 'history-based models,' which were originally discussed in Parikh and Ramanujam (1985, 2003) and Parikh and Krasucki (1992). By contrast, in deontic logic, one of its goals is to prove formally that an agent is obligated to do an action; an agent's knowledge is only informally represented; and the focus is on representing epistemic obligations, that is, what an agent 'ought to know' (Parikh and Pacuit 2006, p. 19).

[22] The case of "Leopold and Loeb" is historic in this debate. Two wealthy recent graduates of the University of Chicago brutally murdered a 14-year-old boy in 1924. At the trial, their defense attorney Clarence Darrow argued that the two young murderers were not responsible for their actions, and the responsibility is "somewhere in the infinite number of [their] ancestors, or in [their] surroundings, or in both."

intelligence (AI). The affinity among the terms such as 'hardwired,' 'software,' 'MCTS' and 'AI' can reveal the theses of this dissertation more vividly.

In March 2016, AlphaGo, an AI program that plays the board game Go, amazed the world by defeating the reigning human world Go champion. Though some AI computer programs had already defeated top human performers in many other fields (for example, in chess by Deep Blue of IBM in 1997 and in TV quiz show by Watson of IBM in 2011), it had not been generally expected that an Al program would defeat a champion human Go player so early. The complexity of the board game Go with 19 x 19 grid is extremely higher than other games such as chess with its 8 x 8 grid. It has been estimated that the total number of possible games for Go is 10^{761}, whereas for chess, 10^{120} (Requote from Jackson 1985, p. 125, originally for chess, Shannon 1950; for Go, Zobrist, 1969). We may (or may not) estimate how big the Go's number 10^{761} is compared to the number 10^{80} which is the estimated number of all the atoms in the universe. So, until AlphaGo's defeat of the human champion in 2016, it had been generally believed that it would take much longer for an AI to beat a human Go champion, after advancing the AI's capacities to calculate all the possibilities (and to cool its machine). Then, suddenly the defeat happened! How was it possible?

AlphaGo has been in development by the Google DeepMind team since 2015. Through adopting a new approach, AlphaGo has become more 'intelligent' than any other AI computer programs playing Go. The 'secret' of this new approach is explicated in a Journal *Nature* article, "Mastering the game of Go with deep neural networks and tree search" (David Silver et al 2016). According to the authors of the DeepMind team, the essence of AlphaGo's algorithm is to combine its neural networks[23] with the Monte Carlo tree search (Ibid. p. 484). Basically, the stunning

[23] For a little further discussion, the neural networks of AlphaGo are convolutional networks "that have achieved unprecedented performance in visual domains: for example, image classification [and] face recognition"

creativity of AlphaGo's algorithm results from the combination of a new kind of machine learning, that is, deep learning and MCTS. AlphaGo often reveals great novel moves that were never imagined or played by human (professional) Go players in the last 2500 years. At first glance, those novel moves seem erroneous (or, silly and crazy), but then later it turns out that those moves are great.

In perfect information zero-sum games such as chess and Go, the traditional well-known winning strategy is the minimax algorithm, which is to minimize the maximum possible loss. However, the minimax algorithm becomes quickly infeasible when the nodes of a game tree increase quickly, as in Go. An alternative strategy is one based on the Monte Carlo tree search. Instead of searching all possibilities exhaustively, the MCTS strategy simulates many, not all, games. At first, simulations are chosen randomly. As simulations go on, each node and simulation accumulate some values. For example, if a node leads to a win more often, that node will have better values. In this way, after many simulations, each node and simulation accumulate the more accurate values, and the players can select the better winning moves, that is, the better optimal decisions[24] (Cameron Browne et al 2012).

(Silver et all 2016, p. 484). AlphaGo's neural networks contain two kinds of networks, the 'policy' networks and the 'value' networks. "The policy networks estimate the probability distribution of moves from a given position; the value networks estimate the "value" of a given board configuration from a given position" (Fu 2016, p. 667). With the expressions of Demis Hassabis, a co-founder of DeepMind, arguably the father of AlphaGo, the "policy networks select the next move to play, ---the value networks predict the winner of the game" (Hassabis 2016).

[24] In the case of AlphaGo, it simulates "thousands of random self-play" with another AlphaGo (Silver et al 2016, p. 486).

Now, relating to evolutionary morality, it is hinted why I am introducing here AlphaGo and MCTS, which may have seemed irrelevant. An aspect of evolutionary morality that I argue is that human morality and moral senses are 'hardwired' to the human mind and body, in a similar way to how MCTS works. Some points made by M. Ruse and E. O. Wilson (1986) are relevant to my current thesis. They argue that "the human brain is not a *tabula rasa*" and "nothing like all-purpose cognition occurred during human evolution" (1986, p. 180). I think these claims can be compared to the situation in which an AI computer program for Go has to calculate all possible 10^{761} moves. As this calculation is infeasible, the human brain cannot think of all possible situations from a blank slate. Conversely, Ruse and Wilson argue that the human brain is not "genetically determined in the strict sense" either (Ibid.), though it may be debatable how strict is strict.

As a synthesis of these two opposite theses, they argue that:

> *The human brain is something in-between: a swift and directed learner that picks up certain bits of information quickly and easily, steers around others, and leans toward a surprisingly few choices out of the vast array that can be imagined* (1986, p. 180).

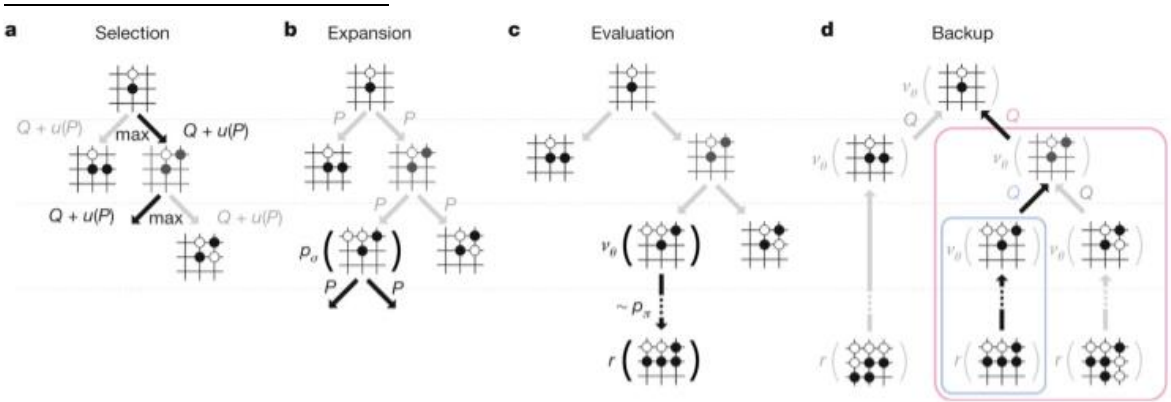


Figure 1: Monte Carlo Tree Search in AlphaGo

Source: Silver et al 2016, p. 486

I emphasize that, like the MCTS strategy, the human brain chooses actions from "surprisingly few choices out of the vast array." These few choices save the time and energy of the human brain when it tries to find the optimal decision.

Ruse and Wilson introduce the term, 'epigenetic rules,' which are "genetically based processes of development that predispose the individual to adopt one or a few forms of behaviors as opposed to others," and are "rooted in the physiological processes leading from the genes to thought and action" (1986, p. 180). Ruse (1995) also uses the expression, 'hardwired,' as similar to the expression, 'rooted.' They summarize that "we think morally because we are subject to appropriate epigenetic rules. These predispose us to think that certain courses of action are right and certain courses of action are wrong" (Ibid.).

To conclude this chapter: AI scientists, since (say) Marvin Minsky, have learned from how the human brain works, in order to develop AI programs. I believe that, conversely, we can now learn from the ways AI programs work, in order to understand the human brain and its moral behaviors. The history of the evolution of AI programs from Deep Blue in 1997, to Watson in 2011, and to AlphaGo in 2016 can be contrasted to the history of the evolution of the human brain for several million years. I argue that morality is social software, and the software program of morality works in the same way as the Monte Carlo tree search. I also argue that these aspects of morality have been developed through the process of evolution. In the next chapter, I will discuss 'evolutionary morality' in more detail.

Chapter 3. Evolutionary Morality, Altruism, and Cooperation

The main subject to be discussed in this chapter is my interpretation of the thesis, "Morality is evolutionary." I will make it clear what I mean by that, among a discussion of a variety of interpretations of scholars and thinkers. More specifically, I first discuss the concept, 'Natural Selection' and its units (or levels). Second, I introduce three cases that demonstrate the concept of natural selection through morality. I then discuss altruism and cooperation, which have been central to the debates about evolutionary morality. Finally, I summarize my interpretation, while discussing Jesse Prinz's 'sentimentalist theory of morality.' The whole purpose of this chapter is to set a 'playground,' where the two concepts, social software and evolutionary morality can 'play' with each other. While I touch on subtle and controversial matters, I will mainly try to summarize the standard views of the major topics in evolutionary moral philosophy.

§3-1. Natural Selection and Its Units (or Levels)

Geological records found in fossils suggest that species morphed into different species over time. The question remains as to how the process of morphing happened. That is, what is the evolutionary mechanism? One answer that the contemporary sciences have embraced is the idea of natural selection. Since Charles Darwin first[25] published his scientific description of it in his *On the Origin of Species* (1859), natural selection has been regarded as a core of the mechanism of evolution.

The process of natural selection includes three components: 1) variation, 2) differential reproduction, and 3) heredity (Lewontin 1970; Godfrey-Smith 2007, 2009; Brandon 2014). An organism obtains variations randomly, and these variations can be passed on to its offspring;

[25] Though it can be traced back to earlier (including ancient) thinkers that introduced the idea.

however, environmental conditions are also varied and not all offspring will have the same genetic variations; so, some environments are harsher for some offspring; while also being beneficial to other offspring with different variations; finally, the offspring with beneficial variations in their environment will survive and can pass on those traits to their descendants.

Darwin's own words on natural selection are also worth re-reading. Note the expression, the 'principle of preservation,' (Darwin, 1859/1872 6th ed., Ch. 4, 'Summary of Chapter', pp. 102-103)[26]:

> *But if variations useful to any organic being ever do occur, assuredly individuals thus characterised will have the best chance of being preserved in the struggle for life; and from the strong principle of inheritance, these will tend to produce offspring similarly characterised. This* ***principle of preservation*** *[emphasis is mine], or the survival of the fittest, I have called Natural Selection. It leads to the improvement of each creature in relation to its organic and inorganic conditions of life; and consequently, in most cases, to what must be regarded as an advance in organisation. Nevertheless, low and simple forms will long endure if well fitted for their simple conditions of life.*

According to Darwin's original view, the term, 'natural selection' is, most of all, about 'preservation.' Darwin wrote that, in "the literal sense of the word, no doubt, natural selection is a false term," but, like other "metaphorical expressions" such as "the attraction of gravity," if we "personify" Nature, then the term 'natural selection' can be used (1859/1872, Ch. 4, p. 63). He

[26] Six different and updated editions of *On the Origin of Species* were published while Darwin was alive: Firstly in 1859 and lastly in 1872. I agree that those differences are not just stylish editorial corrections but subtle philosophical changes of Darwin's view, so that we should be careful when interpret those different editions. Here, I quote from the last, sixth edition of 1872.

expressed his regret for the term in a letter to Charles Lyell in 1860: "Talking of "Natural Selection", if I had to commence de novo, I would have used natural preservation; for I find men like Harvey of Dublin cannot understand me; though he has read the Book twice."[27] Here, I note that the evolutionary mechanism and, as the core of it, natural selection is about 'preservation' (or lack thereof, which is "Extinction" (1859/1872, Ch. 4, p. 85). Now it becomes clearer what I mean by 'morality is evolutionary': I mean that morality is about preservation and extinction. Before moving on to morality directly, let's discuss the units (or levels) of natural selection a bit more.

Then, what is the basic level (or unit) of natural selection? What unit (or level) of an organism does Nature preserve? The two terms, 'unit' and 'level' of natural selection are often used synonymously, though some authors distinguish them (Godfrey-Smith 2009). I think that 'level' may be more useful in a hierarchical understanding of natural selection; whereas 'unit' may be more revelatory in explaining 'morality as social software.' So, here, I use the two terms mostly interchangeably, but at times, distinguishably.

We can consider various levels of natural selection, from genes to species. In a biological hierarchy, the following levels of natural selection are commonly suggested (Sinervo 1997/2013):

Genes[28] → Individuals → Kin → Groups → Species

According to, Richard Dawkins' famous theory of genic selection, 'selfish' genes, not individuals, are the units of natural selection, and higher organisms such as individuals may be regarded as just

[27] Darwin (September 28, 1860), "Darwin, C. R. to Lyell, Charles." Darwin Correspondence Project. Cambridge, UK: Cambridge University Library. Letter 2931. I first learned this fact from the Wikipedia article, "Natural Selection" https://en.wikipedia.org/wiki/Natural_selection.

[28] Lewontin (1970) distinguishes, under the individual level, among molecules, cells, and gametes.

'vehicles' to spread and preserve those genes (Dawkins, 1976/2006). The idea of individual selection seems intuitive: I am the basic unit of natural selection; if I adapt to the environment well, I will survive.

Next, an individual's kin are its relatives who are closely related to the individual (that is, who share some of the same kinds of genes with the individual). An individual helps its kin to promote the preservation of itself and its kin. An ant or a bee, for example, may sacrifice its life in order to preserve its kin under the law of kin selection. How many fellow kin it preserves can be measured by the term, 'inclusive fitness.' W. D. Hamilton (1964) who first defined the term, told a joke that he could sacrifice his own life for the lives of his four cousins who share one-quarter of the same genes with him.[29]

In the level of group selection, the members of a group, unlike kin, do not necessarily share the same genes. Nevertheless, the members interact with one another in order to win in natural selection the competition with other groups.

Species selection is the next. Homo sapiens have survived in the competition with Neanderthals who did not.

Similarly, beyond biology, I think we can imagine a hierarchy of the following levels of natural selection in the human community.

Individuals → Tribes → Countries → Cultures → Civilizations.[30]

[29] Hamilton (1964) gives, in a verbal presentation, that "inclusive fitness may be imagined as the personal fitness which an individual actually expresses in its production of adult offspring," in addition to a strict mathematical definition of "Rij=1+ Rij where Rij is called the inclusive fitness, and Rij the inclusive fitness effect." Still, how to calculate (or define) inclusive fitness is a controversial matter.

[30] I admit that it may seem 'silly' to add 'Civilizations → Planets → Galaxies → and the like' into this hierarchy. But let's recall what the late Hawking (2010) said in a science documentary series, "Into the Universe with Stephen Hawking": "If aliens visit us, the outcome would be much as when Columbus landed in America, which didn't turn out well for the Native Americans." Besides aliens, when humans will move into other planets, say, three hundred years later, the principle of natural selection between them will work, too. Or, in the past, an imaginary civilization,

For the unit of natural selection (what unit does it preserve?), kin selection (relating to genic selection) and group selection has been most discussed in the literature, and multi-level selection (Elliott Sober & David Sloan Wilson, 1998) was a new recent synthesis of some others. I will discuss this topic, the unit of natural selection, further below, relating it to altruism and cooperation. But before that, I introduce some examples.

§3-2. Natural Selection through Morality: Three Cases

The following three cases of natural selection show that morality is closely connected to natural selection, even though they also show that it is debatable whether morality is the cause of natural selection, the effect of it, or otherwise.

Firstly, Darwin's own account in *The Descent of Man* (Darwin, 1871/1874, 2nd ed., Ch. 5, pp. 129-130):

> *Turning now to the social and moral faculties. In order that primeval men --- should become social, they must have acquired the same instinctive feelings ---. They would have felt uneasy when separated from their comrades, for whom they would have felt some degree of love; they would have warned each other of danger, and have given mutual aid in attack or defence. All this implies some degree of sympathy, fidelity, and courage. Such social qualities --- were no doubt acquired by the progenitors of man in a similar manner, namely, through natural selection, aided by inherited habit. When two tribes of primeval man, living in the same*

which once flourished in a twin Earth of another galaxy might have collapsed by the evolutionary mechanism of natural selection, since the intelligent beings of that civilization believed that their God commanded them to offer those intelligent being sacrifices for their religious rituals. This kind of command functions adversely in the process of natural selection; any group that does not cherish the members cannot survive. Cultures and civilizations that have left evidence of human sacrifice in our Mother Earth have disappeared, even if the human sacrifice culture has not been the main cause of their collapses.

country, came into competition, if (other circumstances being equal) the one tribe included a great number of courageous, sympathetic and faithful members, who were always ready to warn each other of danger, to aid and defend each other, this tribe would succeed better and conquer the other. --- fidelity and courage --- disciplined soldiers --- Obedience---. Selfish and contentious people will not cohere, and without coherence nothing can be effected. A tribe rich in the above qualities would spread and be victorious over other tribes ---. Thus the social and moral qualities would tend slowly to advance and be diffused throughout the world.

It is notable that Darwin himself discussed social and moral faculties together, which may reveal that this dissertation, *Morality as Social Software*, follows the tradition of *The Descent of Man*.

Secondly, let us examine Dawkins' discussion in *The Selfish Gene* on Sucker, Cheat, and Grudger birds (1976/2006, Ch. 10). Dawkins supposes a species of bird that is parasitized by a nasty kind of tick. These ticks should be removed immediately, and an individual bird can remove most on its own body, except the top of its head. Now, the birds are divided into three categories, based on their strategies to pull off ticks from their own heads. Suckers groom anybody who needs it; Cheats accept grooming, but never return; and Grudgers only groom strangers (so, those who have never had a chance to betray Grudgers before) and individuals who have previously groomed Grudgers. Dawkins argues that, based on his simulations, the Grudger strategy is an 'evolutionarily stable strategy' against Suckers and Cheats, in the sense that Grudgers invade the population. I note that human morality can be something like the 'evolutionarily stable strategy' of the Grudgers in this example. This Grudger strategy is related to the famous 'tit-for-tat' strategy (I will discuss this topic further in chapter 5, while discussing evolutionary game theory).

Thirdly, lastly, let's review an article from the journal *Science*, "The Coevolution of Parochial Altruism and War" by Choi and Bowles (2007). The authors define altruism as "benefiting fellow group members at a cost to oneself," and parochialism as "hostility toward individuals not of one's own ethnic, racial, or other group"; and they argue, based on their computer simulation, that the combination of the two, parochial altruism "could have evolved if parochialism promoted intergroup hostilities." During the evolutionary process of natural selection, both altruism and parochialism may not survive if they are separated from each other because the payoffs of altruists and parochialists are lower than the alternatives. The computer model of Choi and Bowles simulates possible human conditions in the late Pleistocene and the early Holocene (i.e., from about 125,000 years through 10,000 years until 7000 years ago) with 10 runs of 5000 generations. Their model shows, if combined, parochial altruism can be evolutionarily more stable than others such as tolerant (non-parochial) altruists, parochial non-altruists, and tolerant (non-parochial) non-altruists.[31] I note here again that human morality is something like parochial altruism that has emerged through the evolutionary mechanism of natural selection.

§3-3. Altruism and Cooperation

Altruism and cooperation have been central topics in the discourse on evolutionary morality. Darwin himself related altruistic and cooperative behaviors to morality (1871/1874, 2nd ed., Ch.5, p. 132):

[31] A notable result of the research of Choi and Bowles is that parochial altruism can be evolutionarily more stable than tolerant (non-parochial) altruism and simple altruism. Can this result be interpreted as exhibited by a country that often encourages the people's patriotism (a kind of altruism) by starting aggressive wars against the neighboring countries can be evolutionarily more stable? I mention here that I think this is a critically important question, though I don't go further in this dissertation.

> *A tribe including many members who, from possessing in a high degree the spirit of patriotism, fidelity, obedience, courage, and sympathy, were always ready* ***to aid one another, and to sacrifice themselves for the common good*** *[emphasis is mine], would be victorious over most other tribes; and this would be natural selection. At all times throughout the world tribes have supplanted other tribes; and as morality is one important element in their success, the standard of morality and the number of well-endowed men will thus every-where tend to rise and increase.*

At the core of altruism and cooperation are "to aid one another, and to sacrifice for the common good." Perhaps, altruism and cooperation are not the same as morality of the entirety, but they are the seminal constituents of the republic (not kingdom) of morality. Other constituents may be justice, sharing, fairness, sympathy, empathy and the like.

The two topics, however, have often been the sources of misunderstandings, misuses, and even, at times, miscalculations (relating to inclusive fitness). Some terms that may seem similar but still distinguishable, or must be distinguished, are these:

1) There are many adjectives qualifying each altruism (See Okasha 2003/2013): biological (evolutionary) altruism vs. psychological (vernacular) altruism by (E. Sober & D.S. Wilson 1998); strong vs. weak by (D.S. Wilson 1990); hardcore vs. softcore by (Edward Wilson 1978; we have another Wilson in the dissertation!); reciprocal (Trivers 1971); and short-term vs. long-term altruism;

2) altruism, cooperation, mutualism, and collaboration by (Tomasello 2009); even further,

3) mutualistic cooperation vs. altruism (avoiding the expression, altruistic cooperation) by (Godfrey-Smith 2013).

I do not intend here to draw a perfect and clear-cut map of all these consanguineous terms. Rather, I introduce some concise meanings of necessary terms that are helpful for our further discussion on 'morality as social software.'

I find it useful that we begin by introducing four categories described by David Wilson and Lee Dugatkin (1992, p. 30):

1) Altruism: Decreasing the agent's fitness/ Increasing the fitness of others,
2) Cooperation: Increasing the fitness of both, the agent and others,
3) Selfishness: Increasing the agent's fitness/ Decreasing the fitness of others, and
4) Spite: Decreasing the fitness of both, the agent and others.

According to these distinctions, we can see, above all, the relationship between altruism and cooperation, which is that cooperation is 'not' quite the same as altruism. If I help others by way of cooperation, it is good for me; therefore, my actions are not altruistic. Still, many other authors hold that cooperation is a 'kind' of altruism.

The debate on whether cooperation is altruistic or not leads us naturally to one of the most serious debates in evolutionary theories: how altruism and cooperation have evolved and survived, not become extinct, while they have decreased the agent's fitness. If I live this life altruistically to help other sentient beings without having my own offspring, the altruistic genes within me are likely to become extinct. I wrote in chapter one that "an agent's morality is made up of the

standards that the agent holds regarding what is right or wrong." If we understand this agent, not as an individual, but as a group such as kin, species, tribe, country, culture, civilization, and the like, altruism and cooperation can be understood as helpful for the highest and ultimate moral standard of the group, that is, its survival of the fittest by natural selection.

The flavor of the following thesis by Christine Korsgaard is similar to that of the metaphor of Indra's Net introduced in chapter one:

> "*The primal scene of morality ... is not one in which I do something to you or you do something to me, but one in which we do something together*" (Re-quoted from Tomasello 2009, p. 51).

That is, as I wrote earlier, morality is closely related to the question of how we ought to live together with other human beings and all other sentient beings. That is why our discussion about morality is closely related to altruism and cooperation.

According to Godfrey-Smith's (2013) slightly different distinction from that of Wilson and Dugatkin (1992) above, 'cooperation' can be defined as "helping now with reasonable expectation that you will be repaid later, during your lifetime" and 'altruism' as "giving away resources that will never be repaid to you" (Godfrey-Smith 2013). That is, for 'cooperation', an agent's expected benefit is bigger than the cost; for 'altruism,' an agent's expected benefit is smaller than the cost. Here, I think his phrase, "during your lifetime," is notable. He seems to exclude any religious claim of paying (the cost) and being repaid (the benefit) through many lifetimes such as the concept of Karma. If we limit the time to only 'this' life, I am wondering how long delayed repayment can

be acceptable: I scratch your back 'now,' but if you scratch my back in return when I breathe my last breath, is it an instance of morally acceptable cooperation? Maybe not. Or, still, 'yes.' Perhaps we can develop a mathematical function of morally acceptable cooperation that has variables of time and the amount of repayment. In contrast to cooperation and altruism, the focus of 'mutualism' is on interaction that helps both sides, according to Godfrey-Smith's definition (Ibid.). The term, 'mutualism', can be used for two cases: for immediate mutual gain, and for 'short term sacrifice and longer term gain.' Therefore, using Godfrey-Smith's distinction, there are two kinds: mutualistic cooperation versus altruism. The expression, 'altruistic cooperation' is unnecessary.

§3-4. Michael Tomasello's View on Cooperation

In Tomasello's (2009, 2014, 2016) discussions on evolutionary morality, the concept of 'cooperation' encompasses 'altruism' and 'collaboration.' In his 'taxonomy,' 'altruism' is a form of cooperation, in which "one individual sacrifices for the benefit of another" (2016, p. 1); whereas '(mutualist) collaboration' is "multiple individuals working together for mutual benefit" (2009, p. 2), (which is Godfrey-Smith's 'mutualistic cooperation'). One of Tomasello's primary questions is whether "altruism emerges "naturally" in young children or whether, alternatively, it is somehow imparted by culture (or whether culture perhaps plays some other role)" (Tomasello 2009, p. 2). Naturally, this question can be traced back to the traditional debate "in Western civilization," which "is whether humans are born cooperative and helpful and society later corrupts them (e.g., Rousseau) [that is, Kant], or whether they are born selfish and unhelpful and society teaches them better (e.g., Hobbes)" (p. 3).[32] Tomasello, agreeing with the work of Brian Skyrms (2003, more on chapter five below), argues that humans do not face a Prisoner's Dilemma when building

[32] Similarly, as briefly introduced in chapter one, the traditional debate in China between Mencius and Xunzi can be compared to Kantian (Rousseauian) and Hobbesian, respectively.

human-style collaboration from the ape foundation. “Rather, our scenario is a stag hunt in which everyone prefers to collaborate because of the rewards doing so brings each of us and our compatriots” (p. 54).

Tomasello views that the two forms of cooperation, altruism and collaboration, themselves, are “known as morality,” if understood as some “uniquely human version” (2016, p. 1). Altruistic sacrifice is “based on such ***self-immolating*** [emphasis is mine] motives as compassion, concern, and benevolence”; mutualistic collaboration is “based on such impartial motives as fairness, equity, and justice” (Ibid.). Even if the equation of ‘morality = cooperation’ may seem too simplified, this simplicity is powerful in the sense that it suggests us to imagine some million years of human evolutionary process in which we humans have developed those senses of morality.

Let’s imagine: human ancestors wanted to hunt a bigger stag, instead of a smaller hare; an individual alone was not able to hunt a stag; he needed cooperation with others; once hunting was done, they needed to distribute the stag fairly; and the distribution process must have considered ‘cheaters’ (who did not participate in the hunting for his own benefit) and ‘weakers’ (who were not able to participate, due to, for example, illness or disability), as well as the most valiant warriors. This is just one example of stag hunting. The examples can be enlarged to include larger scale works such as building a city and engaging in a war against a neighboring country. Through this kind of process, it is very likely that we humans have developed the senses of morality, and at the core of it, there is cooperation of altruism and cooperation of collaboration.

§3-5. Thomas Huxley: Ethics against Evolution, or through Evolution?

Thomas Henry Huxley gave the second Romanes Lecture in 1893, and it was published as a book, *Evolution and Ethics* in 1894. In this concise and inspirational book, he discusses various topics

related to evolution and ethics in Western and Indian philosophy and science. Among them, I focus on only one thesis that morality is against evolution, which is contrary to my thesis in this dissertation.[33] While he was a supporter of Darwinian theory of evolution,[34] Huxley did not support the thesis that evolution is the major foundation of human morality. Let's call this thesis the 'evolution-morality' thesis. He argued that it is hard for us "to bring the course of evolution into harmony with even the elementary requirement of the ethical ideal of the justice and the good" (Huxley 1894/2006, p. 58). At the same time, however, some of Huxley's claims seemed to support the evolution-morality thesis according to the audience present at the lecture.

After Huxley's lecture, "a stumbling-block to many" people was a seemingly paradoxical idea that "ethical nature, while born of cosmic nature, is necessarily at enmity with its parent" (1894/2006 in "Preface"). While comparing 'ethical nature' (I understand as ethics) to 'cosmic nature' (I understand as evolution), Huxley argued that the discord between the two is not a paradox, but a "truth." Huxley's other example of an antithesis of cosmic nature is "the horticulture process" that brings the "operation of human energy and intelligence" into existence and maintenance of a garden (p. 11).

> *It is urged that, such being the case, the cosmic process cannot be in antagonism with that horticultural process which is part of itself—I can only reply [to what is urged], that if the conclusion that the two are, antagonistic is logically absurd, I am sorry for logic* (pp. 11-12).

[33] Until recently I perused the entirety of this book for the first time, I had not known that, in this book published in the United Kingdom in 1894, Huxley discussed an essence of Buddhist philosophy (Karma, nirvana, etc.). Interestingly, Huxley seems to have a view that Buddhist thought of Karma, nirvana and the like is a form of evolutionary theory.

[34] Like he famously described himself as "Darwin's Bulldog."

On the contrary, in another place in the book, Huxley's writes that:

> *I have termed this evolution of the feelings out of which the primitive bonds of human society are so largely forged, into the organized and personified sympathy we call conscience, the ethical process* (p. 30).

I think that this seemingly paradoxical or antithetical views of Huxley can be 'sublated' into a synthesis of the evolution-morality thesis, if we interpret the contrary of his two kinds of views as just a 'matter of direction.' That is, I argue that the direction of developing morality, the horticultural process, and another example of his (the process of colonization that makes new colonies in the imperial period) are all towards the direction of, say, 'order 1'; evolution is also towards the direction of, say, 'order 2'; however, the directions of these two, order 1 and 2, are different from each other, though they are still towards the direction of order. They are not opposite to each other; rather both are opposite to the direction of 'disorder.' It seems to me that Huxley thought that he connected evolution to disorder, and morality to order; but he actually connected evolution to 'order 2'; and morality (and horticulture) to 'order 1.' So, I argue that, this discrepancy between what he thought and what he actually did was the main cause of the apparent paradox to the audience. And also, at the core of this paradox is the long-rooted nature-nurture debate, as mentioned in the previous section. The main question in the debate is which one is the major cause, among the two: on the one hand, cosmic nature, evolution, etc.; on the other hand, ethical nature (which is Huxley's own term above), nurture, culture, and the like.

