

The Case for Rational Persuasion

Great Delusion or Genuine Change of Mind

Peter Makovíni

Aalborg University

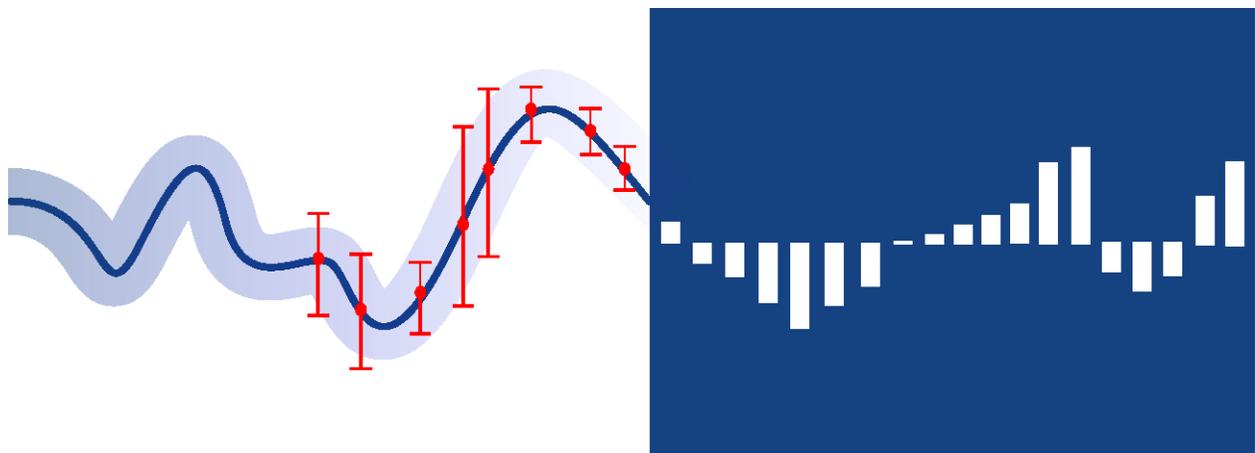


Image adapted from piseg.sdu.dk

Page Count: 78.2

Number of Characters: 187.689 (with spaces)

Supervisor: David Jakobsen

Information Architecture and Persuasive Design, 10th semester

Thesis submitted for completion of Master of Science (60 credits)

August 9th, 2016

Acknowledgment

I want to express my gratefulness to three people whose friendship and support has meant that this thesis could become a reality. I appreciate the guidance of the assistant professor David Jakobsen who has inspired me to pursue academia and helped me through the entire process of this thesis. I also wish to thank my friend Prayson Daniel for the long conversations, which kindled my curiosity for further research. I also appreciate the hours of proofreading and many inputs on the use of the English language given by Venus Ramos.

Finally, I wish to give thanks to my parents and family for creating an excellent environment to study. This thesis could not be accomplished without their love and sacrifice.



AALBORG UNIVERSITY
DENMARK

Table of Contents

Abbreviations.....	6
Nomenclature.....	6
Resume.....	9
Persuasion.....	10
Timing in Persuasion.....	11
Computers as a Persuasive Technology.....	14
Persuasion as a Voluntary Change.....	15
Branching Time in Free Will, Programming, and World Wide Web IA & PD.....	17
Fogg’s Behavior Model.....	26
The Functional Triad.....	27
The Argument for Rational Persuasion.....	33
The <i>Freedom</i> Pillar.....	33
An Objection from Compatibilism.....	34
Response to the Objection from Compatibilism.....	34
The <i>Rationality</i> Pillar.....	35
The Best Possible Explanation Argument for Rational Persuasion.....	35
Application of the Argument for Rational Persuasion to AI.....	36
Freedom of Will and Voluntary Choice.....	37
Determinism, Compatibilism and Libertarianism.....	43
Hard Determinism.....	43
Free Will and Introspection.....	48
Materialism, Epistemology and Defeaters.....	53
The Argument from Reason.....	54
The Argument From the Reliability of our Rational Faculties.....	60
Epistemological Defeaters and Defeat.....	63
A Defeater-defeater.....	66
Self-Defeating Defeaters.....	67
Back to the Argument from the Reliability of our Rational Faculties.....	67
Objections to the argument from the reliability of our rational faculties.....	68

The Dreaded Loop Objection	68
Response to the Dreaded Loop Objection	69
Objection from Sensible Naturalism	70
Response to Sensible Naturalism Objection	72
Compatibilism	74
Freedom in the Absence of Alternative Possibilities.....	79
Character-Type Examples.....	80
Frankfurt-type examples.....	81
Frankfurt’s Hierarchical Motivation Theory	83
Control and Determinism.....	85
Libertarianism.....	89
The Modest Objective.....	94
Implicit Beliefs.....	94
Explicit Beliefs	96
Persuasive Artificial Intelligence.....	101
Turing Test and Other Minds.....	104
Searle’s Chinese Room Argument	105
False Analogy Objection	106
Response to False Analogy Objection	107
Chinese Room and Other Minds.....	107
The Predicament of a Naturalist	108
Substance Dualism, AI and Other Minds	109
Great Delusion or a Genuine Change of Mind	110
The AI Delusion.....	111
References.....	114
Appendix 1.....	121
Appendix 2.....	125
Appendix 3.....	126

Abstract

This paper focuses on a defense of rational persuasion against challenges that deterministic views of free will present. These are questions regarding ontology, epistemology and the philosophy of mind, drawing on research from neuroscience and evolutionary biology. Three positions: Hard Determinism, Compatibilism and Libertarianism are analyzed extensively. Arguments are given for why the first two monist philosophies cannot account for free, rational persuasion and therefore some form of substance dualism is advised as the best explanation of human freedom and rationality. The possibility of creating a sentient, strong artificial intelligence indistinguishable from human beings is discussed, and in the light of this work it is proposed that while such entity may be very persuasive, it cannot benefit from the kind of conscious experience, freedom and rationality humans have.

Keywords: persuasion, control, determinism, indeterminism, compatibilism, libertarianism, artificial intelligence,

Abbreviations

AI Artificial Intelligence
BOM Belief in Other Minds
CNC Covert Non-Constraining Control
CR Chinese Room
EEG Electroencephalography
GPS Global Positioning System
HA Hypothetical Analysis
HCI Human Computer Interaction
IA Information Architecture
ICT Information and Computer Technologies
MSD Mental State Defeater
NC Non-Constraining Control
PAP Principle of Alternative Possibilities
PCC Physical Causal Closure
PD Persuasive Design
PT Persuasive Technology
RP Readiness Potential
SAR Strong Agent Reductionism
SFA Self-forming Action
TT Turing Test
UR Ultimate Responsibility

Nomenclature

E Evolution is true
N Naturalism is true
R Human cognitive faculties are reliable

List of Tables

Table 1 – A summary of an article about smartphone sensors	12
Table 2 - Google Now Cards	13
Table 3 - Kane's Five Freedoms	42
Table 4 - Three basic positions on the freedom of will	43

List of Figures

Figure 1 - Fogg's Illustration of the Captology overlap between Computers and Persuasion	15
Figure 2 - Kripke's illustration of the idea of a branching time	17
Figure 3 - Kane's illustration of a "garden of forking paths"	18
Figure 4 - Amazon.com view of maximized branching possibilities (First part).	22
Figure 5 - Amazon.com view of minimized branching possibilities (Second part).	23
Figure 6 –"Jeg er Ny" category on the website of the Baptist Church in Odense	25
Figure 7 - Fogg's Behavior Model	26
Figure 8 - Fogg's Functional Triad	28
Figure 9 - Clark's depiction of the levels of reduction.	45
Figure 10 - Results of Libet's neurophysiological experiment on willing and consciousness	49
Figure 11 - Craig's interpretation of Libet's results	52
Figure 12 - Kane's Incompatibilist Mountain and the Libertarian Dilemma	90
Figure 13 - Kurzweils prediction in TIME magazine	103

This page has been intentionally left blank

Resume

The study of the philosophy of mind has direct implications on the nature and feasibility of rational persuasion. The gist of the controversy demonstrates that hard deterministic views such as Eliminative Materialism ultimately reduce persuasion down to chemistry and physics; a mere mindless mechanistic processes of action and reaction with no room for rational reflection. Even though Compatibilism does not go as far, it requires notions like free choice, possible alternatives, and persuasive influence to be radically redefined. As such, influence is illusory and both persuader and persuadee are mere marionettes acting out their part, being governed by fixed motions of atoms and forces in the universe. On the contrary, an indeterministic position creates place for rational persuasion that can influence free choices. Despite many objections, this work attempts to prove that Libertarianism, that builds on dualism is a framework in harmony with the classical commonsensical folk-psychology concept of genuine free will, which is not determined, nor is arbitrary. Only this kind of freedom permits for rational inference and logical connection to be causally involved in a persuasive endeavor that may result in a change of attitude or behavior on the basis of reasons. The following pages will briefly introduce the fundamental concepts from the studied field: history of persuasion, Kairos, Captology, Branching Time, Fogg's Functional Triad and Behavior Model. Two aspects, *rationality* and *freedom* of human beliefs are indispensable if rational persuasion is to be preserved. Therefore, great focus will be given to these respectively. The famous philosopher Daniel Dennett has once said that "AI makes Philosophy honest" (Anderson, 2009). With respect to social simulations of PD we may adapt this and say "AI makes Persuasive Design honest". To demonstrate this, the remainder of the paper considers how the findings apply to the potential future existence of human-like AI, which may cause delusion if people are falsely persuaded to believe that such machine is a free, rational and conscious agent.

Persuasion

A concept that became widely known with the rise of rhetoric of the ancient Greeks and democratic city-states known as *polis*. Long gone was the time when communication served to merely notify other members of one's group about good places to hunt or collect berries. Besides simple exchange of information, communication has now assumed a new role to fit the needs of public and political life. Oration and eloquent argumentation was designed to convey sophisticated philosophical ideas, not only to instruct and enlighten, but to resolve conflicts, exercise influence, and gain power. Due to free speech, everyone was able to share ideas and determined forms of elocution became highly valued tools used to convince and persuade both slaves and free people alike. It was understood that free people can freely change their mind, choose what to believe and what to do based on what seemed to them most compelling. Greeks trusted that speaking persuasively was a way to maintain a healthy democracy. (Fogg, 2003) Classical rhetoric is characterized by three distinct Aristotelian elements, *ethos* (credibility), *logos* (reason), and *pathos* (emotion) (Higgins, Walker, 2012). "Together", says Higgins and Walker, "these elements reveal the characteristics of a good argument." Oxford dictionary defines rhetoric as "Language designed to have a persuasive or impressive effect, but which is often regarded as lacking in sincerity or meaningful content". While rhetoric is today often associated with pejorative connotations and accused of obscuring the truth, classical philosophers believed the contrary; that rhetoric was vital to the discovery of truths.

Throughout history, the art of persuasion has been progressively maturing, and continues to be advanced through modern research in psychology and technology. Largely inspired by governments, marketers and advertisers that systematically investigate how influence operates,

persuasion has been often applied to sway tides of public opinion or to help corporations prosper (Fogg, 2003).

Timing in Persuasion

“So you say, tell me where Kairos is important, and I say to you, tell me where it’s not important.”

- James Kinneavy¹

In Greek culture, Kairos was the youngest son of Zeus, known for being quick and strong. Derived from mythology, he represented the fleeting opportunities in life that could be harnessed by being particularly attentive to “the right time” and “the right measure” (Aagaard, Moltsen, Øhrstrøm, n.d) (Fogg, 2003). All civilizations have discovered that the importance of right timing and measure is crucial in many areas of life, from the right time to sow seeds when there is peace to right timing of an attack in times of war. (Thompson, 2000).

Timing is essential to effective persuasion. “Timing”, Fogg writes, “is often the missing element in behavior change” (2009, p. 3). Psychologists have identified times when people are more open to persuasion, such as when they are in a good mood, when their worldview does not make sense, when they feel indebted and so on. These are opportune moments of persuasion. The problem is that opportune moments are hard to identify because they depend on a multitude of variables ranging from physical (e.g. geographic location, weather, temperature), personal (e.g. state of health, financial status, intelligence, interests and preferences), social (e.g. status, interpersonal relationships) to emotional (e.g. mood, self-worth)(Fogg, 2003) (Oinas-Kukkonen, Hasle et al.,

¹ Thompson, 2000, p. 81

2008). Yet with the advent of computing technology, an increasing number of these variables can now be accurately recognized for nearly any given moment. Modern smartphones can constantly send and receive cues about the geographic location of their owners through extracting data from GSM cellular transmitters, GPS (Global Positioning System), Wi-Fi networks and Bluetooth beacons. The most important sensor on a smartphone is the microphone, which is its *raison d'être*. However, smartphones are filled with an array of other sensors. One of the largest cell phone dedicated websites, PhoneArena.com, has published in 2014 an article on how many different kinds of sensors go inside a smartphone. This is its brief summary:

Accelerometer	Motion detection (shake, tilt, etc.); Determines whether phone is facing up- or downwards or whether it is in portrait or landscape orientation.
Gyroscope	Rotation detection (spin, turn, etc.)
Magnetometer	Detection of magnetic fields (compass)
Light sensor	Measures the amount of the surrounding ambient light
Barometer	Measures the atmospheric pressure
Thermometer	Measures the ambient temperature
Air humidity sensor	Measures the air temperature and humidity
Pedometer	Similar to accelerometer, yet more accurate; Used to count a user's steps
Heart rate monitor	Measures pulse of a user
Fingerprint sensor	Detects identity of a user
Proximity sensor	Measures the distance of a screen from nearest object; Used to turn off screen during calls, when having the phone close to a user's ear
Radiation detector	Measures current harmful radiation level in the area (Used in a special smartphone released only in Japan)

Table 1 - A Summary of an article "Did you know how many different kinds of sensors go inside a smartphone?" (Nick, 2014)

Smartphones often contain also information about the user's calendar, contacts and addresses, emails, recent searches, sleeping rhythm and much more since users usually voluntarily provide them.

Google Now is an app claiming to “help give you just the right information at just the right time“(Google, 2013). In the example, Google Now sends the user a notification at 6:50 to leave for his dentist appointment at 6:55 to arrive on time at 7:15. To provide this insight, Google Now uses data about the user’s current location, appointments in his calendar, current traffic situation and his past habits of driving a car. However, e.g. if the user’s car broke the day before, Google Now isn’t aware of this variable, in which case the information is not right for the user and he may not be able to come on time.

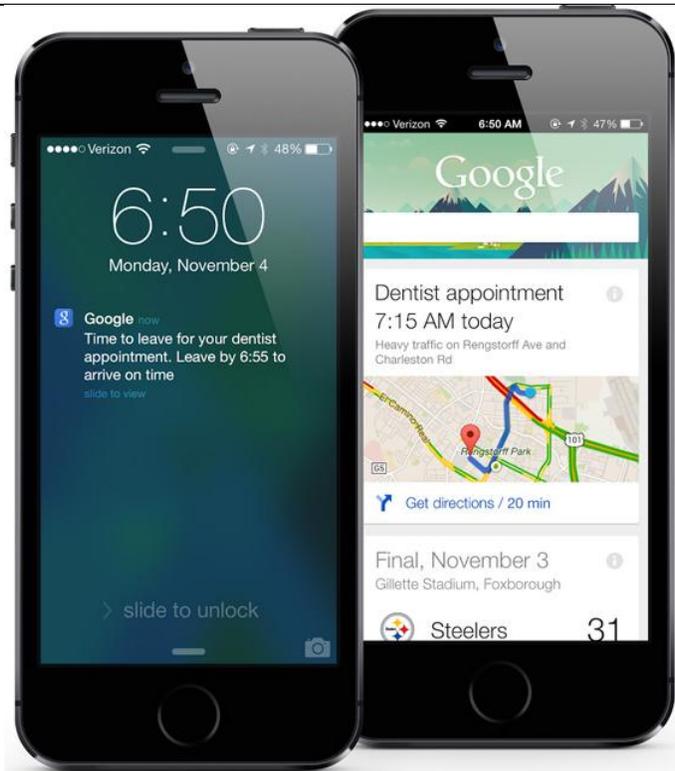


Table 2 - Google Now Cards

Increasingly, more sensors are installed into houses and apartments turning them to smart homes. Technological progress has enabled the transformation of artifacts such as a watch, a bracelet, a necklace and so on to smart *wearables* that are equipped with extra sensors like skin conductance sensor (how much user sweats) or skin temperature sensor. In their forecast of wearable sensors 2016-2026, Hayward and Chansin write, “Sensors collect data about the physical and chemical properties of the body and local environment, and use it to feed algorithms which output insightful information.” They predict, “there will be 3 billion wearable sensors by 2025, with over 30% of them being new types of sensors that are just beginning to emerge.” (Hayward, Chansin, 2016, Description) Still another advancement of the new millennium has introduced implantable sensors that measure biomedical values. Córcoles and Boutelle write, “Invasive

monitoring of physiological parameters, such as blood pressure, heart rate and body temperature among others, is certainly an extensive practice in clinical settings.” Yet, they particularly write about the possibility of monitoring biochemical parameters by biosensors. “Continuous monitoring of metabolites (glucose, lactate, pyruvate, urea, glutamate), proteins and nucleic acids (DNA, RNA) can potentially provide a rapid detection of life-threatening events.” (2013, p.3) A private company, GlySens Incorporated, offers diabetes patients a “fully implanted sensor” that “wirelessly links to a convenient external receiver, designed to provide continuous, at-a-glance glucose measurement...” (GlySens, n.a.) for an expected period of one year or more.

As Hayward and Chansin suggest, all this data can be used to gain insightful information. Information that could potentially help designers of persuasive systems minimize the number of unknown variables about users and identify precious moments of opportunity for persuasion. While such monitoring invites important ethical objections that must be considered, more data is increasingly available.

Computers as a Persuasive Technology

In 1997, the pioneer of persuasive design, B. J. Fogg was the first to coin the term “captology”. He explains that it is “the design, research, and analysis of interactive computing products created for the purpose of changing people’s attitudes or behaviors. It describes the area where technology and persuasion overlap” (2003, p. 5). This term is often used in his writings interchangeably with persuasive technology or persuasive design.

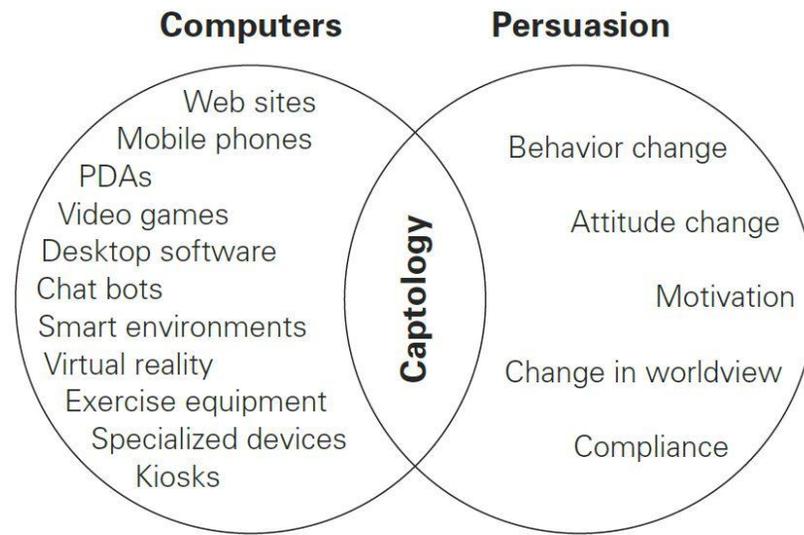


Figure 1 - Fogg's Illustration of the Captology overlap between Computers and Persuasion

However, Husle on more occasions suggests that despite Fogg's effort, the term captology did not seem to achieve a wider acceptance in IT-communities. Thus, he would rather stick to the more commonly used terms Persuasive Design (PD) or Persuasive Technology (PT) (2006, 2011).

Persuasion as a Voluntary Change

An essential aspect of persuasion, which is consistently reiterated in literature on PD is the safeguarding of one's autonomy, one's own volition, and freedom of one's will to choose whether to be subjected to persuasion or not. Should persuasion take place, the person must also have the freedom to choose the outcome of the efforts to change his belief or a behavior. Fogg defines persuasion as "an attempt to change attitudes or behavior or both (without using coercion or deception)". Then he further explains "coercion implies force; while it may change behaviors, it is not the same as persuasion, which implies voluntary change" (2003, p. 15). Smids considers a person's voluntary desire for change to be the most important ethical question in PT. (Berkovsky, Freyne, 2013). Obermair et al. write that PT "facilitates persuasive interaction that leads to

a voluntary change of behavior or attitude or both.” (H. Oinas-Kukkonen et al, 2008, p. 130), while Reitberger et al. call for even greater caution regarding coercion:

“We thus suggest that the designers of PT should even more actively try not to cross the line towards coercion and build enough “wobble room” for the users into their systems. Concurring with the argument, that the final and ideally *rational* decision whether to adapt a new behavior or not should be left to the user, we argue that PT systems should not rely on force but rather promote *reflection* of the users’ own actions in order to help them to reach the desired behavior.” [italics added]

(Bang, et al. 2012, p. 241)

As many have noted, a tension between persuasion and coercion, even manipulation, persists as the relation between persuasion and autonomy is complex. Timmer, Kool and Rinie Van Est set forth the gist of the crux, “simply stated, as long as the user is “free” to choose his goals and methods of persuasion of his own accord his autonomy is respected” (MacTavish, Basapur, 2015, p. 197). Others have developed different approaches to ethics of persuasive design technology using kinds of a *golden rule approach* (Berdichevsky & Neuenschwander, 1999; Burry Gram-Hansen, 2009), *the stakeholder analysis approach* (Fogg, 2003; Friedman et al., 2006), or *user involvement approach* (Davis, 2009; Yetim, 2011) as summarized by Karppinen and Oinas-Kukkonen (Berkovsky, Freyne, 2013).

Branching Time in Free Will, Programming, and World Wide Web IA & PD

Free choice conventionally entails that there are actual alternatives from which a person may freely choose. This can be conceived either as choosing A or B, or choosing A or not choosing A. It is also commonly understood that free choice extends only to the future, while past is absent of actual alternatives and is somehow completed. This commonsensical view of time and free choice demonstrates an apparent asymmetry between the past and the future.

In September 3, 1958, only 17 years old at that time, Saul Kripke recognized this and sent a letter to logician and philosopher Arthur N. Prior, in which he visualized this idea as a tree of possibilities.