When I argue the evolution-morality thesis that evolution is the major foundation of human morality, I assume that both, evolution and morality, are connected to 'order,' unlike what Huxley

'thought,' but rather what he 'did' actually. I will address this order-disorder issue in more detail in the next chapter on entropy, after discussing one of the strongest critiques of evolutionary moral theories by Jesse Prinz.

§3-6. Jesse Prinz's Sentimentalist Critique of Evolutionary Moral Theory

Jesse Prinz in *The Emotional Construction of Morals* (2007) defends a "sentimentalist theory of morality." Prinz's beginning position is not different from one assumption of this dissertation. Prinz argues that "[p]eople who feel uncomfortable with the idea that morality derives from us, should consider some other things that derive from us, such as medicine, governments, and art" (p. 8). This is the exact position with which I began in this dissertation. Unlike my claim, however, Prinz argues that "many philosophers want to find an objective foundation for morality. I don't think such a foundation exists" (p. 164). In this sense, he is not a Kantian who finds the categorical (therefore, unconditional and object) imperative from human reason. In Prinz's sharp contrast between Kantians and Humeans, "[i]f moral imperatives are categorical, sensibility theories are not going to fly" (p. 128). I think the difference between the Kantian and the Humean is 'just' the difference between the foci of the two views in the long tradition of Western ideas since ancient Greece: between logos and pathos; between reason and emotion. The Humean in the tradition of pathos and emotion have also attempted to find a foundation for morality, though not on objective.

Prinz argues that "[a]s a constructive sentimentalist, I think morality is created by us, and, as a relativist, I think different societies create different moralities under different historical conditions" (p. 215). I definitely agree with these views. However, I have a different view from what he states in the following three quotes (Prinz 2007):

[T]here is no use in distinguishing natural and cultural norms, because culture shapes all norms, even when they have a natural foundation.—[N]o evolved norms qualify as moral norms, so, strictly speaking, there is no such thing as an evolutionary ethics (p. 259).

I will concede that we are biologically prone to have certain kinds of values, but I will deny that there is an innate morality --. Our biological predispositions have no authority over values that have a cultural origin, and they can be embellished and overturned under the influence of culture. Moreover, I will argue that our biological predispositions do not qualify as moral rules without cultural elaboration. Morality is artificial all the way down. Taken literally, "evolutionary ethics" is a myth. It is Romanticism reborn as crass scientism, no more plausible than Nietzsche's Übermensch and perhaps no less insidious (pp. 245-246).

If I am right, then moral rules are not innate. Rather, they emerge through interactions between biology and culture (p. 274).

The 'intensity' of Prinz's criticism of the role of biology (and so, evolution) seems to decrease from the first to the third: From 'no evolutionary ethics' through 'some biologically prone values' (though still the evolutionary idea is 'myth'), finally, to 'interaction between biology and culture.' These three quotes show that we can add Prinz into the list of thinkers, like Tomasello and Huxley above, who address the nature-nurture debate.

Do we humans have such moral genes that result in moral universals? Prinz, in "Against Moral Nativism" (written in 2004, first published in 2009,) argues that there is no such biological

morality-generating machinery. If moral nativism is right, and so all human species have the same innate biological machinery to generate moral universals, on Prinz's argument, then we must be able to find moral universals such as "Don't harm innocent people" in (almost) all communities and cultures. But, can we find that? No. Prinz divides the views of the defenders of moral universals into three groups, the *immodest*, the *modest*, and the *minimal*, and then gives arguments against the three, one by one.

Here, Prinz and I agree that morality is constructed by humans, but we don't agree on whether the constructing process is evolutionary or not, and whether the process can be called the foundation. It seems to me that Prinz uses the terms "evolutionary" and "cultural" distinctively, and the term "foundation" strictly. My position is that if human morality is a consequence of human evolutionary and cultural development, we may call it "evolutionary," and we may also call the 'emerging process of human morality' a "foundation." The 'emerging process of human morality' in my terminology may be compared to Prinz's terminology, the process of 'moralization' (2004/2009, p. 16). In other words, in Prinz's thesis that "universal moral rules are the result of convergent cultural evolution" (ibid.), I am specifically interested in what exactly the 'convergent cultural evolution' is. My current hypothesis for my dissertation research is that the 'convergent cultural evolution' is epistemic game-theoretic towards morality as social software.

In the following chapters, we will investigate further whether these differences between Prinz's and mine are just nominal (ones between only the usages) or actual.

Chapter 4. Anti-entropic Morality, Information, and Equilibrium

The word, 'entropy,' to some people, often seems mysterious and difficult to grasp; to other people, the word is doomed to misuse. I think the following anecdote is not only amusing but also revelatory. In the 1940s when Claude Shannon (a founder of modern information theory) asked John von Neumann (a founder of modern game theory) what Shannon's newly discovered formula should be called, von Neumann cleverly replied that:

> *You should call it entropy, for two reasons: In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage* (McIrvine and Tribus 1971; first appeared in Tribus 1964, p. 354).

Von Neumann's clever second reason shows how the word entropy is regarded as mysterious or doomed. (The first reason is also very relevant to a section below.)[35] On the stage of this chapter, entropy plays the central role, along with its siblings such as information, equilibrium, software, and morality playing the supporting roles, together with other 'step-siblings' such as disorder and order. Through the discussions, I hope, "neither of them is shrouded in mystery or rapture beyond our horizon."[36]

[35] The anecdote may seem 'apocryphal.' Regardless of whether the anecdote is true or not, apocryphal stories often keep only the universal and necessary, stripping away the particular and contingent.

[36] With this sentence, I am mimicking Kant's 'accent' adjacent to his epitaph ("the starry heaven above me, the moral law within me") originally written in *Critique of Practical Reason* (1788, in 'Conclusion').

§4-1. A Summary of the Arguments

The single main thesis that I argue in this chapter is as follows:

(1) Main Thesis: Morality as social software is anti-entropic, that is, against entropy.

This main thesis has three connections here:

(a) Foundation (or Background): the main thesis is founded on several well-known, but not always well-recognized, theses (2) to (7) below, which can function as the scientific background for the philosophical thesis;
(b) Argument: the well-known main thesis is argued by several, relatively less-known theses (8) to (13) below, which can function as a series of syllogisms; lastly,
(c) Corollary: the main thesis results in two corollaries.

The (a) foundation part is, 'roughly put,' as follows:

(2) The second law of thermodynamics holds that the total entropy of a system tends to increase over time.

(3) The concept of entropy often translates as 'disorder' in ordinary language. (Whether this "simile" is legitimate is a crucial point in the bitter controversy about entropy.)

(4) So, the second law may be interpreted as stating that the disorder of a system tends to increase over time. (If the simile in (3) above is not legitimate, then neither is this interpretation in (4).)

(5) The inexorable tendency of entropy governs everything in the universe.

(6) Or, for some moment of time, there are exceptions, which are living organisms.

(7) From this ontology of nature, a normative rule can be deduced: organisms such as amoebae and my life do, and should do, resist the law of entropy in order to live lives.

The (b) argument for the main thesis (that morality as social software is anti-entropic) is as follows:

(8) Premise 1: Entropy and information are sibling concepts.

(9) Premise 2: Entropy and equilibrium are also sibling concepts.

(10) Conclusion IE (also, Premise 3) ('I' for 'information'; 'E' for 'equilibrium'): Therefore, information and equilibrium are also sibling concepts.[37]

(11) Premise 4: Information and software are sibling concepts.

(12) Premise 5: Equilibrium and morality are also sibling concepts.

(13) Conclusion MS (also, Premise 6) ('M' for 'morality'; 'S' for 'software'): Therefore, morality and (social) software are sibling concepts, that is, morality as social software.

(14 = 1) Main Thesis: Therefore, morality as social software is anti-entropic, that is, against entropy.

Lastly, from this main thesis, the following (c) corollaries result:

(15) Corollary 1: Morality as social software is a tool for us to fight against the tendency of entropy in the universe.

[37] It is because my siblings are in sibling relationship with each other, if they are not 'numerically' identical to each other. If they are, that means that, in our examples here, information and equilibrium are numerically identical to each other. I do not assume that they are identical. "When we say that two people own the same car, we might mean either they own the same car, e.g., a Honda, or they are joint owners of a single car. ---the former--- means qualitative identity ---, the latter means quantitative or numerical identity" (Fitting and Mendelsohn 1998, p. 140).

(16) Corollary 2: The inevitable clash between the entropic universe and the anti-entropic fighting of organisms against the tendency can be, fundamentally, seen as the origin of the suffering (*Dukkha*, in Pali of Buddhism) of the organisms.

None of these ideas and arguments is uncontroversial. Even thesis (2), the second law of thermodynamics, which seems well-established, is also challenged by some philosophers and scientists. Some people feel enlightened when they first encounter these ideas; on the contrary, some other people feel uncomfortable with incautious usages of terms such as 'disorder' in (3) and (4) above; still, some other people criticize vehemently these ideas for inappropriate applications of concepts in natural science to non-natural science and everyday life.

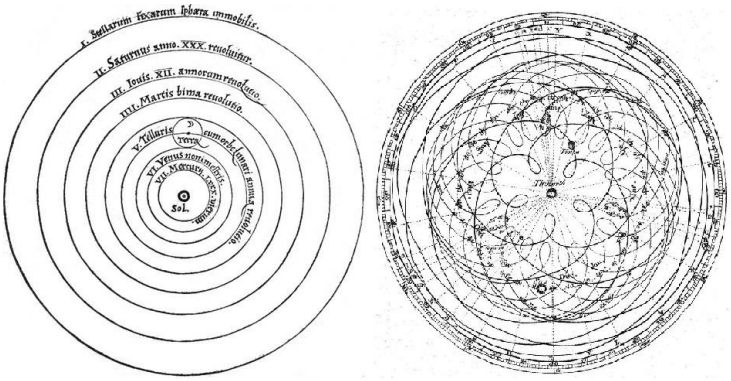
Nevertheless, or therefore, I will not pay equal attention to each of those theses. First, for (a) foundation, since the theses are already well-known, I will briefly 'sketch' them by introducing the major controversies over the pros and cons, and revealing my perspectives. Second, for (b) argument, the conclusion itself that morality is anti-entropic is commonplace in the literature.[38] My emphasis will not be on the conclusion itself, but on the way to draw the conclusion, which is related to the concept of social software. Third, for (c) corollary, I will 'venture' the two corollaries with caution and imagination. While admitting that there may seem to be sheer absurdity and ample lacunas in these arguments, I dare to attempt to draw attention to the novelty of them. I shall now begin in more detail, by discussing the concept of entropy.

[38] For example, Hammond (1985/2005 Eulogy edition) *The Human System from Entropy to Ethics*; Pojman (1990/2017 with Fieser), p. 228. Pojman and Fieser writes: "According to [Geoffrey] Warnock, society has a natural tendency to get worse, an *entropy of social relations* [emphasis is theirs] . --- Morality is antientropic." The authors ascribe the concept of anti-entropic morality to the philosopher, Geoffrey Warnock, but I cannot find any mention of entropy in Warnock's books.

§4-2. Is Entropy Disorder?

The simpler conceptions such as explanation, theory, and principle attract 'beautiful minds' more than the less simple conceptions. Philosophers, mathematicians, scientists, psychologists, economists, and others are more inclined to love the principle expressed in Ockham's (Occam's) Razor: Plurality should not be posited without necessity; often, in somewhat different form, entities are not to be multiplied beyond necessity. This principle of the Scholastic philosopher William of Ockham (1285-1347/49) is a law of parsimony (economy). A modern, common expression of the principle is, I endorse, that, 'the simpler, the better,' which is often found in even commercials these days.[39]

[39] Regarding the beauty of simplicity, I follow, for example, in the history and philosophy of science, the common Kuhnian (*The Structure of Scientific Revolutions* 1962/Second enlarged edition 1970) explanation of why the heliocentric model won rapidly more proponents in the competition with the geocentric model during the period of the Scientific Revolution. It was not just because heliocentrism was considered as more accurate than geocentrism. Geocentrism itself had been able to explain the discrepancies between the Ptolemy theory and the observation, as almost accurately as heliocentrism had. The discrepancies had been accumulated for around 1500 years since Ptolemy. Geocentrism, however, had had to adopt more and more complicated 'epicycles' in order to explain the discrepancies. Geocentrism had had to adopt even epicycles of epicycles when an epicycle had not been able to explain successfully the discrepancies any more that had been used to be explained successfully before. These epicycles were like ad-hoc hypotheses that make a theory look more complex and so, uglier. Heliocentrism did not need these unnecessary epicycles, which accords with Ockham's razor, so that heliocentrism looked simpler and so, more beautiful to the proponents who converted from geocentrism. The following two (still simplified) diagrams can show vividly the contrast between the helio-simplicity (left) with the geo-complexity (right). The helio-one from Copernicus' *On the Revolutions of Heavenly Spheres* (1543); the geo-one is from *Encyclopaedia Britannica*, 1st Edition (1771), by James Ferguson (1710-1776), based on similar diagrams by Giovanni Cassini (1625-1712) and Dr. Roger Long (1680-1770). See also my discussion on beauty in section 7-4.



Heliocentrism Vs. Geocentrism

Figure2: Heliocentrism Vs. Geocentrism
Source: Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Cassini_apparent.jpg

The beauty of any discourse based on the second law of thermodynamics results from the simplicity of the law. The law holds that:

The total entropy of a system tends to increase over time.

More roughly speaking,

Heat flows from a hotter to a colder body.

This proposition is now called the Clausius Statement (1854/1867 English translation, p. 117; in a modern quote, Penrose 2004/2006, p. 689).

Most roughly speaking, the common proverb:

It does no good to cry over spilt milk (Rifkin and Howard 1980, p. 33).

It is claimed that this simple law governs everything in the universe from the beginning to the end of the universe. When I first met this law, perhaps, in a science class in my high school, I was not so amazed by it. It seemed to me to be just one of many insipid laws in science classes that governs only inanimate physical objects such as gas, atoms, and molecules. A decade later when I learned the arcane simplicity that governs everything in the universe, I felt more enlightened.

Most of us are familiar with the concepts such as energy, volume, amount of substance, even if we are the victims of the gulf between the "Two Cultures" (of C. P. Snow). For a pure

homogeneous substance, such as gas, we can conceive its internal energy U, volume V, and the amount of substance n. Likewise, we can treat its entropy S as a function of these parameters (Hołyst and Poniewierski 2012, pp. 57-8).

Though the concept of entropy itself is a complicated mathematical one in thermodynamics, entropy often translates as 'disorder' (the degree of dis-order-ness) in ordinary language (the proposition above (3)). So, the second law may be interpreted as stating that the disorder of a system tends to increase over time (the proposition above (4)). This kind of translation is a "simile," and the legitimacy of a simile depends on how close the similarity of the two things are to each other. A simile, "she runs like a deer," is not accurate if she cannot run gracefully. Since the inception of the concept of entropy in the mid-19th century, the simile of disorder has been accepted widely and used in diverse areas including art critique and discourses on civilization as well as science and philosophy. However, recently some scholars have argued that this simile is inaccurate, or at least needs to be supplemented with some other concepts.

Most authors including Roger Penrose (1989, p. 309) use the English word 'disorder' for entropy, whereas Penrose later (2004/2006, p. 690) changes it into a different word 'randomness.' It seems to me that Penrose's new translation from the mathematical concept to the word 'randomness' in ordinary language is significant, in that the word 'randomness' seems value-neutral, whereas the word 'disorder' does not seem so. One odd thing that I felt when I studied the second law again in a college science class, was that, while we generally regard order as better and more beautiful than disorder, why the universe is going towards disorder, following the second law. The thesis that the universe where I live is doomed to 'bad' disorder made me gloomy. With Penrose's new translation, 'randomness,' we may avoid this kind of value-related interpretation of the second law. The universe that is going towards randomness seems innocent.

The chemist Frank Lambert, by contrast, "decries the use of "disorder" in teaching beginning students about thermodynamic entropy," arguing that "entropy is not disorder, --- not a measure of disorder or chaos, --- not a driving force" (2002, p. 187). On Lambert's claim, the simile to disorder is a "broken crutch" (ibid.): that is, the simile was first used as a visual tool to explain the concept of entropy better by the very fathers, Boltzmann, Helmholtz, Gibbs, and others, and by the teachers in physics and chemistry;

> *But over the years, popular authors have learned that scientists talked about entropy in terms of disorder, and thereby entropy has become a "code word" [emphasis is mine] for the "scientific" interpretation of everything disorderly from drunken parties to dysfunctional personal relationships, and even the decline of society* (Lambert 2002, p. 187).

The physicist Daniel Styer, still, argues that the simile "entropy as disorder" is "inadequate," and another simile, "entropy as freedom," is equally problematic, but "if both are used cautiously and not too literally, then the combination provides considerable insight" into the current subject of understanding entropy with common qualitative concepts (2000, p. 1090).

The physicist-chemist Arieh Ben-Naim also argues that "the identification of entropy with disorder is totally unfounded" (2017, p. 26).[40] A tactic adopted by Ben-Naim and those who criticize the "direct" proportional correlation between entropy and disorder is to show some cases that has the "inverse" proportional correlation between the two: That is, to show a case in which its entropy increases, but its disorder decreases.

[40] https://arxiv.org/abs/1705.02461. Ben-Naim denies even the correlation between entropy and the second law of thermodynamics, which will be briefly discussed below.

I see that Lambert has accumulated the list of "The 36 Science Textbooks That Have Deleted "disorder" From Their Description of the Nature of Entropy."[41] "Deleting disorder" seems to me to be like a religious conversion, or the conversion from geo-centrism to helio-centrism. However, has the conversion been done, as of the year of 2018, or will it be done completely in the near future? My answers are negative. For around 150 years, scholars has connected entropy to disorder, and recently the movement of disconnecting entropy from disorder has started. The simile to disorder has a power to help us understand the concept of entropy very vividly, though not 100 percent accurate. I think the simile, "entropy as disorder" has much more advantages than disadvantages: then, it is more advantages to use the simile to disorder.

Regardless of disorder, randomness, or dispersion, the entropy of a system tends to increase. One of the prominent 'arrows of time,' (which means one-way directional and asymmetrical (Arthur Eddington 1928)), is the thermodynamic arrow by the second law. Events that follow the arrow (that is, the flow) of time on the second law is irreversible. Let's see some common examples of this interpretation. The degree of the disorder (or randomness, or dispersion) of my room increases due to the multiplication of books, copies of papers and articles over time, while I write this dissertation, unless I perform some action to make the room orderly (which means that I perform some action to reduce the entropy). My own body will be dismantled in the long run, so that the degree of the disorder of my body will increase.

What about the scrambled eggs for the omelet in a modern classical example of Leonard Savage's statistics (1954/1972 Second revised edition, pp. 13-4)? The disorder of the eggs increases, once broken and scrambled, which follows the law of entropy. Because of the arrow of

[41] http://entropysite.oxy.edu/

time, if we try to reverse the scrambled eggs, that is, to unscramble them, by making a machine, it is almost impossible. Here, the psychologist Richard Gregory suggests a stunning, original idea. "Chickens can unscramble omelettes: by eating them!" (1981, p. 137), and then they will lay eggs. I completely agree with Daniel Dennett that "Gregory dramatizes this [the law of entropy] with an unforgettable example" (Dennett 1995, p. 70). (This example of scrambled eggs and chickens is revelatory for our discussion on the relationship between entropy and evolution later below.) Let's look at one final vivid example of the irreversibility. Penrose asks us to imagine a glass of water poised on the edge of a table, in order to exemplify the "inexorable increase of entropy" (1989, pp. 304-5). Let me replace the water with red wine, for a better visual effect. Once nudged, fallen, shattered, and splashed, we cannot practically reverse those events, in order to re-make the initial glass of red wine. This is the simple, universal, and cold-hearted law of entropy in the universe.

On the history of this idea, the French engineer Nicolas Carnot (1824) first introduced the idea that is now known as the second law of thermodynamics. Carnot also had a thought (though did not publish) on the first law of thermodynamics (Penrose 2004/2006, p. 692). The first law holds that the total energy of any isolated system is conserved. The form of energy may be changed, for example, from kinetic to thermal (the kinetic energy of a hammer is transformed into the thermal energy of a hot nailhead), but no new energy is created or destroyed. "Whereas the first law is an equality, the second law is an inequality" (Penrose, Ibid.). It seems to me that one reason why the first law has been understood and remembered by more people (who learned the two laws in their schools) is that the concepts in the first, energy and conservation (that is, unchanged), are easier than the concepts in the second, entropy and increase (that is, change). Everybody remembers what energy is; not everybody remember entropy. In addition, regarding Penrose's

novel summary above, I conjecture that people feel more comfortable with equality than inequality, in the deeper level of the human brain.

Whereas Carnot did not enunciate the two laws clearly, it was Rudolf Clausius (1850) who did the enunciation and also introduced the new term, entropy (Clausius 1865). Whereas Clausius did not make the definition of entropy clear, it was the Austrian physicist Ludwig Boltzmann (1877) who did that work. Here is the famous formula of Boltzmann entropy S:

S = k log W, where W is the volume, and k is Boltzmann's constant.[42]

In Europe, in addition to those scientists mentioned just before, many giants including James Maxwell and Lord Kelvin made great contributions to the development of the theory of energy, entropy, and heat. In the US, it was Josiah Willard Gibbs (1839 – 1903) who made a significant contribution similar to those of the European scientists.

§4-3. Entropy versus Information

Entropy and information can be seen as sibling concepts. Let us begin with Norbert Wiener who is the father of cybernetics (and the American prodigy). Wiener is a strong proponent of the interpretation of entropy as disorder and the sibling relationship with information in his two books, *Cybernetics: or Control and Communication in the Animal and the Machine* (1948/1961 Revision) and the more popular version of Cybernetics for lay persons, *The Human Use of Human Beings* (1950/1954).

[42] k= 1.3806504 x 10^{-23} J K^{-1}

According to Wiener,

> *Information is a name for the content of what is exchanged with the outer world as we adjust to it, and make our adjustment felt upon it* (1950/1954, p. 17).

> *Just as entropy is a measure of disorganization, the information carried by a set of messages is a measure of organization. In fact, it is possible to interpret the information carried by a message as essentially the negative of its entropy, and the negative logarithm of its probability* (Ibid. p. 21).

On Wiener's view, organisms in the universe are the only exceptions to the law of entropy. "We ourselves constitute such an island [life] of decreasing entropy, and -- we live among other such islands" (p.40). Well, Wiener's island metaphor seems less persuasive than others such as the Indra's net metaphor. It seems to me that all lives in the universe are not separate islands but may be interconnected to each other just as the Indra's net. In spite of this disagreement, Wiener's idea of decreasing entropy is fascinating. According to Wiener, organisms such as an amoeba and myself do, and should, resist the law of entropy in order to live their lives, while all things that are not organisms follow the law. So, in my interpretation of Wiener's view, a meaning of life is "fighting nature's tendency to degrade the organized and to destroy the meaningful" (p. 17).

§4-4. Maxwell's Demon

Most authors summon Maxwell's Demon when they discuss the close relationship between entropy and information. I follow that general tradition without hesitation. In a thought experiment,

James Clerk Maxwell suggests an imaginary being that is to sort molecules into two categories, hotter (or swifter) ones and colder (or slower) ones. The chief end of the being is, in Maxwell's own writing, to "show that the 2nd Law of Thermodynamics has only a statistical certainty," (requoted from Leff and Rexx 1989/2002 Second edition, p. 5) that is, to show that the law has some limitations.

In the Maxwell's setting, there are two chambers A and B joined by a trapdoor; the chambers are filled with gas at equal temperature; a demon can open or shut the trapdoor rapidly, while watching swift or slow gas molecules that approach the trapdoor. The demon allows "only the swifter molecules to pass from A to B, and only the slower ones to pass from B to A. He will thus, without expenditure of work, raise the temperature of B and lower that of A, in contradiction to the second law of thermodynamics" (Maxwell 1871, requoted from Leff and Rexx 1989/2002, p. 4). A schematic figure follows:

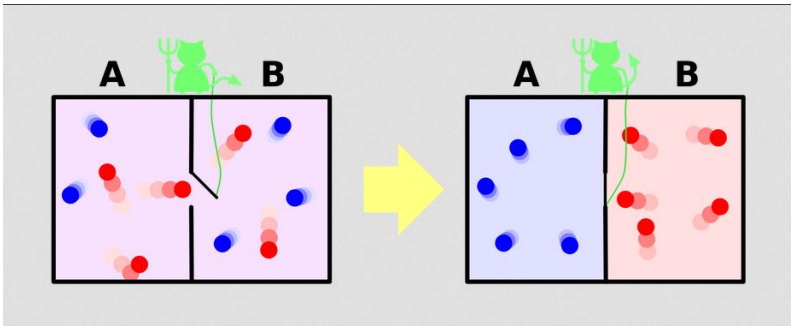


Figure 3: A Schematic figure of Maxwell's Demon (The molecules that have longer trails are hotter ones.) Source: Wikimedia Commons, Creator: User name, Htkym, https://commons.wikimedia.org/w/index.php?curid=1625737

Here, a question raised by this story is: Is this kind of 'molecule-sorting' Demon possible? If yes, one chamber becomes hotter; the other becomes colder, which violates the second law.

First of all, a purely mechanical Demon composed of, for example, ratchets, teeth, wheel, shaft and so on, may not be able to violate the second law, even with entirely frictionless condition (Ibid. P. 154). Second of all, what is more interesting is a 'cognitive' intelligent Demon, who can observe the velocities of the approaching molecules and predict which one is fast or slow. The question given this intelligent demon is:

> *Can the Demon gain information of the molecules, especially their approach velocities, without using as much energy as it gains by sorting the molecules with his trapdoor?* (Ibid. P. 155)

Richard Gregory's specific discussion on this is as follows: (See Gregory 1981, pp. 153-156).

If yes, the intelligent observer Demon can violate the second law. No free lunch! Energy is needed in order to get information. What is crucial here is: is it possible that what the intelligent Demon gets is more than what it loses? If we consider only a closed limited system such as the current setting of the two chambers and the trapdoor, the second law cannot be violated. It is because, for example, the Demon is inside the chamber, so that it has the same temperature as the gas, and therefore cannot receive energy from the molecules. On the contrary, if we consider that the intelligent observer Demon is allowed to receive energy from outside the chambers, it may seem that the Demon violates the second law. It is because it may seem that it creates more than it uses. (This is exactly what is happening in biology, organisms (life-forms), evolution and the like; and this is why it might seem that the law of evolution violates the law of entropy, therefore the law of entropy is false. I will argue below that the law of entropy is not false and discuss further following the discussions on information and equilibrium.)

It was in 1867 that this 'being' first appeared in a letter Maxwell wrote; it was in 1871 that this 'being' was publicly born to the world in a book by Maxwell, *Theory of Heat*; and it was in 1874 that this 'being' was first called 'Demon,' by William Thomson (Lord Kelvin) (Leff and Rex 1989/2002, p. 370).

§4-5. Information versus Software, Entropy versus Equilibrium, and Equilibrium versus Morality

For now, it is evident that there are close relationships between the concepts of information and software, and between the concepts of entropy and equilibrium. The third one between equilibrium and morality is what I address here further.

The morality of a society is a state of dynamic equilibrium that the members of the society have reached. It is in equilibrium because the members have reached some state; and it is dynamic, not static, because the equilibrium is changing, constantly, say, differentially as in calculus of mathematics. I think that many great sages and philosophers have recognized and preached the same point: morality as equilibrium. Aristotle, Confucius, and the Buddha are prominent in the list of those thinkers who support the middle way.

According to the Aristotelian doctrine of the Golden Mean, a moral agent who wants to achieve eudaimonia (happiness) as the ultimate and highest good, should then follow the virtuous midway point between the two extremes (too much and too little), each of which is a vice.[43] The Confucian middle way is argued in *The Doctrine of the Mean* (Zisi 481-402 BCE, the grandson of Confucius) and in some part of *the Analects* (Confucius). It is recorded in Buddhist scriptures that

[43] Aristotle's own example: generosity is the mean between profligacy (wastefulness) and meanness (stinginess-selfish); courage between foolhardiness (rashness) and cowardice; self-respect between vanity and self-abasement; modesty between shyness (bashfulness) and shamelessness.

the Buddha achieved his enlightenment after trying and abandoning some extreme ways, and then following the middle way. Of course, those typical criticisms of Aristotelian doctrine of the Golden Mean can commonly be applied to all these three moral theories. For example, if everybody follows the 'safe' middle way, who would dare to attempt to pioneer the front lines in order to contribute to the progress of history? And also, who decides the so-called, 'middle way?' While admitting that these criticisms are legitimate, I argue that morality as dynamic equilibrium is also valid. The state of morality is like, to use an old-fashioned metaphor, the state of the trembling needle of an analogue scale.

§4-6. A Bold Conjecture: From Cosmology through Axiology to a Meaning of Life

I conclude this chapter with a bold conjecture, relying on Karl Popper's thesis that scientific knowledge grows through the process of raising bold, sometimes seemingly stupid, conjectures and attempting to falsify those conjectures. So, please kill my conjecture, instead of me, if you would.

On the contemporary fashionable cosmology in physics and astronomy, the universe began at the moment of the Big Bang, and has expanded, since then. I don't know whether this theory is true, or some other similar kind such as the theory of multiverse (multiple universes) is true, or not. Still, I think this idea of 'expansion' of the universe is fascinating, and the following picture made by the NASA may show this expansion vividly:

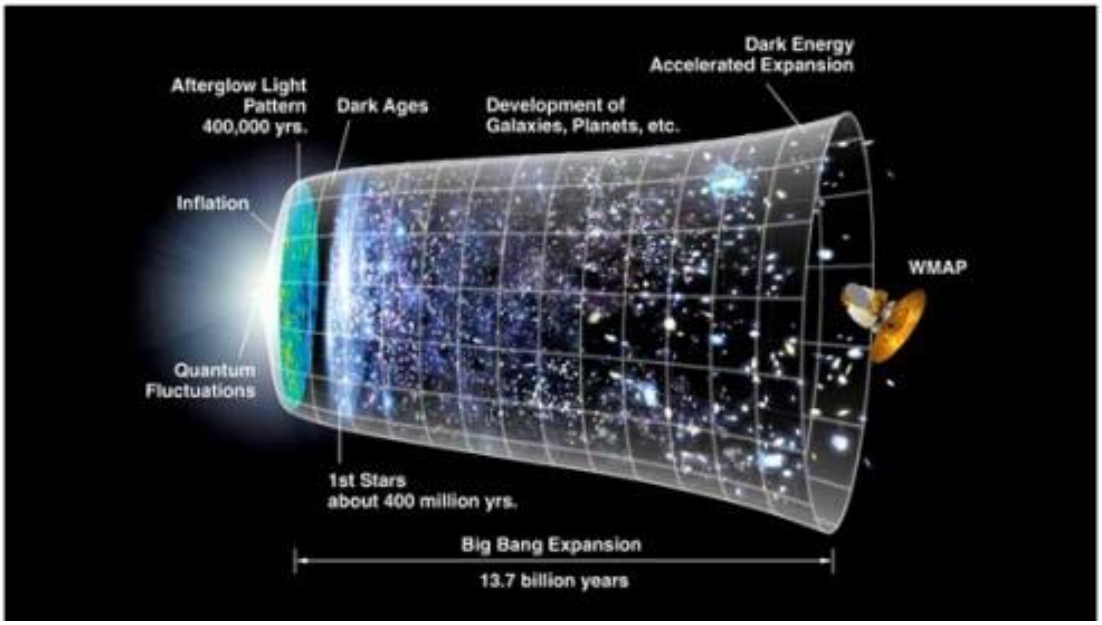


Figure 4: The Big Bang and the Expansion of the Universe

Source: The NASA, https://www.jpl.nasa.gov/infographics/infographic.view.php?id=10824

If we combine the image of this picture with the second law of thermodynamics, we can have a view of the universe that: while the universe has expanded for the last 13.7 billion years since the Big Bang, the entropy of the universe has increased, and will increase to the point of the 'heat death' that is the ultimate end of the universe. When I heard this 'story' for the first time, I was so sad because it is a form of death; later, when I learned that the heat death could come some 'short' time later in the future, say some billion years later, I felt relieved. The age of the Big Bang and the heat death are interesting, but it is nothing to my life, here and now.

So, let's focus on my life here and now. Even though the heat death is far away, the entropy law is still working now, like a headwind. Let's imagine the tube-like image in the picture above as the current of a river. The river has flowed from the (left) Big Bang to the (right) heat death, and will continue to do so, and I am a sentient being like a salmon who has to swim against the current, desperately. A salmon's return trip from the ocean to its home pond in a deep mountain

where it was hatched and grew up, is never idyllic but a struggle, and its final moment of desperate mating is not so romantic, but sublime. Many salmon in a group on the same trip sacrifice their bodies as food for mountain bears, foxes, and crows who have the rare chance to take nutrition from the oceans. Like the salmon, life forms (living organisms) ought to resist the current to generate order from disorder (or randomness, dispersion).[44]

I have argued that human morality as the "order of social software" is a form of resistance to cosmic disorder. Here is a role of axiology including ethics (more discussion of value in general follow below chapter 7). There are many other forms of this resistance: language, culture, religion, and the like. Among them, I would argue that the role of morality is never minimal; human morality itself makes humans human.

In this way, moving from cosmology through axiology, now we can go to the topic of the meaning of life. I never wanted to address this big topic in this small section, without caution. I will touch only one point very briefly in the current context. A rephrase of the expression, 'the meaning of life,' can be 'the purpose of life.' On a view of the entropic cosmology and axiology, a living organism's purpose for its life can be understood as 'struggle for entropy' or 'struggle for existence.' Here, the key concept, I argue, is 'struggle.'

Though philosophy is not history, philosophical investigations often become more fruitful when they include historical investigations. We may call this approach 'genealogical.'[45] I first encountered this idea of struggle when reading Norbert Wiener's (1950, Revision 1954) *The*

[44] Arieh Ben-Naim claims, in his numerous critical works, that this kind of discourse on entropy, life, and the universe is unfounded, because, among other reasons, 1) entropy is not related to the second law of thermodynamics; 2) we don't know what (the definition of) life is; and 3) we don't know how to measure (the entropy) of the universe (since we don't know what the universe is). Well, I admit that his criticism has enough merit to require stricter criteria; nevertheless, I believe we may discuss these concepts of life and the universe. Even if we may not agree on the definition of life, we can still talk about life. One recent book of Ben-Naim on this topic is *Information, Entropy, Life and the Universe: What We Know and What We Do Not Know* (2015).

[45] As an example of genealogical approach, see Prinz, *The Moral Self* (In Production) and a presentation of it at https://www.youtube.com/watch?v=-VEFD4_00MI.

Human Use of Human Beings: Cybernetics and Society. Let's recall what I discuss earlier in this chapter. Without revealing any explicit citation, Wiener wrote that "fighting nature's tendency to degrade the organized and to destroy the meaningful" (p. 17). Before and after Wiener, there are many who express this idea of struggle. Going back through Erwin Schrödinger (1944 *What is Life?),* Alfred North Whitehead (1929 *The Function of Reason*), and Henry Bergson (1907 *Creative Evolution*), we finally meet Ludwig Boltzmann, the very scientist who founded the concept of entropy further (after Clausius first introduced the term entropy).

Boltzmann wrote (1886):

> *"The general struggle for existence of animate beings is not a struggle for raw materials – these, for organisms, are air, water and soil, all abundantly available – nor for energy, which exists in plenty in anybody in the form of heat Q, but of a struggle for entropy S, which becomes available through the transition of energy from the hot sun to the cold earth."*

I interpret Boltzmann's claim as a claim on the purpose of life: Resisting the tendency of entropy in the universe and struggle for entropy. The person who discovered this very enlightening idea of "Struggle for entropy (or disorder)," ironically, committed suicide while suffering from other scholars' criticisms of his works.[46] Let's pay homage to Boltzmann's first revelation of this idea.[47] [48]

[46] Karl Popper, in his autography *Unended Quest* (1982), expresses his view on the tragedy of Boltzmann's suicide.

[47] I think it is an interesting and important question who argues first for this striking interpretation of life, relating to the law of entropy. Let's recall, in the history and philosophy of science, for example, the famous controversy between the Newton disciples and the Leibnitz disciples over who invented calculus first. Such questions and instances of simultaneous discoveries in science are common. From the perspective of the thesis of morality as social software, a society that does not commemorate the first discoverers and inventors, is not 'evolutionarily' competitive against other societies that do commemorate the first.

[48] Then, who is before Boltzmann? Perhaps, Heraclitus is the first, on record, arguing that "strife is justice."

This intrinsic discord between the tendency towards disorder and the struggle for existence (and entropy) of living organisms, I argue, is recognized by many in various ways. Albert Camus in *The Myth of Sisyphus* (1942) describes the intrinsic "clash" between what the universe is and what we ought to do, as "absurdity." I argue that this struggle to resist the cosmic disorder can be viewed as the origin of suffering (*Dukkha*, in Pali of Buddhism), one of the three marks of existence in Buddhism. It may be easy to swim with the current, but not easy to swim against the current. Recall the struggle of the salmon when they encounter water falls in their trips back to home. Finally, morality as social software is what helps us struggle for entropy.[49]

[49] For the record, I would like to add here a brief historical note of how I have developed the ideas discussed in this chapter. It is for the purpose of helping the reader who might be doubtful about the genuine development of these ideas, as well as the reader who might be interested in the history. It was in a chapter of my master thesis in philosophy (Kim 1998, in Korean) entitled "Cyberspace and Karl Popper's World 3," where I first discussed a very nascent form of the ideas of entropy and information (without any connection to equilibrium, evolution, nor morality). While I was writing the thesis, I was 'consciously' aware of only Norbert Wiener's two books (1948 and 1950) on cybernetics, and my discussions owed much to Wiener's great insight. Perhaps, 'unconsciously,' I knew Jeremy Refkin's ideas discussed in his *Entropy* (1980), since I, as a member of an undergraduate physics club, had organized a small conference in 1986 (the entire audience were around 10 fellow students, I remember), and Refkin's book *Entropy* was the topic of the talk of a fellow sophomore speaker, Mr. Kum. However, I had forgotten about Refkin's book and the student conference until my friend Kum reminded me of them when I gave him a hardcopy of my master thesis. (I don't know how much my sub-consciousness remembered it.) In addition, at that time in 1998 I owned the French philosopher Henri Bergson's book *Creative Evolution* (1907), the physicist Erwin Schrödinger's book *What is Life?* (1944) and the French biochemist Jacques Monod's book *Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology* (1970) in my collection, after purchasing them from used book stores. Nevertheless, I did not read them, as is often for dilettante book-collectors. In 2012, more than a decade later, when I published a book in Korean (which is like Amazon Direct Publishing in the US), I included again some discussions on entropy, information, and life, which are just slightly more developed from my 1998 master thesis. And then, now in 2018 here in this dissertation, I am arguing far more developed ideas of entropy and information together with equilibrium, evolution, and morality. While I have been doing research for the dissertation, all those books mentioned above have suddenly popped up together, and some ideas from those books are discussed here. I would say that my 1998 and 2012 publications are 10 to 20 % of the discussions on the topics in this dissertation, while the central thesis (the idea of "from cosmology through axiology to a meaning of life") remains the same. To sum up, the discussion on entropy and more in this dissertation is my first exposition on such topics in English.

Chapter 5. Epistemic Game Theory and Backward Induction

The main subject to be discussed in this chapter is the inference of backward induction. I argue that backward induction is one of the most common and vital kinds of inferences related to morality as social software. I first introduce backward induction intuitively by discussing the centipede game and a dialogue between Alexander the Great and a philosopher. I then argue that Popperian creatures in Dennett's theory are, basically, animals who exercise backward induction, and I introduce some exemplary animals from an Indian tale. I finally discuss the relationship between morality and backward induction. At the end of the chapter, a very brief précis of epistemic logic and epistemic game theory is attached, as an appendix.

§5-1. Intuitively: The Centipede Game and Alexander the Great

Backward induction is an inference, that is, a process of logical thinking. While exercising backward induction, a player of a game in game theory infers backward temporally to decide the best optimal move in the game. The player first starts considering the payoff of the final (or terminal) stage of the game. Then, based on this consideration of the final stage, the player proceeds backward to consider the payoff of the second-to-final stage, in order to decide the best optimal move in the game. And then, proceeding to the third-to-final, and so on, until the player finds the best optimal move. The best one can be found at the very first stage, or at a middle stage.

One of the most common and intuitive examples of backward induction in the literature may be the Centipede Game. The following diagram is an exemplary five-move version of the Centipede Game, which is from Fitting (2011, p. 154) and others, and is a shortened form of Robert Rosenthal's (1981) original 100 move version. (That's why "centi"-pede.)

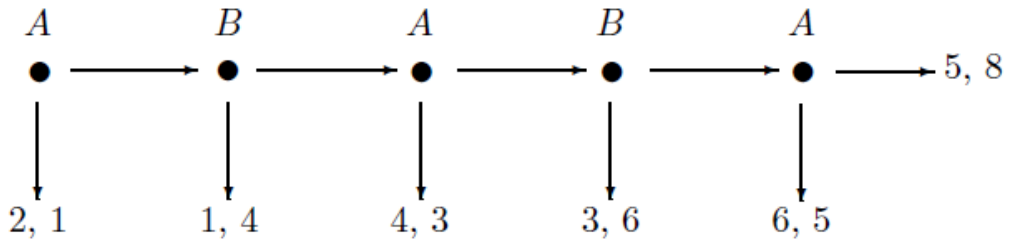


Figure 5: A Five-move Version of the Centipede Game

Player A (Ann) and player B (Bob) play a game together. The goal of it is to maximize their payoffs safely. Ann starts at the first, upper left node, by choosing to move across or down. If Ann chooses to move down, the game stops there: The payoff for Ann is the left number 2, and the payoff for Bob is the right number 1. If Ann chooses to move across, the game continues, and now this time at the second node, Bob chooses to move across or down. If Bob chooses to move down, the game stops there: The payoff for Ann is the left number 1, and payoff for Bob is the right number 4. Note that this time Ann's payoff is only 1, which is less than 2 at the previous (first) stage! In this fashion, the game can go on to the upper right, where Ann's payoff is 5, and Bob's payoff is 8.

Here, the question is what the best optimal move will be, if (1) Ann and Bob are 'rational' players, (2) they know all the payoffs in the beginning of the game, (3) Ann knows that Bob knows all the payoffs, and (4) Bob knows that Ann knows all the payoffs, (5) ad infinitum: that is, all the knowledge is 'common knowledge,' as in Aumann (1995). While exercising backward induction, Ann first imagines which move she will choose at the final (terminal), upper right node. She will choose to move down, because her payoff of moving down, 6, is larger than her payoff of moving across, 5. Now, at the second-to-final node (the fourth node from the left), Bob also does not continue the game, by choosing to move down, because his payoff of moving down 6 is larger

than his next payoff 5 when Ann chooses to move down at the next (final) node. This inference can be going backward to the first, upper left node. Therefore, under backward induction, it may seem the best optimal move for Ann to choose to move down at the very beginning node, and to be satisfied with a meager payoff of 2, which is less than the final payoff of 5. This conclusion, of course, raises lots of questions and problems. I will discuss them further in this chapter. For now, the example of the Centipede Game shows the essential concept of backward induction, intuitively.

In addition, more intuitively than the Centipede Game, I think that the following 'fictitious' dialogue between Alexander the Great and a Greek philosopher can show the essence of backward induction.[50]

Alexander the Great: I am going to conquer all the Greek regions and then the Middle East.

Philosopher: Oh, great, your highness! Then, what would you like to do next?

Alexander: Next, I am going to conquer India.

Philosopher: Great! And then?

Alexander: I am going to conquer China, the Far East, and all other parts of the World.

Philosopher: Oh, really great, your highness! Then, what would you like to do next?

Alexander: (Hmm, taking a moment) And then, I will take a rest in peace.

Philosopher: Your highness, I am taking a rest in peace, here and now.

The inference behind the philosopher's final remark is an instance of backward induction. As in the standard description of backward induction above, the philosopher "first starts considering the

[50] Often, depending on the sources of the story, an Indian Jain or Buddhist monk appears in this fictitious story as the interlocutor in the stage of India. Then, the dialogue must have started from India since the King had already conquered the Greek regions and the Middle East. Historians record that the King died in a palace in Babylon at age 32, without crossing the Indian subcontinent.

payoff of the final (or terminal) stage of the game. Then, based on this consideration of the final stage, the player proceeds backward," quickly, to the initial stage, here and now, that the philosopher is taking a rest in peace.

Finally, once more intuitively and succinctly, I think that the commonplace argument that "since we all die terminally, therefore we don't need to strive to maintain our life," is based on backward induction, though it does not seem to be endorsed by many.

As these paradigmatic examples of backward induction show, backward induction is common in everyday reasoning. Since Zermelo (1913) first formulated backward induction, and more recently, since von Neumann and Morgenstern (1944/1967 3rd Printing, Ch. 15, pp. 112-128) used it to solve the two-person games with perfect information (like chess), backward induction has attracted attention from many kinds of people: philosophers, logicians, mathematicians, economists, psychologists, among others. It is "the oldest idea in game theory" (Aumann 1995, p. 6).[51]

§5-2. Popperian Creatures by Dennett, Popper, and Millikan

5-2-1. Dennett's Five Kinds of Creatures

The concept of the 'Popperian Creature' of Daniel Dennett (1995, 1996) can shed further light on backward induction. Dennett, in his *Darwin's Dangerous Idea* (1995) and *Kinds of Minds* (1996), discusses an evolutionary hierarchy of intellectual progress. He calls the hierarchy the 'Tower of Generate-and-Test,' where there are five kinds of creatures.

On the ground floor of the tower, Dennett proposes, the inhabitants are (1) 'Darwinian creatures,' organisms who are blindly generated and field-tested, and then only the best designed

[51] An interesting phenomenon of backward induction, I observe, is that it is actually an inference of deduction, not induction, nor abduction, though not often emphasized in the literature.

inhabitants survive by natural selection. (See the discussion on natural selection in chapter 3.) The generation, test, and survival of a Darwinian creature is possible randomly, therefore, luckily. On the next level up, as a subset of Darwinian creatures, there are (2) 'Skinnerian creatures,' referencing to the behaviorist psychologist B. F. Skinner. Skinnerian creatures 'blindly' try different responses to the environment until one response is selected by reinforcement. And next time, unlike Darwinian, the Skinnerian creature will choose the reinforced response as its first choice. On the next upper, third, floor, there are (3) 'Popperian creatures,' referencing to the philosopher Sir Karl Popper. A Popperian creature can preselect an action from many options before doing it in the outer environment. A Popperian creature has a filter of 'inner environment,' where its many tryouts (i.e., hypotheses) can be safely tested. On the fourth floor, referencing to the psychologist Richard Gregory, there are (4) 'Gregorian creatures' who import mind tools from the outer cultural environment to construct their better inner environments for better generators and testers. On the fifth floor, finally, Dennett proposes, there are (5) creatures like human beings who can use these mind tools and, most of all, 'language,' in the "structure of deliberate, foresightful generate-and-test known as *science*" (1995, p. 380).

An attribute of backward induction can be demonstrated through the distinction between how Skinnerian creatures and Popperian creatures question themselves. On Dennett's summary of the difference:

> *Skinnerian creatures ask themselves, "What do I do next?" and haven't a clue how to answer until they have taken some hard knocks. Popperian creatures make a big advance by asking themselves, "What should I think about next?" before they ask themselves, "What should I do next?* (Dennett 1995, p. 378; 1996, p. 100)

The strategy of ‘thinking before acting’ can save the creature’s biological life. The cost of the failure of an idea for a Skinnerian can be the death of the creature; by contrast, the cost of the failure of an idea for a Popperian is just the death of the idea. The surviving Popperian creature can continue to try another idea.

5-2-2. Popper’s Evolutionary Epistemology

Dennett’s concept of Popperian creatures is connected to Popper’s ‘evolutionary epistemology’ and similar kinds. Popper argues, among others, in the first Darwin Lecture at Darwin College, in Cambridge (1977, “Natural Selection and the Emergence of Mind”), that there are various stages in the emergence of consciousness. At a possible first stage, Popper proposes, some kinds of centralized warnings evolve: for instance, irritation, discomfort, pain, or fear. These warnings induce “the organism to stop an inadequate movement and to adopt some alternative behavior in its stead before it is too late, before too much damage has been done” (1977). The absence or disregard of a warning signal often leads the organism to death. At a second stage, natural selection favors those organisms who try out before the real movements are executed. “In this way, real trial-and-error behavior may be replaced, or preceded, by imagined or vicarious trial-and-error behavior” (1977). At a third stage, we may consider the evolution of purposeful actions: that is, the aims, goals, or ends of actions. If we start an imagined trial-and-error action, we should necessarily evaluate the end state of the imagined action. Based on this discussion about the three stages, Popper proceeds to argue, “Let our conjectures die in our stead!” which, I believe, is one of the most brilliant ideas in the history of ideas:

> *The evolution of language and --- the products of the human mind allows a further step: the human step. It allows us to dissociate ourselves from our own hypotheses, and to look upon them critically. While an uncritical animal may be eliminated together with its dogmatically held hypotheses, we may formulate our hypotheses, and criticize them. Let our conjectures, our theories, die in our stead! We may still learn to kill our theories instead of killing each other. If natural selection has favored the evolution of mind ---, then it is perhaps more than a utopian dream that may see the victory of --- the rational or the scientific attitude of eliminating our theories, our opinions, by rational criticism, instead of eliminating each other* (1977, "Natural Selection and the Emergence of Mind").

The idea of killing hypotheses, instead of humans, is fascinating for peace. And, backward induction, I argue, is a fascinating way of killing hypotheses.

5-2-3. Millikan's Rationality

Ruth Garett Millikan, in her "Styles and Rationality" (2006), argues that, when discussing the rationality of non-human animals, "being rational is being a Popperian animal," among many interpretations of what it is to be rational. A Popperian animal tries things out in its head, which, Millikan argues, is "quicker and safer than trying them out in the world --- [and] --- than either operant conditioning or natural selection." Millikan suggests, "based on very informal behavioral evidence," that both humans and many higher animals are Popperian. Millikan reports her observation of grey squirrels in her laboratory that show the Popperian behavior of "mental trial and error." Millikan seems to assume that chimps, dolphins, and African grey parrots are Popperian, though those are not in her laboratory.

§5-3. The Cat and the Mouse in an Indian Animal Tale

Now, let's turn from Millikan's Popperian animals to animals in an Indian tale. These Indian animals are anthropomorphic. Harald Wiese (2012) discusses three traditional Indian fables, which, I argue, are related to backward induction. Among them, here I introduce one, "The Cat and the Mouse." The original animal tale is from the grand epic, *Mahabharata*, Book 12 (Trans. Satyamurti 2015, pp. 714-716) and the story is as follows:

> *Around a beautiful banyan tree, there lived a cat, a mouse, flocks of birds, mongooses and other creatures; and in the nearby town, there lived a hunter who came to the tree and set a snare every night, and collected animals caught in it next morning. One night, the careless cat became caught in the snare of net; the mouse saw this, and climbed up on top of the net to eat the meat put by the hunter as bait. That moment, the mouse looked down and saw a mongoose licking his lips; and looked up and saw an owl looking at the mouse.*

How can the mouse escape from this dangerous situation? The mouse suggested to the cat (still caught in the net), "[Hey, Cat, my friend] suppose I climb down to you, and you agree to protect me from the mongoose and the owl. I'll undertake to bite through the bonds that tie you, if you agree not to kill me. I save you, and you save me---how about it" (Satyamurti 2015, p. 715)? They agreed and the mouse "snuggled comfortably on the bosom of the cat." As the mongoose and the owl started to look for other food, the mouse started to gnaw through the strings of the snare. But the mouse did it very, very slowly, while the cat got more and more impatient. After long arguments about honor, respect, friendship and enmity through the night, the mouse finally gnawed

through the last cord next dawn at the moment when they heard the hunter's footstep approaching, therefore, when the danger was identical for both of them.

Wiese (2012) draws the game tree of this animal tale like the diagram below. The first payoff is for the mouse, the second for the cat. The mouse's payoff is 0 when it escapes unharmed; -100 when it is killed. The cat's payoff is 0 when it escapes unharmed; -50 when it is killed by the hunter; -48 when killed by the hunter after eating the mouse. Interestingly in Wiese's game tree below, the value of the mouse's life is 100 for the mouse itself, but only 2 for the cat.

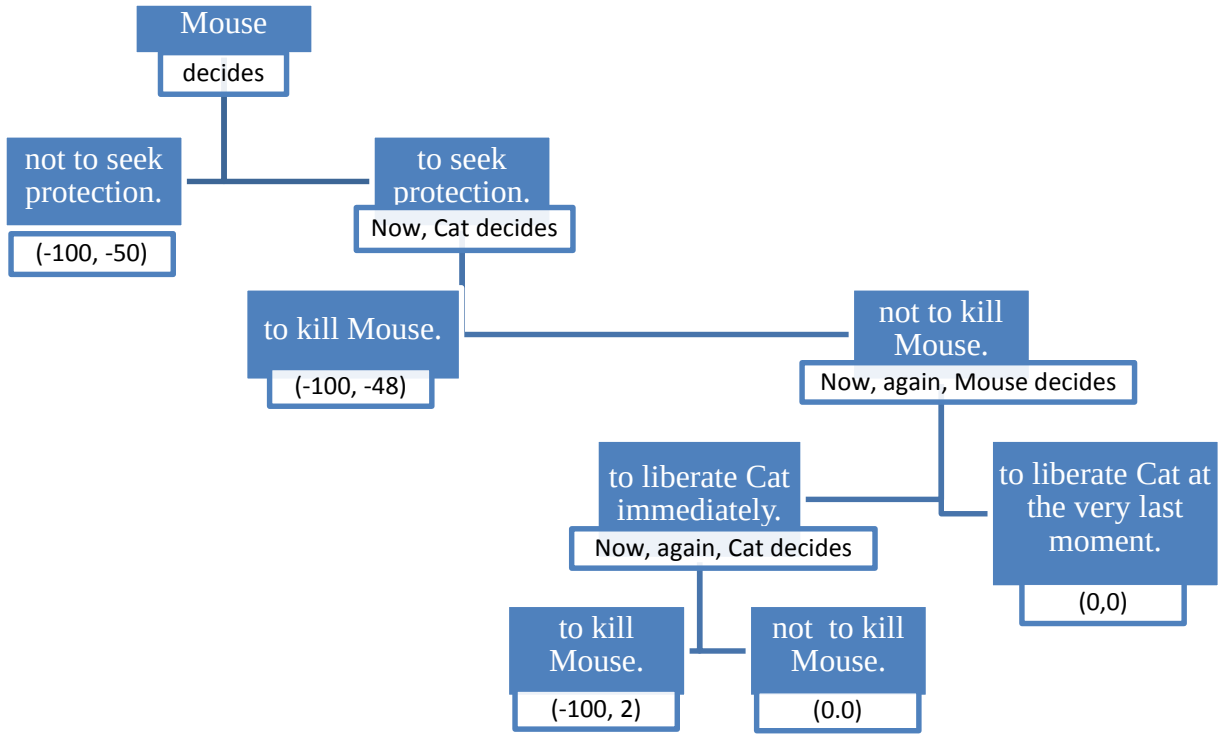


Diagram 2: A Game Tree of an Indian Animal Tale

Source: Wiese 2012, "Backward Induction in Indian Animal Tales"

The mouse's inference starts from the final stage, as customary in backward induction. If the mouse liberates the cat at the very last moment, the cat will have two options: 1) killing the mouse and being caught by the hunter where the payoffs are (-100 for the mouse, -48 for the cat), and 2) letting the mouse go and escaping unharmed where the payoffs are (0, 0). So, the cat will choose the option 2) at the very last moment since 0 > -48. By contrast, if the mouse liberates the cat much earlier than when the hunter approaches, the cat will have another two options: 3) killing the mouse and escaping unharmed where the payoffs are (-100, 2) and 4) letting the mouse go and escaping unharmed where the payoffs are (0, 0). So, the cat will choose the option 3), since 2 > 0. Between the cat's option 2) and 3), the mouse prefers 2) to 3), since the mouse's payoff in 2), 0, is bigger than the payoff in 3), -100. Therefore, at "the second-to-final stage," the mouse chooses the option that it liberates the cat at the very last moment.

Now, going backward, the cat chooses whether the cat kills mouse immediately for just, say, hunger, or the cat lets the mouse gnaw through the cords very slowly. In this fashion, both the mouse and the cat infer that it is in the best interest for each of them (and both of them) not to kill the other or not to let the other die. Literally, "I save you, and you save me."

§5-4. Morality and Backward Induction

Another interesting point of this fable is that it shows a hint of morality based on backward induction. The cat who was impatient at the speed of the mouse's gnawing through the cords asked the mouse to trust the cat, but the mouse replied, giving a lesson in morality:

> *The cat to the mouse: "Why don't you trust me? --- I know I hunted you before, but now we are friends for life. I will always honor and respect you."*

The mouse replied: "Listen to me. --- Between the weak and the strong there can be no real friendship, let alone for life. There are only linked interests. Friendship and enmity are the product of the situation. Unlike the bond between brothers[52]*, neither trust nor sentiment comes into it."* (*Mahabharata*, Book 12, Trans. Satyamurti, 2015, p. 716)

Though we may not completely endorse the mouse's claims, we can easily scent from them some amoralist view, which is similar to, for instance, Thrasymachus' claims in Plato's *Republic*[53] and Machiavelli's claims in *The Prince*[54] in the Western amoralist tradition. Indeed, I do not think it is by accident that the title of Book 12 of *Mahabharata* that contains the story above, is "The Book of Peace," and the sections immediately before the story are about "The Education of the Dharma King." The Indian fable-teller must have had a didactic purpose of teaching the readers moral philosophy (or, at least, political philosophy): mainly, the lesson is that, in order to keep peace, the king (and the people) should understand the situational friendship and enmity between the weak and the strong. What I am emphasizing now is that this understanding is attained through backward induction.

Incidentally, regarding the Western amoralist tradition, it seems to me to be unwieldy and cautious to handle the claims of Thrasymachus and Machiavelli. For example, Thrasymachus' following claims: (See also my discussion in chapter 1)

(a) Justice is nothing else than the advantage of the stronger (338c).

(b) Justice is really the good of another,

[52] It is also worth noting that, in the mouse's 'sermon, there is the bond between brothers, though none between the cat and the mouse. Recall kin selection!

[53] Plato, *Republic*, Book 1.

[54] Niccolo Machiavelli, *The Prince* (1513/1532).

(c) Justice is harmful to the one who obeys and serves.

(d) A just man always gets less than an unjust one.

(e) Injustice, if it is on a large enough scale, is stronger, freer, and more masterly than justice.

(g) Justice is very high-minded simplicity; whereas being unjust is good judgment.

(f) Unjust people are clever and good

(h) Injustice is virtue and wisdom; whereas justice is the opposite (Plato, *The Republic*, Book 1).

Relating to (e), how many parents willingly want to teach their children that, for instance, "if you steal 100 dollars from others and get arrested, you will have to go to jail; whereas if you steal 100 million dollars, you will be likely to be honored"?