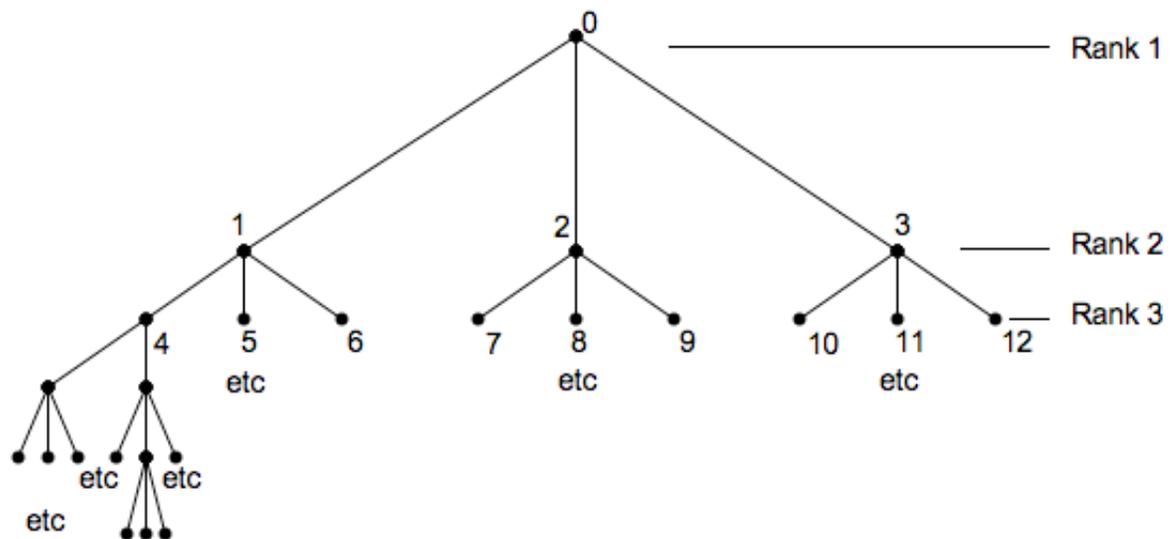


Figure 2 - Kripke's illustration of the idea of a branching time (Ploug, Øhstrøm, 2012)

Kripke explained that branching time presents a model of time in which time is not merely linear, but each moment presents several possibilities and a decision of a current moment will affect the available future possibilities for the next moment as seen in Figure 2. In the point 0, options 1 to 12 represent future moment possibilities, yet in the instant the possibility 1 is actualized, only options 4, 5 and 6 are its feasible successors. Options 2, 3 and 7-12 now remain

merely as not actualized contingencies of the point 0. This was clearly recognized by Prior as the asymmetry between the past and the future was, in his view, central to the notion of indeterminism (Ploug, Øhstrøm, 2012). Ploug and Øhstrøm clarify:

“While recognizing that freedom of choice is very limited, Prior in later writings professed that freedom of choice is real in the sense that the future is something we may to some extent make for ourselves [notes omitted]. Kripke’s notion of branching time leaves room for this understanding of free choice by representing the present as having different possible, alternative futures-the content of the future is not fixed in such a way as to allow for only one possible progression of the world. From the present the world may take different paths into the future depending on, for instance, the choices of agents” (2012, p. 368-369).

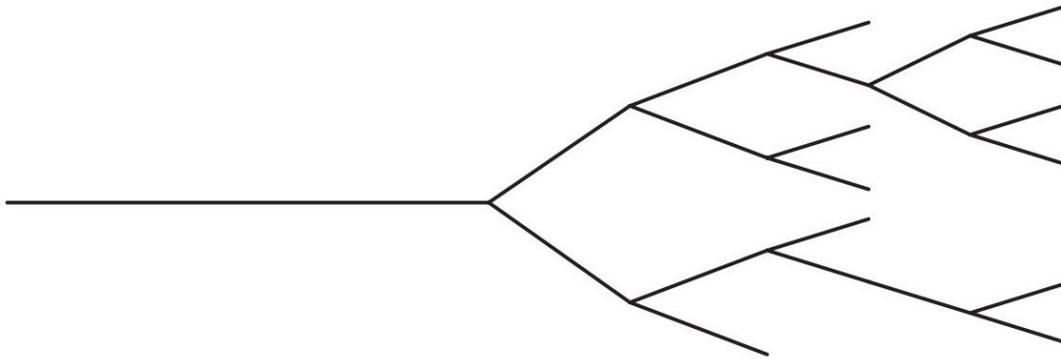


Figure 3 - Kane's illustration of a "garden of forking paths" (2005)

In “A Contemporary Introduction to Free Will” (2005), Robert Kane calls this picture of an open future a “garden of forking paths”. He sees it essential to our understanding of not only free will, but also “to what it means to be a person and to live a human life” (2005, p. 7).

Branching time suggests a notion of “now” or “present time” moving through “the system” (branching tree). The “now” offers an in-time view, while “the system” offers an out of time view.

These different perspectives are ascribed to two different languages or theories of time:

1. A-language – tensed, dynamic, temporal becoming (*Past, Present, Future*), Inside View
2. B-language – tenseless, static, timeless tapestry (*Before, Simultaneous with, After*), Outside View

(Øhrstrøm, Hasle, 1995)

The concept of branching time has proved to be groundbreaking in information and computer technologies (ICT). In programming, the “temporal logic has become an important tool for the analysis of concurrent (parallelistic) programs” (Øhrstrøm, Hasle, 1995, p. 347) and helped in formulating general program properties such as freedom of deadlock, mutual exclusion, fairness, liveness, etc.. Amir Pnueli was among the leading contributors who in 1996 received the Turing Award for “seminal work introducing temporal logic into computing science and for outstanding contributions to program and systems verification” (Hosch, n.d.). Ben-Ari, Pnueli and Manna acknowledge that temporal logic formalism is based on the question involving the underlying structure of time, “The dichotomy is between the linear time approach which considers time to be a linear sequence, and the branching time approach, which adopts a tree structured time, allowing some instants to have more than a single successor.” (1983, p. 207) They perceive the striking resemblance of a similar dichotomy in the field of programming formalisms and in the philosophical question regarding the structure of physical time of Prior, Kripke, Øhrstrøm, Ploug, Hasle and many others; however they have voiced that actually,

“The difference in approaches has very little to do with the philosophical question of the structure of physical time which leads to the metaphysical problems of determinancy versus free will. Instead it is pragmatically based on the choice of the type of programs and properties one wishes to formalize and study” (Ibid., p. 207).

Human computer interaction (HCI) is another field of ICT systems where the A/B theory, language or framework of time has shown very beneficial. It offers two distinct views of information and computer sciences - that of a user who can in one moment occupy only one position (rank) in a program (inside view) and that of a designer who has an overview of the entire program as he designed its every position (outside view). Thus, the A-language is more fitting to describe the user experience while the B-language captures better the overall branching of the whole system. To achieve an optimal user experience, ICT system designers, who have the outside view, ought to always consider the inside view of a user as well. In that way, HCI interaction can be enhanced and the system can come across as more engaging and persuasive for the user. An interaction of a user with a system parallels the idea of a branching time in which “now” is moving through the tree of possibilities.

For example, the online store Amazon.com on their website employs i.a. a combination of persuasive principles such as tunneling and reduction² to enhance their selling technique. The system is designed to maximize sales by deliberately taking advantage of branching time and future possibilities. In the first part, upon the user’s arrival on the website, the amount of future *click* or *interest* possibilities is maximized. The website presents users with browsing and

² These techniques are listed in the list of persuasive principles in Appendix 1

intelligent searching tools, but also a multitude of possibly relevant content like “frequently bought together”, “customers who bought this also bought”, “related to this item”, “customer reviews”, “recently viewed items and featured recommendations”, etc. This first part maximizes branching, or forking, so that the probability that a user will find many options (products) he desires is maximized as well (see figure 4).

When a user finishes with shopping and has filled his online cart with goods, the persuasive focus discreetly shifts to concluding the bargain. This constitutes the second part. The number of future possibilities is reduced to a necessary minimum in order for a user not to get distracted. At this point, the user is presented with a simple website containing only few items, as any diversion may detract his attention from the purchase, and thus threaten the payment completion (see figure 5). The two large yellow buttons “Continue”, unequivocally guide a user on a designated path to close the deal.

The screenshot displays the Amazon product page for an Apple iPhone 5S Space Gray 64GB Unlocked Smartphone (Certified Refurbished). The main product image shows the phone with its home screen. To the right, the product title and price are listed, along with a 'Certified Refurbished' badge. Below the main image, there are sections for 'Frequently Bought Together' (showing the phone with a case and screen protector), 'Customers Who Bought This Item Also Bought' (a grid of related accessories like cases and screen protectors), and 'Related Video Shorts' (a row of small video thumbnails). The 'Sponsored Products Related to This Item' section shows various phone cases and accessories. At the bottom, there is a 'Customer Reviews' section with a star rating and a 'Customer Images' gallery.

Figure 4 - Amazon.com view of maximized branching possibilities (First part).

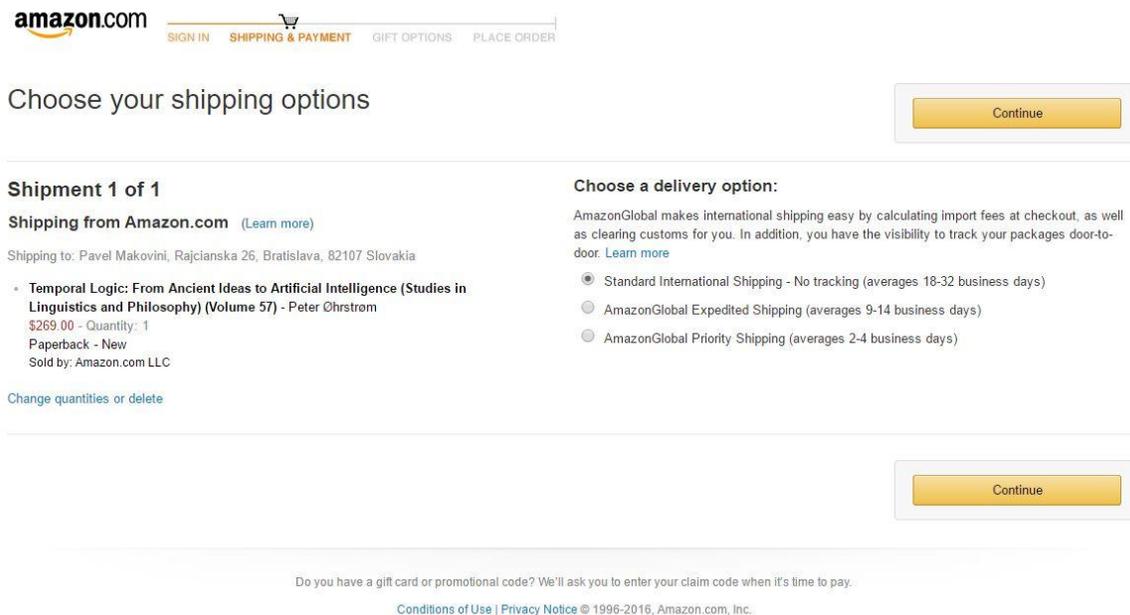


Figure 5 - Amazon.com view of minimized branching possibilities (Second part).

A second example of the branching tree is found in the field of information architecture (IA) and categorization. In their work, Iversen and Pertou agree with Hasle and Christensen who point out that persuasion is “not just the final step, but the entire route or process up to and including the final step that builds persuasion.” (Oinas-Kukkonen et al. (Eds.), 2008, p. 214)³. They draw attention to the importance of categorization, which not just improves usability, but also enhances persuasion. They point to the conclusions of Jesse James Garret, and also Rosenfeld and Morville, who explain, “The way we organize, label, and relate information influences the way people comprehend that information” (Morville, Rosenfeld, 2006, p. 53). In their own

³ Original citation in Hasle, P., Christensen, A.-K.K.: Persuasive Design. In: Kelsey, S., St. Amant, K.: Handbook of Research on Computer-Mediated Communication. IGI Global, Hershey (in print, 2008)

conclusion, Iversen and Pertou define categorization as “inextricably linked with persuasion”. They advise, “When a designer wants to design persuasive software he has to make a suitable categorization in order to strengthen his persuasive intentions. The designer must consider how the categorization he chooses influences the final software and how to categorize in a way that suits his persuasive intentions” (Oinas-Kukkonen et al. (Eds.), 2008, p. 222).

In a case study conducted in cooperation with the Baptist Church in Odense, Denmark, leaders of the church were asked to state the main persuasive purpose, or intention, of their website together with its primary target group. It was agreed that the main target group are all people, regardless of age, nationality or gender, who desire to explore spirituality, metanarratives or are directly in the process of choosing a local church. This group of users may be described by two primary characteristics: (a) users who are interested in religion, (b) users who do not have an extended experience with this particular church – new users. This preference was incorporated in the word *seeker*.

Based on this, a clear persuasive purpose was stated, “The church’s website attempts to persuade *seekers* to visit our service and become an active part of the local church.” Despite the clear persuasive intention, closer analysis of the existing IA of the website revealed that a category dedicated to this target group was missing entirely (see appendix 2). Therefore, to help in achieving this goal, an evident choice in the new IA included a category dedicated to *seekers* labelled “Jeg er ny” (I am new) (see appendix 3). Additional analysis of 53 English church websites along with IA research method called *card sorting* performed on six chosen church members has confirmed that forming this category corresponds with accepted patterns in labelling and categorization among popular contemporary church websites, and that it fits with the structures produced by

member's in the method of closed card sorting.⁴ Creating and proper labeling of a category was further enhanced by a prominent placement of the category in the system. Elevating the category to higher ranks of the branching tree, as opposed to placing it deeper in the hierarchy, increases the probability that a *seeker* will discover and respond to the content tailored for him. Namely, at rank 1 (point 0), the *seeker* is directly aware only of possibilities 1-3. Thus, placing the category "Jeg er Ny" on rank 2 considers his inside perspective in the outside perspective of the whole system. Placing the category on any lower rank of the system would not correspond with the immediate needs of a *seeker* entering the website at rank 1.

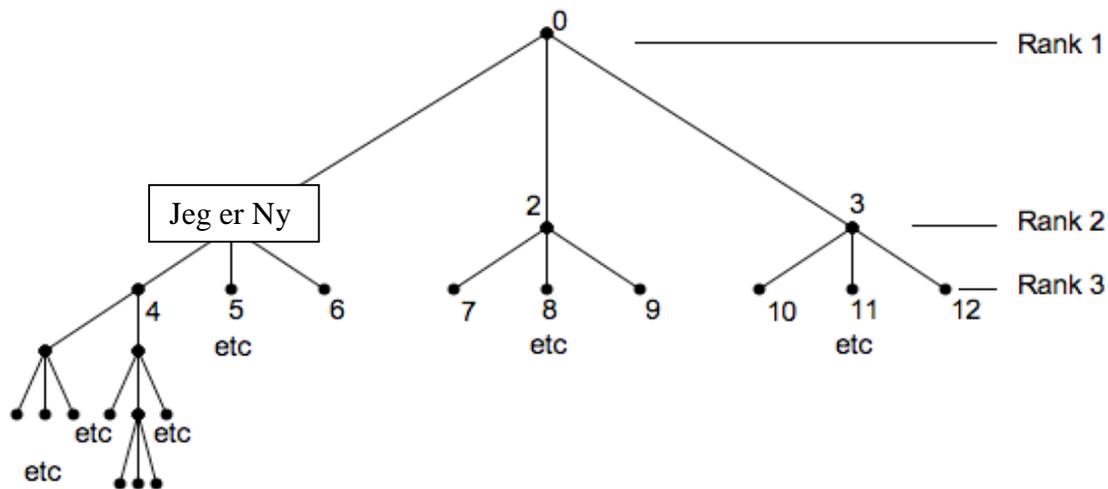


Figure 6—"Jeg er Ny" category on the website of the Baptist Church in Odense from the perspective of Krippel's branching tree.

The concept of branching time appears crucial in the free will discussion, yet it is also clearly beneficial in the design of IA and PD and programming. Greater elaboration of free will,

⁴ Both complete research projects with detailed definition, methodology description, and results can be acquired electronically upon request: makovini.peter@gmail.com

human autonomy and volition will be offered in later sections of this work. First, several models will be presented to provide fundamentals of PD.

Fogg's Behavior Model

Fogg suggests that there are three principal factors generating a behavior: sufficient *motivation*, *ability*, and effective *triggers* to perform the behavior. All three must come together or must “be present at the same time” for the behavior to occur (2009). Only if the combined level of *motivation* and *ability* is above the action line, or activation threshold, effective *triggers* succeed (see Figure 8). Thus, there is a trade-off relationship between motivation and ability.

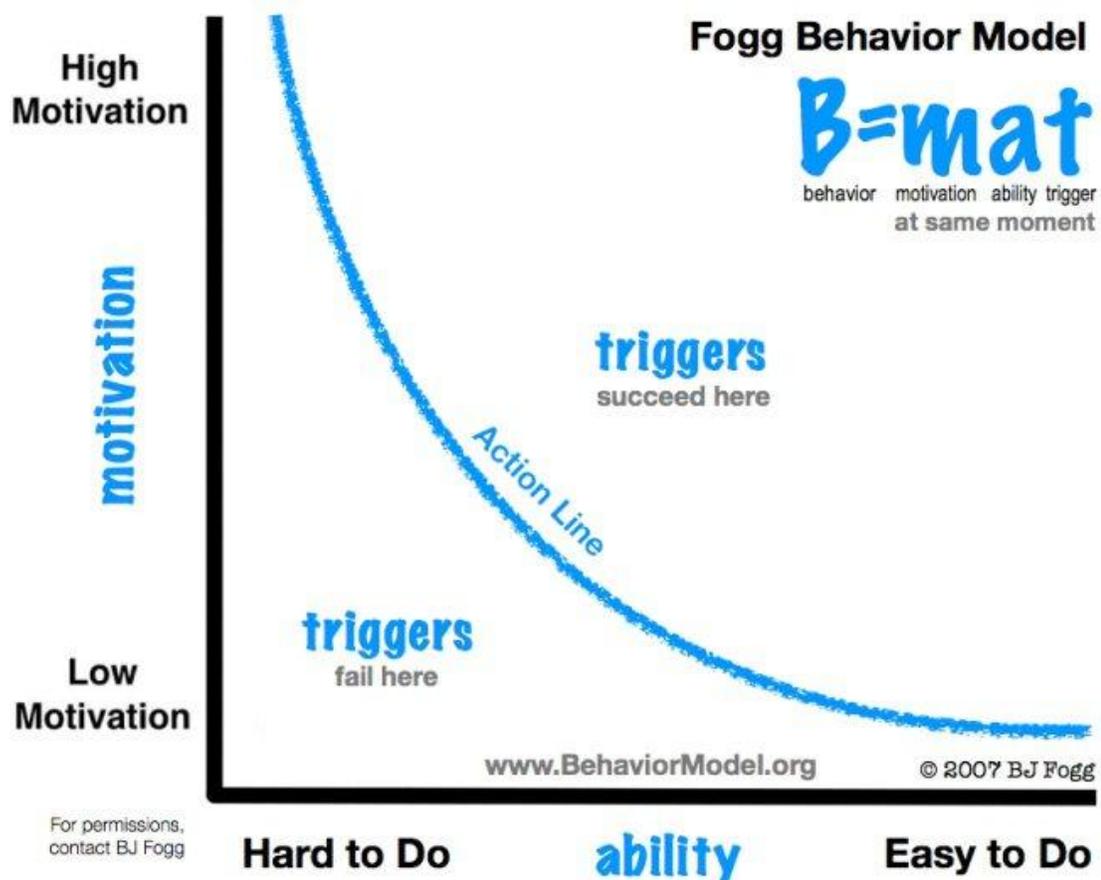


Figure 7 - Fogg's Behavior Model

To help a user have an overall higher motivation, Fogg identifies three specific core motivators, or motivational elements: *sensation* (pleasure/pain), *anticipation* (hope/fear), and *belonging* (social acceptance/rejection). To help a user have an overall higher ability, he recognizes two paths. A “hard path” is persuading people to learn new things and train them to have more skills⁵, or a “better path” is to make a target behavior easier to do, ergo decrease complexity/increase simplicity. Fogg offers, what may be properly called, a “simplicity chain” consisting of six links or elements: time, money, physical effort, brain cycles, social deviance, and non-routine. “If any single link breaks, then the chain fails” (2009, p.5) and simplicity is lost. Depending on the context, effective triggers are classified as *sparks* (insufficient motivation; should be connected with a motivator), *facilitators* (insufficient ability; should assure a user that he has all the necessary resources for a behavior), *signal* (sufficient motivation & ability; should only serve as a reminder or an indication) (Fogg, 2009).

The Functional Triad

Fogg offers three ways or roles on how interactive technologies can operate from the perspective of the user: as tools, as media, and as social actors (2003). Most PD and PT is a mix of these functions.

⁵ Fogg says that real-world products that require people to learn new things routinely fail. This is because people are by natural wiring fundamentally lazy. Simplicity changes behavior (2009).

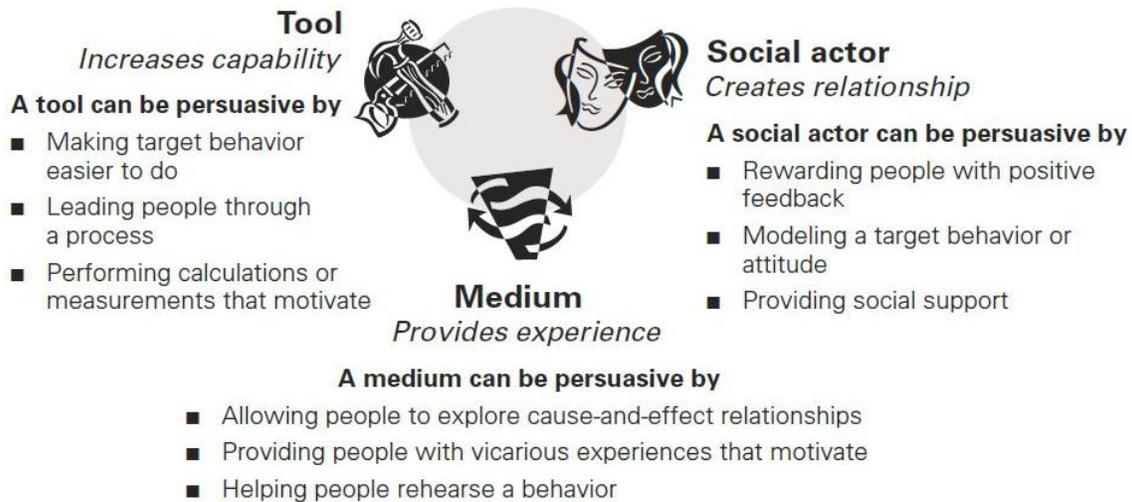


Figure 8 - Fogg's Functional Triad (2003, p. 25)

<p>Tools</p>	<p>Make desired outcome easier to achieve, either by reducing barriers to a behavior such as time, cost or effort, or (and) increasing effectiveness of users to do things virtually impossible without technology (e.g. GPS tracking, accurate self-monitoring). It includes leading human beings through a process. (Basamh et al., 2013)</p> <p>Endomondo is a smartphone app that uses a number of sensors to provide accurate self-monitoring information to users. It is a PD that motivates users to exercise by making it easier to track their progress.</p>	
<p>Media</p>	<p>Shape attitudes and behavior by providing compelling experiences that allow exploring simulated cause-and-effect scenarios, environments and objects.</p> <p>Website of Tesla Motors offers a simple Range Per Charge simulator allowing users to explore how changing various parameters like speed, outdoor temperature, wheel size or A/C ideally affects the range of the car.</p>	

Social Actors	<p>Leverage principles of social influence and social cues in humans. These are primarily physical (e.g. face, eyes), psychological (e.g. humor, empathy), language (e.g. spoken language, language recognition), social dynamics (e.g. cooperation, reciprocity, praise), or social roles (e.g. doctor, teammate, guide) (Fogg, 2003).</p> <p>Siri is Apple’s computer program that operates as an intelligent personal assistant acting like a social agent employing many of the social cues described above.</p>	
---------------	--	---

The functional triad has received wide acceptance, has often been laid out in various journals, and was used as a general framework for PD. Oinas-Kukkonen and Harjumaa called it “the first and most utilized conceptualization of persuasive technology.” (Oinas-Kukkonen et al., 2008, p. 169). It served as basis for other frameworks like persuasive systems design model (PSD) developed by Oinas-Kukkonen and Harjuma.⁶ It appears in works of e.g. Basahm et al. (2003), Zhu (Y. De Kort et al., 2007), Sundar et al., Ferebee and Davis (Bang, Ragnemalm, 2012), Miranda (Berkovsky, Freyene, 2013), Clinkenbeard et al., Zhang-Kennedy et al. (Spagnolli et al., 2014). Yet despite its popularity and contribution to the field, it did not avoid criticism entirely. Adaji and Vassileva write, “the Fogg’s functional triad has been studied extensively over the years and new frameworks have been developed based on this model”, however “as noted by Oinas-Kukkonen

⁶ PSD framework is attached in Appendix 1 for review as it, too, has a great contribution to the field of PD and was utilized in many projects. However, a more detailed treatment of PSD is beyond the scope of this paper.

and Harjumaa, Fogg's framework and principles are too general in terms of designing and evaluating persuasive systems." (Meschtscherjakov et al., 2016, p. 190).

In 2006, Bernardine M.C. Atkinson wrote "Captology: A Critical Review" where he suggests that "on close examination, the triad [...] seems to incorporate several 'categorical' or definitional errors" (Ijsselsteijn et al., 2006, p. 173). First, he questions whether the "tool" belongs to users or designers. If a persuasive objective was established independently of user's intentions, then PT is a tool of the designer, not the user. Only when users freely choose to adopt the course of persuasion to achieve desired behavior, value, or attitude can it be properly called a user's tool. Second, he finds the use of the category "medium/media" to be misplaced as medium is "the means by which something is communicated" (Ibid., p. 174). Fogg's usage of "media" relates primarily to compelling simulated experiences, not just any type of media. Thus, Atkinson believes that *simulations* "more accurately suit the definitional purpose proposed for the 'medium' element of captology's *functional triad*' (p61 – 89 of Fogg, 2003)" (Ibid., p.174). Third, Atkinson strongly disapproves the labeling of computers as "social actors" and deems it inaccurate. He calls attention to the proper usage of language. "Humans are social creatures; computers are machines" (Ibid., p. 175). A machine only has a function of a simulated social presence in its design, thus Atkinson suggests, that the term should be hyphenated (i.e. social-actor)⁷ or better yet called social-simulation since an actor is a man. He quotes his personal communication (2006) with Dr. Mitroy, who stated, "The computer does not exhibit the entire range of responses exhibited by humans acting under free will". Imitative functions of computers are merely mechanistic phenomena and

⁷ Atkinson explains, "This conveys a slightly different meaning to what we know to be a human-only designation, a social actor" (Ibid., p. 175).

“to infer ‘personality’ is an incorrect inference”. (Ibid., p. 175). Atkinson is particularly worried about the enhancements of social cues in creating a sympathetic human/machine dynamic that exploits our innate tendency to make the inference of “sociality” in computers. His final remarks are helpful quoting at length:

“These ‘social cues’ are features normally associated with living creatures: physical features like faces, eyes, voices and voice tonality and the type of language employed; social dynamics, like taking turns offering praise or answering questions; adopting roles such as that of an advisor, doctor, friend, and so forth [note omitted]. Using this concept of social actor uncritically, if we are not careful, will perpetuate an illusion, compound Baudrillard’s Procession of the Simulacra⁸ and cause us to fall victim to Rebe Dubo’s warning that humans continue to adapt to maladaptive situations.⁹ There are many dangers associated with being beguiled into believing we are interacting with genuine personality. But do we interact with computers, do we interact through them or do we simply use them? Fogg [note omitted] says that ‘the computing product is a participant in the interaction’. I beg to differ” (Ibid., p. 176).

⁸ Atkinson seems to misspell the title, which is “Precession of the Simulacra”. Here the main point of reference appears to be to Baudrillard’s three “orders of simulacra”. In his commentary on this work, Tseelon writes, “The first order, that of *imitation* characterized the classical period, presupposes dualism where appearances disguise reality. In the second order, *production*, appearances create an illusion of reality. In the third order, *simulation*, appearances invent reality. No longer concerned with the real, images are reproduced from a model” (Kellner (Ed.), 1994, p. 120).

⁹ In 1965, René Dubos wrote a paper called “Science and Man’s Nature” in which he attempted to show that “while the external environment and the ways of life are being revolutionized by technology, biological man remains fundamentally the same [...]. Outwardly, man makes adjustments to the new conditions of life; inwardly, however, he has so far failed to make true adaptations to them, and this discrepancy creates physiological and psychological conflicts which threaten to become increasingly traumatic” (p. 232).

Here, Atkinson poses a profound question, which however exceeds the borders of what is usually regarded as the study field of IA and PD. This question, crucial for the correct understanding of the nature of human-computer interaction (HCI), extends over to metaphysics and philosophy of mind. Consequently, it should not go unnoticed that Atkinson himself together with Mitroy, Fogg, and others, necessarily argue from their preconceived notions concerning the nature of humans, computers, actors, agents, personality, and free will. Such notions unavoidably affect the employed language.

Repeating Atkinson's first point, users ought to have the freedom to choose whether they want to follow and "adopt [program's] semiotic proffering to achieve desired new behavior, value or attitude" (Ibid, 173). Thus, users should have the possibility to choose whether they want to be persuaded by interacting with a social actor or not. Fogg writes, "Should those who create simulations reveal their biases to the users? I believe they should" if it was designed "to help people make health, financial, and other choices about their lives" (2003, p. 68). Given the increasing trustworthiness of social-simulations¹⁰, and granting a fundamental metaphysical difference between the nature of humans and computer simulations, Fogg and Atkinson may agree that social-simulations ought to reveal their social actor bias, just as any other simulation. To explicate, such a disclaimer may read,

"This is only a social-simulation! Any resemblance of lifelike, animate behavior is a mere illusion. This simulation is not capable of genuine emotions and other social

¹⁰ CNBC. (2016, Mar 16) https://www.youtube.com/watch?v=W0_DPi0PmF0 [Video File].

experience. We are not responsible for any harm that may be caused by an inappropriate handling of the simulation.”

Fogg offers a measure of realism as he recognizes that “revealing bias is not always desirable, practical, or effective. [...] Certainly, designers could—and perhaps should—try to expose users to the assumptions underlying a simulation. But if the product is designed to sell or to promote an ideology, it’s unlikely that creators will risk undermining their effectiveness by admitting to biases, however small” (2003, p. 68). Yet, what if the underlying assumption or ideology is gravely significant to the cause? If a designer believes that there is no fundamental metaphysical difference between humans and social-simulations, then using language such as “social actor” would seem perfectly apt. Such metaphysical foundation represents a point of disagreement that has important implications for the studies of PD. A broader treatment of this subject will follow in later sections of this work.

The Argument for Rational Persuasion

The overall reasoning structure of this work is stated in a form of a twofold formal argument. This argument aims to establish two pillars on which rational persuasion must stand – *freedom* and *rationality*. Therefore, initially two separate arguments are developed concurrently, which show why naturalism cannot account for rational persuasion. Thereafter, substance dualism is presented as the best explanation for *freedom* and *rationality* required for rational persuasion.

The Freedom Pillar

1. Persuasion requires the *freedom to do otherwise* given:
 - a. Persuasion is defined as a *voluntary* change of attitude or behavior free of coercion and deception.
 - b. *Voluntary* change requires existence of freedom of will.

- c. Freedom of will requires the *freedom to do otherwise* (PAP) given exactly the same past and laws of nature (i.e. Branching Time or Garden of Forking Paths).
2. Naturalism is incompatible with *freedom to do otherwise* given:
 - a. Naturalism presumes Materialism, Physicalism and physical causal closure.
 - b. Naturalism denies *efficacious* downward (mental-physical) causation.
 - c. Given (2a) and (2b), Naturalism presumes Determinism.
 - d. Determinism denies the *freedom to do otherwise* given exactly the same past and laws of nature (i.e. Linear Time or sequence of events)
 3. Therefore, persuasion is incompatible with Naturalism.

However, Compatibilists define free will as an *absence of (internal or external) constraints* and through Frankfurt-type examples challenge (3) arguing that freedom of will does not require the freedom to do otherwise (PAP).

An Objection from Compatibilism

4. Persuasion is compatible with Naturalism given:
 - a. Compatibilist-type of freedom falsifies (1c) and requires only *absence of constraints*.
 - b. *Absence of constraints* is compatible with Determinism.
 - c. Because of (4a, 4b) and (2c), Compatibilist-type of freedom is compatible with Naturalism.

Response to the Objection from Compatibilism

5. The Compatibilist-type of freedom is inadequate to account for a notion of free will given:
 - a. Covert non-constraining control (CNC) is by definition *absent of constraints*.
 - b. Because of (5a), Compatibilist-type of freedom is *also* compatible with CNC.
 - c. CNC disqualifies an agent to be an ultimate *source* and *origin* of his ends and purposes.
 - d. Determinism *also* disqualifies an agent to be an ultimate *source* and *origin* of his ends and purposes.
 - e. CNC is similar with Determinism.
 - f. Any definition of free will that accommodates CNC and Determinism is at best questionable.
6. Given (5), (4) may be rejected and (3) remains valid.

The *Rationality* Pillar

7. *Rational* persuasion is incompatible with Naturalism given:
 - a) *Rational* persuasion requires existence of *rational* inference.
 - b) Argument from reason shows that Naturalism is incompatible with *rational* inference.
 - c) Argument from the reliability of our cognitive faculties shows that (even if compatible with *rational* inference) Naturalism & Evolution would offer only highly unreliable *rational* inference.

Since, denial of *freedom* or *rationality* would for obvious reasons be devastating for this paper and all of academia, a best possible explanation must be given to account for these phenomena. Since the answer cannot be found in Naturalism, we are forced to look beyond. The suggested solution lies in some form of substance dualism, and in a basic explanation suggesting a second *agent-substratum* (to that of only material world), which by its nature is *rational* and has the attribute of *volition*. Thus, in certain moments it can make free undetermined decisions (SFAs). This substratum is capable of a different kind of causal relation, which is not entirely subject to the material and physical causal chain. This is referred to as agent-causality. While *prima facie* this account may appear as a mysterious stipulation, reasons will be given for why it is considered here to be the best possible explanation for *freedom*, *rationality*, and consequently *rational* persuasion. A form of best possible explanation argument follows.

The Best Possible Explanation Argument for *Rational* Persuasion

8. *Rational* persuasion is unattainable on Naturalism; its best possible explanation is some form of substance dualism given:
 - a. We are committed to the existence of human *freedom* and *rationality*.
 - i. Naturalism cannot account for human *freedom* and *rationality*.
 - ii. Therefore, Naturalism must be false.
 - b. We possess a properly basic *a priori* experience, viewing self as an undetermined free, *rational* agent (or a mind) that can exercise active power, initiate and redirect causal chains of our surroundings.

- c. Because of (8a, 8b), some form of substance dualism is the best possible explanation of *freedom* and *rationality*.

This argument shows that if Naturalism is true it serves as a defeater for rational persuasion and PD. In order to uphold feasibility of rational persuasion and PD, some form of substance dualism is unavoidable.

Application of the Argument for Rational Persuasion to AI

Reiterating the previous statement that *AI makes Persuasive Design honest*, it may be assumed that an intelligent AI would be the embodiment of an excellent product of Captology. Due to numerous sensors and sophisticated algorithms it would skillfully utilize Kairos using existing PD models of attitude and behavior change such as those described above (Fogg's Behavior Model and Fogg's Functional Triad) and many more. It is assumed, that such a PD simulation would easily pass the Turing Test and would make users believe that it is a free and rational being. However, given the Argument for Rational Persuasion that presupposes substance dualism, this simulation cannot, in principle, be a *free, rational* being. The following reasoning is applied. (i) Humans are *free, rational* agents by the virtue of their agent-substratum. (ii) Human-like AI lacks agent-substratum. (iii) Hence, human-like AI is not a *free, rational* agent. (iv) Therefore, human-like AI can at best be a persuasive delusion.

To keep PD honest and free of deception, the designer's social-simulation bias behind an AI should be revealed to avoid a delusion of a genuine freedom or rationality of such a simulation. The idea of a strong AI removes the possibility to speak loosely about social simulations and social actors in PD. As Atkinson suggests, we do not interact with simulations, they are not participants

in the interaction. Either we interact through them or we simply use them. Such is the honest nature of HCI in PD.

This was the overall argument structure of this paper. A reader may choose to come back to this section to better navigate through the sometimes intricate argument. Now, the premises and conclusions will be elaborated in detail.

Freedom of Will and Voluntary Choice

“Honestly, I cannot understand what people mean when they talk about the freedom of the human will. I have a feeling, for instance, that I will something or other; but what relation this has with freedom I cannot understand at all. I feel that I will to light my pipe and I do it; but how can I connect this up with the idea of freedom? What is behind the act of willing to light the pipe? Another act of willing? Schopenhauer once said: *Der Mensch kann was er will; er kann aber nicht wollen was er will* (Man can do what he will but he cannot will what he wills).”

- Albert Einstein¹¹

The problem of free will and necessity, or determinism, is one of the most difficult and “perhaps the most voluminously debated of all philosophical problems,” (2005, p. 1) says Robert Kane according to a recent history of philosophy. Debates concerning free will point to issues about “crime and punishment, blameworthiness and responsibility, coercion and control, mind and body, necessity and possibility, time and chance, right and wrong” (Ibid., p. 2) etc.. One is forced

¹¹ (Clarke, 2015, p. 84-85)

to question beliefs that to most people appear as *properly basic*¹². Is what we will, choose, and do, determined by the physical universe, physical laws and movements of atoms? Is who we become necessitated by our genes, heredity, birth, upbringing and social conditioning (nature & nurture)? Can our actions be accurately predicted through a sufficient insight into psychology, biology, chemistry, and physics of our bodies and environment? Can a change of mind or an outcome of persuasion affect the course of history by shifting the lane of branching time? Do feasible future alternatives exist or is time, and human life, ultimately linear regardless of our illusion of free choice? These questions have profound implications on the fundamental nature of persuasion and thus PD. Given strong determinism, it appears that persuasion cannot produce an actual difference in a world, where n makes m choose x instead of y . Given determinism m could not do otherwise than choose x , just as n could not do otherwise than persuade m to choose x instead of y . On the contrary, an indeterministic position creates place for rational persuasion that can influence free choices. This would elevate persuasion to a force or (and) a tool that may cause a real change in the world, diverting its course to a different path of a branching time tree (for better or worse).

Peter Clarke observed that an accurate definition is always important when discussing free will (2015). Classic dictionaries offer diverse definitions of free will:

Merriam-Webster: the ability to choose how to act; the ability to make choices that are not controlled by fate or God; voluntary choice or decision; freedom of humans to make choices that are not determined by prior causes or by divine intervention.

¹² A technical term often discussed by Alvin Plantinga. Properly basic beliefs are beliefs that given a persons' cognitive faculties are functioning properly may be rationally accepted without or apart from the evidential support of other propositions. These can be for instance "perceptual beliefs, memory beliefs, beliefs about the mental states of other persons, inductive beliefs and testimonial beliefs" (1993, p. 183).

Oxford: The power of acting without the constraint of necessity or fate; the ability to act at one's own discretion.

Cambridge: The ability to decide what to do independently of any outside influence

Collins: The apparent human ability to make choices that are not externally determined; the doctrine that such human freedom of choice is not illusory; the ability to make a choice without coercion.

The importance of freedom is paramount to the modern age. As described in previous section, it is important to protect the user's autonomy and perceive persuasion as primarily the user's voluntary change, free of coercion and deception. User should be able to pursue whatever goals *he* desires. This may be called a "surface freedom" (Kane, 2005, p.2) or "freedom of *external* constraints"¹³ (Clarke, 2015, p. 85). However, Kane illustrates that a world with this type of free will alone would not seem sufficient to us because,

"In such a world we would have a great deal of everyday freedom to do whatever we wanted, yet our freedom of will would be severely limited. We would be free to act or to choose what we willed, but we would not have the ultimate power over what it is that we willed. Other persons would be pulling the strings, not by coercing or forcing us to do things against our wishes, but by manipulating us into having the wishes they wanted us to have. [...] To some extent, we do live in such a world, where we are free to make choices but may be manipulated into making many of them by advertising, television, spin doctors, salespersons, marketers, and

¹³ A classic definition of compatibilists.

sometimes even by friends, parents, relatives, rivals, or enemies. [...] People feel revulsion at such manipulation and feel demeaned by it when they find out it has been done to them. [It] is demeaning because, when subjected to it, we realize we were not our own persons; and having free will is about being your own person” (2005, p. 2).

This scenario represents a relevant challenge for PD, since it points to a felt need of protecting not only the free volition of the user, but also securing that the ultimate source of volition can be traced back to the user. Thus Kane requests a “deeper freedom” (2005, p.3) or a “freedom from *all* constraints” (Clarke, 2015, p.85). What can be observed is that most dictionary definitions are “heavily loaded one way or the other” (Ibid.).

Several novels like *Brave New World* of Huxley or *Walden Two* of Skinner introduce futuristic societies that live according to their own desires and purposes, but their desires and purposes had been manipulated by others since birth by behavior conditioning or by drugs. Consequently, they can do what they want but what they want is determined by someone or something else. Their wills are determined by factors they do not control (Kane, 2005). The message of these novels has been in recent years made alive by outspoken groups of bloggers, journalists, neuroscientists and seculars such as Sam Harris, philosophers Paul Churchland, and Alexander Rosenberg or the physicist Victor Stenger who completely reject deeper freedom and deem it as an illusion of the human brain (Clarke 2015; Rosenberg, 2011; Reppert, 2003). Following where his materialist conclusions lead him, Rosenberg adopts a view called Eliminative

Reductionism¹⁴ that denies human agency entirely, claiming, “There is no free will, there is no mind distinct from the brain, there is no soul, no self, no person that supposedly inhabits your body...” (2011, p. 147). Some have rather ingeniously pointed out that determinists have lost their mind.

Daniel Dennett makes a distinction between natural determinism and control by other agents. He asserts that nature itself “does not control us” since nature is not an agent (1984, p. 61). Therefore, control by other persons as shown in the novels seem to undermine human freedom. It is objectionable, Dennett argues, because we are used as means to their ends. However, someone like Rosenberg may hardly appreciate such distinction because according to him there ultimately are no minds, souls, selves, agents or persons. Material nature is all there is, mental states or states of intentionality are illusory and thus Dennett’s move may seem to him as a distinction without a difference.

In concluding this chapter, Kane’s framework of human freedoms offers a degree of synoptic categorization. Kane elaborates on the various notions of freedom and offers five meanings that played an important role in historical debates about free will.

<i>Notion</i>	<i>Definition</i>	<i>Example</i>
<i>The Freedom of Self-realization</i>	The <i>power</i> or <i>ability</i> to do what we want or will to do, which entails an absence of external <i>constraints</i> or <i>impediments</i> preventing us from realizing our wants and purposes in action (all <i>surface</i> freedoms).	<i>Social</i> (buy what we want, go where we please, live as we choose, without interference or harassment from others), <i>political</i> (human rights like the freedom of speech, association, and vote)
<i>The Freedom of (Reflective or</i>	The power to <i>understand</i> and reflectively <i>evaluate</i> the reasons	Control of one’s own values, passions and desires as opposed to

¹⁴ Also referred to as Eliminativism or Eliminative Materialism.

<i>Rational) Self-control</i>	and motives one wants to act upon, or should act upon, and to <i>control</i> one's behavior in accordance with such reflectively considered reasons (freedom from internal constraints).	an internal constraint (weakness of will) experienced by drug addicts, the insane or severely retarded. This freedom is often associated with moral responsibility, and higher-order desires.
<i>The Freedom of Self-perfection (capacity self-correction)</i>	The power to understand and appreciate the right reasons for action and to guide one's behavior in accordance with the right reasons.	Knowing the difference between right and wrong as opposed to an utter moral confusion caused by e.g. a violent, sadistic upbringing (see the footnote about JoJo) ¹⁵ .
<i>The Freedom of Self-determination</i>	The power or ability to act <i>of your own free will</i> in the sense of a will (character, motives and purposes) of your own making—a will that you yourself, to some degree, were <i>ultimately responsible</i> for forming.	Responsibility and an ability of a person to ultimately influence whether he becomes a monster (JoJo) or a saint, despite his upbringing, heredity, environment or some other factor (self-determination). This ability does not however need to be available at all times. JoJo's self-determining decisions may have corrupted his will so much that he may no longer be able to do otherwise (self-formation).
<i>The Freedom of Self-formation</i>	The power to form one's own will in a manner that is undetermined by one's past by virtue of <i>will-setting</i> or <i>self-forming</i> actions (SFAs) over which one has plural voluntary control.	