Similarly, the famous fox and lion passage of Machiavelli: (*The Prince*, Ch. 18)

[Prince,] You must know there are two ways of contesting, the one by the law, the other by force; the first method is proper to men, the second to beasts; but because the first is frequently not sufficient, it is necessary to have recourse to the second. Therefore, it is necessary for a prince to understand how to avail himself of the beast and the man.

A prince, therefore, being compelled knowingly to adopt the beast, ought to choose the fox and the lion; because the lion cannot defend himself against snares and the fox cannot defend himself against wolves. Therefore, it is necessary to be a fox to discover the snares and a lion to terrify the wolves. Those who rely simply on the lion do not understand what they are about.

Therefore, a wise lord cannot, nor ought he to, keep faith when such observance may be turned against him, and when the reasons that caused him to pledge it exist no longer.

Here, being a fox is also being a Popperian. Let's recall that the title of the chapter 18 of *The Prince* is "Concerning the Way in which Princes should Keep Faith." The superficial format of Machiavelli's book is, like *Mahabharata*, to teach a prince (of the Medici), though the book is, actually, a serious work on socio-political philosophy.

I think the common interest in all these claims of the three (the mouse, Thrasymachus and Machiavelli) is that morality is not that simple: Morality must consider both sides of the coin, and so must consider 'morality as social software,' consequently.

§5-5. The Stag Hunt Game and the Prisoner's Dilemma

Two famous games in epistemic game theory, the Stag Hunt and the Prisoner's Dilemma show more clearly the epistemic game-theoretic aspect of morality. Brian Skyrms in *The Stag Hunt and the Evolution of Social Structure* (2003) argues that one simple exemplary game that can better explain the central problem of the social contract is not the Prisoner's Dilemma, but the stag hunt. He believes that the emphasis on the Prisoner's Dilemma is "misplaced," and the "most appropriate choice is --- the stag hunt" (Skyrms 2003, xii). Rousseau, in *Discourse in Equality* (1755), contrasted hunting the hare with the deer: the risk of hunting the hare without cooperation is small, but the reward is also small; the risk of hunting the stag without the certainty of cooperation is large, but the reward is much greater. Skyrms, regarding Rousseau's contrast as a prototypical story, argues that "the key to the evolution of cooperation, collective action, and social structure is correlation. Correlation of interactions allows the evolution of cooperative social structure that

would otherwise be impossible" (2003, pp. xii-xiii). In his argument, the three main components of this correlation are (1) location: the effect of interactions with neighbors, (2) signals: the exchange of signals prior to an interaction, and (3) association: the interactions in an evolving social network (2003, p. xiii).

The stag hunt game is about how to gain collectively. Then, the next task should be about how to divide the hunted stag justly. Skyrms notes that a bargaining game can deal with this task of distribute justice. Skyrms argues that the two main principles of distributive justice are: (Skyrms 2003, p. 18)

> *Optimality: A distribution is not just if under an alternative distribution all recipients would be better off.*
>
> *Equity: If the position of the recipients is symmetric, then the distribution should be symmetric.*

Skyrms thinks that these two principles are the "two most uncontroversial requirements" (ibid.) in Nash's (1950) axiomatic treatment of bargaining and are shared by other axiomatic treatments such as Kalai and Smorodinsky (1975). Skyrms' interpretation is that Kalai and Smorodinsky "disagree with Nash's theory in less symmetric bargaining situations, but agree with Nash in divide-the-dollar" (Skyrms 2003, p. 18). Stambaugh (2017) finds that the exact conditions under which the two solutions to the bargaining problem by Nash, and Kalai and Smorodinsky, coincide.

§5-6. Why I should be Nice to Others: a Backward Answer through the Centipede Game

We can use the framework of epistemic game theory to investigate some cases of epistemic game-theoretic morality. I introduce a sample question, "Why ought I to be nice to others?." And then, I

introduce two kinds of answers: one is well-known and common; the other is, hopefully, new and creative.

In the first sentence of this dissertation, I ask an ultimate question, "How ought I to live?" And now here, I introduce a possible answer, "nicely to others," with the help of game theory. (Here 'nice' may mean just 'nice,' and below can be defined more specifically, as in Axelrod's *Evolution of Cooperation*.) Then, we may ask further, "Why ought I to be nice to others?." I think this question has two kinds of answers: forward and backward. A forward answer based on the iterated Prisoner's Dilemma is well-known and common to us: I ought to be nice to others, because, in the future, those people will be good for me. That is, good for my 'future' fitness, survival, success, enlightenment and the like. A backward answer is, relatively, not well-known to us: I ought to be nice to others, because, in the past, they were good to me. That is, good for my 'present' fitness, survival, success, enlightenment and the like: so now, I ought to pay others back. These two 'folk' ethics (or forms of wisdom) can be reinforced through a game-theoretic framework. [55]

Since Robert Axelrod's monumental work, *Evolution of Cooperation* (1984/2006 Revised), it has been well-known that the strategy of tit-for-tat (which "starts with cooperation, and thereafter do what the other player did on the previous move," (p. viii)) is the strongest form in many game situations of the iterated Prisoner's Dilemma. And, as the title of a BBC documentary program *Nice Guys Finish First* (1986) (which is largely based on Axelrod's *Evolution of Cooperation* and is presented by Richard Dawkins) suggests, an essential property of tit-for-tat is the property of "being nice." Axelrod defines "being nice" as "never being the first to defect" (1984/2006, p. 33). In Axelrod's first computer tournament for the study of effective choice in the iterated Prisoner's

[55] The forward answer looks like more a form of wisdom; the backward answer, by contrast, looks like more an ethic. However, I take, in this dissertation, a position that wisdom cannot be distinguished from ethics, distinctively. Something good due to wisdom can also be good ethically.

Dilemma, tit-for-tat won, defeating thirteen other programs (entries). In the second tournament even when other programs knew the winning of tit-for-tat in the first tournament, tit-for-tat achieved the highest average score of the sixty-two entries.

Tit-for-tat strategy showed that there was a substantial correlation between being nice and being well in the tournaments. Axelrod, based on the analysis of the data from these tournaments, finds four successful properties of a decision rule: (1984/2006, p. 20)

> *[1] avoidance of unnecessary conflict by cooperating as long as the other player does,*
>
> *[2] provocability [possibility of retaliation or punishment] in the face of an "uncalled-for" [exactly what this means is not precisely determined (p.44), but loosely "unwanted"] defection by the other,*
>
> *[3] forgiveness after responding to a provocation, and*
>
> *[4] clarity of behavior so that the other player can adapt to your pattern of action.*

Axelrod argues that the results from the tournaments demonstrate that "cooperation can indeed emerge in a world of egoists without central authority," and in a nutshell, "the evolution of cooperation requires that individuals have a *sufficiently large chance to meet again* [emphasis is mine] so that they have a stake in their future interaction" (p. 20).

I interpret Axelrod's analysis above as a very good "forward" answer to the question, "Why ought I to be nice to others?." Then, analogously, I conjecture that, if the Axelrod-type analysis is reversed in the backward way around, and if we construct an epistemic game-theoretic analysis carefully, we will be able to reinforce the backward answer. For example, the fact that two players meet in the present can be strong evidence that they had a *'sufficiently large chance to meet again'*

[emphasized above] in the past. In an Axelrod-type analysis, if a game is played a known finite number of times, the players have no incentive to cooperate, and the mutual defection on the first move is typical, like in the Centipede Game of a typical backward induction. By contrast, if a game is played an *indefinite number of times* (and in most realistic settings, the players cannot be sure when the last interaction between them will take place), on an Axelrod-type analysis, cooperation can emerge.

Furthermore, exerting our slight imagination, we may regard the *indefinite number of times* as *infinite times*. 'Infinity' is akin to 'convergence' in calculus. Then, such scientific thesis about selfish genes as Richard Dawkins', and such religious teaching about Karma (which is based on the concept of reincarnation), may well be explained by the infinite number of interactions of players (or beings). A gene can be cooperative with other genes, and a human or sentient being can be cooperative with other beings, because they will meet some time in the infinite time of the universe. If that infinity converges into one moment, the distinction among the past, the present, and the future is not significant. On this "imagination," just as we ought to be nice to others for the future, so we ought to be nice to others due to the past. Again, these are all my, hopefully "imaginative," conjectures.[56]

I conclude this chapter by recalling some main theses in this dissertation: if morality is social software that has been developed through the evolutionary process of individuals and groups, that process must have included moral reasoning, and one of the main components of the moral reasoning must have been backward induction that is most used in epistemic game theory.

[56] And also, I conjecture that this backward answer will be different from "what, Scanlon (1998) argues, we owe to each other." Scanlon is Kantian, and so contractualistic; my possible backward answer will possibly be epistemic game-theoretic.

§5-7. Appendix: A Précis of Epistemic Logic and Epistemic Game Theory

5-7-1. Epistemic Concepts: Common Knowledge, Backward Induction, and Logical Omniscience

We examine some essential epistemic concepts in epistemic logic such as common knowledge, backward induction, and logical omniscience. Roughly speaking, an informal definition of each concept can be described as follows:

- Common knowledge: In a group G if a proposition is common knowledge, then the proposition is true; everybody in G knows it; everybody in G knows that everybody in G knows it; and so on, ad infinitum.
- Backward induction: "[b]ackward induction concludes X is true throughout a game by showing that X is true at terminal game states, and also showing X is true at a state provided some transition from that state takes the game to another state at which X is true. Thus, one works backward from terminal states to encompass the entire game tree" (Fitting 2011, p. 152).
- Logical omniscience: the problem of logical omniscience occurs when the believer cannot reproduce all the theorems that belong to the underlying logic; it occurs when the believer is not aware of some logical equivalence.

We note that the concept of common knowledge was discovered and formalized by, "independently," Lewis (1969), Aumann (1976), Schiffer (1972), and even Nozick (in his doctoral dissertation 1963). The topic of the hypothesis that rationality is common knowledge is also interesting. Aumann (1995) seems to believe that common knowledge of rationality is related to

backward induction. The claims such as "morality is rationality," "rationality is common knowledge," "common knowledge is related to backward induction." are "inspirational," since they may conclude that "morality is based on backward induction."[57]

5-7-2. Propositional Dynamic Epistemic Logic

Melvin Fitting in his "Reasoning about Games" (2011) discusses propositional dynamic epistemic logic (PDL+E), which is a fusion of epistemic logic (E) and propositional dynamic logic (PDL), relating it to game theory. Fitting first describes the axiomatics and semantics described in Schmidt et al. (2002, 2008). For epistemic logic, the 'basic settings' are common, and there are three additional axiom schemes: (Fitting 2011, p. 145)

- *E-1* $K_A X \supset X$, *Factivity*
- *E-2* $K_A X \supset K_A K_A X$, *Positive Introspection*
- *E-3* $\neg K_A X \supset K_A \neg K_A X$, *Negative Introspection*

- And, semantics is the familiar Kripke/Hintikka possible world semantics.

"Propositional dynamic logic (PDL) is a logic of non-deterministic *actions*--the action 'go to a store and buy milk' could be executed in many ways since a choice of store is unspecified. The

[57] In addition, I have just a conjecture for now, but it seems to me that there may be some fruitful lessons from the relationship between epistemic logic and the de re/de dicto distinction. Those key epistemic conceptions in this sub-section have been actively discussed in the discipline of logic, game theory, and computer science. Besides those disciplines, however, we have epistemology and philosophy of language, in which the de re/de dicto distinction is much investigated. Regarding this issue, the stepping stones of our research will be Mendelsohn's "Referential/attributive: a scope interpretation" (2008), "Sinn and Bedeutung with Scope" (2012), and Fitting and Mendelsohn, *First-Order Modal Logic* (1998), especially, section 9.2 "Scope" and 9.3 "Predicate abstract."

formula [α]X is intended to express that X will be true after action α is executed" (Fitting 2011, p. 145). Then, there are a standard PDL axiom system and rules.

PDL and E can be fused, axiomatically and semantically. In addition, there are three interaction conditions between PDL and E: No Learning, Perfect Recall, and Reasoning Ability. Based on this setting, Fitting gives a detailed analysis of the Centipede game and argues that "PDL+E can be a useful basis for the logical investigation of game theory" (p. 143).

5-7-3. The Contrast between Classical Game Theory and Epistemic Game Theory

Traditionally, game theory has been distinguished from decision theory, in that game theory considers other players' actions. However, if these epistemic concepts discussed above are embedded in game theory (that is, epistemic logic is concerned), the distinction between game theory and decision theory becomes less clear.

Epistemic game theory is contrasted with classical game theory in some respects (see Pacuit and Roy in "Epistemic Foundations of Game Theory" (2015)). Classical game theory includes components such as players, feasible options (actions or strategies), and players' preferences. In classical game theory, a 'solution concept' is important to denote a set of practical recommendations about what the players should do in a game. The famous Nash equilibrium is an example of a solution concept. By contrast, in epistemic game theory, epistemic components such as common knowledge, backward induction, and logical omniscience play important roles, in addition to the traditional component such as players, actions, outcomes, and preferences. These epistemic ingredients can "bring back the theory of decision making in games to its decision-theoretic roots" (Pacuit and Roy, 2015).

Chapter 6. The Less We Know, the More Rational and Moral We Are

The main subject to be discussed in this chapter is a combination of a series of ideas: that is, (1) if we know less, then it is sometimes possible that we become more rational; (2) being rational is being moral; 'therefore,' (3) if we know less, it is sometimes possible that we become more moral. The first idea of (1) above is discussed in Parikh (2017), and I develop it further by adding the ideas of (2) and (3). I first discuss two simple cases, cake cutting and a tiger as a motivational introduction to this chapter. I then discuss John Rawls's 'veil of ignorance' in the 'original position.' I finally discuss a verse of T.S. Eliot's poem "April," comparing it to Norbert Wiener's thought of information.

§6-1. Motivation: Cake Cutting and the Tiger in the Bathroom

6-1-1. The first case: cake cutting procedure, again.

I discussed, in chapter 2, cake cutting procedure as an exemplary case of social software. Let's see cake cutting here from a different perspective. N participants are to divide a cake. They assume that if each gets an equal amount (say, size) of the cake, the division is fair. If one participant is chosen to divide the cake by *n* pieces, and he is allowed to take the last piece after all other n-1 participants take theirs, what strategy can be the most rational one for him? The cutting person will try to divide the cake equally as far as he can, because he does not know which piece will remain last for him. His ignorance of his piece makes him rational, so that the division will become more moral, if fair and equal division is moral and ethical. So, the less we know, the more rational and moral we are.

In the following section, from this motivation, we discuss an aspect of John Rawls's theory of justice. Rawls himself briefly discusses the fair cake cutting procedure (1971/1999 Revised, p. 74, pages numbers are of 1999 edition).

6-1-2. The second case: the tiger in the bathroom is unknown to a person.

Parikh discusses, in his paper titled "An Epistemic Generalization of Rationalizability" (2017), an 'amusingly' scary situation where there is a tiger in the bathroom of your apartment. Let's suppose I know that there is a tiger in your bathroom and you need to use it now.

(1) If you do not know about the tiger, you will proceed to the bathroom;

(2) If you know about it, you will proceed to your neighbor's apartment and ask if you can use his bathroom.

The proposition, "a tiger is in the bathroom" can be true or false, and you have two possible actions. Parikh's model payoff matrix is as follows (Parikh 2017, p. 3):

	Tiger	No Tiger
Use Own Bathroom	(a) -20,000	(b) 10
Neighbor Bathroom	(c) -5	(d) -5

Table 1: A Payoff Matrix of "the Tiger in the Bathroom"

Source: Parikh 2017

(a) If you proceed to your bathroom without knowing about the tiger, you are very likely to be killed by the tiger, and your payoff is -20,000.

(b) If you proceed to your bathroom and there is no tiger, you enjoy the convenience payoff of 10.

(c), and (d) If you proceed to your neighbor's bathroom, you have to pay some "social cost," so your payoffs are -5.

Relating to the current subject of this chapter, the following strategy S1 and S2 are of interest to us:

S1: You proceed to your bathroom.

S2: You proceed to your neighbor's bathroom.

According to the payoff matrix above, S1 (your bathroom) is dominant over S2 (your neighbor's), if you do not know about the tiger; whereas S2 (neighbor's) is dominant over S1 (yours) if you know about the tiger. On Parikh's summary (2017):

> *The strategy of using your own bathroom was rationalizable for you before you knew about the tiger. But once you know about the tiger, that strategy is not rationalizable. There are fewer rationalizable strategies when we know more*. (p. 6)

> *The less you know, the more rational you are!* (p. 5)

I here note that we need more sophisticated discussions on the classical Aristotelian conversion, obversion, and contraposition for the issue of 'less and/or more'; and also, on the rationalizability

as a solution concept in game theory. Nevertheless, I do not go further on those issues now by claiming that this is enough for a motivational introduction.

§6-2. John Rawls's Veil of Ignorance in the Original Position

John Rawls argues, in his book *A Theory of Justice* (1971/1999), for 'justice as fairness,' which is a social contract conception of justice. In the theory, the concepts of the 'original position' and the 'veil of ignorance' in the original position are essential to construct the conception of justice as fairness. I will argue in the following that Rawls's argument for the veil of ignorance is a great exemplary case of 'the less we know, the more rational and moral we are'.

Rawls's concept of the 'original position' "corresponds to the state of nature in the traditional theory of the social contract" (1971/1999, p. 11). It is natural to imagine an early state where human beings began to live together and to think about rules that they should follow in order to maintain their community. The community may be called a society, state, polis (in Ancient Greece), country or similar. This imagined early state is called the 'state of nature' by those traditional social contract theorists including Hobbes, Rousseau, and Locke, and sometimes even by Kant, though Kant may not be regarded as a social contract theorist. This state of nature, and the original position too, Rawls argues, should not be thought of as an "actual historical state of affairs," nor a "primitive condition of culture"; rather, it should be understood as a "purely hypothetical situation" to construct and agree on a certain conception of justice (Ibid.). Whether the state of nature existed actually and historically or not is an interesting question. Of course, in the early period of human history, there certainly were not gatherings like the modern congress or senate. Nevertheless, humans' early processes of building morality, laws, and nations can be regarded as the state of nature, if we see the period from the contemporary viewpoint of, e.g., collective intelligence and

big data. The state of nature may also be compared to the fictional situation described in William Golding's novel, *Lord of the Flies* (1954). The state where a group of boys stranded on a deserted island in the Pacific Ocean attempt to govern themselves can be an exemplar of the state of nature, provided that all apparent discrepancies are muted.

Now, parties to the original position, Rawls argues, do not know certain kinds of particular facts. If the parties were to know these particulars, it would be tempting for them to exploit social and natural circumstances to the parties' own advantage. In order to nullify this advantage, Rawls suggests that the parties are situated behind a veil of ignorance. Some particular facts in Rawls's list of ignorance are as follows:

> *First of all, no one knows his place in society, his class position or social status; nor does he know his fortune in the distribution of natural assets and abilities, his intelligence and strength, and the like. Nor, again, does anyone know his conception of the good, the particulars of his rational plan of life, or even the special features of his psychology such as his aversion to risk or liability to optimism or pessimism. ---- [T]he parties do not know the particular circumstances of their own society. That is, they do not know its economic or political situation, or the level of civilization and culture it has been able to achieve. The persons in the original position have no information as to which generation they belong* (1971/1999, p. 118).

Though parties to the original position do not know most particular facts, they know, at least, the only particular facts that "their society is subject to the circumstances of justice and whatever this implies" (1971/1999, p. 119). So, it is not a situation of absolute ignorance. If they do not know

even these particulars, I think all these discussions are pointless. Rawls suggests, by contrast, that parties know the general facts about their community. For instance, they understand political affairs, the principles of economic theories, the basis of social organization, the laws of human psychology and the like (Ibid.).

It may be argued that (1) what the parties know and do not know are not clear, (2) that Rawls suggests that too many particular facts are not known to the parties, and (3) that, therefore, it is hard to understand the concept of the original position. I do not think that these raised difficulties are legitimate. If I imagine the original position in Rawls's favor, the original position is exactly the same world as human beings live in now, except that we humans become blind and ignorant when our interests are involved. For example, we can assume that parties to the original position know all kinds of laws discovered in logic (laws of identity, non-contradiction (and/or paraconsistency), excluded middle, Modus Ponens, Modus Tollens, etc.), but as soon as these laws are involved in a person's interest, that person becomes ignorant of what kinds of person he is and will be. The person does not know whether the person's parents are rich, and what the person's religion, race, and nationality are. As a thought experiment, it is easy to think of this kind of imaginary position where the switch for a person's interest can be turned on or off automatically.

From the standpoint of the original position, Rawls postulates, the parties are rational:

> *"They assume that they normally prefer more primary social goods rather than less.--- It is rational for the parties to suppose that they do want a larger share. --- even though the parties are deprived of information about their particular ends, they have enough knowledge to rank the alternatives. They know that in general they must try to protect their liberties, widen their*

opportunities, and enlarge their means for promoting their aims. --- no longer guesswork. They can make a rational decision in the ordinary sense" (1971/1999, p. 123)

Unlike Rawlsian rationality in the original position, it may be rational for a person, for example, from a saint's viewpoint, to decrease his property by donating most of it to charities, or to reduce his rights of freedom and equality by alienating them to others. Rawls's rationality, by contrast, is like the "standard in economic theory," and "must be interpreted as far as possible in the narrow sense," "of taking the most effective means to given ends" (1971/1999, p. 12). Rawls seems to depend highly on the economic notion of rationality by referring to Amartya Sen's *Collective Choice and Social Welfare* (1970), and Kenneth Arrow's *Social Choice and Individual Values* (1963), (1971/1999, p. 116, note 9, and p. 124, note 14).

I regard the standard notion of rationality in economic theory as a form of teleological theory. Then, is the Rawlsian concept of rationality teleological, therefore, utilitarian (which is the most prominent version of teleology)? This question is interesting, because Rawls's theory of justice, since its inception, has been widely accepted as a strong criticism of utilitarianism which had gained widespread acceptance in Anglo-American ethics at the time. To give a quick answer, I argue that Rawls adopts the core concept of teleological rationality and elaborates it, and then provides an alternative moral theory to utilitarianism.

More specifically, following the (so-called) 'main stream' discourse of ethics, Rawls assumes that the right and the good are the two main concepts of ethics, and that the "structure of an ethical theory is -- determined by how it defines and connects these two basic notions" (1971/1999, p. 21). These ideas are so amazingly simple that they may over-simplify the discourse of ethics. Nevertheless, they have the beauty of simplicity. Setting aside the 'over-simplification' for another

occasion, I focus here on the beauty. Rawls, following William Frankena's definition of teleological theories in *Ethics* (1963), understands them as: "[1, (numbering is mine)] the good is defined independently from the right, and then [2] the right is defined as that which maximizes the good" (1971/1999, pp. 21-22). I am now talking about Rawlsian rationality influenced by the teleological notion of rationality. Rawls admits that teleology has a "deep intuitive appeal" because it embodies the "idea of rationality," which is "maximizing something" and, in ethics, "maximizing the good" (1971/1999, p. 22).

So far, it may seem that Rawls admits that his theory, 'justice as fairness', is a teleological theory. However, he explicitly denies the two tenets of teleological theories mentioned above, [1] and [2]. He argues that his theory of 'justice as fairness' is a deontological theory, "one that either does not specify the good independently from the right [which is denying [1]], or does not interpret the right as maximizing the good [which is denying [2]]" (1971/1999, p. 26). How can Rawls resolve this apparent contradiction in his theory? How can his theory (which is based on the teleological concept of rationality) be a deontological theory (which is the counterpart of the teleological)? It seems to me that for Rawls's position to resolve this contradiction it is to be hovering between deontology and teleology: that is, he argues that "[a]ll ethical doctrines [including deontological] worth our attention take consequences into account in judging rightness. One which did not would simply be irrational, crazy" (Ibid.). He writes, "crazy." I understand this claim in a way that the rightness of an action in any moral theory cannot be free from the consequence of the action, regardless of whether the theory is (seemingly) noble Kantian deontology or (seemingly) mundane utilitarianism. I see that Rawls may have the view that, for example, even the Kantian moral theory of categorical imperative, which is categorized as a typical form of non-consequentialist theory, is also a form of consequentialist theory, in which the

rightness of an action is decided by its consequence. I think that the Rawlsian way of resolution may not have the taste of clear-cut distinction, but it still has a pragmatic merit.

Rawls develops further the concept of rationality of the parties to the original position by adding special assumptions. In the original position, behind the veil of ignorance, "a rational individual does not suffer from envy" (1971/1999, p. 124). This envy-free individual does not accept a loss of social goods for himself even though others have less; they do not feel unhappy when others have more social goods. So, the rationality in the original position is "mutually disinterested rationality," that is, "put in terms of a game, -- , they strive for as high an absolute score as possible. They do not wish a high or a low score for their opponents, nor do they seek to maximize or minimize the difference between their successes and those of others" (p. 125).

Now, the parties who are equipped with ignorance and rationality in the original position, Rawls argues, choose more moral options for the community they construct when they agree on the conception of justice. That is, "the combination of mutual disinterest and the veil of ignorance achieves much the same purpose as benevolence," and "has the merits of simplicity and clarity -- - insuring the effects of what are --- morally more attractive assumptions" (1971/1999, pp. 128-129). Rawls argues for the two moral principles agreed upon by the parties to the original position:

> *First: each person is to have an equal right to the most extensive scheme of equal basic liberties compatible with a similar scheme of liberties for others.*
>
> *Second: social and economic inequalities are to be arranged so that they are both (a) reasonably expected to be to everyone's advantage, and (b) attached to positions and offices open to all.* (1971/1999, p. 53)

The first principle is about equal liberty, and (b) of the second principle is about fair equality of opportunity, both of which are less debatable. What is of most interest to our current discussion is (a) of the second principle. (a) is called the 'difference principle,' and can be rephrased as follows:

> *The Difference Principle: social and economic inequalities are to be arranged so that they improve the "expectations of the least advantaged members of society"* (p. 65).

For the main purpose of this chapter, I am most interested in and will focus on how this difference principle (which is "the more moral we are" clause) is deduced from the combination of the veil of ignorance (which is "the less we know" clause) and the mutually disinterested rationality (which is "the more rational we are" clause).

Rawls adopts the "maxmin" rule in order to improve the goods of the "least advantaged members of society." "The term "maxmin" means the *maximum minimorum*; and the rule directs our attention to the worst that can happen under any proposed course of action" (p. 133, note19). For example, let's see the following three decisions in the table below.[58]

Decisions	Circumstances		
	C1	C2	C3
D1	4	8	12
D2	2	7	14
D3	5	6	8

Table 2: An Example of the Maximin Rule of Rawls

Source: Rawls 1971/1999, p. 133, note 19

[58] For clarity, I slightly changed decision 1's minimum value from Rawls' -7 into 4, and decision 2's minimum value from Rawls' -8 into 2. No negative values are needed now.

The maxmin rule requires us to choose decision 3, since its minimum payoff, 5, is larger than decision 1's minimum, 4, and decision 2's minimum, 2.

A person can become a "least advantaged member of society" from the view point of, for example, religion, race, nationality, disabilities, financial situation, or sexual orientation. This list can be enlarged to include all kinds of conflicts of society. However, the person cannot know, in Rawls' theory, whether he will be one of the least advantaged, because of the veil of ignorance. Therefore, by exercising the maxmin rule of rationality, the person will choose an option in which even if he will be one of the least advantaged, he will be safe. Some may argue that a certain brave person is wagering that he will not be a least advantaged member, and if he will be, he is willing to suffer from being a least advantaged. However, in the Rawlsian thought experiment of the original position, I think this kind of seemingly courageous action becomes 'foolhardy' which is an extreme that violates the Aristotelian doctrine of the Golden Mean. The agreement on justice in the original position is final and irrevocable, and if the person loses, he loses everything including his life.

To sum up, parties to the original position are so mutually disinterestedly rational thanks to the veil of ignorance that they agree on a more moral conception of justice by exercising the maxmin rule. We see that Rawls's moral theory is connected to an idea in the Western tradition of moral philosophy: the idea is that "being rational is being moral" (rationality is morality), which may be traced back to Aristotle, Aquinas, and Kant.

§6-3. Norbert Wiener's Information and T.S. Eliot's "April"

Relating to "the less we know," Norbert Wiener in his book, *The Human Use of Human Beings* (1950/1954 Revision) introduces us to an intriguing insight:

[T]he more probable the message, the less information it gives (p. 21).

In chapter 4, I discussed the relationship between entropy and information, I think, 'tenaciously.' What we can learn from that discussion is helpful for the current context.

According to Wiener,

Information is a name for the content of what is exchanged with the outer world as we adjust to it, and make our adjustment felt upon it (p. 17).

Just as entropy is a measure of disorganization, the information carried by a set of messages is a measure of organization. In fact, it is possible to interpret the information carried by a message as essentially the negative of its entropy, and the negative logarithm of its probability. That is, the more probable the message, the less information it gives. Clichés, for example, are less illuminating than great poems (p. 21).

Wiener's thesis, "the more probable the message, the less information it gives" reminds us of, in philosophy of science, Karl Popper's demarcation criterion of falsifiability between the scientific and the unscientific: The more falsifiable a statement is, the more information it gives.

Wiener's thesis also may explain the popularity of the verse, "April is the cruellest month," from T.S. Eliot's poem, "The Waste Land" (1922); and may explain why it has not disappeared into oblivion and is still commonly sung. The beginning lines of the poem are as follows:

APRIL is the cruellest month, breeding

Lilacs out of the dead land, mixing

Memory and desire, stirring

Dull roots with spring rain.

Winter kept us warm, covering

Earth in forgetful snow, feeding

A little life with dried tubers.