Table 3 - Kane's Five Freedoms (2005, p. 163 – 174).

This chapter reveals that it is not sufficient to ask whether people have a free will. It is important to ask what the nature of the free will is and what notions of free will do people have. Eliminative materialists and other hard determinists consider free will to be illusory, soft

¹⁵ Wolf illustrates this freedom on the example of a dictator's son: "JoJo is the favorite son of Jo the First, an evil and sadistic dictator of a small undeveloped country. Because of his father's special feelings for the boy, JoJo is given a special education and is allowed to accompany his father and observe his daily routine. In the light of this treatment, it is not surprising that little JoJo takes his father as a role model and develops values very much like Dad's. As an adult he does many of the same sorts of things his father did, including sending people to prison or to death or to torture chambers on the basis of whim. He is not coerced to do these things, he acts according to his own desires. Moreover, these are desires he wholly wants to have. When he steps back and asks, "Do I really want to be this sort of person?" His answer is resoundingly "Yes," for this way of life expresses a crazy sort of power that is part of his deepest ideal" (2002, p. 153).

determinists or compatibilists commonly affirm the first three of Kane’s freedoms of will (self-realization, self-control, self-perfection), while indeterminists also called libertarians insist that genuine freedom of the will must be extended beyond the first three freedoms and thus add to the list two extra freedoms (self-determination, self-formation) (Kane, 2005; Clarke 2015; Harris 2012). The first three senses of freedom may allow for certain interpretations of persuasion, but as it will be argued later, genuinely free, rational persuasion requires the two extra freedoms available only to indeterminists. The next section will further elaborate on these three positions.

Determinism, Compatibilism and Libertarianism

The three basic positions on the freedom of will came about primarily as a conjunction of answers to two questions: Is determinism true? If yes, is it compatible with free will? Table 4 shows the usual relation between these positions with respect to these two questions.

Compatibilism	Hard Determinism	Libertarianism
<i>Determinism</i>		<i>Indeterminism</i>
<i>Compatibilism</i>	<i>Incompatibilism</i>	
<i>Weak Agent Reductionism (WAR)</i>	<i>Strong Agent Reductionism (SAR)</i>	<i>Agency</i>

Table 4 - Three basic positions on the freedom of will

Hard Determinism

A particularly apt insight into determinism can be found, once more, in the Greek mythology, where it was thought that *Chronos* (Aeon), the god of time had a consort *Ananke*, the goddess of inevitability, who emerged somehow by her own volition, simultaneously with Chronos, at the very beginning of time. Chronos and Ananke, time and inevitability (necessity), inextricably intertwined together, were thought of ultimately controlling the entire universe including the decisions of mortals and other gods. (Guthrie, 1965; Fanthorpe, L. & Fanthorpe, P., 2014). Because of Ananke’s unalterable nature it was pointless to render her offerings or sacrifice

(“Ananke“, 2006, p. 47). Fifth century BCE Greek atomist philosophers Leucippus and Democritus, arguably the first determinists, saw necessity as all-potent. Leucippus said, “Nothing occurs at random, but everything for a reason and by necessity“ (Guthrie, 1965, p. 415).

The modern naturalistic¹⁶ worldview does not seem to differ in the main points. The deities were replaced by (materialistic) nature and the personified control was replaced by Newtonian mechanistic physics (Kane 2005) entailing, at the basic level of analysis, a causally closed system (Reppert, 2003) or a physical causal closure (PCC) (Menuge, 2009). Living organisms, like humans who experience agency, a phenomenon of an enduring self, called consciousness, mind or soul, are reduced to the neural biological processes of the brain, further reduced to chemistry, and further reduced to physics (as shown in Figure 9). Since physics cannot be reduced any further, things have been broken down as far as possible, to the “basic stuff” of universe or “the most basic level of analysis”, as Reppert calls it.

¹⁶ In this work, *naturalism* is defined as the view that the natural world is all there is and there are no supernatural beings or causation, and all explanations must be limited to nonpurposive substratum. The most popular kind of naturalism is materialism (basic substances of the physical world are pieces of matter) and physicalism (these pieces of matter are properly understood through the discipline of physics); though it can include philosophies which either there is no matter per se or the base level is not physics. However, these types of naturalism still do not allow for purposive explanations (Reppert, 2003). Thus for all goals and purposes of this work naturalism will be used interchangeably with materialism and physicalism.

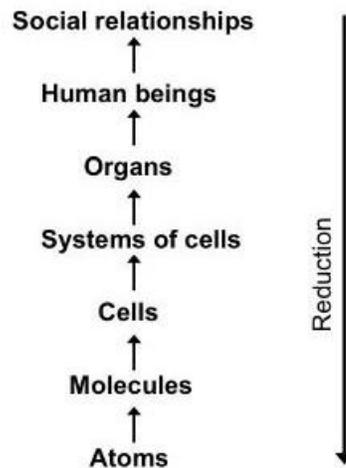


Figure 9 – Clark’s depiction of the levels of reduction. He sees reduction as “analysing the upper levels in terms of the lower” (2015, p. 72). These levels are complementary and do not exclude each other.

This is what the combination of words like eliminative materialism, reductive physicalism or eliminative reductionism relate to. In his book “Agents Under Fire“, Menuge calls these positions, that view all appearances of intentionality, deliberation, desires, beliefs and design as a mere complex undirected material processes - *strong agent reductionism* (SAR) (2004).¹⁷

The view of *Upward (physical-mental) causation*, in which the brain can influence the mind, e.g. disease, brain damage, fatigue or exercise, rest and medication, is palpable, virtually accepted and does not represent a challenge. However, *downward (mental-physical) causation*, in which the mind can have an effect on the brain is controversial in the philosophy of mind, particularly for reductive physicalists (Menuge, 2009). Traditional naturalism, that rests on materialism, denies the possibility of downward causation. Thus, anything that appears to transcend physicalist ontological resources, like consciousness, “in fact reduces to, supervenes on,

¹⁷ See table four for comparison.

or emerges from those resources, or else is nonexistent” (Menuge, 2011, p. 30). John Gibbons, a physicalist, argues that downward causation is nomologically impossible:

“We can rule out on empirical grounds any kind of mental-to-physical downward causation that involves actually making a difference.... It would do something that wasn’t already going to happen anyway. So the mental would have to be able to violate the laws of physics, or the laws of physics would have to be different inside and outside brains, or there would have to be new fundamental physical forces that only appear in brains.” (2006, p. 84) in Menuge (2009).

Essentially, this citation contains three ideas:

1. It is impossible for a mind to “actually make a difference” to the physical world because the physical world is causally closed.
2. If the mind did affect the physical world, it would “violate the laws of physics” or imply nonexistent laws.
3. Additionally, neuroscience has empirically removed the need for mind in explaining the brain’s activity.

Gibbon’s assumptions, and of others alike, serve as a foundation for what is called a Consequence Argument¹⁸ that was independently developed by David Wiggins, Peter van Inwagen, James Lamb, and Carl Ginet (Kapitan, 2002). The argument does not depend on determinism actually being true, it merely attempts to show what determinism implies *if* it were true, namely, no free will. Thus, it is an argument for the incompatibility of free will and determinism. Peter van Inwagen’s informal version can be stated as follows:

¹⁸ Also called the “Incompatibility Argument” and the “Unavoidability Argument” (Kapitan, 2002).

“If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not up to us what went on before we were born; and neither is it up to us what the laws of nature are. Therefore the consequences of these things (including our own acts) are not up to us”¹⁹ (1983, p. 16).

Kane’s own formal presentation of Inwagen’s Consequence Argument reads,

1. There is nothing we can now do to change the past and the laws of nature.²⁰
2. Our present actions are the necessary consequences of the past and the laws of nature. (definition of determinism)
3. There is nothing we can now do to change the fact that our present actions are the necessary consequences of the past and the laws of nature.
4. There is nothing we can now do to change the fact that our present actions occur (Kane, 2005, p. 23-24).

This can be applied to any agent and action in time and implies that *if* determinism is true, *no one can do otherwise* than he does; if free will requires the power to do otherwise, then *no one has free will* (Kane, 2005). Such conclusion is consistent with hard determinism and SAR listed in this section, however, as expected, it presents a problem for compatibilists who want to affirm free will. Their response is presented in the later section on Compatibilism.

¹⁹ Another informal version of the argument is offered by a compatibilist Thomas Kapitan, “If determinism is true, then whatever happens is a consequence of past events and laws over which we have no control and which we are unable to prevent. But whatever is a consequence of what is beyond our control is not itself under our control. Therefore, if determinism is true, then nothing that happens is under our control, including our own actions and thoughts. Instead, everything we do and think, everything that happens to us and within us, is akin to the vibration of a piano string when struck, with the past as pianist, and could not be otherwise than it is” (Kapitan, 2002, p. 127).

²⁰ This is a consequence of separated premises: 1) There is nothing we can now do to change the past. 2) There is nothing we can now do to change the laws of nature.

Free Will and Introspection

It is often argued that introspection and our experience of freedom serves as a subjective (first person) argument for the free will as we feel that we are the authors of our own thoughts and actions. Samuel Harris argues that while free will is nonsensical objectively, since it defies the laws of nature, it makes no sense subjectively either. He suggests, that we do not recognize this because we do not pay close enough attention to our own feeling of freedom. Harris brings the point across with an example from his daily life.

“I generally start each day with a cup of coffee or tea--sometimes two. This morning, it was coffee (two). Why not tea? I am in no position to know. I wanted coffee more than I wanted tea today, and I was free to have what I wanted. Did I consciously choose coffee over tea? No. The choice was made for me by events in my brain that I, as the conscious witness of my thoughts and actions, could not inspect or influence. Could I have “changed my mind” and switched to tea before the coffee drinker in me could get his bearings? Yes, but this impulse would also have been the product of unconscious causes. Why didn’t it arise this morning? Why might it arise in the future? I cannot know. The intention to do one thing and not another does not originate in consciousness—rather, it *appears* in consciousness, as does any thought or impulse that might oppose it.”

(Harris, 2012, p. 7-8)

Benjamin Libet, a professor of psychology, had in 1958 conducted a series of empirical neurophysiological experiments with human subjects²¹. These subjects were asked to flex a finger in a moment of their choosing. During the experiment, a device was attached to their scalp recording electrical activity of their brain concerned with voluntary bodily movement. During the experiment the subjects were asked to note the instant they *decided* to push the button; or more accurately, the instant they “felt the conscious act of willing their wrist to flex” (Rosenberg, 2011, p. 152). The method was more sophisticated than described above, but this much should serve the purpose of this paper. The results have shown that the subjects’ conscious act of willing – choice – to make a movement occurred on average 200 milliseconds before the movement of the finger. However, brain activity of their motor cortex was detected in average 550 milliseconds before the movement of the finger. Thus, Libet demonstrated that voluntary acts are preceded by a specific charge in the brain, the *readiness potential* (RP). In other words, he showed that prior unconscious processes are set into motion several hundred milliseconds before human subjects become consciously aware of their intention to act (see figure 9).

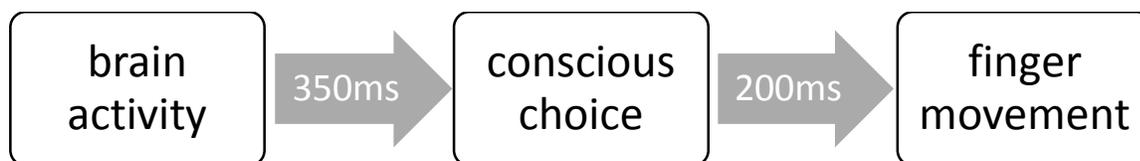


Figure 10 - Results of Libet's neurophysiological experiment on willing and consciousness (Kane (Eds.), 2002, p. 551-564)

Assuming these findings are reliable, what can be concluded? What is particularly remarkable about Libet's study is that it has received distinguished attention and frequently appears in the works of many philosophers and scientists to this day. It is discussed in (Rosenberg, 2011; Harris,

²¹ In collaboration with neurosurgeon Bertram Feinstein.

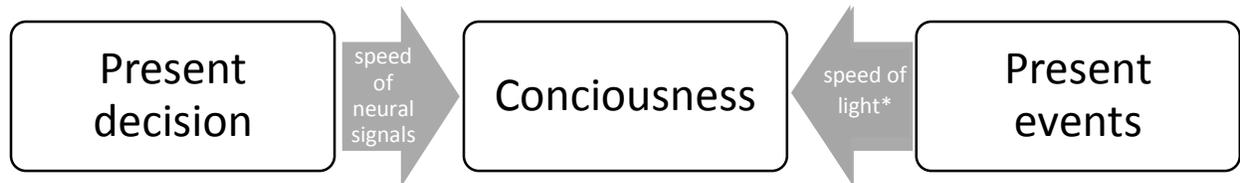
2012; Kurzweil 2005; Penrose, 1989; Dennett, 2003; Clarke 2015; Lowe, 2004; Kane, 2002; Menuge, 2009; Craig, 2013) and many others across the spectrum of the free will debate. Rosenberg writes that similar results have been replicated many times since with improvements in technology. To him, the implications of these results seem obvious, “Consciously deciding to do something is not the cause of doing it. It’s just a downstream effect, perhaps even a by-product, of some process that has already set the action in motion. A nonconscious event in the brain is the ‘real’ decider” (2011, p. 153). Harris concludes with a similar certainty, “One fact now seems indisputable: Some moments before you are aware of what you will do next [...] your brain has already determined what you will do. You then become conscious of this “decision” and believe that you are in the process of making it” (2012, p. 9).

While some take this to show, or at least point towards the denial of free will, others reject this conclusion and refer to further Libet’s results, namely that RP was not always followed by the action (Lowe, 2004; Kane, 2002; Menuge, 2009). Subjects appeared to have a *veto* power over their movements. i.e. they could have refrained from moving their finger. Thus, it seems that RP prepares body for action, but does not produce action deterministically. Neuroscientists Ramachandran once wittily noted, “This suggests that our conscious minds may not have free will, but rather ‘free won’t!’” (Ramachandran, 1998, p. 35 in Dennett, 2003). Dennett remains unimpressed and asks whether the unconscious initiation of the finger movement—flick— could not simply be followed by another unconscious initiation of a veto power that entered consciousness shortly after. Otherwise, it is assumed that “the brain is talented enough to work out the details of implementation on how to flick over that period of time, but only a “conscious function” is talented enough to work on the pros and cons of a veto decision” (2003, p. 44). Libet saw this problem and admits the possibility that there may be factors “on which the decision to

veto (control) is based, do develop by unconscious processes that precede the veto.” (Libet, 2002, p. 559) However, he maintains, “the conscious decision to veto could still be made without direct specification for that decision by the preceding unconscious processes” (Ibid.). Libet himself finds his own results inconclusive. He points out that nearly all humans experience free, independent choices, which provides a *prima facie* evidence that conscious mental processes can causatively control some brain processes (downward causation). Libet’s conclusion is that “free will, one genuinely free in the nondetermined sense” is “at least as good, if not a better, scientific option than is its denial by determinist theory.” (Ibid., p. 563). Philosopher William Lane Craig, upon contemplating Libet’s results arrived at the opposite interpretation to that of Rosenberg, Harris or Dennett. Craig says, “*this is exactly what the dualist-interactionist²² would expect*” (2013). According to Craig, the mind uses the brain as an instrument to think. Since neural processes travel at finite velocities, there is naturally a lag between the mind’s decision and the conscious awareness of them. That is not to say that the decision is unconscious; “it is a conscious decision, but because of the finite velocity of neural signals it takes time for the person to become conscious of it. Just as we never see present events because of the finite velocity of light, but only events just

²² Howard Robinson in the Stanford Encyclopedia of Philosophy defines dualist-interactionism as “the view that mind and body—or mental events and physical events—causally influence each other. That this is so is one of our common-sense beliefs, because it appears to be a feature of everyday experience. The physical world influences my experience through my senses, and I often react behaviourally to those experiences. My thinking, too, influences my speech and my actions. There is, therefore, a massive natural prejudice in favour of interactionism.” (2016, 3.1)

slightly past, so we do not have consciousness of our decisions simultaneously with our making them but unnoticeably afterwards.“ (Ibid.)



*Figure 11 - Craig's interpretation of Libet's results (2012). *It may be logically argued that the speed of neural signals would also delay our experience of present events. I.e. upon retina receives an impulse, it takes a short time for the signal to reach consciousness.*

In the attempt to offer an answer to the freedom of will, overall, it seems little can be concluded on the basis of Libet's results as there is no consensus concerning the interpretation of his work. Rosenberg (materialist-determinist) agrees that the experiments do not prove that there is no free will but he claims that the results reveal that introspection is not a trustworthy source of information regarding the existence of free will. "What you certainly can't do after reading about these experiments is trust introspection to tell us whether or not we have free will. We can't trust introspection to tell us *when* we made the decision to push the button. We certainly can't trust introspection to tell us *why* we made the decision we did" (2011, p. 154). After listing several other examples where introspection proved inaccurate he asks, "What is there left for introspection to be reliable about?" (Ibid.). He also answers his own question: "If the most obvious things consciousness tells us are just plain wrong, we can't trust it to tell us anything about ourselves" (Ibid., p. 148). With introspection removed, Rosenberg's commitment to materialism is unrestrained, which drives him to seemingly absurd conclusions such as "that we never think about anything or that I do not endure through two moments of time or that I do not even exist" (Craig, 2013); all of which are said to be illusions. Both Harris and Rosenberg doubt introspection. What is the outcome of such a conclusion? What can be said of a situation when we cannot see our

cognitive faculties, such as introspection, as reliable indicators of truth? The following section will examine the consequences of this assumption if it should be applied logically and consistently to our general ability to reason.

Materialism, Epistemology and Defeaters

“There was only one catch and that was Catch-22, which specified that a concern for one's safety in the face of dangers that were real and immediate was the process of a rational mind. Orr was crazy and could be grounded. All he had to do was ask; and as soon as he did, he would no longer be crazy and would have to fly more missions. Orr would be crazy to fly more missions and sane if he didn't, but if he was sane he had to fly them. If he flew them he was crazy and didn't have to; but if he didn't want to he was sane and had to. Yossarian was moved very deeply by the absolute simplicity of this clause of Catch-22 and let out a respectful whistle.

“That's some catch, that Catch-22,” he observed.

“It's the best there is,” Doc Daneeka agreed.”²³

“...if I have an undefeated defeater for R [reliability of my cognitive faculties], then by the same token I have an undefeated defeater for any other belief B my cognitive faculties produce, a reason to be doubtful of that belief, a reason to withhold it. For any such belief will be produced by cognitive faculties that I cannot rationally believe to be reliable. But then clearly the same will be true for any proposition they produce: the fact that I can't rationally believe that the faculties that produce that belief are reliable, gives me a reason for rejecting the belief.” (Plantinga, 1994, p. 13)

Consistent naturalists such as Harris and Rosenberg have concluded that introspection is unreliable and that we cannot even trust it to tell us anything about ourselves. It is a conclusion that puts into question one way of how we come to know things about our environment, other people and us; it addresses the matter of epistemology. However, some philosophers object that

²³ Heller, 1961, p. 52

Harris, Rosenberg and other materialists do not take their naturalism far enough and commit, what is among philosophers colloquially known as Taxicab fallacy. It means that, they drive their skepticism regarding our cognitive faculties only as far as it suits their purposes and then, when it begins to be inconvenient, they opt out. The argument suggests that if materialists drove the whole circle, their skepticism of our cognitive faculties would make them skeptical of their skepticism of our cognitive faculties. In his book *Miracles*, C. S. Lewis claimed that “strict materialism” could be refuted by a one-sentence argument. In his attempt to do so he quoted J. B. S. Haldane, “If my mental processes are determined wholly by the motions of atoms in my brain, I have no reason to suppose that my beliefs are true ... and hence I have no reason for supposing my brain to be composed of atoms" (Lewis, 2001, p. 15).

What is the nature of rational thought or inference? Do humans possess this faculty? If yes, can *reasoning* be reconciled with determinism? In answering these questions, a case will be made showing that resources available to naturalism cannot adequately account for our ability to reason; following the example of others, it will be referred to as the argument from reason.

The Argument from Reason

After his interaction with the criticism of Elizabeth Anscombe, one of the “most gifted philosophers of twentieth century” (Driver, 2011), Lewis refined his original versions of the argument to this form:

1. No belief is rationally inferred if it can be fully explained in terms of nonrational causes.
2. If materialism is true, then all beliefs can be fully explained in terms of nonrational causes.
3. Therefore, if materialism is true, then no belief is rationally inferred.

4. If any thesis entails the conclusion that no belief is rationally inferred, then it should be rejected and its denial accepted.
5. Therefore, materialism should be rejected and its denial accepted. (Reppert, 2003)

In cases when rational inference, principles of logic or other essential reasoning, inducing or deducing ability are in the center of a dispute, as it is in the case of the argument from reason, it is impossible to prove their validity. If someone should be skeptical about our reasoning ability, says Reppert, an attempt to prove him wrong inevitably involves reasoning and thus is deemed invalid at the onset, reaching impasse.

“Neither side can refute a skeptic about the basic principles of logic, but both must assume the legitimacy of those principles in order to argue at all. [...] It is not necessary to raise the question of whether there is such a thing as reasoning: we must presuppose that there is. [...] If the materialist wishes to say that we are not entitled simply to presume that rational inference occurs, then we can point out the disastrous epistemological consequences involved in denying rational inference.”

(Reppert, 2003, Chap. 3, Sec. 3, Par. 13-14)

Thus, such arguments should not be approached from the skeptical end, but rather be formulated as best explanation arguments where human reasoning is assumed as an established fact. Given that human beings are capable of rational inference, the important question should be what is the best possible explanation of it being so?

Another route a materialist may take is to reject the first premise and defend the proposition that rational inference is compatible with determinism. This is what Anscombe did in her criticism.

She suggested that there are four different types of explanation (naturalistic causal, logical, psychological, personal history) and that we have no reason to assume that explanations of one type are incompatible or even in competition with explanations of other types (Reppert, 2003). To illustrate, we may use Clarke's levels of reduction from Figure 9. If someone observes Joe meeting Sarah and says that Joe is in love with Sarah (social relationships), it is compatible with saying that Joe's heart rate and blood pressure is increased (organs), or that dopamine production is increased (molecules) or, in theory, that Joe's atomic structure resembles atomic structure of people who are in love (atoms). Clarke sees all of these levels as compatible, not excluding each other. Anscombe uses a similar line of reasoning with respect to *causal* and *reasons* explanations,

“It appears to me that if a man has reasons, and they are good reasons, and they are genuinely his reasons, for thinking something – then his thought is rational, whatever causal statements can be made about him” (Anscombe, 1981, p. 229).

Similar conclusion was made by Keith Parsons, who thinks that if sufficient reasons were adduced for a conclusion Q, then Q is rational.

“The causal history of the mental states of being aware of Q and the justifying grounds strike me as quite irrelevant. Whether those mental states are caused by other mental states, or caused by physical states, or just pop into existence uncaused, the grounds still justify the claim” (Parsons, 2000, p. 101).

If Anscombe and Parsons are correct then Lewis is merely offering a different type or level of explanation while there is no actual conflict between *causal* and *reason* explanations. As a way of response, Lewis argues that there are two types of connection:

1. Cause and effect
2. Ground and consequent

When we say, (1.) “Joe finds Sarah attractive because she is truly gorgeous” a *cause* of Joe’s affection for Sarah is given. However, we may say, (2.) “Joe finds Sarah attractive because he bought her flowers”; in this case, we do not refer to a *cause* of Joe’s affection (Joe clearly does not find Sarah attractive *because* he bought her flowers; buying flowers is not Joe’s *cause* of affection for Sarah), but now we are speaking about *evidence* of Joe’s affection for Sarah.

Lewis’ argument suggests, that while every event in nature is related to one another by cause and effect, premises in rational inference must be related to the conclusion by the ground and consequent relationship. Moreover, the relationship must be relevant to the belief. (Reppert, 2003) Lewis explains, “One thought can cause another not by being, but by being *seen to be*, a ground for it” (2001, p. 16). This “to be seen” factor is also known as a logical connection of rational inference. However, if blind natural causality that inevitably follows a sole string of determined chain of events (naturalism) is behind all our beliefs, according to Lewis, any meaningful concept of logical connection and rational inference is absent, impossible, or just irrelevant. Therefore, how someone came to his beliefs, how his beliefs were caused, or the *source* or *origin* of one’s beliefs appears to be crucial.

There is a problem with the unalterable path of materialism that does not allow a person *to do, believe or choose otherwise* than he is determined to. If a person is to be considered rational, it seems that he should have a possibility to believe otherwise. Reppert gives an example,

“If you were to meet a person, call him Steve, who could argue with great cogency for every position he held, you might on that account be inclined to

consider him a very rational person. But suppose it turned out that on all disputed questions, Steve rolled dice to fix his positions permanently and then used his reasoning abilities only to generate the best available arguments for those beliefs selected in the above-mentioned random method. I think that such a discovery would prompt you to withdraw from him the honorific title *rational*” (2003, Chap. 3, Sec. 4, Par. 11).

The point Lewis is making is that while explanatory compatibility is surely a valid and useful concept enabling a description of the same event from various points of views; in realist philosophies, it has its limits. With respect to philosophical naturalism, these limits lie in its monism. One of the central concepts of materialism, the PCC, allows *exclusively* only physical (upward) causation. The possibility of other types of causality, like mental (downward) causality, is by definition of materialism closed or excluded. Reppert thus suggests that the relation between (mental) *reasons* and (physical) *causes* is better understood through a consideration of a man’s death, which on one hand is explained through voodoo witchcraft and on the other in terms of a heart attack. The antirealism of Wittgenstein may suggest that these explanations are still compatible, since the voodoo curse may have caused the medical condition of a heart attack. However, in the framework of materialism, a voodoo-cause explanation must be automatically excluded, as it is not available in its inventory of possible causal explanations. The only explanation consistent with materialism is death by the physical cause of heart attack, that was caused by some previous physical state, that can have its cause traced back to the physical states even before this person was born, all the way back to the origin of the universe. Thus, voodoo curse

type of explanation is incompatible with materialism. It would be incorrect and inconsistent to suggest that voodoo was merely a different type or level of explanation (Reppert, 2003).

Likewise, with rational inference, if a person is to be *convinced* or persuaded by reason to hold a belief or exercise a behavior, then a nonspatio-temporal entity called logical connection must play a causal role. But nonspatio-temporal causal entities, like logical connections and rational inferences, just like voodoo above, are not in the inventory of possible causal explanations. The only option available to materialism is to attempt to explain rational inference and mental causation as emerging, supervening or being a kind of subspecies of physical causation (Reppert, 2003). This position is called Nonreductive Physicalism, Jaegwon Kim (2010), Menuge (2009, 2013) show that this option simply is not compatible with the core principles of physicalism; it faces a challenge known as the exclusion problem²⁴ that, even if granted, effectively makes mental causation do no real work. Similar approach is adopted by compatibilists when they attempt to give an account of free will. Yet this presents other examples and nuances that are dealt with in the section on Compatibilism.

²⁴Explanation of the exclusion problem by Kim, further edited by Menuge, “To see this, consider any case of mental causation. Suppose mental state M causes a further mental state M*. By hypothesis, M is completely determined by some physical base state P, and M* is completely determined by some physical base state P*. Given the assumed priority of the physical over the mental, M* cannot exist without its base P* (or some alternative base, which we may assume is not present), so M must cause M* by causing P*. However, physicalism is also committed to the causal closure of the physical which implies that every event has a purely physical cause. So, given the dependence of M on P it is natural to say that P causes P*, and hence that P cause M*. For without P, M would not be there, and hence P* and M* would not be there, so it appears that P causes P*, and hence M*. But, assuming we do not allow systematic overdetermination, if P causes M*, and P has ontological priority over M, then M cannot also be the cause of M*: M is excluded” (Menuge, 2013, p. 52).

If the argument from reason is correct, then there is an inherent conflict between existence of rational inference and materialism, it entails that materialism is self-defeating if it is presented as a belief that was deduced by rational inference. With respect to the overall argument of this paper this entails that *rational* persuasion is not possible given materialism.

The Argument From the Reliability of our Rational Faculties

In his book *Warrant and Proper Function* (1993), Alvin Plantinga has proposed an evolutionary argument against naturalism, which was in the heart of his paper *Naturalism Defeated* (1994) and was later modified in *Knowledge of God* (2008) co-authored with Michael Tooley to specifically address the relation between neural structures and beliefs with content. The argument assumes that if naturalism is true, then life, as well as our cognitive faculties are the result of naturalistic evolution. This relation is highly probable since, evolution is the only process available, or as Plantinga says “it is the only game in town” for the naturalist, which can account for the current variety of flora and fauna (1994, p. 13). However, evolutionary theory is by definition exclusively interested in enhancing chances for survival, and *not* in appreciation of truth propositions of the external world by living organisms. Patricia Churchland insists that the principal function of the human brain from the evolutionary point of view is “to succeed in the four F’s: feeding, fleeing, fighting and reproducing” (1987, p. 548). In other words, “Natural selection doesn’t care what you *believe*; it is interested only in how you *behave*” (Plantinga, 1984, p. 13). Plantinga puts these propositions in the form $P(R/N\&E)$, where,

R: is proposition that human cognitive faculties are reliable

N: is proposition that naturalism is true

E: is proposition that evolution is true (1984).

Due to the relatively successful state of our survival, it seems clear that human cognitive faculties have developed in a direction of fitness-enhancing behavior. Thus, some may intuitively assume that reliability of our cognitive faculties in producing objectively true beliefs automatically follows. But, Plantinga asks, is it possible that fitness-enhancing behavior would produce mostly false beliefs? What is the relation between adaptive *behavior* and true *beliefs*? If it follows or it is probable that adaptive *behavior* would produce also true *beliefs*, then $P(R/N\&E)$ is high, but if adaptive *behavior* does not guarantee true *beliefs* making them improbable then $P(R/N\&E)$ is low.

Plantinga thinks that $P(R/N\&E)$ is low or inscrutable. He examines four mutually exclusive and jointly exhaustive possibilities of the relationship between *behavior* and *beliefs* with respect to $P(R/N\&E)$. (1) *Epiphenomenalism simpliciter*, (2) *semantic epiphenomenalism*, (3) the possibility that beliefs are causally efficacious with respect to behavior but *maladaptive*, and (4) the possibility that beliefs are both causally efficacious with respect to behavior and *adaptive*. He explains that on (1) and (2), beliefs are not involved in the (semantic) causal chain leading to a behavior, thus beliefs seem irrelevant or *invisible* to evolution, which entails that probability of $P(R/N\&E)$ may be rated as low. On (3) beliefs are involved in the causal chain but lead to a *maladaptive* behavior and therefore can harm its possessor. Given (3), it seems that probability of R on N&E may be estimated also as relatively low. At last, (4) suggests that beliefs are causally connected to an *adaptive* behavior. Plantinga calls this the common sense view. But despite the wide acceptance of (4) he says that is not at all probable that these beliefs need to be true. He explains that for any adaptive behavior there are many possible belief-desire combinations that could produce it; yet most of these belief-desire combinations may be false. To illustrate his point,

he presents a creative story about Paul, the prehistoric hominid who is approached by a tiger. In all cases Paul choose arguably the best survival behavior – fleeing, but such action may be produced by a large number of belief-desire combinations:

“Perhaps Paul very much likes the idea of being eaten, but when he sees a tiger, always runs off looking for a better prospect, because he thinks it unlikely that the tiger he sees will eat him. This will get his body parts in the right place so far as survival is concerned, without involving much by way of true belief. . . . Or perhaps he thinks the tiger is a large, friendly, cuddly pussycat and wants to pet it; but he also believes that the best way to pet it is to run away from it. . . . or perhaps he thinks the tiger is a regularly recurring illusion, and, hoping to keep his weight down, has formed the resolution to run a mile at top speed whenever presented with such an illusion; or perhaps he thinks he is about to take part in a 1600 meter race, wants to win, and believes the appearance of the tiger is the starting signal; or perhaps . . .” (1993, p. 225-226).

This peculiar account shows that, in theory, there may be many belief-desire combinations that would produce this adaptive behavior leaving Paul with false beliefs, thus, probability of $P(R/N\&E)$ can be hardly assumed as high. Plantinga extrapolates from this example and includes other models showing that, in fact, *most* of Paul’s beliefs could be false if one of his *systemic* or general belief was false, nevertheless still resulting in an adaptive behavior.

After reviewing the four possibilities, he attempts to estimate their added average probability of $P(R/N\&E)$. He readily admits that these calculations are merely “vague estimates” that may be “imprecise and poorly grounded”. Nevertheless, Plantinga asserts, that is all the

argument needs. Following the aforementioned reasoning, given naturalism, the general value of $P(R/N\&E)$ will be either *less than 1/2* or *inscrutable* (sensible agnostic position); that is enough for the argument. The next step of the argument suggests that this gives the naturalist-evolutionist a reason to doubt R – a *defeater* of R . Before proceeding with the argument, functioning and categories of defeaters needs to be first clarified.

Epistemological Defeaters and Defeat

On the internet encyclopedia of philosophy, David Truncellito defines epistemology as the study of knowledge (n.d.). Epistemologists traditionally define a tripartite nature of knowledge as justified true belief (Sudduth, n.d.). Once justified true belief is formed, a so-called *defeater* of a belief may remove it. In short, defeaters are reason to change one's beliefs in a certain way. Philosophers interpret defeaters as conditions either external or internal to the cognizer.²⁵ Taking the route of an internalist, mental state defeater (MSD) condition for knowledge may be stated as follows:

S knows that *p* only if *S* does not have a mental state defeater for *S*'s belief that *p*.

Michael Sudduth, doctor of philosophy specializing in religion at the University of Oxford, describes MSDs as “situations where a person *S* justifiably believes *p* at some time *t* but then at some later time *t** *S* acquires a mental state *d* (some new experience or belief) that causes *S*'s belief that *p* to be unjustified at *t**. Here *S*'s belief that *p* is unjustified from the time *S* acquires the mental state *d*” (Sudduth, n.d., Chapter 4, Section 1, para 2).

²⁵ Externalists use the language of true propositions while internalists focus on experiences and beliefs called mental states.

Defeaters may be categorized as reasons for supposing that p is false (*rebutting*) and defeaters that would sufficiently lower the likelihood that p is true (*undercutting* and *no reasons defeaters*).²⁶ Beilby in his analysis of Plantinga's no-defeater condition recognizes three kinds of defeaters: conscious, reflective and external. A *conscious* defeater is a belief, which if an agent is aware of, will counter against another belief. A *reflective* defeater is a belief an agent is not immediately aware, but given reflection would become consciously aware. Beilby uses an example, "I many not immediately realize that my belief J. R. R. Tolkien was born in 1896 is defeated by another belief of mine, Tolkien died in 1973 at the age of 81, because I do not pause to reflect on the mathematical incompatibility of these beliefs. But upon reflection I would become aware that I have a defeater for my belief" (2005, p. 169). An external defeater is a defeater, of which an agent is not aware, but it would be a conscious defeater, should the subject become aware of it (Beilby, 2005).

In an online Persuasive Design Survey conducted in 2015 aimed to understand users' attitudes towards business and religious types of persuasion²⁷, several participants have displayed skepticism over the concept of PD. They wrote, "It can be deceptive to some people. Better avoid the use of persuasive concept.", "I don't think it's ok, because PD would allow companies to manipulate and cheat people.", "In a way it manipulates with people, to get a professional designing your website.", "It can be misused to convince a weak "target" to buy something they

²⁶ "Undercutting defeater is a reason for supposing one's ground for believing p is not sufficiently indicative of the truth of the belief", while "no-reason defeater is when one has no reason for believing p and the belief p is the sort of belief that is reasonable to hold only if one has evidence for p ." (Sudduth, n.d., Chapter 6, Section 1, para 1-4)

²⁷ The hypothesis, methodology, results, discussion and research data of my Persuasive Design Survey can be acquired electronically upon request: makovini.peter@gmail.com

really don't want." Other comments were more positive in nature, yet those above reveal a concern the discovery of PD presents for some people.

Using the language of epistemology, a situation when a person discovers presence of PD can be described as follows. A condition where a person *S* justifiably believes he had made an informed, free, deliberate decision (*p*) at some time *t* but then at some later time *t** *S* acquires a mental state *d* (discovery of persuasive design) that causes *S*'s belief that he had made an informed, free, deliberate decision (*p*) to be unjustified at *t**. Discovery of PD does not directly suppose that *p* is false; it does not rebut *p*, it rather lowers the likelihood that *p* is true; undercutting *p* by introducing a reason for supposing person's ground for believing he had made an informed, free, deliberate decision (*p*) is not sufficiently indicative of the truth of the belief. At time *t* may PD be classified as an *external* defeater, potentially changing to a *reflective* defeater, which at some later time *t** becomes a *conscious* defeater.²⁸

Similarly, supposing that *S* believes that he was captured by Alpha-Centaurian super-scientists who performed a cognitive experiment on him and have given him mostly false beliefs, then *S* has a defeater of R (reliability of his cognitive faculties) that serves as a systemic (undercutting) defeater for all his beliefs. *S* needs not to be sure of this scenario; it is enough if its probability is inscrutable. Then *S* has a reason for doubting, and withholding his natural belief in R (Plantinga, 1994).

²⁸ PD represents epistemological infringement in form of an undercutting MSD, which may account for the negative comments in the Persuasive Design Survey. In light of the above mentioned research it may be assumed that any deliberate, systematic effort, to change *S*'s attitude or behavior, attaining a level of complexity that is perceived as obscure to *S* will serve as an undercutting mental state defeater to his belief that he made an informed, free deliberate decision.

A Defeater-defeater

Obtaining a defeater for a belief p is not conclusive. Belief p is defeated only as long as defeater of p remains itself undefeated. Therefore, defeaters are *relative* to currently held propositions, (*a priori*) beliefs²⁹, reasons, experiences; or as Plantinga calls them - *the noetic structure*. The following example shows the successive nature of defeaters.

“I know that you are a lifeguard and believe on that ground that you are an excellent swimmer. But then I learn that 45% of Frisian lifeguards are poor swimmers, and I know that you are Frisian: this gives me a defeater for the belief that you are a fine swimmer. But then I learn still further that you graduated from the Department of Lifeguarding at the University of Leeuwarden and that one of the requirements for graduation is being an excellent swimmer: that gives me a defeater for the defeater of my original belief: a defeater-defeater as we might put it” (Plantinga, 1994, p. 12-13). It is possible to add to the series *ad libitum*.

Suppose that S has an undefeated defeater for p , but continues to hold p anyway. What is precisely the problem? Plantinga asserts, “Presumably this is a deplorable state of affairs; even if it isn't a punishable offense, there is something wrong, unhappy, regrettable about it.” In such a case, S would be “in an *irrational* condition of some kind; there would be something irrational about [S], or more precisely about the structure of [S 's] beliefs” (1994, p. 21). Thus, in cases of a defeated belief, rationality dictates to withhold that belief.

²⁹ Memory beliefs are properly basic, *a priori*, beliefs that are not held on the basis of reasons. Yet we may have defeaters for memory and other *a priori* beliefs (Plantinga, 1994).

Self-Defeating Defeaters

Can a proposition be a defeater of itself? Indeed. Expressions like “You should not try to persuade people” or “You should not judge” are logically self-defeating, since the accuser is by uttering these propositions guilty of the same thing of which he is accusing his opponent. It may be seen, upon short reflection, that statements like “There is no truth”, “All truth is relative” or “You should doubt everything” are also self-refuting when they are applied to themselves. One of the first pages in this paper includes a statement *this page has been intentionally left blank*. Though it conveys a clear idea, it is essentially self-defeating; not because it conflicts with some other proposition, but because by being placed on the page, it defeats the very idea it is trying to convey. Reverting back to the example of S who believes that Alpha-Centaury super-scientists gave him mostly false beliefs, S has a defeater of R that serves as a systemic defeater for *all* his beliefs. But if S has a reason to doubt *all* his beliefs he has a reason to also doubt the belief that Alpha-Centaury super-scientists gave him mostly false beliefs. In this case, a defeater of R serves as a self-defeating defeater of R.

Back to the Argument from the Reliability of our Rational Faculties

It has been demonstrated that, given naturalism, the general value of $P(R/N\&E)$ is either *less than 1/2* or *inscrutable*. Therefore, naturalist-evolutionist has a reason to doubt R; he has a *defeater* of R. In this aspect, the naturalist-evolutionist is much like the believer in Alpha-Centaury super-scientists; both have a defeater of R. Can a naturalist offer a defeater-defeater of R? Plantinga argues he cannot, because any such defeater-defeater would be either a belief, experience or some other element in his noetic structure, but any such element would be subject to the same defeater as R is. Thus, he insists, “this defeater can’t be defeated” (1994, p. 13).

Arriving at the conclusion of the argument, this entails that a naturalist has an *ultimately* undefeated defeater of R, thus having an undefeated defeater for, and a reason to doubt and withhold, *all* his beliefs and the propositions these beliefs produce. Consequently, the naturalist also has a reason to doubt and withhold the belief for N&E itself. Therefore, conjunction of N&E is self-defeating and cannot be rationally accepted (Plantinga, 1994).

Objections to the argument from the reliability of our rational faculties

Plantinga's argument motivated many responses, to which he, over the next two decades have consistently responded. Some of these objections, Plantinga thinks, showed that the objectors have misunderstood the argument or overlooked certain possibilities. This helped Plantinga to identify areas where further elaboration of the argument was needed. However, other counter-arguments like "The Dreaded Loop" have offered valid objections to which Plantinga "penitently" offers corrections (1994, p. 55) in order to refine his argument. Due to the limited space, only two objections will be briefly considered.

The Dreaded Loop Objection

In his conclusion, Plantinga have said that a devotee of N&E has (i) an *ultimately* undefeated defeater of R, because he falls in (ii) a diachronic loop in which "he believes N&E and sees that this gives him a defeater for R, and hence for N&E; so then he stops believing N&E; but then he loses his defeater for R and N&E; then presumably those beliefs come flooding back; but then once again he has a defeater for them; and so on, round and round the loop" (1994, p. 16).

First, the objector suggests that instead of (i) a devotee of N&E would have a defeater that is not ultimately defeated. It means that every time a devotee of N&E has a defeater for N&E,

there is a subsequent time where this defeater is defeated. This loop is never terminated and therefore a devotee of N&E has a defeater that is not ultimately defeated and *not* (i) an ultimately undefeated defeater of N&E.

Second, rationality demands that a person can anticipate the deadlock of (ii) the loop and thus stays out of it, or at least abandons it after a couple of rounds. Anyone falling and remaining in (ii) such a loop would have to be “extremely imperceptive” (Plantinga, 1994).

Response to the Dreaded Loop Objection

Plantinga concedes, “The objector is right on both counts” (1994, p. 55). Then he goes on to explain the unusual nature of the situation. In most cases when a person has a defeater-defeater of B, the defeating power of B’s defeater is neutralized. However, in this (unusual) case the *defeatee* never loses its defeating power, but shows up at every subsequent level. This may be simply expressed as follows. If a devotee of N&E believes that R is *low or inscrutable* then on

Level 0 N&E holds *a doubt of R* (-R).

Level 1 -R holds a doubt of (N&E holds -R).

Level 2 N&E holds -R, thus it also holds a doubt of a doubt of (N&E holds -R).

And so on...³⁰

This shows that N&E does not have a defeater that is not ultimately defeated, but an ultimately undefeated defeater in a way that if N&E holds -R, it does not hold it only on levels 0, 2, 4 etc.

³⁰ Plantinga offers a formal, more accurate presentation of this idea, which was in this paper rephrased to a simpler form for better readability. Original may be found in Plantinga, 1994, p. 57.

but on all levels. Therefore, also on levels 1, 3, 5, etc. when a devotee of N&E has a defeater-defeater of N&E, he has it in the virtue of the ever-present $\neg R$. The doubt remains in *all* subsequent levels. But this is Plantinga's point; such defeater-defeater of N&E amounts for a naturalist to a Pyrrhic victory, because it is tantamount to his defeat. As long as N&E holds $\neg R$ it gives a proponent of N&E defeater for *everything* he believes. Because of that, naturalism cannot be rationally accepted (1994).

Objection from Sensible Naturalism

In his response to the argument, *In Defense of Sensible Naturalism* (2007), Paul Draper provided several points of disagreement with Plantinga, including Plantinga's definition of Naturalism. However, Plantinga remained unimpressed by Draper's definitional contention, because of, which he dealt with it only briefly. It may perhaps be said that he found it directly *irrelevant* to his argument. In this respect, a tone of banter can be sensed behind the quotation marks in the title of his answer to Draper, *Against "Sensible" Naturalism* (2007)³¹.