I may not be able to fathom the deep connotation of the verse, and it may be beyond the topic of this dissertation. What I want to emphasize here is that, if Eliot had written the verse such that, "April is the most beautiful," it would not have been so popular. In the northern hemisphere, April is usually a beautiful month in spring; and in April most beings are waking up from the winter sleep. Flowers are blooming; butterflies are mating. In this kind of atmosphere, as soon as a poet sings, "April is the most beautiful," the verse becomes a cliché, which is "less illuminating than great poems" (Wiener 1950/1954, p. 27).

Chapter 7. No Naturalistic Fallacy on Morality as Social Software

"Flower"

Until I called its name

it had been

nothing more than a gesture.

When I called its name,

it came to me

and became a flower.

Chun-su Kim (1922-2004, Poet, South Korea)

(First released in 1952)

The main thesis to be argued in this chapter is that the naturalistic fallacy and its 'living ancestor,' Hume's Law, can be resolved from the viewpoint of morality as social software that we have developed so far. The naturalistic fallacy is committed, on G. E. Moore's accusation, when one draws X's goodness from any of its natural properties; and, an 'ought' (value), on Hume's contention, cannot be deduced from 'is' (fact). Since this dissertation project is, essentially, based on the evolutionary concept of morality, the project may not be completely immune to the criticisms from the naturalistic fallacy and Hume's Law. That is, it may be critiqued that the construction of evolutionary morality as social software here is not enough to explain why we ought to follow 'that' morality just constructed. It seems inevitable, therefore, that I should provide sound replies to these kinds of criticisms before concluding the dissertation. I hope that the discussion in this chapter instills a bit more philosophical rigor into the theses of the dissertation.

I first discuss Moore's initial open question argument (which cannot be separated from his naturalistic fallacy), some prominent objections to his open question argument, and possible responses to those objections. I then introduce R. M. Hare's modified version of the open question argument, and argue that Hare's version is not valid, either. Through these discussions, I will argue that, those open question arguments of Moore's and Hare's 'in the tradition of analytic philosophy,' are not strong enough to show that moral naturalism is wrong: I will argue that the naturalistic fallacy is not a genuine fallacy. While reinforcing this point, I will argue further that the gap in Hume's Law between 'is' (fact) and 'ought' (value) can be filled with the concept of morality as social software and related distinctions among several kinds of 'oughts.'

§7-1. G. E. Moore's Open Question Argument

Since the first publication of Moore's *Principia Ethica* in 1903, already "much ink has been spilled" over this topic and "a lot of fire has been directed at" (Lenman 2006) the open question argument. Still, "[a]fter one hundred years of commentary, Moore's views --- are poorly understood (Piervincenzi 2007, p. v)."

I think that one way of summarizing G. E. Moore's claims and project for moral non-naturalism in *Principia Ethica* (1903) may be as follows:

(1) Moore argues that the term 'good' is simple, indefinable, and (therefore) non-natural.

(2) On Moore's contention, it is an error to attempt to define the term 'good' in connection with natural properties such as 'pleasure' and 'survival': analytic reductionism in the moral realm is erroneous.

(3) Moore calls this error the 'Naturalistic Fallacy.'

(4) The argument that Moore constructs when he presents these claims, has come to be known as the 'Open Question Argument.'

(5) Moore ultimately relies on intuition.

I begin, more specifically, by introducing two common forms of summaries of Moore's open question argument in Chapter 1 of *Principia Ethica*: firstly, Bernard Baumrin's (1968) summary of the structure of the open question argument; secondly, a summary by Roger Hancock (1960) and his followers who accept the 'Hancock-style' summary of the open question argument.

Firstly, Baumrin (1968, p. 79) analyzes Moore's argument as follows:

Premise (A): 'Good' denotes either: (1) a simple property, (2) a complex property, or (3) nothing at all.

P (B): 'Good' is either definable or indefinable. And,

P (C): If 'good' is definable, it must denote a complex, for only complexes are definable.

P (D): If 'good' is indefinable, it must denote: (1) a simple, or (2) nothing at all.

P (E): It does not denote a complex.

Conclusion (F): Therefore, it is indefinable (from Premise B & the contraposition of C).

P (G): It does denote something.

Conclusion (H): Therefore, it does denote a simple property (from D, E & F).

Secondly, in the following Hancock-style summary, which reconstructs the open question argument, 'property-identity' statements play the central roles (Hancock 1960, Horgan and

Timmons 1992, Vessel 2004/2009):[59]

(Premise 0: If moral naturalism is true, the following is valid.)

P1: For every natural property P, if the property of P (for example, pleasure) is identical with the property of 'good', then P and 'good' designate the same properties. (e.g. pleasure and good must designate the same property.)

P 2: But whenever we ask, "Is P good?," we can see that P and 'good' designate two distinct properties.

P3: If P and 'good' designate two distinct properties, the question "Is P good?" is not closed, that is, it is open. (e.g. To the question 'Is pleasure pleasure?,' we can answer "Yes!" obviously. Therefore, this question is a closed question; by contrast, to the question 'Is pleasure good?,' the answer is not obvious. We need to reflect on it for some time, no matter how short it is. Therefore, the question 'Is pleasure good?' is open.)

Conclusion 4: P and 'good' designate two distinct properties.

C 5: Therefore, there is no natural property that has the same property as 'good.'

(C6: therefore, moral naturalism is not true.)

The two forms of summaries above can be made clearer if we recall Moore's two analogies with the concept 'yellow' (1903, Ch. 1, Sec. 10), and with the relationship between the whole and the parts (Sec. 20). Moore argues that defining 'good' is impossible in the same way as defining a simple concept 'yellow.' Only complex objects can be defined (premise C above) by listing their

[59] Some philosophers such as Vessel (2004/2009) argue that the following summary of the open question argument is not correct: It is not the exact form of argumentation that Moore presents in *Principia Ethica*. I will discuss this point further below while discussing 'Objections from *A Posteriori* Identities.'

parts and the relationships between those parts. For example, it is possible to define the concept, 'horse,' (Sec. 7 and 8) because a horse has many different properties and qualities, all of which can be listed. On the contrary, 'yellow' and 'good' are not complex. 'Yellow' cannot be defined in terms of any other concept. Defining 'yellow,' for example, by describing that it is the color of visible light with a wavelength of 570 - 590 nanometer (nm) is not successful, since, on Moore's contention, the wavelength is not the color. The notion of yellow cannot be explained fully to anyone who does not already know it. Since 'good' is not a complex notion in the same way that 'yellow' is not, 'good' is indefinable in the same way that 'yellow' is indefinable (Conclusion F above). If 'good' is not complex, it must be either simple or meaningless. Moore contends that the idea that 'good' is meaningless can be rejected (Premise G), since everyone understands the question "Is pleasure good?" as distinct and meaningful. Therefore, 'good' denotes a simple notion (Conclusion H).

To sum up, in the simplest form, Moore argues with the open question argument (OQE) that:

OQE: Since, for example, the question "Is pleasure good?" is an open question, 'pleasure' and 'good' are not the same.

We understand that a question is open if it is not closed; a question is closed if, for example, "most any semantically competent speaker who considers the question carefully, and who properly brings his semantic competence to bear on the question, will judge both that the answer to the question is obviously 'yes' (or obviously 'no')" (Horgan and Timmons 1992, p. 161). When an analysis of a moral term is proposed in a reductive way, we can put the analysis in an interrogative sentence. If we can doubt (i.e., 'ask intelligibly or ask without stupidity') about how to answer the question,

then the question is open. This openness suggests that the initial reductive analysis fails. Moore himself seems to think that the term 'open' is interchangeable with the terms 'significant', 'intelligible,' and perhaps, 'doubtful.'

In this regard, the open question argument can function as a kind of 'test' of whether an analysis expresses sameness of meaning. The open question 'test' can function like the Kantian categorical imperative 'test' to check whether something is such-and-such (whether meaningful for Moore; whether moral for Kant). Furthermore, since similar questions with other natural or metaphysical terms are also open, Moore seems to argue that the entire project of analytic reductionism in the realm of ethics is erroneous. In short, Moore seems to view that 'good' is an irreducible, *sui generis*, non-natural property.

This view is closely related to Moore's worries about the naturalistic fallacy, which consists in identifying the simple non-natural notion of good with some other natural notion. Ethics, Moore argues, "aims at discovering what are those other properties belonging to all things which are good," but "far too many philosophers have thought that when they named those other properties --- these properties--- are absolutely and entirely the same with goodness" (1903, Sec. 10). Moore proposes to call this view of "far too many philosophers" the naturalistic fallacy.

§7-2. Objections to the Open Question Argument and Responses to the Objections

In this section, I discuss some common objections to Moore's open question argument, before discussing the naturalistic fallacy further. There are various objections to Moore's open question argument. Moore himself in his later work "A Reply to My Critics" (1942, p. 582; See Ridge, 2003/2014) admits that: in *Principia Ethica,* "I did not give any tenable explanation of what I meant by saying that "good" was not a natural property." My discussions in this section and the

next section of R.M. Hare's modification will follow the 'analytic tradition' of philosophy, more than other parts of this dissertation.

7-2-1. *A Posteriori* Identities

An objection from *a posteriori* property identities (See Lenman 2006; Ridge 2003/2014) focuses on Moore's failure to consider the distinction between 'denotation' and 'meaning.' Two terms can denote the same property though they are not equivalent in meaning. Let's consider the following two biconditionals:

(1) *x is good iff x is N* (e.g., sweet).

(2) *x is water iff x is* H_2O.

On Moore's contention with his open question argument, 'good' and N in (1) are not the same, because the question, "Is N good?" is open. (If 'good' and N in (1) are the same, the question cannot be open.) By contrast, on this objection with a counterexample in (2), though 'water' does not mean the same as 'H_2O, 'being water' and 'being H_2O' are considered to be the same identical property in chemistry. They 'denote' the same physical entity, and the property identity 'water = H_2O' is a necessary and *a posteriori* truth. The question, "I know it is H_2O, but is it water?" is still open to a speaker who does not know the identity relation 'water = H_2O.' Analogously, on this objection, the same argument goes with 'goodness' and, say, 'sweet.' Though the question, "I know it is sweet but is it good?" is open to a speaker who does not know the identity relation, 'goodness = sweetness,' it is still possible that goodness and sweetness may well be the same

property. So, Moore is wrong, on this objection, and Moore's argument can be simply dismissed by the claim that he has a pretty 'idiosyncratic' idea of 'definition.'[60]

By contrast, Horgan and Timmons (1992) who embrace the force of Moore's open question arguments against naturalism, respond to this objection by arguing that, to put it roughly, "The objection is valid for water, but not for goodness." They find that the analogy between 'goodness' and 'water' in the objection is untenable. On our planet, if we know that water is H_2O, then the question at issue is closed; if we don't know that identity, the question open; and this situation applies in the same way on another twin earth. Horgan and Timmons argue, however, that the situation about goodness is quite different from water. They suppose a 'moral twin earth' where the use of evaluative terms on the planet is the exact opposite of our use. Certainly, in this moral twin earth, the question "Is pleasure good?" is open, and Moore's open question argument still works. It seems to me that Horgan and Timmons believe that natural kind terms such as 'water' and 'H_2O' are, on Kripke's terminology, rigid designators; but moral terms such as 'good' and 'right' are not.[61]

One more response: Vessel (2004/2009) presents a newer and more unique response to this kind of 'objection based on property identity.' He argues that the Hancock-style construction of Moore's open question argument, in which property identity statements play prominent roles in the central premises, is not true to the Moorean texts. Rather, Vessel argues that Moore is interested in the meanings of certain interpretations of the predicates 'x is good' and 'x is pleasant,' as well as certain others. Vessel suggests that if we interpret the term 'pleasure' in Moore's questions as

[60] This objection is a form of *reductio ad absurdum* (See Strandberg 2004, p. 185). If the open question argument were valid, it could refute the hypothesis that, for example, even if two terms are used differently, they can refer to the same property. But the argument cannot refute the hypothesis, for example, that, two terms, water and H_2O are used differently, but refer to the same thing. Therefore, the argument is not valid. QED.

[61] The rigidity of moral terms is another controversial and interesting topic, which I do not pursue further here.

the ‘nominative counterpart’ of the predicate ‘x is pleasant,’ that is, the term ‘pleasure’ is used to refer to all the pleasant things, then the open question argument may be better defended to the objection based on property identity. I agree with Vessel that this kind of interpretation is more “charitable” towards Moore’s initial thoughts and does not nullify the lesson from Frege’s distinction between sense and reference.

7-2-2. Some Objections Related to the Paradox of Analysis, and Analyticity

Some objections to the open question argument are related to the paradox of analysis, which is a paradox that concerns how an analysis can be both correct and informative. Consider an analysis of the form ‘A is B,’ where A is the *analysandum* and B the *analysans*. Then, either (1) ‘A’ and ‘B’ have the same meaning, in which case the analysis expresses a trivial identity; or else (2) they do not have the same meaning, in which case the analysis is incorrect. A response to the paradox is Frege’s distinction between sense and reference. A different form of the paradox of analysis: To give a correct analysis of the meaning of, say, a concept, the philosopher must in some sense already understand or know the meaning of the concept. But if he already knows the meaning of the concept, how can his work be significant?

According to an objection related to the paradox of analysis, the general form of Moore’s open question argument is simply a particular instance of the paradox of analysis. A lesson from 20th century analytic philosophy is that competent speakers can ‘legitimately’ question analyticities (See van Roojen 2004/2013). That is, if there are given analyticities (or definitions) such as “N is good.” and “Water is H_2O.,” a competent speaker can ask questions such as “Is N good?” and “Is water H_2O?” about the analyticities (definitions). “The mere fact that a speaker can doubt a candidate analysis may not tell against that analysis” (Ibid.). In a simpler way relating to

the current context, the openness of the question on which Moore heavily relies, becomes less significant (Smith 1994, pp. 37-39). The openness is just an instance of the paradox of analysis. This kind of objection is, I think, essentially and analogously, like the biblical saying, "He who is without sin among you, let him be the first to throw a stone at her." If every analysis (definition) has, in principle, such an aspect of the paradox of analysis, Moore's criticism based on the openness may not be valid.

Another objection to the open question argument is related to the problem of "unobvious analyticity" (Lewis 1989, p. 129; van Roojen 2013). This objection is similar to the former objection above, but still, I argue, they are distinct from each other. Even if a moral term such as 'good' and a natural term such as 'sweet' are analytically equivalent, that analyticity may not be known to a competent speaker obviously, so that the analyticity can be questioned by the speaker. This unobviousness is well illustrated by an analogy with grammar, using Gilbert Ryle's (1949) seminal distinction between knowledge-how and knowledge-that. An identity could be analytic because competent speakers are tacitly aware of the analyticity, though they may not recognize the adequacy of the analyticity. Competent speakers can follow certain rules without knowing what the rules are. So, the question, "I know the sentence violates this rule but is it ungrammatical?" might seem to be open in the same way as in the moral case of Moore's argument (Ridge 2003/2014; Strandberg 2004, p. 188). Regarding the distinction between knowledge-how and knowledge-that, the openness in question results from the level of knowledge-how. The opponents of the open question argument may criticize Moore for ignoring the level of knowledge-that. However, interestingly and conversely, the proponents of Moore's argument may defend it by saying that that's exactly what Moore actually intended: That is, Moore argued simply of the level of knowledge-how.

§7-3. Non-cognitivists and R.M. Hare

7-3-1. Non-cognitivists

Non-cognitivists contend that moral statements have no truth values. According to non-cognitivists, when moral statements are spoken, 'non-cognitive attitudes' such as emotion (Ayer and Stevenson), desire (Carnap), commendation (Hare), approval, or disapproval, are expressed. Non-cognitivists, while agreeing with Moore's claim that moral terms do not refer to natural properties, argue that Moore's mistake is that he neglects another option: Perhaps, moral terms do not refer to expressions at all. The open question is always open, non-cognitivists argue, not because the moral term (e.g. good) in question 'wrongly' refers to a natural term (e.g. pleasure), but because the moral term does not refer to any property. Thus, non-cognitivists argue that moral terms in the open questions do not function to represent any property or meaning; rather, moral terms function completely differently: expressing an emotion, prescribing something, or taking a stand. Non-cognitivists have taken advantage of various modified forms of open question arguments to reinforce their metaethical theories such as emotivism (Ayer and Stevenson) and prescriptivism (Hare). I will focus here on only Hare's development of the open question argument.

7-3-2. R. M. Hare's Open Question Argument

Hare argues in his *Language of Morals* (1952) that when we call something 'good,' we 'commend' it. To avoid many weak points of Moore's refutation of ethical naturalism, Hare tries to improve Moore's argument or, justly to Hare, to produce new ones. According to Stojanović (1963, p. 264), since we can find so many new things in Hare's argument, we may call it "Hare's argument" instead of simply "Hare's improvement" on Moore. Lange (1966, p. 244) also observed that, for some philosophers, "it seems to be taken for granted that Hare has successfully cooked the goose [of

ethical naturalism] which Moore had [---] already prepared for the oven," though Lange himself did not take it for granted. Hare's argument, like Moore's, is intended to work against all sorts of naturalism. Hare thinks it can be shown that naturalism is in principle fallacious and untenable.

One way of summarizing Hare's main arguments against ethical naturalism follows (See Walker 1973, p. 45; originally, Hare 1952, Ch. 5):

(A) The basic thesis of naturalism: For all things of a certain kind X, the phrase 'a good X' means the same as 'a C X' (where C is some natural characteristic such as sweet and juicy).

(B) Therefore, the sentence (1) "A C X is a good X." means the same as the sentence

(2) "A C X is a C X." (by the substitution rule)

(C) Sentence (1) is typically used to commend X for being C.

(D) Sentence (2) is analytic, and therefore cannot be used to commend.

(E) Therefore, sentence (1) does not mean the same as sentence (2).

(F) Therefore, proposition (B) is false.

(G) Therefore, proposition (A), the basic thesis of naturalism is false.

Hare (1952, Ch. 5) re-illustrates this argument by using an example of a strawberry. Suppose that 'a good strawberry' means 'a strawberry that is sweet and juicy.' Now, by saying that "A strawberry that is sweet and juicy is a good strawberry," we can typically commend 'a good strawberry.' On the contrary, by saying that, after replacing the synonyms, "A strawberry that is sweet and juicy is a strawberry that is sweet and juicy," we cannot commend them, because this sentence is analytic. This impossibility shows that the initial supposition was wrong. Therefore, the thesis of ethical naturalism is wrong.

Hare, here using the strawberry example intentionally instead of, for example, cannibalism, so he avoids moral examples. He wants to make it clear that the current issue of the open question argument is not related to any morals, but to "the general characteristics of value-words." I see, indeed and incidentally, that 'sweetness' is a very good example of a value concept. The human tongue has developed the inclination to like a sweet taste through the evolutionary process, because un-sweet (or bitter) plants and minerals in nature are very likely to be poisonous. Nevertheless, both Western and Eastern medicine have recommended that humans should not eat too much sweet food. Sweetness may not be goodness, so, the term 'sweetness' may not be more eligible than the term, e.g., 'cannibalism' as an example of a value-term.

One merit of Hare's argument is that it is not vulnerable to a famous objection to Moore's argument. According to this objection, if Moore's refutation of the naturalistic definition of ethical terms were to be valid, it would invalidate any definition whatsoever. But this objection does not hold in Hare's case. It is because, firstly, Hare, unlike Moore, does not argue in terms of 'meaningfulness' of the question "Is an A which is C good?." And, it is also because, secondly, the complete parallel between "An A which is C is good." and a typical definition of the moral term, is not possible for Hare's case. Hare shows that "An A which is C is good." is also used to give a recommendation.

A naturalistic objection (Walker 1973) to Hare's argument is that, in proposition (B) of the summary above (that is, (B) The sentence (1) "A C X is a good X." means the same as (2) "A C X is a C X."), the phrase 'means the same as' is ambiguous; it can mean either (i) 'is necessarily true if and only if' or (ii) 'can be used to perform the same functions as'; Walker calls the meaning (i) 'locutionary identity,' and the meaning (ii) 'illocutionary identity.' If the phrase means (ii), then Hare's argument is invalid, so that the basic thesis of naturalism remains plausible.

More specifically, this objection is to the thesis that the sentence "A C X is a good X." itself is analytic. This objection is similar to the objection to Moore's argument above, where the analogy between the unobviousness of the concept 'analyticity' and grammar, is discussed. Now, this time, let's think of a strawberry expert who accepts that a good strawberry is sweet and juicy. If we assume that sentences can be only either 'synthetic' or 'analytic,' the sentence "A good strawberry is sweet and juicy" becomes analytically true for our strawberry expert.

This analyticity of the sentence "A C X is a good X." (e.g. "A sweet and juicy strawberry is a good strawberry.") can attack the lacuna between above propositions (C) and (D) of Hare's argument, since Proposition (D) rests upon a suppressed premise proposition (C*) (Walker 1973, p. 46).

(C) Sentence (1) "A C X is a good X." is typically used to commend X for being C; whereas

(D) Sentence (2) "A C X is a C X." is analytic, and therefore cannot be used to commend.

(C) No analytic sentence can be used to commend, and no sentence that can be used to commend can be analytic.*

The naturalist, however, may claim that proposition (C*) is in fact false, by giving the counter-example of a sentence which is both analytic and can be used to commend. The sentence "A sweet and juicy strawberry is a good strawberry" is analytic for our expert, and yet can be used to commend a strawberry. When the expert says that "These strawberries are good, because they are sweet and juicy," that is a recommendation of his. If proposition (C*) is false, then proposition (C)

also becomes false. I have argued so far, following the line of Walker's (1973) argument, that Hare's open question argument is invalid.

The above refutation of Hare's open question argument is just one kind. In addition, the essence of Hare's thesis itself is doubtful. Though Hare takes it for granted that commendatory use of moral terms is also part of their meanings, some others doubt that, as well as doubting whether ethical judgments are empirically verifiable or falsifiable (Stojanović 1963, p. 267). I will not go further into this 'analytic' analysis of open question arguments, concluding that it may be enough for my dissertation, and that the analytic criticisms of open question arguments are not conclusive. Rather, I turn to my original thesis that the naturalistic fallacy can be resolved from the viewpoint of morality as social software.

§7-4. Rescuing Value from the Naturalistic Fallacy and Hume's Law

The naturalistic fallacy is 'notorious.' Or, 'famous,' if you would. The naturalistic fallacy is poorly named (Williams 1985/2006, p. 121; Ridge 2003/2014); and, it is even worse than misnamed, I argue, that it is not a genuine fallacy. On Moore's accusation, the naturalistic fallacy is committed when we infer in the following way:

> X is natural property-like (e.g., X is sweet).
>
> ---
>
> Therefore, X is good.

This form of argumentation violates, exactly, Hume's law: 'value' conclusions (that contain 'ought,' evaluative, or prescriptive terms) cannot be deduced from solely 'fact' premises (that

contain ‘is,’ non-evaluative, or descriptive terms). As a simpler form, “no ‘ought’ from an ‘is’” (Hare 1981. P. 16; Putnam 2002, p. 28, p. 149). The naming of the naturalistic fallacy, actually, is not quite appropriate. As Moore himself argues, for example, divine command theories that find their moral foundations from God, which is a supernatural or non-natural entity, are also guilty of committing the naturalistic fallacy. The essential point of Moore’s various arguments is that, drawing moral value from any other property or fact, regardless of whether it is natural, non-natural, or supernatural, is to fail. And, this is exactly what Hume’s law claims.

Now, in order to reinforce my defense of morality as social software from the criticism of the naturalistic fallacy, I give a critique of Hume’s Law, an 18th century ancestor of the 20th century naturalistic fallacy. David Hume in Book 3 of *A Treatise of Human Nature* (1739/1978) criticizes the careless mistake of transitioning from ‘is’ to ‘ought,’ which is often found in moral systems. Let’s recall directly his monumental passage: [emphases are Hume’s]

> *I cannot forbear adding --- an observation---. In every system of morality, --- I have always remark’d, that the author proceeds for some time in the ordinary way of reasoning, and establishes the being of a God, or makes observations concerning human affairs; when of a sudden I am surpriz’d to find, that instead of the usual copulations of propositions,* ***is****, and* ***is not****, I meet with no proposition that is not connected with an* ***ought****, or an* ***ought not****. --- this* ***ought****, or* ***ought not****, expresses some new relation or affirmation ---;--- how this new relation can be a deduction from others ---. But as authors do not commonly use this precaution, I shall presume to recommend it to the readers; and am persuaded, that this small attention wou’d subvert all the vulgar systems of morality, and let us see, that the*

distinction of vice and virtue is not founded merely on the relations of objects, nor is perceiv'd by reason (Hume, 1739/1978: III.i.i).

This short passage has been "subversive," I think, in the sense, at least, that it has opened the way in which reason yields to emotion in the realm of moral discourse. Besides Hume's own moral theory, for example, Ayer's emotivism in the 20th century and Prinz's sentimentalist moral theory in the 21st century, both of which highly rely on emotions in their theories, are highly influenced by Hume's distinction between 'is' (fact) and 'ought' (value). These moral theories attempt to fill, with emotions, the gap between facts and values. I here use the is/ought distinction and the fact/value distinction, interchangeably.

Let's consider the following form of argumentation with the example of sugar and sweetness, to clarify this gap between facts and values:

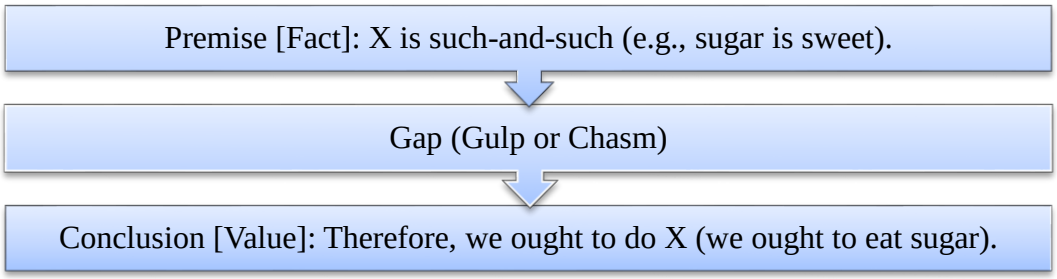


Diagram 3: The Gap between Facts and Values

There is a gap, gulp, or chasm between the fact premise and the value conclusion. Once premises are facts or believed to be facts, the instances of the fact premises may include a long list of various natural, non-natural, and supernatural properties, depending on the arguer's 'imagination' and 'ulterior motive': the fact premises are about, for example, natural medicine, vegetarianism,

environmentalism, homosexuality, racism, cannibalism, and 'so far too much more extreme' instances that I leave to the reader's imagination. The common form of argumentation using these value concepts are usually given this way: "There is (or is not) such-and-such a fact (often, found in nature), therefore, we ought (or ought not) to do that." This form of argumentation may be understood as an enthymeme that misses some premises, and those missing premises are possibly those that follow in the below diagram.

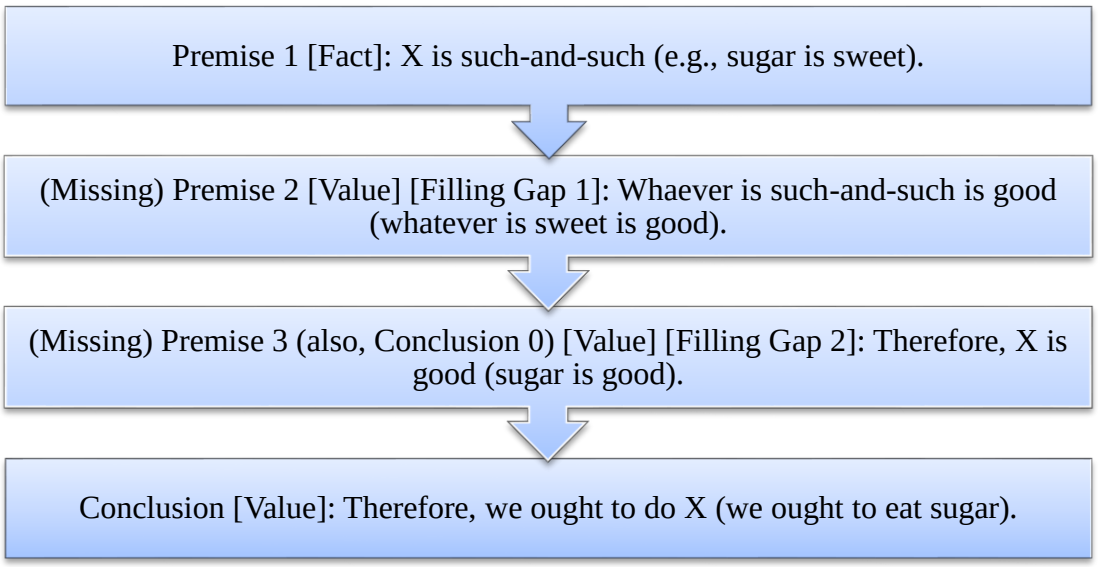


Diagram 4: The Gap between Facts and Values, and Missing Premises

The common essence of the criticism of Hume and Moore is that any attempt to fill the gap with such [Filling Gap 1] and [2] above, is never successful. Regarding their accusations that I interpret, any recovered missing premise used to fill the gap, such as premise 2 and 3 above, cannot avoid the question by the ultimate judge, "Then, are you a fact or a value?," and no matter what the answer is, the gap cannot be filled. On the contrary, some moral philosophers attempt to fill the gap with emotions that can be viewed as a different 'category' from the categories of facts and

values, ontologically. My argument based on the concept of morality as social software adopts a strategy different from those that use emotions in that way. To put this in rough terms now (more specifically later): there is no such gap between facts and values because the category of facts is different from the category of values; the problem of the fact/value dichotomy is a 'pseudo-problem'."

It is a really interesting and helpful insight that "a distinction is not a dichotomy," which Hilary Putnam discusses in *The Collapse of the Fact/Value Dichotomy and Other Essays* (2002, p. 9). Putnam interprets that John Dewey's real target of his criticisms "throughout virtually all of his long and exemplary career" "was not the idea that, ---, it might help to draw a distinction (say, between "fact" and "values"); rather his target was what he called [say] the fact/value "dualism"." Here, 'dualism' corresponds to 'dichotomy.' It is one thing to distinguish A from B; it is quite another to say that, in the universe, there are only two kinds, A and B, as the counterpart to each other. I think that this concept of "distinguishing between a distinction and a dichotomy" (Bernstein 2005, p. 253) is quite remarkable. Distinctions are "harmless" (Putnam 2002, p. 2) and "innocent" (p. 11); by contrast, dichotomies are "absolute" (p. 2) and "metaphysical" (p. 11).

One of the strategies that Putnam adopts in his book above in order to "collapse" the fact/value dichotomy is to connect it to the famous analytic-synthetic dichotomy, genealogically. Though I do not repeat his entire discussion here, let me give a précis of it briefly. In Putnam's discussion, the two dichotomies in their very inceptions had a close connection to each other in Hume's thought in the mid-18th century; Hume's initial claim of the dichotomy between "matters of fact" and "relations of ideas" was grounded by his claim that "no "ought" from an "is"," when an "is" judgment describes a "matter of fact"; this initial claim of Hume's affected Kant's development of the analytic-synthetic distinction (and also, in order to resolve a problem Hume raised, Kant

attempted to establish 'synthetic a priori' judgments); Kant's analytic-synthetic distinction became one of the pillars used by the logical positivist in the 20th century to establish their distinction between the meaningful and the meaningless; and finally, this analytic-synthetic distinction collapsed under Quine's attack in his famous essay, "Two Dogmas of Empiricism," in 1950s, which was around 300 years after Hume first developed these thoughts (Putnam 2002, Ch. 1 and 2). In this précis, I rephrase, with awe, what Putnam discusses: I think it is 300 years of a quite impressive history. Now, Putnam argues that the fact/value "dichotomy collapses in a way that is entirely analogous with the collapse of the analytic-synthetic dichotomy" (p. 8). If the fact/value dichotomy is untenable, then what kind of relationship between the two is possibly there? Putnam's 'choice' of the terminology is "entanglement": there is the "entanglement of facts and values" (Ch. 2). The word 'cruel,' is, Putnam argues, a good example that may show what he means by "entanglement." "Cruel simply ignores the supposed fact/value dichotomy and cheerfully allows itself to be used sometimes for a normative purpose and sometimes as a descriptive term" (p. 35).

On the one hand, unlike Putnam's genealogical strategy and, on the other hand, in a way similar to his analogical strategy, my argument to resolve the fact/value dichotomy will substantially rely on the methodology of "argument from analogy." I will address various kinds of dichotomies and contrasts, and, based on analogies and contrasts found, I will resolve the fact/value dichotomy.

7-4-1. The mind-body dichotomy

I think the discussions on the traditional mind-body problem shed some light upon the discussions of the fact/value dichotomy: the fact/value dichotomy can be analogous to the mind-body dichotomy. The traditional mind-body problem raises the question of how the mental mind is

connected to the physical body, where there are only two kinds of substances, the mental and the physical. All attempts to resolve the mind-body problem, I would say, have not been completely successful. Regardless of whether it is the pineal gland (Descartes), the brain, or qualia, which connect mind and body, all those attempted solutions have had to encounter the ultimate question that, "Then, is it itself the mental or the physical?," and their answers to this question have seemed to blur.

7-4-2. The human/God dichotomy, the human/Buddha Dichotomy, and the human/artificial intelligence dichotomy

I think this kind of difficulty found in the mind-body problem is common in the idea of dichotomy, the idea that there are only two kinds of such in the universe. Let me mention briefly another, the human being/God dichotomy, without mentioning many others such as the dichotomy between human being and soon-to-be human-like artificial intelligence. According to the typical Judeo-Christian-Islamic Western concept of God and human being, a human cannot become God, no matter how much he exerts himself. There is a chasm between the two, like that between the mental and the physical. By contrast, according to some Eastern religions (or philosophical thoughts) such as Buddhism, there is no such chasm between a common sentient being and ideal Buddha. A sentient being such as a human can become a Buddha (which literally means an "enlightened being"), as she exerts herself. Actually, according to the standard tenets of Buddhism, a sentient being's purpose of life itself is to practice in order to become a Buddha (or, to be enlightened to know that she is already a Buddha, having the Buddhahood inside, if you would like). In short, according to the Judeo-Christian-Islamic Western concept, there is an ontological gap between a human being and God; By contrast, for example, in Buddhism, there is no such gap between a

sentient being and Buddha, and that is why the historic figure, Siddhārtha Gautama, is regarded as a great revolutionary thinker. He resolves, or 'sublates' in a dialectical jargon, the gap between the two 'universes.'

Analogously, I argue that our current concern with the fact/value dichotomy can be resolved through a way similar to that of Buddhism. The fact/value dichotomy is 'definitely' similar to the sentient being/Buddha dichotomy, 'moderately' similar to the mind/body dichotomy, and 'never' similar to the human being/God dichotomy.

Here, comparing value judgments to fact statements is a "category mistake," since the category of value is not an independent substance or entity such as "gold or water" (Pojman 1990/ 2017 with Fieser, p. 230). I would argue that value judgments do not separately exist, say, in the Platonic world of forms or in a Kantian world of noumena. Rather, I would say, value judgments are something 'attached' to fact statements, 'afterward.' The common saying that "fact is fact and value is value and never the twain shall meet," (Putnam 2002, p. viii), which is often 'mistakenly' used to strengthen the fact/value dichotomy, I argue, is just a truism: the saying does not mention anything worthwhile; and also, the saying commits the category mistake: the category of fact statements and that of value judgments are different from each other, so that there is no way to intelligibly say that they do not go together. In short, I would argue that Hume's Law itself is a contention based on a category mistake.

7-4-3. The dichotomy between beauty and lack of that

In order to buttress my thesis that there is no such thing as value separate from fact, let me add two more contrasts, which perhaps I do not need to call 'dichotomies': 'beauty, or lack of that' and

'cleanness, or lack thereof.' A value judgment, 'She is beautiful'[62] can be made, on a typical evolutionary aesthetic theory, when the speaker recognizes consciously (or feels unconsciously) the symmetries of her body (face, breast, arms, legs, and the like). Through the process of evolution (say, deduction and induction, and, on contemporary popular terminologies, inferences similar to "learning from big data" and "deep machine learning"), humans (and our genes) have learned that: humans having more symmetric bodies are healthier (symmetry = health); healthier humans are more likely to give birth to healthier offspring (health = fecundity); and further, healthier individuals are more likely to have more productivity (fecundity = productivity). To sum up, with this kind of evolutionary aesthetic explanation, the final value judgment can be made, in the following order:

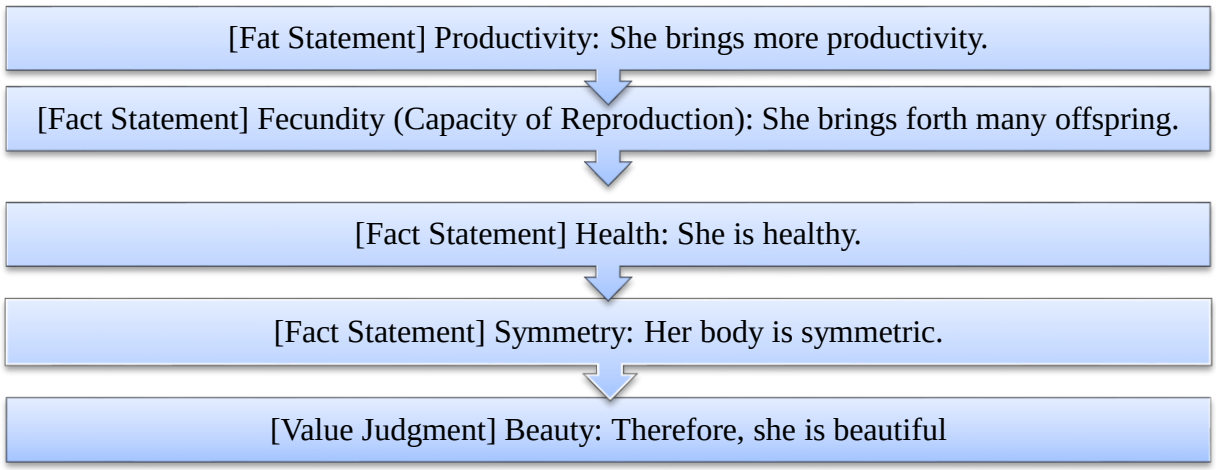


Diagram 5: From Productivity through Fecundity to Beauty

I think we can easily confirm the logic of this, say, 'algorithm,' as soon as we imagine a person who pledges not to have his/her own baby. Insofar as he/she pledges not to do so, (like fathers and

[62] Of course, we can develop the same kind of argument, using the sentence "Is he beautiful?"

nuns in Catholic Church, and monks (bhikkhu) and nuns (bhikkhuni) in some sects of Buddhism,) the value judgment of bodily beauty is not supposed to be applied to him/her any longer. The chain reasoning from productivity through fecundity to beauty can be broken. (Nevertheless, I am not claiming that they feel 'anesthetic aesthetically' in the real world of our everyday life.) Similarly, the human judgment of beauty is not applied to animals, like in the Ancient Chinese philosopher, Zhuangzi's (c. 369 BCE – c. 286 BCE) book:

> *Lady Li and Lady Mao were beauties in the eyes of men, but when fish saw them they swam down to the depths, when birds saw them they flew high, when deer saw them they bolted away at a gallop. Which of these four knows what is truly beautiful in the world?* (*Zhuangzi*, Ch. 2, 2.17)

If the value judgment of beauty is absolute, why do those fish, birds, and deer escape quickly from the beauty?

This human preference for symmetry can be extended from the human body to many other things such as painting, music, architecture, and Maxwell's equations in physics. (Yes, that Maxwell whom we discussed in chapter 4 regarding the Demon.) The mathematical symbols of the four equations of Maxwell's "convey to physicists a proportion, elegance, order and simplicity that is beautiful" (Chopra and Dexter 2008, p. 82).[63] However, unlike Chopra and Dexter who argue that the beauty of equations "are not based on their visual aspect, but on a deeper aesthetic quality" (p. 83), I think that the visual symmetric aspect is also one characteristic that constitutes beauty. Especially, the apparent symmetry of the four equations, (both differential and integral,)

[63] Chopra and Dexter discuss the beauty of Maxell's equations in the context of the aesthetics of computer programming code.

is so conspicuous that even someone who does not know anything about those mathematical symbols can notice the symmetry.[64] And many students in electrodynamics often feel beauty from the symmetry, even though it is not directly related to anyone's generative power.

The beauty of Maxell's equations, of course, may be explained by any other factors or any other aesthetics theories. Here, I am saying neither that evolutionary aesthetic theories are the only accurate ones, nor that this kind of explanation is the only accurate evolutionary aesthetic theory. Rather, I am just saying that understanding beauty in this way can show that the value judgment of beauty is not separated from the fact statements.

7-4-4. The dichotomies between cleanness and a lack of it, and between health and a lack of it

One last contrast: 'clean' and 'dirty' related to 'healthful' and 'unhealthful.' It is often regarded that 'clean' is connected to 'healthful'; 'dirty' to 'unhealthful.' But, is it always true? The digestive tract (gastrointestinal tract) of humans is a series of hollow organs composed of the mouth, esophagus, stomach, small intestine, large intestine, and anus. Here, the question is, 'Inside the digestive tract, is it clean or dirty?' What makes us think that, for instance, the mouth is clean, but

[64] One differential form of the equations can be written as follows:

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}$$

$$\nabla \cdot \mathbf{B} = 0$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

$$\nabla \times \mathbf{B} = \mu_0\mathbf{J} + \mu_0\varepsilon_0\frac{\partial \mathbf{E}}{\partial t}$$

Figure 6: A Differential Form of Maxwell's Equations

the other end is dirty? Through the 'metamorphosis' of the food that we eat, is it clean inside the stomach, but dirty inside the tube of the large intestine? One 'wondrous'[65] thing, to my eyes, that we learn from contemporary sciences and the philosophical reflections on them is that there is no absolute cleanness connected to being healthy, inside the digestive tract (See, e.g. Rhodes 2013, *The Human Microbiome: Ethical, Legal and Social Concerns*). Rather, on my interpretation and terminology, there is just a 'dynamic equilibrium' inside the digestive tract. For example, inside the large intestine, is a great ecosystem where there are various microorganisms, bacterium, and viruses living together. Seemingly, they are 'good, bad, or ugly,' 'clean or dirty,' and 'healthful or unhealthful,' but, according to contemporary discoveries and philosophical interpretations of them, those organisms just live together, often peacefully and some other times belligerently. Again, in my terminology, the 'dynamic equilibrium' is breached when the harmonious balance among the seemingly "good, bad, or ugly" is not sustained. Then, the clean state of the ecosystem becomes dirty, and the healthy human individual who 'temporarily shares the physical space (the body) with all those organisms'[66] gets sick.

I think that this kind of interpretation of 'clean,' 'healthful,' and 'healthy,' (which is originally found in the traditional Eastern (Indian Ayurvedic, Chinese, and Korean) philosophy of medicine, and re-discovered by the contemporary science and philosophy of human microbiome) can illuminate the current attempt to resolve the fact/value dichotomy. There are no independent value judgments such as 'my large intestine is dirty'; 'your mouth is clean (so, lovable), but not the other

[65] It has long been claimed that philosophy begins in wonder, since Plato and Aristotle. Let's recall Plato: "This wondering: this is where philosophy begins and nowhere else" (*Theaetetus*, 155d); and Aristotle: "It is through wonder that men now begin and originally began to philosophize; wondering in the first place at obvious perplexities, and then by gradual progression raising questions about the greater matters too, e.g. about the changes of the moon and of the sun, about the stars and about the origin of the universe. Now he who wonders and is perplexed feels that he is ignorant" (*Metaphysics*, Book 1, 982b, 12-21).

[66] In this regard, the individual may not be the 'owner' of the 'space and time,' that is, his body and life. I think this is another wondrous topic, but I don't go further here and now.

end'; and 'my stomach is absolutely healthful, so I am healthy.' Rather, there is only the 'dynamic equilibrium' inside those digestive organs.

These value judgments on cleanness and health are not far from value judgments on morality. Let's recall our discussion on entropy and morality in chapter 3. There, I argued that, after comparing the concepts of entropy, information, and equilibrium, anti-entropic morality is a state of dynamic equilibrium. I further introduced in chapter 3, as examples of morality as dynamic equilibrium, Aristotelian ethics of the Golden Mean for 'eudaimonia'(happiness), Confucius' moral philosophy of the Middle Way, and Buddha's Middle Way. These three great thinkers all point to the same state: the state of dynamic equilibrium of morality. I here add, as examples of dynamic equilibrium, the value terms, 'clean,' 'healthful,' 'healthy' as well as 'beautiful.'

7-4-5. An 'observation' as the final remark of this section:

I started this section, 'Rescuing value,' by introducing Hume's Law. Now, let me finish this section, by mentioning, briefly, the affinity between Hume's philosophy and the Buddhist philosophy. Those who are familiar with both philosophical thoughts may 'feel' some communality in them: Among others, I would say, Hume's philosophy and the Buddhist philosophy are subversive. That affinity can be no wonder, if Hume had some opportunity to encounter Buddhist philosophical views. Alison Gopnik, the psychologist-philosopher at UC Berkeley shows, carefully, that the encounter between the Western and Eastern philosophical views was possible when Hume met Jesuit missionary scholars at the Royal College of La Flèche between 1723 and 1740. Two articles: Gopnik's (2009) "Could David Hume Have Known about Buddhism? Charles Francois Dolu, the Royal College of La Flèche, and the Global Jesuit Intellectual Network," in *Hume Studies*; and, for lay people, "How an 18th-Century Philosopher Helped Solve My Midlife Crisis: David Hume,

The Buddha, and a search for the Eastern roots of the Western Enlightenment" in *The Atlantic* (2015 October). The word in the latter's title 'enlightenment' is notable.[67]

It is often regarded that *The Heart Sutra* crystalizes the entire lessons of Buddhism.[68] Let's savor some verses of *The Heart Sutra*, comparing them to Hume's philosophy and my discussions of beauty, cleanness, and health in this section.

Sariputra,

all Dharmas are marked with emptiness;

they are neither born nor destroyed,

neither impure nor pure,

[67] I don't have scholarship to examine whether this encounter happened, and such an examination may be beyond this dissertation. If the encounter really happened, it may seem surprising. However, on the contrary, it seems very possible, if we recall the historical fact that, already in 6th century or earlier, foreign traders from the city of Rome visited the East Asia. Trans-Eurasian trade through the Steppe Route and later the Silk Road is itself the history of human spread. Traders from Rome arrived at the capital of the Kingdom of *Silla* (57 BCE – 935 ADE), located at the Southern part of the Korean peninsula. To the other worlds including the Middle East, *Silla* was known as the Kingdom of Gold, like Eldorado. Those traders perhaps pursued profit margins by trading their products for gold of the Kingdom. In addition, *Silla* was also the Kingdom of Buddhism. The entire nation was under the absolute Buddhist culture after the kingdom's official approval of Buddhism in 527 ADE. The Buddhist culture and philosophy of *Silla*, (e.g. the monk, Wonhyo's Buddhist philosophy) together with those of *Baekje*, a neighboring country in the peninsula, affected Japan's nascent Buddhism. Then, it may be unnatural to infer that those traders and caravans from Rome and other Western regions traded their products for only gold and silk: some of their products must have been traded for Buddhist enlightenment, I imagine. Leibnitz in the 17th century constructed the binary system, 1 and 0, which is a foundation of the contemporary digital civilization, when he was inspired by the ancient Chinese book, the *I ching* (or *Book of Changes*), introduced to Europe by the Jesuit missionaries. John Locke, unlike Thomas Hobbes, held the view that the citizens have the right to resist authority: if a king is bad, the people are entitled to replace the king. We have evidence that Locke developed this 'radical' view through his encounter with Confucian political philosophy introduced to Europe at that time, and we know that the founding fathers of the U.S. read Locke. That means that Confucian political philosophy is subsumed within the constitution of the U.S. If we view the history of the philosophical interaction between the West and the East in this regard, the strict dichotomy between the two may not be so tenable. And finally, Hume's contact with Buddhist views may be no wonder.

[68] The title of the scripture means, etymologically, 'The Heart of the Perfection of Wisdom' (Prajñā-pāramitā-hṛdaya). We may have some sense of how much *The Heart Sutra* encapsulates the entire Buddhist philosophy, if we compare the amount of 'information' contained in *The Heart Sutra* and that in all other Buddhist canon. *The Heart Sutra* is a very short scripture: it can be 1 page-long; a translation into English contains around 300 words; and, a translation into Chinese contains 260 words. By contrast, the entire Pali or Sanskrit canon (Tripitaka) may be 15000 or 20000 page-long, perhaps.

neither increasing nor decreasing.

Therefore, in emptiness, no form, no feelings, perceptions, impulses, consciousness;

no eye, ear, nose, tongue, body, mind;

no color, sound, smell, taste, touch.[69]

I don't know whether Dharmas are "neither born nor destroyed," but I have argued that they are "neither impure nor pure, neither increasing nor decreasing," and there is no such thing as absolute beauty.

§7-5. Resolving the Gap, Using the Distinction between 'Ought Practical' and 'Ought Moral'

In the preceding section, the focus of my argument to rescue 'value' was that there is no absolute value judgments separate from fact statements. In this section, I argue for another way to help us resolve the problem, "No "ought" from an "is"." If there were such a gap, I argue that we may fill the gap with something, by analyzing various kinds of 'oughts' and other concepts.

When we consider Hume's claim that 'ought' cannot be derived from 'is,' we may ask what kind of 'ought' this is. Actually, there are many kinds of 'oughts' discussed and studied in the literature.[70] Here, it is not my intention to give a thorough review of those discussions and studies

[69] Again, I don't have scholarship to translate these verses from the Pali text nor the Sanskrit. I have some understanding in Korean, Ancient Chinese, and English. This translation is my 'concoction' of, mainly, the following two translations and some others, for my current context.

-The Stupid Way: About Zen Buddhism by "Peter," http://www.zen.ie/heartsutra.html.

-Kwan Um School of Zen, https://kwanumzen.org/resources-collection/2017/9/6/heart-sutra-in-english.

[70] For instance, Zimmerman (1996, Ch. 1) discusses an anatomy of various kinds of 'oughts,' suggesting the following diagram (See the next page footnote). Here, each right node's subspecies are omitted: e.g., 'nonmoral' ought is also divided into 'binding' and 'nonbinding,' and the like. When "a stranger approaches me on a street corner and politely asks me for a match" (p. 8), my action of giving the stranger a match can be, e.g., in the sense of 'nonbinding, prima facie, obligation-to' obligation, or in some other sense.

on the anatomy of 'ought.' Rather, I introduce the distinction between 'ought P (for practical)' and 'ought M (for moral),' and argue that Hume can be right in the sense of 'ought M,' but, not right in the sense of 'ought P.'

'Ought P' represents the practical ought that is needed for an agent (a group agent of a society or an individual agent) to maintain itself better. For example, the 'ought' in the expression, "In this country, all drivers ought to drive on the right side of the road" is 'ought P.' By contrast, 'ought M' represents the moral ought for an agent. For example, the 'ought' in the expression, "We humans ought to help other humans in need" is 'ought M.' Both 'ought P' and 'ought M' are still 'deontic,' in the sense that they are 'binding,' and so suggesting 'duty,' which is connected to the etymological meaning of the ancient Greek word, 'deon (δεον),' "that which is binding."

I argue that 'ought P' can be derived from 'is,' whereas 'ought M' cannot. There is no logical problem in the leap, for instance, from the following fact statement to the value judgment, using 'ought P':

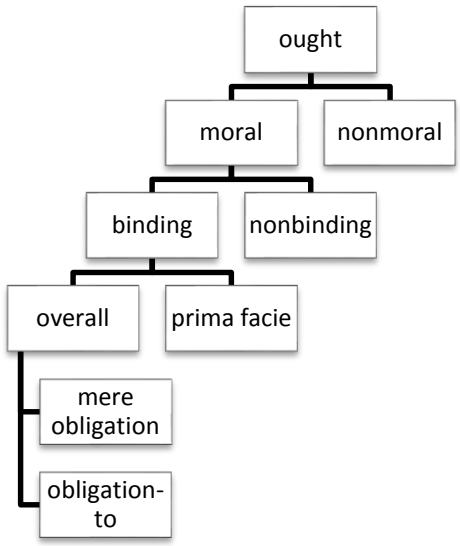


Diagram 6: A Chart of Many Kinds of Obligations
Source: Zimmerman 1996, The Concept of Moral Obligation, p. 10

In this country, all drivers drive on the right side of the road.

Therefore, in this country, all drivers ‘ought to’ drive on the right side of the road.

Moral agents including individual agents and group agents (e.g., societies) need to fulfill ‘ought P,’ in order to maintain their continuity. Those moral agents sometimes make use of ‘ought M’ of individuals, in order to secure certain kinds of ‘ought M.’ Two cases: a Marine Corps veteran resorts to his sense of ‘ought M,’ consciously or unconsciously, to save a fellow citizen when an elderly lady falls into a subway track, believing that he contributes to the ‘ought P’ of the society by his action (which can be ‘courageous,’ but still may seem ‘foolhardy,’ the opposite of ‘cowardly,’ on the Aristotelian doctrine of the Golden Mean); nations, especially during wartimes, resort to the people’s senses of ‘ought M’ to recruit soldiers, advertising that the people’s services will contribute to the ‘ought P’ of the nations.

In order to fill any gap, we should measure the distance between the two ends of the gap. If the gap, for instance, between the rich 1% and the poor 99 % has been exacerbated worldwide in this early 21st century, and if we want to mitigate it (though cannot fix it), we should first know how wide the gap is. Analogously, between ‘fact’ and value, or between ‘is’ and ‘ought,’ what gradations are there? I think the gradations may be explained with the following diagram:

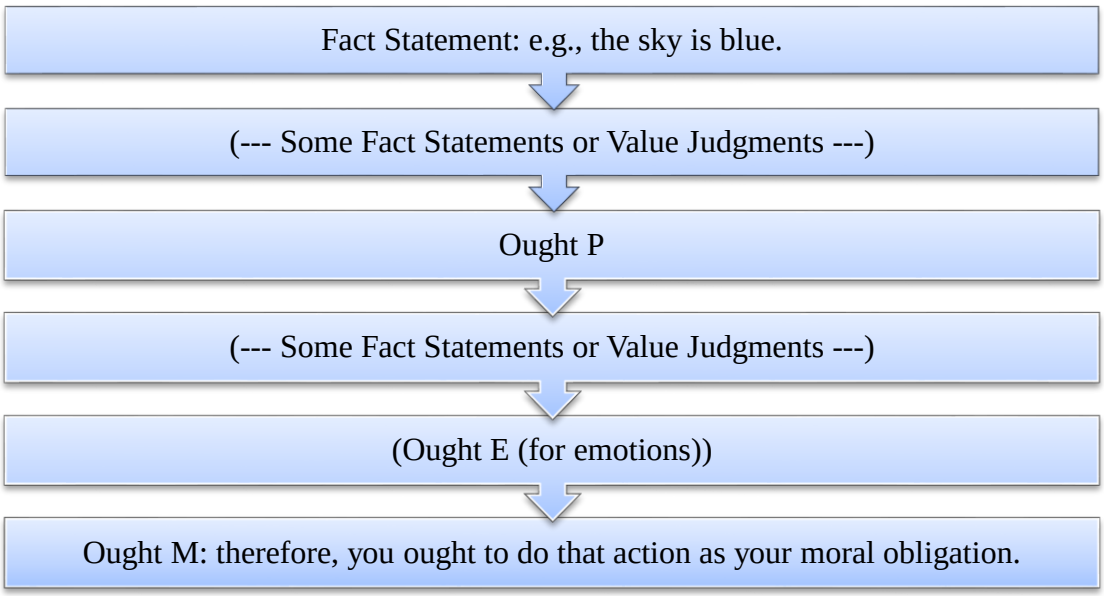


Diagram 7: From Fact Statements through Ought P and Ought E to Ought M

A moral agent deduces a conclusion of 'ought M' as her moral obligation, from a 'serene' fact statement, then through a series of fact statements and value judgments, including some 'ought P.'

In this process, philosophers relying on emotions, such as Ayer of emotivism and Prinz of sentimentalist moral theories fill the gap with emotions. When a nation recruits young soldiers, commercials often attempt to stimulate the young people emotionally: in such a way as this, "our enemies are invading us. Your country is waiting for your patriotism." We may call these Ayer-Prinz-like statements 'ought E (for emotion).' 'Ought E' can function as an intermediary between, among other oughts, 'ought P' and 'ought M.'

In contrast to 'emotion philosophers,' Putnam attempts to resolve the problem of fact/value dichotomy by introducing a new terminology, 'entanglement,' as we discussed above. Some other philosophers claim that moral judgments 'supervene' on factual statements, by introducing

Jaegwon Kim's famous concept 'supervenience,' which I don't discuss further here.[71] One more: I think, J. P. Sartre's famous claim, "existence precedes essence" can also be connected to our current discussion, even if not directly. By my contention, existence corresponds to value; essence corresponds to fact, though again, I don't discuss it further here.

In this regard, there is no such thing as value separated from fact. Rather, there is just something attached to various fact statements, and we humans call 'that something' various names: right or wrong, good or evil (or bad), beautiful or lack of that, clean or lack thereof, healthful or not, healthy or not, and the like. The ways of 'being attached' are also named in various terms: entangle, supervene, precede, and 'attach' (in my inadvertent usage). I have argued in this dissertation that this process of attaching value can be enlightened by the concept of morality as social software that includes the characteristics of evolutionary, anti-entropic, epistemic game-theoretic, cooperative, altruistic and the like. It is often 'preached' in Buddhism that the enlightenment is a process for a sentient being to become aware that there is no outside Buddhahood separated from the sentient being; that a sentient being is already a Buddha. Analogously, there is no independent value separated from fact; rather, we become aware that the 'value-hood' are already in the 'fact-hood,' and the 'value-hood' are realized as the forms of many 'oughts.' We may call these realizations 'moral sockets,' if 'moral genes' is too strong. Finally, we humans, not anything else, have called what has 'emerged' through all these processes, 'morality.'

[71] See Prinz's discussion on Blackburn's supervenience (Prinz 2009. pp. 4, 109-110, 144-152).

§7-6. How to Get Angry with the Reactionaries in the World, While Swimming without the Life Vest of the Naturalistic Fallacy

It seems to me that the naturalistic fallacy is a very powerful weapon[72] to attack some reactionary views based on naturalism. Or, passively, the fallacy is a life vest that saves us from the absurdities of those reactionary views. I mean, among others, some (if not all) claims of those views such as racism, sexism, speciesism, phrenology and social Darwinism in the 19th century, and (again, some claims of) sociobiology in the 20th century. The list of those views can be elongated to include many tragic examples from history. If we endorse the naturalistic fallacy unlike my argument in this dissertation, then we can easily dismiss many reactionary arguments.

In order to confirm the power of the naturalistic fallacy, let's consider an example of vegetarianism (or veganism): a very milder one, compared to others, I think.[73] A commonplace argument for eating meat can go as follows:

---

[72] I endorse the view that the emergence of philosophy in the West, around 580 BCE in the cities of Asia Minor, was related to the severe class struggle between the rich and the poor. On this view, philosophy was a much stronger 'weapon' to persuade other fellow citizens than swords. I never forget what Bertrand Russell describes and quotes in the beginning of *A History of Western Philosophy* (1945/1972, p.24):

> *"Thales was a native of Miletus, in Asia Minor, a flourishing commercial city, in which there was a large slave population, and a bitter class struggle between the rich and poor among the free population. "At Miletus the people were at first victorious and murdered the wives and children of the aristocrats; the aristocrats prevailed and burned their opponents alive, lighting up the open spaces of the city with live torches." (Rostovtseff, History of the Ancient World, Vol. 1, p. 204) Similar conditions prevailed in most of the Greek cities of Asia Minor at the time of Thales."*

The historian, Rostovtseff's description is appalling. If the people of the cities in Asia Minor lost the struggle, they had to lose their lives and families. They must have been desperate and found that philosophy was a stronger weapon than others such as swords and spears. I see that this view of philosophy has been inherited through the history of philosophy, and the recent culmination of the view is Karl Marx's criticism in his "Theses on Feuerbach" (1845, XI) that "The philosophers have only interpreted the world, in various ways. The point, however, is to change it." Since I endorse this view of philosophy, I am a citizen of the city of Miletus.