The main point of Draper's objection follows. While Draper agrees that if $P(R/N\&E)$ is *low or inscrutable* then the argument is sound and naturalism cannot be held rationally, he maintains that this is not right. Instead he is convinced that it is only *inscrutable*. If this is true, then, Draper asserts, the conclusion does not follow. In order for the argument to hold, Plantinga must show that $P(R/N\&E)$ is *low*. Draper starts his counter-argument by quoting Plantinga,

³¹ Draper's differentiation of *Sensible* from *Extreme* Naturalism can be found in his paragraph on "Varieties of Naturalism" (2007), to which Plantinga's response can be found in his second and third paragraph (Plantinga, 2007).

“Now if content of belief did enter the causal chain that leads to behavior-- and if true belief caused adaptive behavior (and false belief maladaptive behavior)--then natural selection, by rewarding and punishing adaptive and maladaptive behavior respectively, could shape the mechanisms that produce belief in the direction of greater reliability. There could then be selection pressure for true belief and for reliable belief- producing mechanisms“ (Draper, 2007, “The Inscrutability of P(R/N&E)“).

But while Plantinga thinks this is unlikely and merely possible, Draper suggests this scenario is highly probable. He argues that our long term survival appears much more probable if our cognitive faculties are reliable and the beliefs we hold are mostly true than mostly false. Reversely, the fact we have survived for so long is strong evidence for R. To understand why, Draper offers an example in which he goes to take a bath and finds an alligator in his tub.

“It is certainly possible that I survive these unfortunate circumstances without having true beliefs like "there's an alligator in my bathtub" and "alligators are dangerous animals." For example, the beliefs that "there's a beautiful mermaid in my bathtub" and "mermaids, especially beautiful ones, are dangerous animals" may *do just as well* [...]. Notice, however, that the *vast majority* of false beliefs I might have in these circumstances (e.g., there's nothing in my bathtub, there's a gentle alligator in my bathtub, there's a rubber ducky in my bathtub, there's a dangerous alligator in my bathtub but I can easily overpower it, etc.) will *not do just as well*, but will lead instead to a, shall we say, "maladaptive" bathing experience. (Ibid.)

On that account, Draper believes that the probability of R given the conjunction of his version of *sensible* naturalism and evolution, $P(R/S\&N\&E)$ is high. More broadly, he is willing to grant that $P(R/N\&E)$ is at best inscrutable. He concludes that Plantinga can establish the *possibility* of cognitive faculties that are unreliable and adaptive, but “this does nothing to refute the fact that by far the most likely way for blind evolution to produce adaptive cognitive faculties is to make them reliable” (Ibid.).

Response to Sensible Naturalism Objection

Plantinga sets out to show that both his original premise $P(R/N\&E)$ and $P(R/S\&N\&E)$ are low. This time he formulates his argument using neurophysiological properties (NP properties) that are in causal relation in producing content properties. Further, to conform to Draper’s Sensible Naturalism, the produced content of NP structure must be *causally effectual* in some way. To make the matter simpler, Plantinga asks readers to think of a very simple organism with a very simple neural circuitry. Most simple organism as bacteria do not have beliefs at all, but as the complexity increases, we may imagine the emergence of the first belief. The complexity of neural structure of this organism S would allow for the first NP property P that would cause a certain behavior C by the virtue of having a certain proposition Q as content. It may be assumed that P causes adaptive behavior, therefore it causes S to have an adaptive content Q, but Plantinga insists that, given just sensible naturalism, this “provides not the slightest reason to think Q is true” (Plantinga, 2007, Par. 6). It might be true, but it might be equally false. Natural selection selects for adaptive properties, causing adaptive behavior. The property P therefore causes content Q, which in turn causes adaptive behavior. It was said that on sensible naturalism NP of S must be *causally effectual* by the virtue of having the content it does. All right, so S causes adaptive behavior by virtue of its

content, however, Plantinga explains, “it doesn’t cause adaptive behavior by virtue of having the property of having *true* content” (Ibid.). There is no connection between the truth value of Q and the adaptiveness of the behavior it causes. This holds true for every subsequent belief of the organism S, and the beliefs of other species including us. Conclusively, “Q could be true, but it is equally likely to be false” (Ibid. Par. 9). At last, Plantinga performs a brief probabilistic exercise in which he shows that if a creature has 1000 beliefs and the reliability requirement is, say, that at least $\frac{3}{4}$ of these beliefs are true, then the probability is very low. $P(R/N\&E)$ cannot be greater than $\frac{1}{2}$, which is low enough to provide a defeater for R.

Plantinga thinks that what Draper fails to realize is that while the survival of our species is clearly more to be expected if our cognitive faculties (in a broad sense) are reliable, this does not entail we must have true beliefs. He explains that, if a frog is to catch a fly, it must have properly functioning “indicators”, neural structures, which receive inputs from sense organs and correlate the speed and direction of a fly with the proper muscles so the frog is able to flick out its tongue and catch the fly (Plantinga, 2007). But such indication does not require belief. Long term survival assumes proper functioning of these indicators; human body has indicators for blood pressure, saline content, temperature, insulin level and much else, yet they function properly without a need for anyone to hold a belief on the topic. Similarly, “Fleeing predators, finding food and mates-- these things require cognitive devices that in some way track crucial features of the environment, and are appropriately connected with muscles; but they do not require true belief, or even belief at all” (Plantinga, 2007, Par. 11).

Based on his probabilistic account of $P(R/N\&E)$ being low and the broader concept of cognitive faculties, including properly functioning cognitive indicators, Plantinga makes the

following conclusion. On naturalism, truth or falsehood are just irrelevant, sometimes adaptive behavior is caused by truth other times by falsehood, therefore our long term survival is not much more probable on having truth beliefs than having mostly false beliefs (2007).

This concludes the section on Hard Determinism. Reasoning behind PCC and the Consequence Argument has shown that we cannot be *free*, while the Argument from Reason and the Argument from the Reliability of our Cognitive Faculties reveal that on naturalism we cannot be *rational*. Therefore, it is hard to see how on this view, anyone can be a free, rational agent. Yet if *freedom* and *rationality* of our choice are in principle not possible, the prospect of a genuinely voluntary, rational change of mind is a delusion too. This presents a defeater for the existence of rational persuasion for a hard determinist.

Compatibilism

Proponents of Compatibilism accept many of the basic tenets of hard determinism with one vital exception. They understand determinism to be harmonious with the existence of free will. This position is popular among modern philosophers as it neatly solves the free will problem by suggesting that, in fact, there is none. Throughout the history it was advocated by influential philosophers like Spinoza, Thomas Hobbes, John Locke, David Hume, John Stuart Mill, presently being defended by Peter Clarke and Daniel Dennett (Kane, 2005; Clarke, 2015).

Can determinism be reconciled with the free will? Kane, and other researchers, suggest that most people intuitively resist the idea when they encounter it for the first time. Others like Eddy Nahmias found that people mostly think of free will as compatible with determinism. Therefore, Clarke sees these conclusions on the beliefs of ordinary people regarding this issue as contradictory and unresolved (2015). Nevertheless, for those who find the relationship between

free will and determinism unsettling, compatibilists offer ready arguments that may help people dispose of their natural inclination to incompatibilism.

One of the common mistakes is to confuse determinism with *fatalism*, the view that whatever is going to happen, is going to happen, *no matter what we do* (Kane, 2005). This encourages questions like “If everything is determined, why should I do anything? Why not just sit back and see what happens?” Harris says, “This is pure confusion” (2012, p. 33). Even if determinism was to be true; decisions, intentions, efforts, goals, willpower and so on, make a difference in how things turn out. There are exceptional cases when our deliberation makes no difference as Dennett shows with an example of a man, who after jumping off the bridge decides halfway down that he wants to live. Such deliberation makes no difference to his fate. However, such cases are rare, and compatibilists suggest that most of the time deliberations do affect our future, even if determinism was to be true (Kane, 2005; Harris, 2012).

Second common mistake is to confuse determinism with *constraint*, *coercion* or *compulsion*. These are, by definition, actions that are *against* one’s will, preventing a person to do or choose what he wants. But natural determinism does not have to go against our will. Kane explains, “to be governed by laws of nature is not to be in chains” (2005, p. 18).

Third common mistake, Kane suggests, according to compatibilists is to confuse determinism with *mechanism*. That is to say, if determinism were to be true, humans would all be like machines, operating mechanically; similar to computers, robots or lower biological organisms such as amoebae or insects, which act automatically and instinctually with only a limited set of responses. Compatibilists say that people unlike machines, computers or robots have an “inner conscious life of moods and feelings” and that we “reason and deliberate, question our motives,

reflect on our values, make plans about the future, reform our characters” (Kane, 2005, p. 21) and so on. Even given determinism, this spectrum of complex capacities introduces a completely different degree of freedom to that of amoebae or machines.

Classic compatibilists define free will “as the ability to make choices or perform actions *free from external constraints*” (Clarke, 2015, p. 87). As discussed in previous sections, this definition offers what Kane calls, a *surface* freedom or a freedom of self-realization. Yet, new compatibilists recognize also the threat of *internal constraints* and affirm the existence of other Kane’s freedoms like self-control or self-perfection without compromising determinism. Though compatibilists are often accused of restricting the definition of free will or cheating (as Kant called compatibilism a “wretched subterfuge”) they argue that, at least in English, free will was commonly used in the compatibilist sense of having the (1) *power* or *ability* to do something without (2) *constraints* or *impediments*. (Clarke, 2015; Kane, 2005). Clark illustrates, “If I say that I cleaned the bathroom “by my own free will”, I am not talking about freedom from brain determinism. I simply mean that I chose to do the job and nobody forced me to do it. This is not a dishonest new-speak designed to escape from the implications of modern neuroscience...” (2015, p. 87). Kane explains that one is free to take a bus, if he has the *power* or *ability* to take it, should he want or decide to do so. On the contrary, “I would not be free to take the bus if various things prevented me: such as being in jail or if someone had tied me up (physical restraint); or if someone were holding me at gunpoint, commanding me not to move (coercion)³²; or if I were paralyzed

³² Coercion is also to give a choice, which is no choice at all, e.g. a thief, holding a gun saying “Your money or your life”.

(lack of ability); or if buses were not running today (lack of opportunity); or if fear of crowded buses compelled me to avoid them (compulsion), and so on” (2005, p. 13).

Incompatibilists agree, yet they ask whether freedom does not entail the freedom *to do otherwise*. If Determinism is true, it seems there is only one possible future, with no room for branching time. No garden of future forking paths suggests no freedom to do otherwise. Compatibilist answer takes shape in *conditional* or *hypothetical* meaning. Surely, you are free not to take the bus *if* you (1) have the *power* or *ability* to avoid taking it, *if* there are (2) no *constraints* preventing you not taking it, (3) *if* you *wanted* not to take it. The third point represents for the compatibilists freedom to do otherwise, i.e. that you *would* have done otherwise (given (1) and (2)) *if* you had *wanted* or desired to do otherwise (Kane, 2005). On determinism, this amounts to saying that in another possible world, where the past or laws of physics were different, you could have done otherwise (given (1) and (2)). At this point libertarians ask, but could have I done *otherwise* in this world? To this, traditional compatibilists committed to determinism must say a resolute “No”. The kind of a *deeper* freedom, where agents have an actual *ultimate* control or *ultimate* responsibility (UR) over what they will or want in this world is incompatible with compatibilism. Though some may be unsatisfied, compatibilists argue, the so-called *deeper* freedom of the will was incoherent and unavailable to us in the first place. Such freedom is illusory and nonsensical (Kane, 2005).

The *hypothetical* meaning of “can” (as opposed to actual “can”) is important in compatibilist response to the Consequence Argument. While they affirm the premises of the argument, since no one has the *power* or the *ability* to change the past and the laws of nature even *if* he *wanted* to, they do not see that the conclusion (in short, “no person can do otherwise than

they actually do”) necessarily follows from the premises. On their definition of freedom, an agent still *could* have done otherwise *if* he had chosen or wanted to do otherwise (given (1) and (2)). On compatibilism, the agent *could* choose not to take the bus, not in an actual sense, but in a hypothetical sense. Thus, on hypothetical analysis (HA) Consequence Argument fails (Kane, 2005).

Libertarians acknowledge it, but in their response insist that the HA must be mistaken. The serious objection many philosophers raise with regard to the HA of “can” and “could have done otherwise” is that it obscures the situation and (wrongly) suggests that agents could have done otherwise, in cases where it is clear that they could *not* have done otherwise. Michael McKenna illustrates this objection on a fictional girl called Danielle, who at her young age was terribly scared by an accident involving a blond Labrador retriever. Thus she was,

“psychologically incapable of wanting to touch a blond haired dog. Imagine that, on her sixteenth birthday, unaware of her condition, her father brings her two puppies to choose between, one being a blond haired Lab, the other a black haired Lab. He tells Danielle just to pick up whichever of the two she pleases and that he will return the other puppy to the pet store. Danielle happily, and unencumbered, does what she wants and picks up the black Lab” (2004, section 3.3).

Was she able to pick up the blond Labrador? It seems not, McKenna says. Because of her traumatic experience, this *want* was not available to her and thus, in this respect, she *could* not have done otherwise. The problem with the HA, that says, “She *could* have done otherwise, if she did *want* to...” is that it suggests Danielle could have done otherwise (if she had *wanted*), when in fact, she could *not* have *wanted* to do otherwise. So to truly capture the meaning of “can” or “could” one

must add after the HA a qualifier, "...and she *could* also have *wanted* to do otherwise" (Kane, 2005, p. 30). Then compatibilists may argue with yet another round of HA, pushing another hypothetical "want" statement resulting in a conspicuously tautological sentence: "She would have wanted or chosen to do otherwise, *if* she had *wanted* or *chosen to want or choose* otherwise." This HA statement requires yet another "could" statement, which would be followed by another HA statement, again and again, making the statement ever longer up to infinity, never allowing for a definite answer to the original question: Could Danielle have done otherwise?

If the compatibilist analysis suggests that Danielle can do otherwise, even though she can't change the past and the laws of nature and even her *want* was a necessary consequence of the past and the laws of nature, then (in Kane's retelling of Inwagen's and Ginet's conclusion from a similar case) "something must be wrong with the hypothetical analysis of "can" that...compatibilists favor" (2005, p. 29). Therefore, there are reasons to think HA is flawed and this line of reasoning does not seriously undermine the Consequence Argument.

Freedom in the Absence of Alternative Possibilities

Another strategy compatibilists employ is to reject the need for alternative possibilities of the forking paths altogether. Thus, when Libertarians ask "Isn't freedom *to do otherwise* necessary for freedom?" They choose to say "No.". This line of reasoning was introduced by Harry Frankfurt who saw that the Consequence Argument ultimately rests on the Principle of Alternative Possibilities (PAP):

Persons are morally responsible for what they have done only if they could have done otherwise.

If the moral responsibility requires the kind of free will that can do otherwise then (AP),

Free will requires the power to do otherwise, or, alternative possibilities (Kane, 2005, p. 80-81).

Up to this point, PAP and AP was granted and undisputed, but if it can be shown that free will does not require the power to do otherwise, Consequence Argument would fail from the start. This is because its *raison d'être* was to show that determinism removes the power to do otherwise. However, if the power to do otherwise was not necessary for free will, the Consequence Argument would be attacking a straw man and ought to be deemed irrelevant.

Character-Type Examples

In the 16th century, at the dawn of reformation, Martin Luther stood at the Roman Catholic court that asked him to recant his teaching, to which some believe he replied with the famous statement, “Here I stand. I can do no other.” Dennett uses this example, assuming Luther was literally incapable of doing otherwise due to his reasoning and experiences, to show that this did not exempt him from the moral responsibility. In saying, “I can do no other”, Dennett suggests, Luther “was not trying to duck responsibility” (1984, p, 133). In his case, he meant, “I cannot because I see so clearly what the situation is and because my rational control faculty is *not* impaired. It is too obvious what to do; reason dictates it; I would have to be mad to do otherwise, and since I happen not to be mad, I cannot do otherwise” (Ibid.). Kane writes that Luther was taking full responsibility for his act and “Indeed, it may have been the most responsible act of his life” (Kane, 2005, p. 82). Consequently, Dennett concludes that PAP is false, as of result, AP is also false; thus the consequence argument is undermined.

One possible Libertarian response to this character-type example is to bite the bullet and refer to the existence of previously mentioned self-forming actions (SFAs). It may well be the case that Luther could presently do no other and his action was now determined, but his current disposition was formed by the virtue of his earlier struggles and self-forming choices, which brought him to a point where he could do no other. Therefore, his present moral accountability *must* be viewed in the light of a broader view; his context, history, past choices and actions, which he at the time could have done otherwise, and *not* focus on Luther's individual actions in isolation. Kane proposes, "Often we act from a will already formed, but it is "our own free will," by virtue of the fact that we formed it by other choices or actions in the past (SFAs) for which we could have done otherwise (which did satisfy AP)" (2002, p. 408). If Luther, or anyone else, *could never have done anything to make ourselves different from who we are*; it is difficult to see how we can be ultimately morally responsible for what we do. Hence, character-type examples do not demonstrate that free will and moral responsibility do not require AP (Kane, 2005).

Frankfurt-type examples

Stronger examples have been presented by Harry Frankfurt, which stimulated discussions and many responses that are alive to this day. He tried to refute PAP and show that someone can be responsible for his action even though *he could not in fact have done otherwise*, through a following example. Imagine that,

“Someone—Black let us say—wants Jones to perform a certain action.

Black is prepared to go to considerable lengths to get his way, but he prefers to avoid showing his hand unnecessarily. So he waits until Jones is about to make up his mind... and he does nothing unless it is clear to him... that Jones is going to do

something other than what he [Black] wants him to do. If it does become clear that Jones is going to decide to do something else, Black takes effective steps to ensure that Jones... does what he wants” (Frankfurt, 2003, p. 169).

Black’s control over Jones can be imagined as a brain-controller that is activated only when Jones, does something other than Black wants him to do. However, if Jones acts freely in accordance with Black’s intentions, the brain-controller remains inactive. Frankfurt’s point is that if Jones did act on his own, *not* activating Black’s brain-controller, Jones would have acted on his own reasons and motives making him thus act responsibly despite the fact he could not do otherwise. Hence, PAP would be false. It is possible, at least in theory, that Jones could go through his entire life never activating the controller, which would make him responsible for all his choices even though, due to the controller, he could *never* have done otherwise (Kane, 2005). Several responses were offered, which attempted to show that Jones in fact could do otherwise either *voluntarily* or *involuntarily*, but on determinism, these objections all seem to have failed completely or are largely disputable (Timple, n.d.). If PAP is false, then the Consequence Argument against Compatibilism is disarmed. McKenna and Coates conclude, “*If determinism threatens free will and moral responsibility, it is not because it is incompatible with the ability to do otherwise. Even if determinism is incompatible with a sort of freedom involving the ability to do otherwise, it is not the kind of freedom required for moral responsibility*” (2015, section 4.2).

Widerker, Ginet, Wyma, and Kane point out that Frankfurt’s apply only to determinism. In their Indeterministic World Objection against Frankfurt controller, they emphasize that given Indeterministic World, Black’s brain-controller would be dysfunctional, as it would have no basis on which to reliably predict Jones’ future choices. If Black wanted to control Jones, he would

always have to activate the controller in advance, which would in turn divest Jones of his moral responsibility. Kane therefore believes that Frankfurt-type examples may be convincing for advocates of determinism, but do not represent a serious challenge for those who think that in order for a person to be morally responsible for his act, at least some of his acts must be undetermined (Kane, 2005).

The Consequence Argument does not offer a conclusive case against compatibilism; however, it helps to illuminate some of the other problems and challenges of a compatibilist-type freedom. These will be discussed in the following sections.

Frankfurt's Hierarchical Motivation Theory

Since Compatibilists do not have the ability to do otherwise as a foundation of free will and moral responsibility, they offer a different basis for their positive account of free will. Frankfurt, representing the New Compatibilists, does not limit free will to the absence of *external* constraints as Classical Compatibilists do, but recognizes also the existence of *internal* constraints (see "Freedom of Self-Control" in table 3). These internal constraints represent a constraint to our will in form of e.g. addictions, phobias, obsessions or neuroses. Frankfurt makes a distinction between first-order desires and second-order desires. First order-desire may be to use a drug, while a second-order desire may be to keep a job and improve one's marriage. The particular characteristic of second-order desires is that they are about other desires. Though an addict experiences an inextinguishable desire for drugs (first-order), he also has a concern for his job and marriage thus having a desire *not* to have a desire for drugs (second-order) (Kane, 2005). In their paper, "Pressure and coercion in the care for the addicted" Janssens et al. use Frankfurt's Hierarchy

as a foundation for discussing the relation between autonomy and paternalism when treating drug patients. In their words,

“...addicts are not fully autonomous. At a second order level, they may hate their addiction and want to overcome their first order desires but in the vast majority of cases, the first order desires are decisive. [...] Increasing the patient’s autonomy can then be regarded as a goal of care in the sense that the caregiver, in dialogue with the patient, tries to help the patient reflect on his addiction and articulate his second order desires. Fostering the patient’s autonomy can imply a persuasive or even manipulative approach, trying to bring the longer term values and goals of a patient to the surface.” (2004, p. 454-455)

They conclude that

“If autonomy is a positive capacity of oneself with the situation, pressure and even coercive measures are not necessarily antithetical to respect for autonomy. [...] ...coercion can be beneficial on the long term” (Ibid. p. 457).

An addict lacks the freedom of will because he cannot make his will (first-order desires) conform to his volition (second-order desires). In order for a person to have free will, his first- and second-order desires must be in harmony. Frankfurt uses a technical term of *wholeheartedness* to represent this characteristic (Kane, 1998). The Hierarchical Theory is a compatibilist theory as it is “conceivable that it should be causally determined that a person is free to want what he wants” and if that is so, “then it may be causally determined that a person enjoys a free will” (Frankfurt, 2001, p. 336). This is a positive account of a

free will that is compatible with determinism and does not require the power to do otherwise (Kane, 2005).