[73] I would not like to call meat-eating people reactionaries. That designation may be, I believe, too over-used in Western countries in the year of 2018. I am a meat-eating person, too. Here, I just take advantage of the simplicity related to the vegetarianism debate, for my current context.

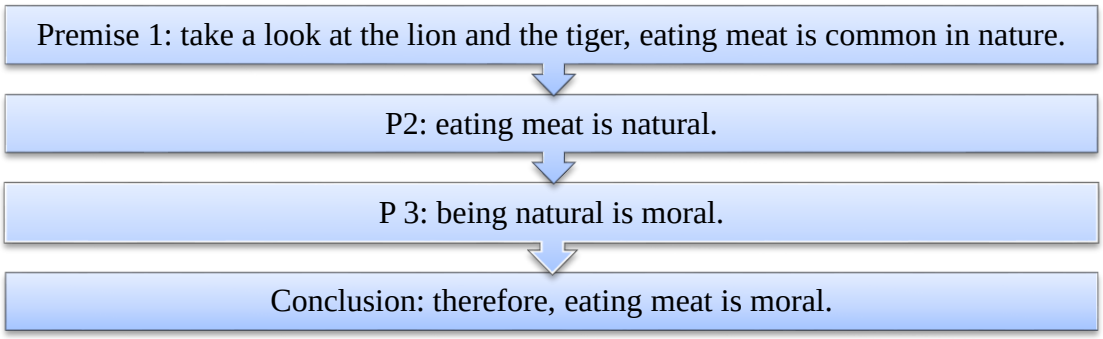


Diagram 8: The Naturalistic Argument for Eating Meat

The conclusion can be interpreted in many ways: for instance, eating meat does not violate any moral rule; or, in a stronger way, we ought to eat meat in order to take good care of our health. A vegetarianism activist, on the contrary, can easily refute this argument by pointing out that this argument commits the naturalistic fallacy due to premise 3 (being natural is moral), provided that the activist endorses the naturalistic fallacy. If he does not, he should find some other counter-arguments.

So, if we throw away the weapon of the naturalistic fallacy, it may seem that we have to surrender to some 'ostensibly' attractive claims of those reactionary theories. I confess that, when my understanding of moral discourses was very limited decades ago, some of those claims that I encountered seemed surprisingly easy and attractive. Now, I would argue that we should not, and ought not to surrender to them, and we don't need to do so, since we have another weapon of morality as social software. This understanding of morality as social software that I have developed so far includes the characteristics of evolutionary, anti-entropic, epistemic game-theoretic, cooperative, altruistic and the like. I argue that these features of morality, when those reactionary views run amok, can help to curb them.

Chapter 8. Concluding Remarks

I conclude the dissertation by mentioning the topic of ethical relativism, which is seemingly necessary but has been neglected so far; and some future research topics, which have been recognized, but not developed in my dissertation research, and still look promising. And finally, some last words.

§8-1. Ethical Relativism and the Direction of History

The topic of ethical relativism in moral discourses is, I think, everything, and therefore, nothing. It is everything because any talk of morality, ultimately, should deal with the question: how fundamentally different moral systems can be compatible or compete with one another. I think that since it is everything, the authors of many basic textbooks on ethics and philosophy spend much of their energy discussing the topic of ethical relativism (e.g., Pojman 1990/2017, Rachels 1986/2011, Rauhut 2003/2010, Lackey 1989, Williams 1972/2005, 1985, Prinz 2007). Conversely, I think, for the same reason that ethical relativism is 'ultimate and fundamental,' this topic should be nothing in moral discourses. Ethical relativism, I think in figurative speaking, is like 'air,' which exists everywhere around us without the need to be recognized; ethical relativism is not like 'aether,' which was postulated as the medium for the propagation of light, instead of a vacuum, but was proven not to exist by the famous Michelson-Morley experiment in 1887, and ultimately resulted in Einstein's theory of relativity in 1905. Following this thread of "everything, therefore, nothing," I mention briefly my view of ethical relativism to conclude the dissertation.

When it comes to 'ethical' relativism, there is no way not to deal with 'cultural' relativism at the same time. And, when we discuss these two concepts, I think, John Ladd's distinction between

the ‘diversity thesis’ and the ‘dependency thesis’ is useful (1973/1985, p. 3). Let’ start with this distinction:

> *The diversity thesis: throughout the world and through history there has always been an irreducible diversity of cultural patterns, institutions, economy, language, personality, and so on, as well as a diversity of moral beliefs, rules, and practices.*
>
> *The dependency thesis: the moral beliefs, rules, and practices of a society are necessarily and invariably dependent for their validity on other facets of the culture - for example, its institutions, its economy.*

Cultural relativism is related to the diversity thesis. Cultural relativism asserts plainly that different cultures have different moralities (moral standards, codes, or principles). It is a simple ‘description’ or observation in anthropology and sociology. It is not possible (in our world, not in an imaginary twin Earth) that this description is false: it is undoubtedly true. By contrast, ethical relativism is related to the dependency thesis. Ethical relativism asserts that the validity of the morality in a culture (or society) depends on (or is relative to) the acceptance of the culture (or society). Ethical relativism is like a ‘prescription’ in moral philosophy.[74] It is possible that the assertion itself and its consequences such as, “all moralities in all cultures are equally valid” can be false, and indeed, many people claim that this consequence is false, with which I do not agree.[75]

[74] I am relying on the common contrast between prescriptive moral philosophy and descriptive sociology and anthropology.

[75] Here, there is a little more complicate ‘taxonomy.’ Some (e.g., Pojman) claim that both of the diversity and dependency thesis are ingredients of ‘conventional’ ethical relativism. In this case, the conventional ethical relativism can be called the ‘broad-sensed’ ethical relativism, while only the dependency thesis is called the ‘narrow-sensed’ ethical relativism. And, cultural relativism is a subset of (‘broad-sensed’ conventional) ethical relativism.

The framework of morality as social software that has been developed in this dissertation research inevitably implies, to some degree, ethical relativism. If morality is created by human beings, it can vary from place to place, and time to time. Ethical relativism, nevertheless, does not necessarily imply that we don't have morality: That is, ethical relativism does not necessarily imply ethical nihilism. I have argued that the morality of an individual moral agent or a group of moral agents is a state of dynamic equilibrium, and the state is like, as a metaphor, the state of the trembling notch of an analogue scale. Though the notch is trembling, it is not the case that the notch does not point to a specific number of weight: the notch always points to a state. Analogously, while trembling, the morality of an individual agent or a group of agents always points to a certain state (again, like a differential state in calculus).

This concept of morality as dynamic equilibrium is crucial in my arguments for ethical relativism in two ways. Firstly, we can still compare one moral rule to another, talk about tolerance and respect for other moral rules, and criticize moral rules both our own and others', though we endorse ethical relativism, without resorting to ethical absolutism (or moral objectivism). A

Some others (e.g. Rachels) claim that both theses are ingredients of 'cultural relativism,' so that ethical relativism (of the dependency thesis) is a subset of cultural relativism. It is not my intent to be involved in this complication in this dissertation. For my current purpose, it may be enough to distinguish the concept of descriptive cultural relativism and prescriptive ('narrow-sensed') ethical relativism. (See the diagrams below.)

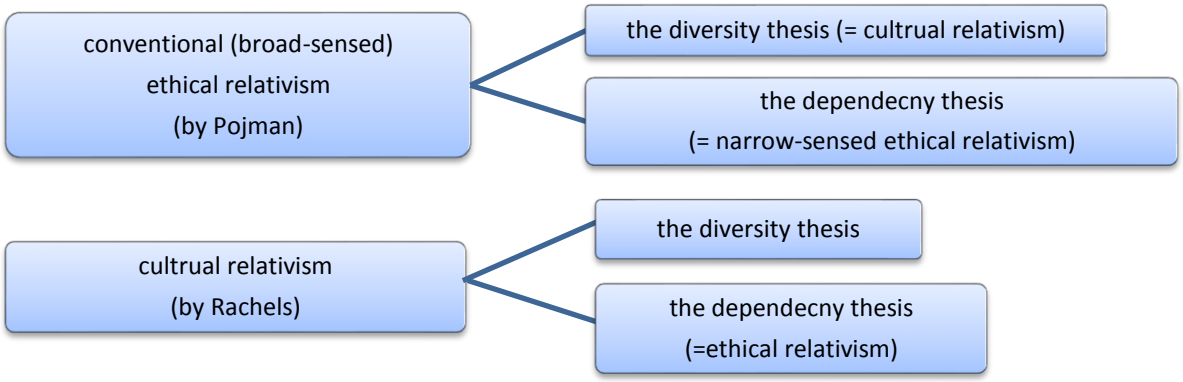


Diagram 9: A Taxonomy of Ethical Relativism and Cultural Relativism

version of ethical absolutism (or moral objectivism) states that there is only one 'universal,' absolute, and objective set of moral standards that everyone should always follow everywhere, which I assume is definitely false. Secondly, though we endorse ethical relativism, we don't need to worry about the attacks from moral nihilism (or amoralism), which is the antipode of ethical absolutism (or moral objectivism). A version of moral nihilism (or amoralism) states that there is no such thing as morality, and even if there are moral facts and truths, we cannot know them. With these two crucial points (i.e. possible tolerance and impossible moral nihilism), again, we achieve a middle way by avoiding the two extremes.

One of the natural consequences of ethical relativism is that we should be tolerant and respectful of the moral standards of other people and societies. It seems to me that, together with endorsing ethical relativism, the virtue of tolerance and respect for others has become one of the main ingredients of global ethics in the early 21st century. But, how far ought we to be tolerant and respectful of others? What is the difference between hate speech and legitimate political freedom of speech? Logically, just as we cannot respect others to an extreme (then, we lose our identity), so we cannot tolerate others to an extreme (then, again, we lose our identity). We should find the optimal level of tolerance and respect.

Regarding the optimal level, how far ought we to be tolerant and respectful of the moral beliefs of others? Let's consider the discussions by Prinz (2007) and Williams (1972/2005). Prinz starts his discussion on 'moral relativism' (his term for what I refer to as 'ethical relativism') quite dramatically by quoting from Tobias Schneebaum's memoir about his experience living among cannibals in the remote Amazon jungle. The title of that chapter of Prinz's book is "Dining with Cannibals," which seems quite, I would say, 'ambivalent.' Can we or should we be tolerant and respectful of our cannibal companions at our dinner tables? Certainly not, in the early 21st century.

Prinz writes, “To call cannibalism immoral is an understatement. It is morally monstrous” (2007, p. 173), while acknowledging, “Yet cannibalism has been practiced by cultures all over the world. It is, by some accounts, the default cultural practice, rather than the exception.”

Prinz’s defense of ethical relativism is, like his moral theory previously discussed in section 3-6, based on emotions: “Moral relativism is a straightforward consequence of the sensibility theory” (2007, p. 173). He introduces the distinction between ‘descriptive moral relativism’ and ‘metaethical relativism.’ His version of descriptive moral relativism, which holds that “some people have fundamentally different values” (Ibid.), corresponds to the diversity thesis above. His version of metaethical relativism, which holds that “The truth conditions of a moral judgment depend on the context in which that judgement is formed,” (pp. 173-174) corresponds to the dependency thesis above. Prinz sees that the “master-argument” for the metaphysical thesis is as follows: (p. 174)

Premise1: Descriptive relativism is true.

Premise 2: If descriptive relativism is true, then metaethical relativism is true.

Conclusion: Therefore, metaethical relativism is true.

Premise 1 is true obviously, as we discussed above. Whether premise 2 is true or not is crucial to the validity of this argument. Prinz argues that premise 2 is defensible if the sensibility theory that he favors is true. On Prinz’s view, to sum up, people have different moral sentiments toward the same things (for example, different sentiments toward cannibalism and meat-eating); and the differences in people’s sentiments “entails” a difference in moral facts, “if moral rightness and wrongness depend, metaphysically, on the sentiments people have”; therefore, metaethical

relativism can be derived from descriptive relativism (pp. 174-175). This 'metaethical relativism' is the 'general' ethical relativism that I discuss in this section, and now, is defended by emotions.

Bernard Williams' view on ethical relativism in his *Ethics: An Introduction to Ethics* (1972/2005) is different from Prinz's and mine. Williams criticized a view that "was popular with some liberal colonialists, notably British administrators in places (such as West Africa)" (p. 20). According to the view, those administrators following ethical relativism, had no business to interfere with the culture of the aboriginal tribes, even though the culture included 'human sacrifice.' On Williams' terminology, relativism is "the anthropologist's heresy, possibly the most absurd view to have been advanced even in moral philosophy," and the three components of relativism are 1) our dependency thesis above ("'right' means --- 'right for a given society'"), 2) a functional understanding of 'right for a given society,' and 3) no condemnation and interference with others (p. 20). It seems to me that Williams did not endorse ethical relativism, by rejecting 3) no condemnation and interference with others, especially, by criticizing the administrators for not interfering with the practice of human sacrifice. I absolutely agree with Williams that human sacrifice must not be allowed in the modern world; but I think that the rule 3) not to condemn or interfere, should be observed mostly, if not always; and finally, I embrace the theses of ethical relativism without hesitation in 2018; it seems to me quite odd that a renowned Anglophone moral philosopher criticized ethical relativism in 1972. I think, in 2018, that human sacrifice should never be tolerable; by contrast, eating beef and pork can be tolerable when Hindus eat pork, but not beef, whereas Muslims eat beef, but not pork (See Lackey 1990, p. 83).

Then, what makes me think this way? The answer could be related to the concept of 'the direction of history.' I argue that the human history moves towards a certain direction driven by a

certain force. I am cautious about claiming that the direction is towards 'progress'; I would not like to claim that there is even 'moral progress' in history, so humans in the 21st century are more moral than humans in the 5th century BCE. Rather, I argue that there is a direction in history. I cannot discuss here the grand theme on the direction of history in the philosophy of history, by considering Herodotus' *The Histories* in Ancient Greece and Sima Qian's *Records of the Grand Historian* in ancient China. Rather, I would introduce one point of the direction of history.

Daniel Nettle (2010) in his "Dying young and living fast: variation in life history across English neighborhoods" published in the journal, *Behavioral Ecology*, argues that English women have their first birth later and fewer children, as the quality of the women's neighborhood environment becomes more affluent.

More specifically, the number of years of good health an English woman (excluding those living in Scotland, Wales, and Northern Ireland) expects, decreases rapidly from (approximately) 70 to 50, as the quality of the woman's neighborhood environment decreases from 10 to 1: the worse the neighborhood quality is, the less the healthy life expectancy is.

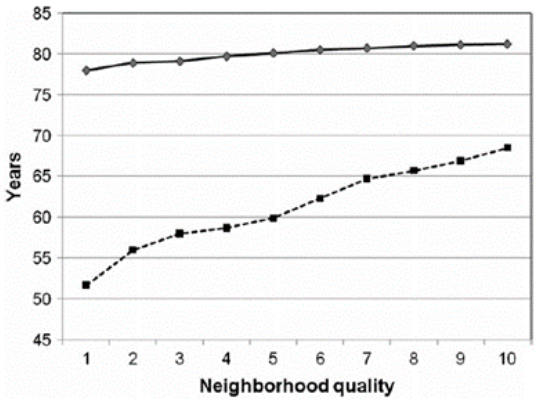


Figure 7: Female Total Life Expectancy (solid line) and Healthy Life Expectancy (dashed line); Source: Nettle 2010, "Dying young and living fast: variation in life history across English neighborhoods"

When they are expected to die young, they follow a "fast life-history strategy of early reproduction": the earlier age when they have their first birth, and the more children they have.

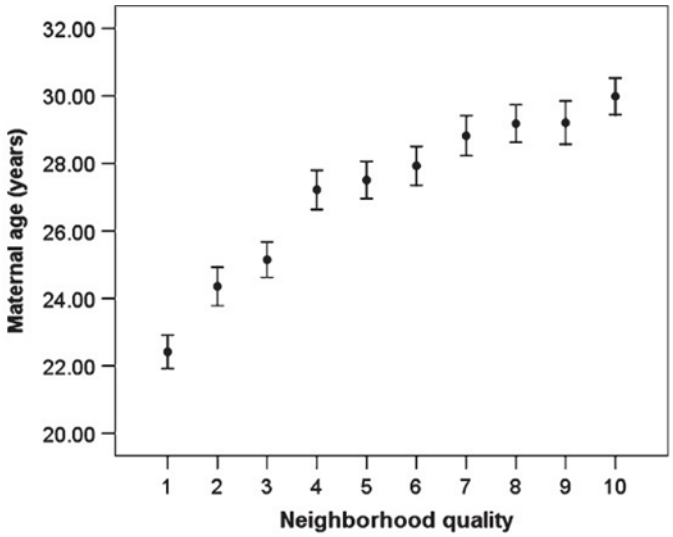


Figure 8: Maternal Age versus Neighborhood Quality

Source: Ibid. (Nettle 2010)

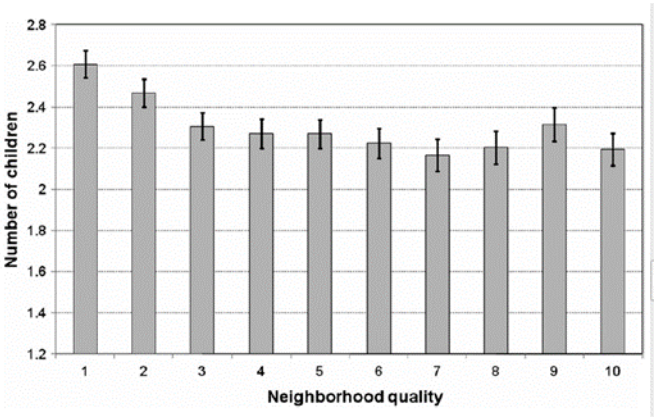


Figure 9: Number of Children versus Neighborhood Quality

Source: Ibid. (Nettle 2010)

I interpret this research result as a clue showing that there is a direction of history driven by a certain force. I understand that, as the quality of women's lives becomes more affluent in history, women want to have fewer children (and realize themselves with other professions). While this research is 'synchronic' among English women, I can introduce a 'diachronic' record among Korean women. My two grandmothers who were born around the year of 1900 gave birth to, respectively, 11 babies and raised 8 of 11 to adulthood, and gave birth to 7 babies and raised 5 of 7 to adulthood; my mother, born in the 1940s, gave birth to 3 and raised all to adulthood; my two sisters-in-law, both born in the 1970s, gave birth to only one child each. As the economy has developed in the country, women have found meaning in their life, from not just raising children, but from others. I argue that this is a direction of history.

Let's see an example of women's suffrage. J. S. Mill completed his manuscript of *The Subjection of Women* in 1861 and first published it in 1869 to advocate legal and social equality for women, including women's right to vote. In the United Kingdom, women got the right to vote partially in 1918 and fully in 1928.[76] In the United States, suffrage was granted to women in 1920. In countries around the globe, though earlier or later, women have gotten the right to vote throughout the 20th century, and recently in 2015, women in Saudi Arabia. As of now in 2018, any 'foolhardy' politician in the world who attempts to repeal women's suffrage in the country will lose his voters in the speed of light. Women's right to vote is now irreversible, and this is, I argue, a direction of history.

On animal welfare and rights, in 1789 when Jeremy Bentham asked questions, "The question is not *Can they reason?* or *Can they talk?* but *Can they suffer?*" (p. 144), he might not have had many supporters. In 2018, 229 years later, he has many advocates and followers. When Aristotle,

[76] Here, dates are from the Wikipedia entry, "Women's Suffrage," https://en.wikipedia.org/wiki/Women%27s_suffrage

the great moral philosopher 2400 years ago, advocated natural slavery in his *Politics*, he may not have had many critics. But, in 2018, no sane moral philosopher may advocate natural slavery. When a moral philosopher in 2018 tries to advocate the welfare of Artificial Intelligence (AI), and asks, "the question is, *Can they suffer?*," perhaps she does not have many supporters: The current 'slavery' between human beings and computers (AI) is accepted by many humans without any guilty conscience. What will happen in 229 years (the time gap between Bentham and us now) when many AI programs will have already reached the so-called 'superintelligence' (Bostrom 2014) or 'singularity' (Kurzweil 2006, Chalmers 2010, 2012)? When AI programs reach the state of superintelligence or the singularity, the faculty of AI programs will have surpassed most human faculty (as the AlphaGo has shown in the board game, Go); in addition, perhaps, I imagine, AI programs will enslave human beings, or sit under the Bodhi Tree in order to attain enlightenment, after being tired of all mundane things. Let me curb my imagination here. My point with these three examples (animal right, slavery, and AI) is the same: there is a certain irreversible direction of history driven by a certain force.

For one last case that supports my argument for a direction of history: the Cultural Map of the World Values Survey. The surveys have been done six times from 1981 to 2015 among the peoples of countries worldwide, asking peoples' beliefs in various value topics such as religion, politics, family, and happiness. Based on the survey results, intriguing global cultural maps have been drawn. The following 'Inglehart-Welzel Cultural Map' is the most recent one from 2015. In this map, values are divided into two dimensions of four categories: the vertical axis represents the ratio of traditional values vs. secular-rational values; the horizontal axis represents the ratio of survival values vs. self-expression.[77]

[77] "Traditional values emphasize the importance of religion, parent-child ties, deference to authority and traditional family values," so that "people who embrace these values also reject divorce, abortion, euthanasia and

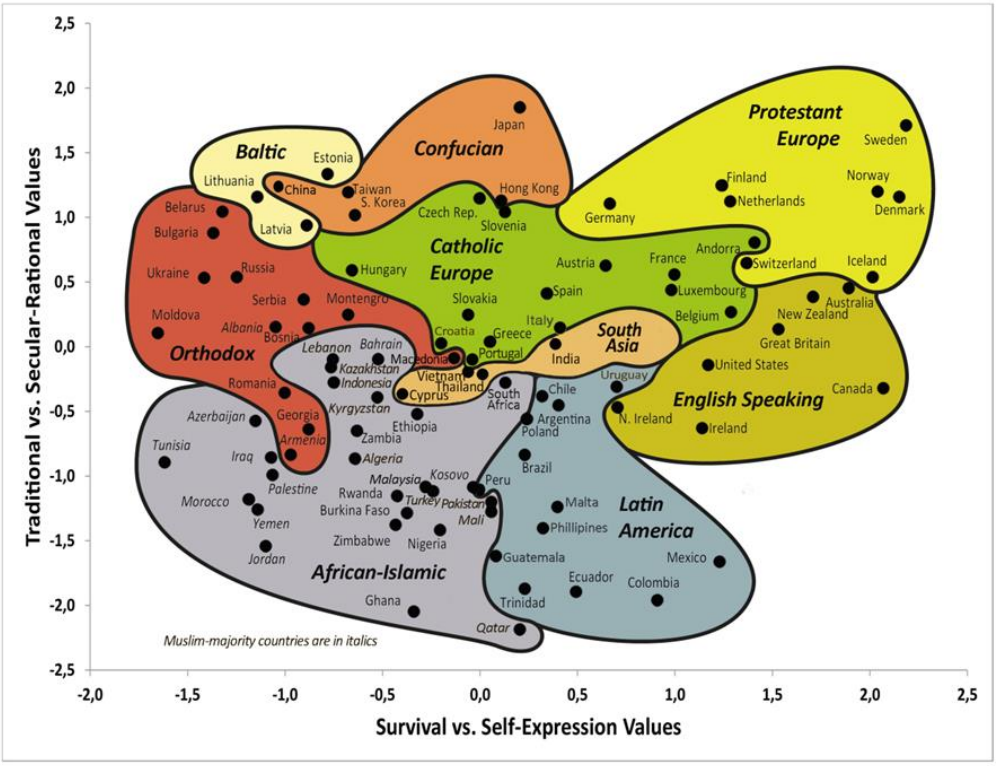


Figure 10: World Cultural Map

Source: The World Values Survey Site, http://www.worldvaluessurvey.org

The global cultural map shows the locations of societies based on their scores on these two dimensions. Moving upward means the shift from traditional to secular-rational; moving rightward means the shift from survival to self–expression. In the map, if we compare the six locations of a society in the six surveys, now we can have a line; if we see the six locations in a moving picture (as the authors show), we can have a dynamic movement of locations. I dare to say that this line

suicide." By contrast, secular-rational values are the opposite preferences. "Survival values place emphasis on economic and physical security. It is linked with a relatively ethnocentric outlook and low levels of trust and tolerance." By contrast, "Self-expression values give high priority to environmental protection, growing tolerance of foreigners, gays and lesbians and gender equality, and rising demands for participation in decision-making in economic and political life." http://www.worldvaluessurvey.org/WVSContents.jsp "Findings and Insights."

and movement is a direction of history. As time goes on, and as the condition of a society, such as the economy, changes, the beliefs of the members of that society move toward a certain direction by certain driving forces.

Finally, I do not think that it is reasonable to imagine that the great ancient sages such as Socrates, the Buddha, and Confucius did not consider me, a feeble sentient being who will come 2500 years later. While Socrates 'corrupted' the youth at the Ancient Agora of Athens, the Buddha travelled with his disciples to preach by mendicancy in humid India, and Confucius travelled with his disciples to share his virtue ethics and the art of harmoniously living with others through the yellow sands of windy China, I believe they all must have considered the entire human race including the future beings who will arrive 2500 years later. Then, it must be reasonable for us, in the year of 2018, to consider the sentient beings who will follow in the next 300 years, if not 2500 years.

I have so far discussed a direction of history in order to answer the question, "How far ought we to be tolerant and respectful of moral beliefs of others?" relating to ethical relativism. When we look 300 years ahead, it seems certain that a culture practicing human sacrifice will not be widespread among human societies. Therefore, now, we ought not to be tolerant of human sacrifice. By contrast, while I am almost certain that eating beef or pork will be replaced by many alternatives that are being developed (such as insect food and lab-grown (cultured, in vitro) meat) within the next 300 years, I am not certain that will happen within next 30 years. Therefore, I can be tolerant of meat eating by others and myself. In this way, we can swim with ethical relativism, which is not dreadful, but inevitable.

§8.2 Future Research

The following topics are some future research projects that I have found intriguing but did not follow up while doing my current research. I write them here as if to carve them on a stone.

-**Teaching Artificial Intelligence (AI) morality as social software**: I have already argued for this topic in various sections of the dissertation. One conjecture: when coding morality into AI, the axiomatic approach that provides several axioms of moral rules for AI (which can be compared with the continental 'civil law' system) may be harder than the approach that can be compared with the Anglo-American 'common law' system. As soon as axioms are coded, the AI may find contradictions immediately.

-The early intellectual interaction between the originators of the concept of entropy and the concept of evolution in the mid-19th century: the concepts of evolution and entropy, which are central to this dissertation, were developed in a similar time and place, mainly in the mid-19th century in Scotland, England, and broadly Europe. Those scientists and philosophers of the two ideas of evolution and entropy that I have discussed in this dissertation must have been aware of the others' works: Darwin, Huxley, Maxwell, the Lord Kelvin (William Thomson), Boltzmann (in Austria), and others. Especially, I conjecture that, in the volatile Scottish weather, there must have been deep intellectual interactions between the two groups, since there are many of those originators were Scottish.

- How many times should we forgive? Answer: 2.6 times: as we discuss in chapter 5, the tit-for-tat strategy was the strongest one in Axelrod's many game situations of the iterated Prisoner's

Dilemma. The tit-for-tat strategy starts with cooperation, and thereafter do what the other player did on the previous move. Then, we may have questions such as: "How many times does the tit-for-tat forgive?," "How many times does the tit-for-two-tats forgive?" (The strategy that defects only when the other player has defected twice in a row.) And generally, "How many times should we forgive, in the general life setting, in order to produce the best outcome for the moral agent herself, or for both of herself and her opponent?" I think both theoretical (mathematical) and empirical research into this topic may be possible. Recall Jesus Christ's statement on forgiveness: we should forgive the sinner as often as seventy times seven. Hongwu (1328-1398), the founding emperor of China's Ming dynasty held the view that, for the first defect, forgive it with patience; for the second one, forgive it with tolerance; but for the third one, revenge it. Together with the philosopher Miranda Fricker's reflective works on blame and forgiveness, my game-theoretic approaches, again, both theoretical (mathematical) and empirical, will be a contribution. The assumption that I have in mind is that not all of us can become a religious saint; we are a victim of our insatiable desire. Temporarily, I have a conjecture that the answer might be 2.6 times of forgiveness, based on some empirical observations.

§8.3 The Last Words

If your views are different from mine, we can "agree to disagree" (Aumann 1976). We can make progress by suggesting bold conjectures and refuting them (Popper 1962), and by killing hypotheses instead of lives (Popper 1977). Based on the research in this dissertation, now I have my own answer to a question that I raised in section 1.1. It is not morally right: Do not burn yourself; rather, be a flower.

Bibliography

Aristotle *Metaphysics.*

Aristotle *Nicomachean Ethics*.

Aristotle *Politics*.

Arrow, Kenneth J. (1963) *Social Choice and Individual Values*, Hoboken: John Wiley & Sons.

Ashford, Elizabeth and Tim Mulgan (2012) "Contractualism," *Stanford Encyclopedia of Philosophy*.

Aumann, Robert (1976) "Agreeing to disagree," *Annals of Statistics* 4, pp. 1236-1239.