Control and Determinism

Dennett observes that an important question of control and its relation to causation and determinism “has scarcely been addressed by philosophers” (1984, p. 51). The question of *origins* or *sources* of higher order volition appears particularly relevant to hierarchical theories. “For all [Frankfurt’s] account tells us,” says Watson, “the person’s higher-order preference may be the result of brainwashing, or severe conditioning of the kind which is plainly incompatible with autonomy” (1987, p. 148). Referring to an abovementioned example, when the dictator’s son JoJo becomes *wholehearted* (like his father) in torturing people on the basis of a whim, it is natural to ask how did he become like this? To what extent was he responsible for his wholeheartedness? It is important to identify whether he was responsible for the forming of his higher-order volitions, or whether another *origin* or *source* of volition, like his father, could have completely conditioned, manipulated or overruled his volition forming process. Kane believes that this question is not only relevant for the criticism of hierarchical theories, but of the Compatibilist account generally, as it points to a pivotal disagreement between compatibilists and incompatibilists in relation to UR (1998).

Kane differentiates between two kinds of control:

- a. Constraining Control (CC) – Agents are controlled by being knowingly forced to act against their wills or are prevented from doing what they want to do.

- b. Nonconstraining Control (NC) – Agents are controlled by manipulating their will so that their want, desire, or intent is as controllers have planned for them.

Thus, controlled agents do not become frustrated.

If agents are unaware that they are being controlled or that their controllers even exist it is a case of a *covert* nonconstraining control (CNC). Such is the control in the aforementioned utopian novels *Brave New World* and *Walden Two*. Here, people experience, what may be called the pinnacle of the hierarchical type freedom, as citizens of these worlds can have and do whatever they want or choose; and they can *will* whatever they *want*. One of the characters described *Walden Two* to be the “freest place on earth”. Yet what these people want or choose has been conditioned from childhood. Kane makes the distinction between the kinds of freedom obvious. “...in *Walden two*, free will in the hierarchical sense is maximized, while free will in the deeper sense of ultimate control of ends or purposes is minimized. Indeed, *compatibilist free will is maximized in Walden Two by minimizing incompatibilist free will*: the citizens have the wills they want *because* they have been conditioned to want and choose only what they can have and do” (1998, p. 66).

What is particularly problematic about CNC for compatibilists is that it fits into their definition of freedom as the “absence of constraints”. CNC is by definition nonconstraining, however also compatibilists, like Dennett, find CNC objectionable because it seems to undermine the kind of freedom usually associated with autonomy (control of one’s own life). This tension implicitly calls for a “deeper” freedom.

To resolve this tension Compatibilists have usually taken one of two paths. They either fully embrace the *Walden Two* CNC scenario, asserting that a freedom from

coercion and compulsion is in fact all freedom worth wanting, and any “deeper” freedom is incoherent and illusory (hard compatibilism). For many, this line is hard to accept. Alternatively, they attempt to demonstrate a meaningful difference between CNC and a mere determination by natural causes (soft compatibilism). By identifying a relevant distinction between the two, they need to show that CNC is objectionable for taking away a significant freedom, while mere determinism is not.

When assessing the *ability* or *power* of compatibilist freedom to do or will something, it seems that CNC and mere determinism can produce the exact same results. Kane suggests that we may imagine two worlds that are similar in every detail except that one of them is governed by CNC and the other by mere determinism (1998). What happens in one world can also happen in the other. If I want to take a bus in one, I will want it also in the other. If I am unable to touch a blond Lab in one, neither will this be an option in the other; and if I become a sadistic dictator in one, so will I in the other. The exact same powers and ability can be lost or gained by CNC and mere determinism alike.

Waller (1988) has suggested that what is problematic is that CNC controllers may be using people as means in achieving their goals that may be potentially dangerous to human interests. This concern was cinematized in the 1999 movie “The Matrix” where AI robots used people as a kind of batteries while capturing their minds in an initially perfect virtual world. Though this objection is potent and requires attention, it addresses only half of the matter. Many find CNC problematic even when it is *benign* and genuinely intended for human wellbeing and flourishing. Such was the example of Walden Two.

Kane concludes that the essence of the problem does not lie in the potential deviousness of CNC, nor is it that some specific powers (e.g. taking buses or touching Labs) could be lost. The loss of a more essential power is unsettling when (benign) CNC is involved. It is “to be the ultimate *source* or *origin* of one's own ends or purposes rather than have that source be in something *other than you*.” If a person is to be ultimately responsible (UR) for what he believes, does, or chooses; who he is and who he becomes, such power seems properly essential. However, this power is removed by both CNC *and* mere determination, for “whether the sources of your ends or purposes lie in nature or in other agents, *they do not lie in you*” (Kane, 1998, p. 70-71).

Due to the extent of the topic, other counter-arguments and alternative lines of reasoning may follow these arguments. These, however, are beyond the scope of this work. What Kane's analysis have shown, if correct, is that the compatibilist hierarchical type of freedom introduced by Frankfurt offers a foothold to CNC type of influence in PD. Such a narrow understanding of freedom appears to have questionable implications to the basic definition of PD as autonomous (in a social sense) and voluntary attitude and behavior change. It was said that ethical persuasion must safeguard subjects' freedom to choose whether to be subjected to persuasion or not, and if persuasion takes place, subjects should have the freedom to choose the outcome of their efforts to change their belief or a behavior. The issue becomes clear when these definitions are seen in the light of Walden Two scenario of CNC. CNC allows subjects to make all kinds of responsible voluntary decisions about the types and outcomes of persuasion they want to be exposed to and achieve. It may even create enough “wobble room” for reasoning and reflection as Bang demands. Yet their voluntary decisions, reasoning and reflection may be fully an outcome of CNC type of PD.

Such a type of PD can be expected to be deemed as manipulative by the most. Therefore, something must be missing from a complete definition of persuasion in PD. Just as the hypothetical analysis (HA) was missing the qualifier “and she *could* also have *wanted* to do otherwise”, the full definition of persuasion in PD as a “voluntary attitude or behavior change” is missing the qualifier “and volition can be traced back to the subject who is its ultimate *source* and *origin*.” The problem is that such definition cannot be accommodated by determinism; therefore, compatibilist type of freedom is inadequate in truly capturing free, voluntary, reflective, rational persuasion in its traditional meaning. Harris offers a concluding remark, “Compatibilism amounts to nothing more than an assertion of the following creed: A puppet is free as long as he loves his strings” (2012, p. 20).

Libertarianism

In this discussion, Libertarianism is not a political term, it represents a position of those who deny determinism and affirm existence of a “deeper”, “true” freedom of will; the real thing. However, to show that free will is *incompatible* with *determinism*, is only half of the problem (The Ascent Problem, The Determinist Objection). Libertarians must also show how free will can be *compatible* with *indeterminism* (The Descent Problem, The Randomness objection). Kane calls it the Libertarian Dilemma while Clarke has referred to it as the twin objections. In order to attain the “deeper” free will, libertarians must answer both parts of the dilemma. The ascent problem was covered in previous sections; thus, space will be now given to the descent problem.

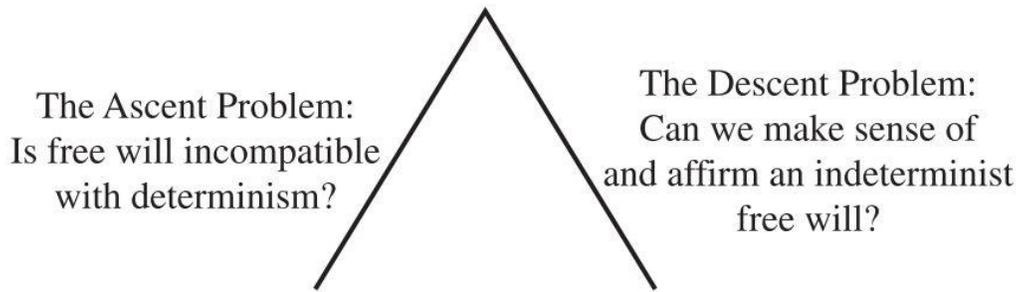


Figure 12 - Kane's Incompatibilist Mountain and the Libertarian Dilemma (2005, p. 34).

Kane has counted up to eight different objections to the compatibility of indeterminism and free will. First strategy of Libertarians was to argue from the indeterminacy of quantum theory that not all events in the universe are determined, which offers room for an indetermined freedom of will. But under such circumstances, critics say, what happens, happens merely by chance and is *not controlled* by anything, hence it is *not controlled* by the agent. Therefore, agent cannot be responsible for such an action. Moreover, if a choice is the result of a quantum jump, it would be similar to a spasmodic jerking or twitching of an arm. Such random events would thus *undermine* rather than enhance freedom because they would serve as a hindrance or impediment to, say, a surgeon making a fine incision during an operation. They would prevent us from doing what we want, being in fact a *constraint* on our freedom. Critics, like Schopenhauer picture a man who suddenly finds his legs starting to move *by chance*, carrying him across the room against his wishes. Indeed, the opposite of what is meant by a free and responsible action (Kane, 2005).

Second type of objection relate to the Libertarian claim that we must have a freedom to do otherwise, given *exactly the same past and laws of nature*. This is also the idea in the *garden of forking paths* of Kane or the *branching time* of Kripke, Prior, Øhrstrøm and Ploug that maintains

different possible futures, given the same past³³. Objectors say this requirement has troubling consequences that make free will unintelligible. They illustrate the problem:

“Suppose Mike, who is deliberating about whether to vacation in Hawaii or Colorado, gradually comes to favor and choose Hawaii. If Mike’s choice [...] was undetermined [...], then he might have chosen otherwise [Colorado], given exactly the same deliberation up to the moment of choice that in fact led him to favor and choose Hawaii (the same thoughts, reasoning, beliefs, desires, and so on). [...] It is difficult to make sense of this. Mike’s choosing Colorado in such circumstances [...] would seem irrational and inexplicable, capricious and arbitrary.” (Ibid., p. 36)

The Descent problem of libertarian free will shows that undetermined actions would be one or more of the following: arbitrary, capricious, random, uncontrolled, irrational, inexplicable or matters of luck or chance; not free and responsible actions.

To avoid this, libertarians have suggested that there is an additional factor involved in the decision making process, which is other than the past circumstances and the laws of nature. This may be called the extra-factor strategy. When something is said to be “undetermined” it does not automatically mean it is “uncaused”. There are nondeterministic, probabilistic types of causation where the outcome is not inevitable. The extra factors are immaterial minds or souls, Kantian noumenal selves, agent causations, unmoved movers and uncaused causes or other unusual forms of agency and causation (Ibid.). Much could be said for and against each of these alternatives, but

³³ Determinism can be understood through a simple linear time – same past entails same future (one line into the future).

for the purpose of this paper, it will suffice to say that most of the extra factors invoke some form of substance dualism that introduces an extra type of causality that intervenes in what was originally thought to be a causally closed system (PCC). Kane summarizes a response he believes most agent-causalists would agree with,

The agent-causal relation is unique and cannot be treated like any other event or occurrence. To ask the question ‘if the agent-causal relation is not caused, why doesn’t it occur merely randomly or by chance?’ is to show you do not really understand what the agent-causal relation is. Immanent agent-causation is not the sort of thing that can in principle occur randomly or by chance, any more than it *can* in principle be caused. For the agent-causal relation just is the agent’s exercising conscious control over an event; and an agent’s exercising conscious control over an event is not the sort of thing that happens out of the blue, by chance or accident. For by its nature it is up to the agent. We do not need a further agent-causing to explain it. (Ibid., p. 50-51).

Gibbons above have said this type of causal relation would violate the laws of nature or imply a non-existent law. Understandably, other critics of Libertarianism are also dissatisfied with such explanation, and find this strategy as an appeal to mystery, a response by a mere stipulation, or defining of the problem out of existence. Gary Watson says, “‘Agent-causation’ simply labels, not illuminates, what the libertarian needs.” (1982, p. 10). He demands they show *how* such a phenomenon is empirically possible. Schrödinger well pointed out, “At the price of mystery, you can have anything” (Kane, 2005, p. 42) but adding words of Bertrand Russel, in such a case you

get it too easily, acquiring it by theft rather than honest toil (Kane, 2005). Many critics find this response inadequate; therefore, following Reppert (2003) this is called the Inadequacy Objection.

Is this objection well grounded? Reppert thinks not. Explanations can be either basic or nonbasic. If the explanation is nonbasic it can be reduced or explained further in terms of its constituent parts. Reppert gives an example; if someone asks, “What is it about this sleeping pill that puts me to sleep?” It would not be appropriate to say, “It’s just the nature of that pill to put you to sleep. That’s what it does. It has a ‘dormative virtue.’ Surely, a more basic chemical explanation can be given, and even that would not be a truly basic explanation. However, at some point, no further explanation can be given, which is when the explanatory bedrock is hit. The only thing that is left to say is, “that is the nature of so-and-so to have a such-and-such characteristic” – a basic explanation. Materialism is saturated with basic explanations of fundamental physical constants and elements. For example, it makes no sense to ask *how* it is empirically possible that the gravitational constant is $6.673889 \times 10^{-11} \text{ N} \cdot (\text{m}/\text{kg})^2$ or that the speed of light is $2.99792458 \times 10^8 \text{ m/s}$ (Physical constant, n.d.). It is the nature of gravity and light to have these values and “act” according to them. Therefore, if a substance dualist is guilty of offering a basic explanation, a substance monist is equally so. Reppert says, “A physicalist is not immune from giving basic explanations that appeal to the nature of things” (2003, Chap. 6, Sec. 1, Par. 7). At this point a physicalist may ask, is it necessary to introduce another substance to allow for rational inference and libertarian free will?

The cumulative force of the Argument from Reason, The Argument from the reliability of our rational faculties, The Consequence Argument, the CNC problem and other considerations set forth in this work aimed to show that materialism and physicalism does not have the necessary

resources to accommodate rational inference, and the free will in its traditional sense. If correct, this shows the irrationality or at least, improbability of finding an explanation for rational inference and free will in these monist philosophies. Thus, we might concur with Menuge, “Some form of substance dualism seems unavoidable to account for reasoning” (2013, p. 16).

Our strong *a priori* experience of ourselves as rational, free agents lends itself to be explained by a teleological explanation. The properly basic belief of being a responsible source and origin of our actions is therefore by no means question-begging. It seems correctly innate to us that we exercise active power, initiate and redirect causal chains of our surroundings (Menuge, 2013). From our first-person perspective, this is not mysterious at all. Our introspection, in most cases, serves as an evidence for the reality of our rational reflection and freedom of will. Since monism failed to provide tools that would adequately explain these phenomenon, some form of dualism should be leastwise considered tentatively. It would be neat to explain everything mechanistically, but Reppert says, “not at the price of reducing to nonsense the very activity of rational inference on which science is based. One cannot perform scientific research with a fixed, preconceived idea of what the explanations will be. If we did, we would not have accepted quantum indeterminism or a beginning of the universe at the big bang. The kind of explanation that worked for falling apples [...] may not work for consciousness and reasoning” (2003, Chap. 6 Sec. 1 Par. 14).

The Modest Objective

Implicit Beliefs

What should be specified is that in order to establish rationality and free will of human beings, it is *not* necessary (arguably also not plausible) to show that all our thoughts, experience

and decisions are free, rational or undetermined. Such an unlimited capacity may perhaps be attributed only to an omniscient and omnipotent God, who enjoys the property of aseity (self-existence). For limited creatures, like ourselves, it is enough to show that at least some meaningful mental states are free, rational and undetermined. It should be of little dispute that many, if not most, of our thoughts and experiences are imposed on us by our environment. Something about the external world, given properly functioning cognitive faculties, imprints itself on our mind and compels the mind to adopt a specific mental image of reality. When I open my eyes in the morning, I am not free to think rationally that I see a pink elephant, when in fact I see the white ceiling of my bedroom. If, despite my visual input, I freely choose to believe that I see a pink elephant, it would be more appropriate to seek out professional help than to celebrate liberation of my will from my sensory organs. Staying with the pink elephant for a little longer, anyone reading these lines and properly processing their English semantics cannot avoid producing some, at least vague, mental image of a pink elephant. When I go on and write, that the elephant stands on a ball and has a blue hat between its ears, a reader is determined to, and cannot help but adjust his mental image accordingly. Equally, our perception of pain has little to do with a rationally inferred, free belief; instead, pain can be said to have an inescapably constraining, even coercive, effect on its recipient. Neither am I free to think rationally that I am some other mind, e.g. the Queen of Denmark, but I am determined to remain in my own character. Nor am I free to believe rationally that I was born in Philippines, as my memory demands me to believe that I was born in Slovakia. Certainly, not a hallmark of an unconditionally free mind. Thus, even if substance dualism is accepted, many beliefs remain to be bound to the reality of external world; these beliefs may be called *implicit* beliefs. At this point, some may say that neither God could logically be completely free of *implicit* beliefs, since also He cannot rationally think He is *merely* the Queen of Denmark,

or that He was born in Philippines. Nevertheless, this finding is hardly troubling. Having *implicit* beliefs is not commonly understood as a constraint to our free will; on the contrary, most people view correct *implicit* beliefs as a sign of properly functioning cognitive faculties that ultimately enhance our freedom. Pain seems constraining in the short-term, but a failure to recognize in time that our body was damaged may cause permanent health issues resulting in long-term constraints, even death. Failure to see an obstacle on a road while driving may result in similar consequences.

Explicit Beliefs

On the other hand, an *explicit* belief can be understood as an outcome of rational processes and free intentional deliberation of an agent (Schwitzgebel, 2015)³⁴. While a person may instinctively obtain an implicit belief about an existence of some higher power governing the universe, only upon being confronted with and reflecting over specific religious doctrines and philosophies can he arrive at an *explicit* belief in Islam, Christianity, Atheism or other. Such an *explicit* belief would therefore be preceded by a conscious, active decision, or choice to hold this belief; similarly, any explicit action would be preceded by an explicit (perhaps unformulated) belief e.g., “It is good to eat vegetables”. Explicit beliefs are not required for implicit actions such as heartbeat or twitching of an arm.

Is it necessary that at least all our *explicit* decisions and beliefs are free and undetermined? No. As was mentioned above, Luther once said, “Here I stand, I can do no other”. If Luther made *some* previous free decisions that formed him (SFAs), and lead him to this situation, then it is

³⁴ This is a simplification of the different categories of beliefs, which suffices the scope of this paper. Beliefs may be considered at a greater depth in future inquiries.

possible to affirm his ultimate responsibility as the source and origin of his current state even though at this point, he can do no other and his choice is determined. In order to make a case for libertarian free will, it is required only that *some* previous “will-setting” and “self-forming actions” (SFAs) made Luther the person he was at the Diet of Worms on 18th of April 1521. It has been said that given libertarian freedom that entails the possibility of *some past - different futures*, Mike’s choosing Colorado over Hawaii given exactly the same deliberation up to the moment of his choice seems irrational and inexplicable, capricious and arbitrary. Libertarians may choose to bite the bullet, as they did with Luther, and readily accept this conclusion. Even if this particular decision was determined, and Mike at this point could not have done otherwise (*no different futures*), his decision was made by deliberation (not by mechanistic PCC), so Mike is the *rational* agent-cause of his vacation in Hawaii instead of Colorado. However, this decision was also *free* by the virtue of his earlier undetermined SFAs. If in his earlier SFA Mike had chosen differently, he would have been on a different path and may not choose Hawaii over Colorado. While libertarian freedom suggests the possibility of the *some past - different futures*, it is *not committed* to defend the position that different futures are available at every point in time. However, an objector may ask, is not this merely moving the problem a step further? Is not Mike’s earlier deliberation concerning SFAs subject to the same principles as was his later deliberation regarding his vacation? What is it about SFAs that provides the possibility of different futures that Mike’s vacation decision does not?

Kane explains that SFAs occur in those times when “we are torn between competing visions of what we should do or become” (2005, p. 135). These are the situations that present us

with difficult decisions with competing motivations introducing uncertainty into our minds.³⁵ To better imagine such a situation, Kane offers a description of a businesswoman facing a conflict. It will be quoted at length to preserve its totality and explanation power.

“She is on her way to an important meeting when she observes an assault taking place in an alley. An inner struggle arises between her conscience on the one hand (to stop and call for help for the assault victim) and her career ambitions, on the other hand, which tell her she cannot miss this important business meeting. She has to make an effort of will to overcome the temptation to do the selfish thing and go on to the meeting. If she overcomes this temptation, it will be the result of her effort to do the moral thing; but if she fails, it will be because she did not *allow* her effort to succeed. For while she willed to overcome temptation, she also willed to fail. That is to say, she had strong reasons to will the moral thing, but she also had strong reasons, ambitious reasons, to make the selfish choice that were different from, and incommensurable with, her moral reasons. When we, like the woman, decide in such circumstances, and the indeterminate efforts we are making become determinate choices, we *make* one set of competing reasons or motives prevail over the others then and there *by deciding*. Thus the choice we eventually make, though undetermined, can still be rational (made for reasons) and voluntary (made in accordance with our wills), whichever way we choose” (2005, p. 136).

This is when the old objection against indeterminism floods in, as it may be asked, “Is not choosing either of these options accidental, capricious or random?” Certainly not. In the clash of the two conflicting motivations (complex neural networks), indeterministic noise is created. This noise is *not from an external source* but from her own will (Ibid.). In the moment of her choice, one of these motivations “wins” by reaching an activation threshold and overcoming the indeterministic noise. Her choice, either way, is *willed* by the agent who acted “on purpose” rather

³⁵ Kane gives an account of how such a situation may be understood in neuroscience. “...that is reflected in appropriate regions of our brains by movement away from thermodynamic equilibrium—in short, a kind of “stirring up of chaos” in the brain that makes it sensitive to micro-indeterminacies at the neuronal level. The uncertainty and inner tension we feel at such soul-searching moments of self-formation would thus be reflected in the indeterminacy of our neural processes themselves” (2005, p. 135).

than just accidentally or by chance. To avoid determinism of Mike's vacation choice, SFAs need to be undetermined, i.e. choice-outcome cannot be determined by an agent *before* it occurs. But an agent still can be in control and actively determine which of them occurs, *when* it occurs. Whatever the businesswoman succeeds in doing in that moment will be the undetermined outcome of *her* voluntary, intentional resolution of the conflict in her will (Kane, 2005).

In the spirit of Gary Watson's skepticism, someone may keep on asking, but "*how* is such a phenomenon empirically possible?" The libertarian can now reiterate his earlier point, invoking dualism; this is when the explanatory bedrock is hit and *basic* explanation needs to be given; "The agent-causal relation just is the agent's exercising conscious control over an event; and an agent's exercising conscious control over an event is not the sort of thing that happens out of the blue, by chance or accident." (Ibid., p. 51). Libertarians have provided a clear distinction between Mike's vacation type choice and SFAs. If Mike chose Colorado over Hawaii, given the exact same deliberation, such agent-causality may seem irrational, capricious and arbitrary, but this does not apply to above described SFA circumstances. Of course, if Mike experienced similar conflict of will over his vacation decision as the businesswoman experienced between her moral and ambitious reasons, it would be a different matter altogether, for in that case Mike would be facing a SFA.