Aumann, Robert (1995) "Backward Induction and Common Knowledge of Rationality," *Games and Economic Behavior* 8, pp. 6-19.

Aumann, Robert (1999a) "Interactive epistemology I: Knowledge," *International Journal of Game Theory* 28(3), pp. 263-300.

Aumann, Robert (1999b) "Interactive epistemology II: Probability," *International Journal of Game Theory* 28(3), pp. 301-314.

Axelrod, Robert (1980a) "Effective Choice in the Prisoner's Dilemma," *Journal of Conflict Resolution* 24, pp. 3-25.

Axelrod, Robert (1980b) "More Effective Choice in the Prisoner's Dilemma," *Journal of Conflict Resolution* 24, pp. 379-403.

Axelrod, Robert (1981) "The Emergence of Cooperation among Egoists," *American Political Science Review* 75, pp. 306-18.

Axelrod, Robert (1984/2006 Revised edition) *The Evolution of Cooperation*, New York: Basic Books.

Axelrod, Robert and William D. Hamilton (1981) "The Evolution of Cooperation," *Science* 211, pp. 1390-96.

Barney, Rachel (2004/2011 Revision) "Callicles and Thrasymachus," *Stanford Encyclopedia of Philosophy*.

Başkent, Can (2017) "A Non-classical Logical Approach to Social Software," in Başkent, Moss, and Ramanujam (2017).

Başkent, Can, Lawrence S. Moss and Ramaswamy Ramanujam (2017) *Rohit Parikh on Logic, Language and Society*, Cham, Switzerland: Springer.

Başkent, Can, Lose Olde Loohuis, and Rohit Parikh (2012) "On Knowledge and Obligation," *Episteme*, Volume 9, no. 2, 171-188.

Baumrin, Bernard H. (1968) "Is There a Naturalistic Fallacy?," *American Philosophical Quarterly* 5 (2), pp. 79-89.

BBC Horizon documentary program (1986) *Nice Guys Finish First*, presented by Richard Dawkins.

Ben-Naim, Arieh (2015) *Information, Entropy, Life and the Universe: What We Know and What We Do Not Know*, World Scientific Publishing Company.

Ben-Naim, Arieh (2017) "Can entropy be defined for, and the Second Law applied to living systems?," Cornel University Library, https://arxiv.org/abs/1705.02461, May 6.

Bentham, Jeremy (1789) *An Introduction to the Principles of Morals and Legislation*.

Bergson, Henri (1907) *Creative Evolution*, originally in French.

Bernstein, Richard J. (2005) "The Pragmatic Turn: The Entanglement of Fact and Value," in Yemima Ben-Menahem (2005), ed. *Hilary Putnam*.

Boltzmann, Ludwig (1877/2015 English translation) "On the Relationship between the Second Fundamental Theorem of the Mechanical Theory of Heat and Probability Calculations Regarding the Conditions for Thermal Equilibrium," translated by Kim Sharp and Franz Matschinsky, E.R. Johnson Foundation, University of Pennsylvania.

Boltzmann, Ludwig (1886). *The Second Law of Thermodynamics*. In B. McGinness, ed. (1974), *Ludwig Boltzmann: Theoretical physics and philosophical problems: Select Writings*. Dordrecht, Netherlands: D. Reidel.

Bostrom, Nick (2014) *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press.

Brams, S. J. and A. D. Taylor (1996) *Fair Division: From Cake-Cutting to Dispute-Resolution*, Cambridge: Cambridge University Press.

Brams, Steven (2005) "Fair Division," in Barry R. Weingast and Donald Wittman, eds. *Oxford Handbook of Political Economy*, Oxford: Oxford University Press.

Brandon, Robert (2014) "Natural Selection," *Stanford Encyclopedia of Philosophy*.

Browne, Cameron et al (2012) "A Survey of Monte Carlo Tree Search Methods," *IEEE Transactions on Computational Intelligence and AI in Games*, Vol. 4, No. 1, March.

Brundage, Miles, and Avin Shahar et al. (2018) *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*.

Camus, Albert (1942) *The Myth of Sisyphus*, originally in French.

Carnot, Nicolas (1824) *Reflections on the Motive Power of Fire* (originally in French, *Réflexions sur la puissance motrice du feu et sur les machines propres à développer cette puissance*) Paris.

Chalmers, David (2010) "The Singularity: A Philosophical Analysis," *Journal of Consciousness Studies*, Vol. 17, Numbers 9-10, pp.7-65.

Chalmers, David (2012) "The Singularity: A Reply to Commentators," *Journal of Consciousness Studies*, Vol. 19, Numbers 7-8, pp.141-167.

Choi, Jung-Kyoo and Samuel Bowles (2007) "The Coevolution of Parochial Altruism and War," *Science* 318 (5850), October 26, pp. 636-640.

Chopra, Samir and Scott D. Dexter (2008) *Decoding Liberation: The Promise of Free and Open Source Software*, New York: Routledge.

Clausius, Rudolf (1850/1851 English translation) "On the Moving Force of Heat, and the Laws regarding the Nature of Heat itself which are deducible therefrom," *Philosophical Magazine and Journal of Science*, 4th. 2 (VIII): 1–21; 102–119, originally in German "Ueber Die Bewegende Kraft Der Wärme Und Die Gesetze, Welche Sich Daraus Für Die Wärmelehre Selbst Ableiten Lassen," *Annalen der Physik*. 79, pp. 368–397, 500–524.

Clausius, Rudolf (1854/1867 English translation) The Mechanical Theory of Heat: with its Applications to the Steam Engine and to Physical Properties of Bodies, London: John van Voorst, originally in German, "Über eine veränderte Form des zweiten Hauptsatzes der mechanischen Wärmetheorie," *Annalen der Physik*. Poggendoff. xciii, pp. 481–506.

Copernicus, Nicolaus (1543) *On the Revolutions of Heavenly Spheres.*

Cudd, Ann (2000/2012 Revision) "Contractarianism," *Stanford Encyclopedia of Philosophy*.

Darwin, Charles (1859/6th ed. 1872) *On the Origin of Species, Darwin Online*, http://darwin-online.org.uk/.

Darwin, Charles (1871/2nd ed. 1874) *The Descent of Man, Darwin Online*, http://darwin-online.org.uk/.

Darwin, Charles (September 28, 1860). "Darwin, C. R. to Lyell, Charles," *Darwin Correspondence Project*, Cambridge, UK: Cambridge University Library. Letter 2931, http://www.darwinproject.ac.uk/letter/?docId=letters/DCP-LETT-2931.xml;query=2931;brand=default.

Dawkins, Richard (1976/2006) *The Selfish Gene: 30th Anniversary Edition*, New York: Oxford University Press.

Dennett, Daniel (1995) *Darwin's Dangerous Idea*, New York: Simon & Schuster.

Dennett, Daniel (1996) *Kinds of Minds*, New York: Basic Books.

Eddington, Arthur (1928) *The Nature of the Physical World*, Cambridge University Press.

Eijck, Jan van and Rineke (L.C.) Verbrugge (2009) eds. *Discourses on Social Software*, Amsterdam University Press.

Eijck, Jan van and Rineke (L.C.) Verbrugge (2014) "Formal Approaches to Social Procedure," *Stanford Encyclopedia of Philosophy*.

Eijck, Jan van and Rineke Verbrugge (2012) *Games, Actions, and Social Software: Multidisciplinary Aspects*, Springer.

Eijck, Jan van and Rohit Parikh (2009) "What is Social Software?," in Jan van Eijck and Rineke (L.C.) Verbrugge (2009) (eds.) *Discourses on Social Software*, Chapter 3, Amsterdam University Press.

Eliot, T. S. (1922) *The Waste Land*.

Fitting, Melvin (1990/2nd ed. 1996) *First-Order Logic and Automated Theorem Proving*, Springer Verlag.

Fitting, Melvin (2011) "Reasoning About Games," *Studia Logica* Volume 99, Issue 1-3, pp. 143-169, originally first published as, Melvin Fitting (2010), "TR-2010002: Reasoning About Games," *Computer Science Technical Reports*, Paper 338, The Graduate Center, CUNY.

Fitting, Melvin and Richard L. Mendelsohn (1998) *First-order Modal Logic*, Dordrecht: Kluwer Academic Publishers.

Frankena, William (1963/1973 2nd ed.) *Ethics*, Upper Saddle River: Prentice Hall.

Fu, Michael C. (2016) "AlphaGo and Monte Carlo Tree Search: The Simulation Optimization Perspective," *Proceedings of the 2016 Winter Simulation Conference*.

Fukuyama, Francis (1995) *Trust: The Social Virtues and the Creation of Prosperity*, Free Press.

Gansberg, Martin (1964) "37 Who Saw Murder Didn't Call the Police," *New York Times*, March 27.

Gauthier, David (1986) *Morals by Agreement*, Oxford: Oxford University Press.

Godfrey-Smith, Peter (2007) "Conditions for Evolution by Natural Selection," *The Journal of Philosophy*, Volume CIV, No. 10, October 2007.

Godfrey-Smith, Peter (2009) *Darwinian Populations and Natural Selection*, New York: Oxford University Press.

Godfrey-Smith, Peter (2013) *Philosophy of Biology*, Princeton: Princeton University Press.

Golding, William (1954) *Lord of the Flies*, New York: The Berkley Publishing Group.

Gopnik, Alison (2009) ″Could David Hume Have Known about Buddhism? Charles Francois Dolu, the Royal College of La Flèche, and the Global Jesuit Intellectual Network," *Hume Studies*, Vol. 35, No. 1 & 2, pp. 5-28.

Gopnik, Alison (2015) "How an 18th-Century Philosopher Helped Solve My Midlife Crisis: David Hume, The Buddha, and a search for the Eastern roots of the Western Enlightenment" in *The Atlantic*, October.

Gregory, Richard (1981) *Mind in Science: A History of Explanations of Psychology and Physics*, Cambridge University Press.

Hamilton, W. D. (1964) "The Genetical Evolution of Social Behaviour I," *Journal of Theoretical Biology*, Volume7, Issue 1, pp. 1-16.

Hammond, Dick E. (1985/2005 Eulogy edition) *The Human System from Entropy to Ethic*, San Marcos: Minuteman Press (previously printed by Prentice-Hall, Inc.)

Hancock, Roger (1960) "The Refutation of Naturalism in Moore and Hare," *Journal of Philosophy* 57 (10), pp. 326-334.

Hare, R. M. (1952) *Language of Morals*, Oxford: Clarendon.

Hassabis, Demis (2016) "AlphaGo: Using Machine Learning to Master the Ancient Game of Go," https://www.blog.google/topics/machine-learning/alphago-machine-learning-game-go/

Hawking, Stephen (2010) "Into the Universe with Stephen Hawking," A BBC Science Documentary.

Hobbes, Thomas (1651) *Leviathan*.

Horgan, Terence and Mark Timmons (1992) "Troubles for New Moral Semantics: The 'Open Question Argument' Revived," *Philosophical Papers*, Vol. XXI, No.3.

Hume, David (1739/1888) *A Treatise of Human Nature*, L.A. Selby-Bigge (ed.), Oxford: Clarendon.

Huxley, Thomas Henry (1894) *Evolution and Ethics, and Other Essays*, Echo Library.

Jackson, Philip C. (1974/ 1985 2nd ed.) *Introduction to Artificial Intelligence: Second, Enlarged Edition* (Dover Books on Mathematics), Dover Publication.

Kalai, E. and M. Smorodinsky (1975) "Other Solutions to Nash's Bargaining Problem," *Econometrica* 43, pp. 513-518.

Kim, Chun-su (1952), "Flower," originally in *Drawing of Flower*, in Korean. Trans. Jong-Gil Kim (1998), *The snow falling on Chagall's village: selected poems*, Ithaca, N.Y.: East Asia Program, Cornell University.

Kim, Jongjin (1998) "Cyberspace and Karl Popper's World 3," Master Thesis, Korea University, Seoul (in Korean).

Kuhn, Thomas S. (1962/Second enlarged edition 1970) *The Structure of Scientific Revolutions*, Chicago: The University of Chicago Press.

Kurzweil, Ray (2006) *The Singularity Is Near: When Humans Transcend Biology*, Penguin Books.

Lackey, Douglas (1989) *God, Immortality, Ethics: A Concise Introduction to Philosophy*, Wadsworth Publishing Company.

Lambert, Frank L. (2002) "Disorder - A Cracked Crutch for Supporting Entropy Discussions," *Journal of Chemical Education*, Vol. 79, Number 2, February.

Lange, John F. (1966) "R. M. Hare's Reformulation of the Open Question," *Mind*, New Series, Vol. 75, No. 298 (Apr.), pp.244-247.

Leff, Harvey S. and Andrew F. Rex (1989/2002 Second edition) ed. *Maxwell's Demon 2: Entropy, Classical and Quantum Information, [sic] Computing*, Institute of Physics Publishing.

Lenman, James (2006) "Moral Naturalism," *Stanford Encyclopedia of Philosophy*.

Lewis, David (1969) *Convention: A Philosophical Study*, Cambridge: Harvard University Press.

Lewis, David K. (1989) "Dispositional Theories of Value," *Proceedings of the Aristotelian Society*, 63 (Supplement), pp. 113–137.

Lewontin, R.C. (1970), "The Unit of Selection," *Annual Review of Ecology and Systematics* 1, pp. 1-18.

Machiavelli, Niccolo (1513/1532) *The Prince*.

Marx, Karl (1845) "Theses on Feuerbach."

Maynard Smith, John (1974) "The Theory of Games and the Evolution of Animal Conflict," *Journal of Theoretical Biology* 47, pp. 209-221.

Maynard Smith, John (1978) "The Evolution of Behavior," *Scientific American* 239, pp. 176-192.

Maynard Smith, John (1982) *Evolution and the Theory of Games*, Cambridge: Cambridge University Press.

McIrvine, Edward C. and Myron Tribus (1971), "Energy and information," *Scientific American*, 224 (September 1971).

Mendelsohn, Richard L. (2008) "Referential/Attributive: A Scope Interpretation," *Philosophical Studies* 147 (2), pp. 167-191.

Mendelsohn, Richard L. (2012) "*Sinn* and *Bedeutuing* with Scope," *Journal of Philosophy* 109 (1-2), pp. 175-203.

Millikan, Ruth Garett (2006) "Styles of Rationality," in *Rational Animals?*, Susan Hurley and Matthew Nudds eds., Oxford: Oxford University Press.

Monod, Jacques (1970) *Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology*, originally in French.

Moore, G. E. (1903/1988) *Principia Ethica*, Amherst: Prometheus Books.

Moore, G. E. (1942) "A Reply to My Critics," in Schilpp (1942) pp. 533-678.

Moore, J (1992) "Implementation, Contracts, and Renegotiation in Environments with Complete Information," in J.-J. Laffont, ed. *Advances in Economic Theory- 6th World Congress*, Volume I, Cambridge: Cambridge University Press.

Nash, John (1950) "The Bargaining Problem," *Econometrica* 18, pp. 155-162.

Nettle, Daniel (2010) "Dying young and living fast: variation in life history across English neighborhoods," *Behavioral Ecology*, Vol. 21, Issue 2, 1 March 2010, pp. 387-395 (Advance Access publication 27 January 2010).

Neumann, John von, and O. Morgenstern (1944/1967 3rd Printing) *Theory of Games and Economic Behavior*, Princeton University Press.

Nozick, Robert (1963), *The Normative Theory of Individual Choice*, Ph.D. dissertation, Princeton University.

Okasha, Samir (2003/2013) "Biological Altruism," Stanford Encyclopedia of Philosophy.

Old Testament, 1 Kings 3:16-28.

Pacuit, Eric (2005) *Topics in Social Software: Information in Strategic Situations*, Ph.D. Dissertation, New York: The Graduate Center, City University of New York.

Pacuit, Eric and Olivier Roy (2015) "Epistemic Foundations of Game Theory," *Stanford Encyclopedia of Philosophy*.

Pacuit, Eric and Rohit Parikh (2006) "Social Interaction, Knowledge, and Social Software," in Dina Goldin, Scott Smolka and Peter Wegner eds., *Interactive Computation: The New Paradigm*, Springer-Verlag.

Pacuit, Eric, Rohit Parikh and Eva Cogan (2006) "The Logic of Knowledge Based Obligation," *Synthese* 149, pp. 311-341. Originally presented at DALT 2004.

Parikh, Rohit (1990) "Recent Issues in Reasoning about Knowledge," *TARK '90 Proceedings of the 3rd Conference on Theoretical Aspects of Reasoning about Knowledge*, pp. 3-10.

Parikh, Rohit (1995) "Language as Social Software" (abstract), in *International Congress on Logic, Methodology and Philosophy of Science*, p. 415.

Parikh, Rohit (2001) "Language as Social Software," in Juliet Floyd and Sanford Shieh, eds. *Future Pasts: The Analytic Tradition in Twentieth Century Philosophy*, pp. 339-350, Oxford University Press, Oxford.

Parikh, Rohit (2002-1) "Social Software," *Synthese* 132, pp. 187-211, Kluwer Academic Publishers.

Parikh, Rohit (2002-2) "Towards a Theory of Social Software," in *Proceedings of DEON 2002,* pp. 187-211, September.

Parikh, Rohit (2003) "Levels of Knowledge, Games and Group Action," *Research in Economics* Volume 57, Number 3, pp. 267-281.

Parikh, Rohit (2014) "What is Social Software?," Presentation File in the October Workshop of the CUNY Social and Political Philosophy Working Group (SPP).

Parikh, Rohit (2017), "An Epistemic Generalization of Rationalizability," in a talk of *University Seminar on Logic, Probability, and Games*, March 24, Columbia University.

Parikh, Rohit and Paul Krasucki (1990) "Communication, Consensus and Knowledge," *Journal of Economic Theory*, 52, pp. 178-189.

Parikh, Rohit and Paul Krasucki (1992) "Levels of Knowledge in Distributed Computing," *Sadhana- Proc. Ind. Acad. Sci*., 17, pp. 167-191.

Parikh, Rohit and R. Ramanujam (1985) "Distributed Processes and the Logic of Knowledge," in *Logic of Programs*, volume 193 of *Lecture Notes in Computer Science*, pp. 256-268, Springer.

Parikh, Rohit and Ramaswamy Ramanujam (2003) "A Knowledge based Semantics of Messages," *Journal of Logic, Language and Information*, Volume 12, Number 4, pp. 453-467.

Parikh, Rohit, Çagil Tasdemir and Andreas Witzel (2013) "The Power of Knowledge in Games," *International Game Theory Review*.

Pauly, Marc (2001) *Logic for Social Software*, ILLC Dissertation Series 2001-10, University of Amsterdam.

Pauly, Marc (2005) "Changing the rules of play," *Topoi* 24, pp. 209-220.

Penrose, Roger (1989) *The Emperor's New Mind: Concerning Computer, Minds, and the Laws of Physics*, Oxford University Press.

Penrose, Roger (2004/2006) *The Road to Reality: A Complete Guide to the Laws of the Universe*, New York: Alfred A. Knopf.

Piervincenzi, William (2007) *G.E. Moore's Naturalistic Fallacy and Open Question Argument Reconsidered*, Ph.D. Dissertation, University of Rochester.

Plato *Theaetetus.*

Plato, *Georgias*.

Plato, *The Republic*.

Pojman, Louis P. (1990/2017 8th edition with James Fieser) *Ethics: Discovering Right and Wrong*, Boston: Cengage Learning.

Popper, Karl (1962) *Conjectures and Refutations: the Growth of Scientific Knowledge*, Basic Books.

Popper, Karl (1977) "Natural Section and the Emergence of Mind," First Darwin Lecture, Darwin College, in Cambridge, UK.

Popper, Karl (1982) *Unended Quest*, Open Court.

Priest, Graham (1987/2006 2nd (Extended) edition) *In contradiction: A Study of the Transconsistent*, Oxford: Oxford University Press; First edition, Martinus Nijhoff, 1987.

Priest, Graham (2014) *One: Being an Investigation into the Unity of Reality and of its Parts, including the Singular Object which is Nothingness*, Oxford: Oxford University Press.

Prinz, Jesse (2004/2009) "Against Moral Nativism" In Dominic Murphy and Michael Bishop (eds.) *Stich and His Critics* (2009), Chichester: Wiley-Blackwell.

Prinz, Jesse (2012) "Singularity and Inevitable Doom," *Journal of Consciousness Studies*, Vol. 19, Numbers 7-8, pp. 77-86.

Prinz, Jesse (In production) *The Moral Self*, New York: Oxford University Press.

Prinz, Jesse J. (2007) *The Emotional Construction of Morals*, Oxford: Oxford University Press.

Putnam, Hilary (2002) *The Collapse of the Fact/Value Dichotomy and Other Essays*, Cambridge: Harvard University Press.

Rachels, James (1986/2011 7th edition with Stuart Rachels) *The Elements of Moral Philosophy*, New York: McGraw-Hill Higher Education.

Rauhut, Nils Ch. (2003/2010 3rd ed.) *Ultimate Questions: Thinking about Philosophy*, Pearson.

Rawls, John (1971/1998 Revised ed.) *A Theory of Justice*, Cambridge, Massachusetts: Harvard University Press.

Refkin, Jeremy and Ted Howard (1980) *Entropy: A New World View*, New York: Viking Press.

Rhodes, Rosamond, Nada Gligorov, and Abraham Paul Schwab (2013) eds. *The Human Microbiome: Ethical, Legal and Social Concerns*, Oxford University Press.

Ridge, Michael (2003/Revision 2014) "Moral Non-Naturalism," *Stanford Encyclopedia of Philosophy*.

Rosenthal, Robert W. (1981) "Games of Perfect Information, Predatory Pricing and the Chain-Store Paradox," *Journal of Economic Theory* 25, pp. 92-100.

Rousseau, J. (1755/1984) *A Discourse on Inequality*, Trans. M. Cranston, New York: Penguin Books.

Ruse, Michael (1986) "Evolutionary Ethics: A Phoenix Arisen" *Zygon* 21 (1), pp. 95-112.

Ruse, Michael (1986) *Taking Darwin Seriously: A Naturalistic Approach to Philosophy*, Oxford: Oxford University Press.

Ruse, Michael (1995), "Evolution and Ethics: the Sociobiological Approach," in Louis P. Pojman Ed., *Ethical Theory: Classical and Contemporary Readings*, pp. 91-122, Wadsworth Publishing Company.

Ruse, Michael and Edward O. Wilson (1986) "Moral Philosophy as Applied Science," *Philosophy*, Vol. 61, No. 236, April, pp. 173-192.

Russell, Bertand (1945/1972) *A History of Western Philosophy*, New York: A Touchstone Book published by Simon and Schuster.

Ryle, Gilbert (1949) *The Concept of Mind*, Chicago: The University of Chicago Press.

Satyamurti, Carole (2015) *Mahabharata: A Modern Retelling*, New York: W. W. Norton & Company.

Savage, Leonard J. (1954/Second revised edition 1972), *The Foundations of Statistics*, Dover Publications.

Scanlon, Thomas (1998) *What We Owe to Each Other*, Cambridge: Harvard University Press.

Schiffer, Stephen (1972) *Meaning*, Oxford: Oxford University Press.

Schilpp, P.A., (ed.) (1942) *The Philosophy of G.E. Moore*, The Library of Living Philosophers, Evanston: Northwestern University.

Schmidt, R. A., and D. Tishkovsky (2002) "Combining dynamic logic with doxastic modal logics," In P. Balbiani, N.-Y. Suzuki, F. Wolter, and M. Zakharyaschev, (eds.), *Advances in Modal Logic*, volume 4, pp. 371–391, London: King's College Publications.

Schmidt, R. A., and D. Tishkovsky (2008) "On combinations of propositional dynamic logic and doxastic modal logics," *Journal of Logic, Language and Information* 17(1), pp. 109–129.

Schneebaum, Tobias (1969) *Keep the River on Your Right*, New York: Grove Press.

Schrödinger, Erwin (1944) *What is Life?*, Cambridge University Press.

Sen, Amartya (1970/2017 Expanded ed.) *Collective Choice and Social Welfare*, Cambridge, Massachusetts: Harvard University Press.

Shannon, Claude E. (1950) "Programming a Computer for Playing Chess," *Philosophical Magazine*, Ser.7, Vol. 41, No. 314 - March 1950.

Silver, David et al. (2016) "Mastering the game of Go with deep neural networks and tree search," *Nature*, volume 529, pp. 484–489 (28 January 2016).

Sinervo, Barry (1997/2013) "Ch. 4: Levels of Selections," in his *Behavioral Ecology* class, http://bio.research.ucsc.edu/~barrylab/classes/animal_behavior/BEHAVIOR.HTM.

Skyrms, Brian (1996) *Evolution of the Social Contract*, Cambridge: Cambridge University Press.

Skyrms, Brian (2003) *The Stag Hunt and the Evolution of Social Structure*, Cambridge: Cambridge University Press.

Smith, Adam (1776) *Wealth of Nations*.

Smith, Michael (1994) *The Moral Problem*, Oxford: Blackwell.

Sober, Elliott & David Sloan Wilson (1998) *Unto Others*, Cambridge, Massachusetts: Harvard University Press.

Stambaugh, Todd (2013/2017) "Coincidence of Two Solutions for Nash's Bargaining Problem," Economics Letters, Volume 157, August 2017, pp. 148-151. Originally presented in the International Game Theory Conference (2013), Stony Brook.

Steinhaus, H (1949) "Sur la division progmatique," *Econometrika* (Supplement) 17, pp. 315-319.

Stojanović, Svetozar (1963) “Hare's Argument Against Ethical Naturalism,” *Mind*, New Series, Vol. 72, No. 286 (Apr.) pp. 264-267.

Strandberg, Caj (2004) “In Defence of the Open Question Argument,” *The Journal of Ethics*, Vol. 8, No. 2, pp. 179-196.

Styer, Daniel F (2000) “Insight into Entropy,” *American Journal of Physics*, 68 (12), December.

Tomasello, Michael (2008) “How are humans unique?,” *New York Times*, May 25.

Tomasello, Michael (2009) *Why We Cooperate*, Cambridge: MIT Press.

Tomasello, Michael (2014) “The Ultra-social Animal,” *European Journal of Social Psychology*, 44, pp. 187-194.

Tomasello, Michael (2016) *A Natural History of Human Morality*, Harvard University Press.

Tribus, Myron (1964) ‘‘Information theory and thermodynamics,’’ in *Heat Transfer, Thermodynamics, and Education: Boelter Anniversary Volume*, edited by Harold A. Johnson, New York: McGraw-Hill, pp. 348–368.

Trivers, Robert L. (1971) “The Evolution of Reciprocal Altruism,” *The Quarterly Review of Biology*, Vol. 46, No. 1, March, pp. 35-57.

Van Roojen, Mark (2004/Revision 2013) “Moral Cognitivism vs. Non-Cognitivism,” *Stanford Encyclopedia of Philosophy*.

Vessel, Jean-Paul (2004/2009) “Moore’s Open Question Maneuvering,” American Philosophical Association Colloquium, Pacific Division Meeting, March 2009.

Walker, Jeremy (1973) “A Naturalist Reply to Hare,” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, Vol. 24, No. 1 (Jan.), pp. 45-51.

Wallach, Wendell and Colin Allen (2009) *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press.

Whitehead, Alfred North (1929) *The Function of Reason*, Princeton University Press.

Wiener, Norbert (1948/1961 2nd edition/1962 2nd printing) *Cybernetics: or control and communication in the animal and the machine*, Cambridge: MIT Press.

Wiener, Norbert (1950/1954 Revision) *The Human Use of Human Beings: Cybernetics and Society*, Boston: Houghton Mifflin Company.

Wiese, Harald (2012) “Backward Induction in Indian Animal Tales” *International Journal of Hindu Studies* 16, 1, pp. 93-103.

Williams, Bernard (1972/2005) *Morality: An Introduction to Ethics*, Cambridge: Cambridge University Press.

Williams, Bernard (1985) *Ethics and the Limits of Philosophy*, Harvard University Press.

Williams, Bernard (1985/2006) *Ethics and the Limits of Philosophy*, Cambridge: Harvard University Press.

Wilson, David Sloan & Lee A. Dugatkin (1992), "Altruism: Contemporary Debates," in *Keywords in Evolutionary Biology*, E. F. Keller and E. A. Lloyd (eds.), Cambridge, MA: Harvard University Press.

Wilson, David Sloan (1990) 'Weak Altruism, Strong Group Selection,' *Oikos* 59, pp. 135–48.

Wilson, Edward O. (1975/2000) *Sociobiology: The New Synthesis, Twenty-Fifth Anniversary Edition*, Cambridge: Belknap Press of Harvard University Press.

Wilson, Edward O. (1978/2004 2nd ed.) *On Human Nature*, Cambridge, MA: Harvard University Press.

Zermelo, E. (1913), "Ueber eine Anwendung der Mengenlehre auf die Theorie des Schachspiels," in Proceedings of the Fifth International Congress of Mathematicians, Cambridge, 1912, Vol. II, pp. 501-504. Cambridge: Cambridge University Press.

Zhuangzi *The Zhuangzi*.

Zobrist, Albert L. (1969) "A model of visual organisation for the game Go," In *Proceedings of the Spring Joint Computer Conference*, volume 34, pp. 103–112.

Online Resources:

"Natural Selection," in *Wikipedia*, https://en.wikipedia.org/wiki/Natural_selection.

The Stupid Way: About Zen Buddhism by "Peter," http://www.zen.ie/heartsutra.html.

Kwan Um School of Zen, https://kwanumzen.org/resources-collection/2017/9/6/heart-sutra-in-english.

https://physics.nist.gov/cuu/Units/meter.html.

https://physics.nist.gov/cuu/Units/second.html.

https://commons.wikimedia.org/w/index.php?curid=1625737.

https://www.jpl.nasa.gov/infographics/infographic.view.php?id=10824.

https://www.youtube.com/watch?v=-VEFD4_00MI.

http://entropysite.oxy.edu/