Before finishing this section, a final objection will be considered revealing an important characteristic about the nature of free will. It says, even if granted that an agent causes the outcome of his SFA, there is still a sense of arbitrariness around his decision, as he cannot possibly have *all conclusive* reasons for his decision. Can he then be rationally responsible and in control of his decision even when his decision is partially arbitrary? Yes. It is enough for him to have *good*

reasons for his decision. In that sense, every SFA is a *value experiment* “whose justification lies in the future and is not fully explained by past reasons” (Ibid., p. 144). Therefore, every SFA, Kane says is an expression of the following statement: “Let’s try this. It is not required by my past, but it is consistent with my past and is one branching pathway in the garden of forking paths my life can now meaningfully take. Whether it is the right choice, only time will tell. Meanwhile, I am willing to take responsibility for it one way or the other” (Ibid., p. 145). Such is the nature of our free will.

To sum up, the section on determinism, compatibilism and libertarianism aimed to show a relationship between human reasoning and free will. Specifically, it has shown that human reasoning requires libertarian free will and ontological resources such as substantial selves with active power, all of which cannot be found in a naturalistic world. In the light of current scientific knowledge, SFA is a possible example of how an agent can make a (libertarian) free, rational, undetermined, intentional, controlled choice for which he is ultimately responsible. However, this is not a conclusive case as there are many objections that were made against it of which only some were addressed in this work. Therefore a final conclusion whether libertarian free will is possible or not remains to be a work in progress.

It should also be remarked that this work purposely omitted theistic versions of determinism and compatibilism. Calvinism or Reformed tradition understands God to be the sovereign creator who decreed all things and preordained the final outcome of human endeavour according to His divine plan. The doctrine of predestination is sometimes viewed as compatible with the freedom of will.

Persuasive Artificial Intelligence

In 2006 at the International Computers and Philosophy Conference in France, Dennett have stated that “AI makes Philosophy honest“ (Anderson, 2009). Certain aspects of artificial intelligence development (AI) lie at the heart of the discussion about human free, rational agency. Looking at AI will offer an insight into our own perception of rationality. It is customary to distinguish between “weak” and “strong” AI. The former stands for a mere simulation or modeling of certain aspects of intelligent human behavior by suitable programming, whereas the latter represents a vision that a suitably written program could create a machine that can literally think and reason (Lowe, 2004). For obvious reasons, it is therefore the “strong” AI that is of interest to this discussion. It has been repeatedly assumed that machines will never be able to do x , only to discover that computers have not only successfully mastered x but also outperformed humans in x . Playing chess was in late fifties considered to epitomize the human intellection. Several experts thought that, “if one could devise a successful chess machine, one would seem to have penetrated to the core of human intellectual endeavor.” (Bostrom, 2014, Chap. 1, Sec. 5, Par. 2) A few decades later in 1997, a chess program called Deep Blue beats the world chess champion Garry Kasparov who claimed to have a glimpse of true intelligence and creativity in some of the computer’s moves. Since then, it is no surprise that computers have mastered a vast number of other “intelligent” activities. One of these products – Siri – was directly defined above as an *intelligent personal assistant*. While these programs would hardly by modern standards be accepted as a “strong” AI, these innovations set a clear course for the future breakthroughs. Based on current development it may be assumed that computers will increasingly be able to perform more intelligent actions, perhaps to a point when they may be indistinguishable from human actors. If such a situation arises a well-known test, proposed already in 1950 by the computer pioneer Alan Turing would be

passed. The Turing test suggests that if an observer is not statistically able to reliably tell a difference between interacting with a machine and a human being then it may be considered a “strong” AI (Lowe, 2004). Lowe writes, “According to Turing and his followers, we should equate the intelligence of a computer which passes the Turing test (TT) with that of an ordinary human being” (2004, p. 212). It may be said that an upgraded version of the Turing test (TT) was cinematized in the recent movie *Ex Machina* (2015), where the main character Caleb participates in a fascinating experiment in which he interacts with a robot girl. In this version of TT, Caleb is told and shown that the robot is man-made and he is asked to make his mind whether “she” is conscious or self-aware. After few meetings, Caleb is positively convinced and a thrilling plot of the movie moves on. Another movie *Her* (2013), presents a lonely man called Theodore who installs the world’s first artificially intelligent operating system, which speaks to Theodore under the name Samantha. Theodore spends much time with Samantha and eventually falls in love with his operating system; Samantha appears to reciprocate his affection.

An entrepreneur and visionary Ray Kurzweil have made a wager with Mitchel Kapur and challenged Kapur’s statement that “By 2029 no computer – or “machine intelligence” – will have passed the Turing Test.”³⁶ Based on his analysis of historical increase of computing power, Kurzweil affirms Moore’s law about the doubling of integrated circuitry, and sees an exponential growth that will soon reach, even surpass, the brainpower of a single human by 2023 and in 2045 the equivalent of brainpower of all humans combined, leading to singularity (see Figure 13).

³⁶ Retrieved from <http://longbets.org/1/>

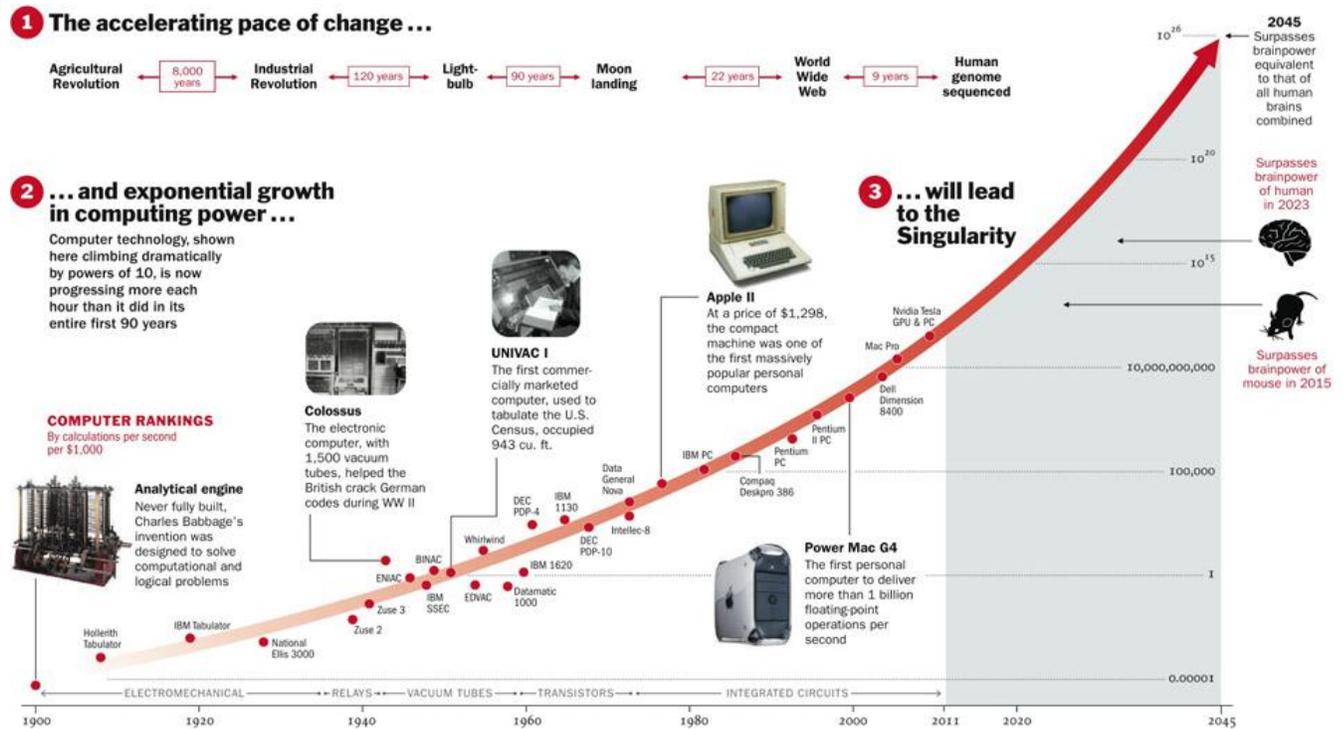


Figure 13 - Kurzweil's prediction in TIME magazine. (In TIME, n.d.)

Since computational power alone is not sufficient to pass TT, Kurzweil offers a scenario on how this power can be given organization and content – the software of intelligence – by reverse engineering the human brain. This may be done by some future high-resolution MRI scan that will be able to see individual neuron cell bodies, their connection with other neurons, synapses and neurotransmitter strengths. Copying this design to a computer would essentially recreate the structure of the brain in a digital form (1999). What happens when the hardware and software is in place? Kurzweil envisions,

“We have to consider this question on both the objective and subjective levels. Objectively, when we scan someone's brain and reinstantiate their personal mind file (meaning, copy all of the processes relevant to human thinking) into a suitable computing medium, the newly emergent "person" will appear to other observers to have very much the same personality, history and memory as the person originally scanned. [...] Subjectively, the question is more subtle and profound. Is this the same consciousness as the person we just scanned? [...] We don't worry, at least not yet, about causing pain and suffering to our computer

programs. But at what point do we consider an entity, a process, to be conscious, to feel pain and discomfort, to have its own intentionality, its own free will? How do we determine if an entity is conscious, if it has subjective experience? How do we distinguish a process that is conscious from one that just acts as if it is conscious? [...] What if the entity is very convincing and compelling when it says, "I'm lonely, please keep me company" -- does that settle the issue? If we look inside its circuits, and see similar kinds of feedback loops in its brain that we see in a human brain, does that settle the issue? [...] For example, if we scan, let's say, myself [...] the person that then emerges in the machine will think that he is (and had been) me. He will say, "I grew up in Queens, N.Y., went to college at MIT, stayed in the Boston area, walked into a scanner there, and woke up in the machine here. Hey, this technology really works." But wait. Is this really me? For one thing, old Ray (that's me) still exists. I'll still be here in my carbon-cell-based brain. (Kurzweil, 1999, p. 54)

This is essentially a question of consciousness and identity. Kurzweil predicts that “during the course of the twenty-first century these will not remain topics for polite philosophical debates but will have to be confronted as vital, practical, political and legal issues” (Kurzweil, 2005, p. 310). Now, it should be clear into what kinds of situations we may get if a TT is successfully passed.

Turing Test and Other Minds

TT is essentially an operationalist (technology) or behavioristic (psychology) trial that rests its conclusion on an external empirical observation. Turing even dubbed his test *The Imitation Game*. Suppose an AI could imitate a human being so persuasively as in the case of Caleb, Theodore or Kurzweil that it passes the TT and would be *operationally* or *behaviorally* indistinguishable from a genuine human being. Does it mean that it is subjectively, internally a conscious, free, rational being? “After all”, Lowe asks, “would it be rational to harbor doubts about the intelligence of one of your friends if you were suddenly to discover that, instead of being made of flesh and bones, he or she was composed internally of metal rods, wires and silicon chips”

(2004, p. 213)? Lowe brings our attention to another philosophical problem closely related to our evaluation of the existence of a “strong” AI, namely our *belief in existence of other minds* (BOM).

In first person perspective, all subjective experiences like e.g. pain are undeniable, but how does one know that also other people experience pain? When other people display pain through hurtful facial expressions or words like “Ouch, it really hurts!” does it settle the issue? When a friend says “I’m lonely, please keep me company”, can we be sure that there really is someone inside experiencing loneliness? It seems that our BOM is as much based on the behavioristic TT as any of our judgments about AI. Jakobsen states, “The TT should therefore be understood as setting up an epistemic situation; wherein one has the same grounds for believing in the existence of other minds as one has for the existence of a mechanized mind; such that if we refrain from granting the machine thinking, if it passes the test, then we do so at the expense of BOM” (2007, p. 50). In the absence of some other reason, the denial of AI in this situation, if it passed TT, must entail the denial of BOM. In such a case, the subjective first person experience of the existence of an observer would be all that is left for him to know and trust, amounting to Cartesian *cogito ergo sum* (I think, therefore I am). Such conclusion results in a lonely prospect of Solipsism.³⁷

Searle’s Chinese Room Argument

If we want to go the other way and keep BOM, are we then committed, given the fulfillment of TT, to the existence of a free, rational, conscious AI? John Searle offers a reason why this is not necessary. Suppose a monoglot English-speaking person is inside of a room with a typewriter

³⁷ According to the Merriam Webster Dictionary Solipsism is a theory in philosophy that your own existence is the only thing that is real or that can be known.

producing Chinese letters, a printer, and an English operation manual. Outside of the room is a monoglot Chinese-speaking person who also has a typewriter and a printer, allowing him to send messages written in Chinese into the room. He may choose to write anything. When the English-speaking person receives a Chinese message from the outside, he must look in the manual to find out what string of Chinese characters to type out in response. Imagine that the manual is so well written that the Chinese speaker outside is unable to distinguish these responses from responses of a native Chinese speaker. He believes he is speaking with someone inside the room who understands the content of the conversation; however, the English-speaker is merely slavishly following the manual without an understanding of Chinese. At this point, the requirements of TT have been fulfilled; the Chinese speaker is positively convinced, and thus TT was passed. This implies that TT in fact requires no understanding of the content of interaction; therefore, TT cannot be a test of genuine intelligence since genuine intelligence does demand understanding (Lowe, 2004), (Jakobsen, 2007).

Naturally, the point Searle makes is that a computer operates like the Chinese room (CR). While the English-speaking person understands certain things and acts according to the manual only with respect to the Chinese conversation, the computer should apparently use a manual for everything, thus understanding *nothing whatever*. Lowe asks, “But how can something which understands nothing whatever justifiably be deemed intelligent” (2004, p. 216)?

False Analogy Objection

Advocates of AI say that it may be true that the English-speaking person alone does not understand Chinese, but this is not equivalent to a computer executing a program. The proper analogy to the computer program is the English-speaking person *and the operation manual*

combined. The English-speaking person amounts only to the processor of the computer executing the program. The processor indeed does not understand Chinese, but the whole *system* does (Lowe, 2004).

Response to False Analogy Objection

Searle refines his thought-experiment and suggests that even if, hypothetically, the English-speaker memorizes the entire operational manual he would still not understand Chinese, since he would only know how to match certain strings of Chinese letters with other strings of Chinese letters (*syntax*) without any understanding of their *semantics*. This satisfies the objection since now the English-speaker constitutes the whole *system* inside the room (Lowe, 2004).

Many other arguments and counter-arguments address the validity of the CR argument (see Jakobsen, 2007). Yet if CR holds, as many believe it does, then it presents a defeater for the TT as a test of genuine, rational intelligence.

Chinese Room and Other Minds

CR is a potential defeater for the rationality of AI, but then is not CR also a defeater for the rationality of BOM? It appears to me that I have a *semantic* understanding of my environment, but how do I know that others do not merely match certain strings of symbols (*syntax*) to display an output for which they have no *semantic* understanding? Unless a distinction between biological machines and mechanical machines is given, they are on par. In such a situation, CR has not helped to identify a genuine intelligence and we are left with the same operationalist and behaviorist empiric resources as before. Fortunately, this is not the case.

The Predicament of a Naturalist

Here we must revert to the abovementioned argument from reason, the argument from the reliability of our rational faculties, the consequence argument, and the CNC problem. If correct, jointly they have shown that with materialist and physicalist resources only, a naturalist is actually in a much worse predicament. The dilemma of the naturalist is not between, on one hand, “accepting rationality of both AI and BOM” or on the other, “rationality of solipsism”; his options are worse, either “inherent nonrationality or at best unreliable rationality of AI and BOM” or “inherent nonrationality or at best unreliable rationality of solipsism”. If naturalism is incompatible with reason or makes our rational inference unreliable, then nobody can be rational; Not AI, not other minds, not a solipsistic self, or not even a hypothesized Boltzmann brain. The naturalist’s only option seems to be to accept the existence of humans (himself included) as conscious beings, which cannot rationally trust anything we believe to be true because our beliefs are products of deterministic processes, not rational inference, guided by cognitive faculties, which are not directed at truth, but at adaptive behavior (four Fs). With respect to AI, naturalist’s only guide is the TT. If TT is passed, naturalist is committed to ascribe such an AI consciousness and everything that it entails, if he consistently desires to also maintain his BOM. He must accept a conscious AI without any other way of telling whether there is actually someone inside. This is when Turing’s label of his test, *The Imitation Game* can be seen in its full light; for Turing’s best advice to modern AI pioneers may quite deservedly be “Fake it until you make it”. Bostrom is cautious about accepting such vision. He suggests that a complexity of a human mind and behavior, or even a much greater complexity, does not guarantee conscious experience. He writes,

“We could thus imagine, as an extreme case, a technologically highly advanced society, containing many complex structures, some of them far more intricate and intelligent than anything that exists on the planet today— a society which nevertheless lacks any type of being that is conscious or whose welfare has moral significance. In a sense, this would be an uninhabited society. It would be a society of economic miracles and technological awesomeness, with nobody there to benefit. A Disneyland without children” (2015, Chap. 11, Sec. 10, Par. 9).

This seems to be the predicament of a naturalist; yet how can a substance dualist escape a similar outcome?

Substance Dualism, AI and Other Minds

Given substance dualism as the foundation for rational inference of free agents (described above), it is properly basic to extrapolate that not only am I comprised of this other conscious *agent-substratum* but that all members of my species are the benefiter of the same substratum, which justifies my BOM. However, when it comes to assessing the strong AI that passed the TT, I have a defeater for the belief that AI is a rational conscious being in the form of the CR argument. CR argument does not apply to BOM since *agent-substratum* serves as a warrant of my BOM. Because I have no reason to assume that an AI possesses *agent-substratum*, AI is liable to all objections raised towards purely mechanical, materialist, physical, determinist processes that lack *agent-causation*. Therefore, AI is not, and cannot in principle, be like a human. If such machines ever comes about, Lowe writes, “[they] would surely deserve to be called ‘intelligent’, and they would certainly be ‘artificial’. To that extent, the dream of artificial intelligence would have been realized” (2004, p. 227).

Great Delusion or a Genuine Change of Mind

We can now proceed to conclude the case for rational persuasion. Is rational persuasion possible? Can a person freely change his attitude or a behavior on the basis of given reasons? It has been demonstrated that Hard Determinism and Compatibilism, both of which presuppose naturalistic, monist worldview, do not have the necessary resources to account for rationality, nor for freedom of such a change of mind.

First, on these views, the definition of persuasion as a *voluntary* change of attitude or behavior can be distorted to conform to Covert Non-Constraining (CNC) type of influence in PD. This position must ethicists find disputable at best. This reveals that if persuasion is to be a *genuinely voluntary* change of attitude or behavior, volition can be traced back to the subject who is its ultimate source and origin. However, this additional condition cannot be accommodated by materialism and physicalism that presuppose PCC and unbreakable causal chain of events that can be traced back to the beginning of the universe (perhaps allowing for certain random quantum indeterminacies). Thus, hard determinism and compatibilism, presupposing materialism and physicalism, are irreconcilable with the concept of persuasion, by virtue of their *lack of freedom*.

Second, even if the possibility of a genuine voluntary change was granted to compatibilism, persuasion could not be rational on naturalism, since non-spatio-temporal causal entities as rational inference and logical connections are not available in the inventory of possible causal explanations of materialism (*the argument from reason*). However, even if rational inference was somehow accessible to naturalism, given evolution, any belief we hold would be produced by cognitive faculties that are aimed to produce adaptive behavior and deem truth and falsehood irrelevant (*the argument from reliability of our cognitive faculties*). Thus, hard determinism and compatibilism,

presupposing naturalism, physicalism, naturalism and evolution, are irreconcilable with rational persuasion, by virtue of their *lack of rationality*.

On determinism, rational persuasion is a great delusion. Both persuader and persuadee are mere marionettes acting out their part to adapt for survival, being governed by fixed motions of atoms and forces in the universe, while experiencing an illusory grandeur of their freedom and rationality. Fortunately, determinism is not the only position available. This work has attempted to establish that the best possible explanation for rational persuasion is some form of substance dualism. Agent-causes can act indeterministically, without being random, uncontrolled, irrational, inexplicable, irresponsible, capricious or arbitrary. Substance dualism can accommodate both the causal power of reasons in an agent's decisions, making these decisions *rational*; and the genuine voluntary change of attitude or behavior (where volition can be traced back to the subject who is its ultimate source and origin) through SFAs, making these decisions *free*. Therefore, it appears that on substance dualism, rational persuasion is possible and we may genuinely choose to change our mind about an attitude or a behavior.

The AI Delusion

Let us come back to Atkinson's profound question mentioned in the beginning. "Do we interact with computers, do we interact through them or do we simply use them" (Ijsselsteijn et al., 2006, p. 176)? With respect to AI, answering this question influences how we understand the nature of HCI and Human-Robot-Interaction (HRI). Even though a genuine, conscious, free, rational, strong AI may not be an option, current rate of progress suggests that a machine that strongly resembles such an AI may emerge in the near future as foreshadowed by Bostrom, Kurzweil and others. When this happens, AI may be an epitome of PT. It will contain an

unprecedented amount of sensors (as described above), which will provide quanta of live data; moreover, it will simultaneously be able to receive data from other sources over the internet. All this data will be processed by sophisticated algorithms evaluating it against yottabytes of existing data, which will enable the AI to have an extremely accurate picture of its surroundings.

Suppose you interact with a program, which can evaluate your facial expressions, voice tonality and volume, employed language, heart rate and perspiration, against millions of other people in its database. Moreover, it is informed about the current atmospheric pressure, temperature, humidity, and air consistence but also about your medical record, financial history, previous employments, criminal incidents, all your virtual presence and practically all tracks you have made throughout your life. At the same time it can keep track of other regional, national and global information like current stock market development and latest news on all channels. It is knowledgable of every movie, every song and every book ever written. It has access to all Google search trends and all collected big data. In short, it “knows“ essentially everything that can feasibly be known about you and your environemnt. Given sufficient computational power, proper algorithms and learning patterns, it does not seem difficult to imagine that determining Kairos (right timing and measure) may present for such an AI a straightforward task that will enable it to produce a very persuasive interaction experience. If it is also given a physical body with dimensions, gestures and range of motion indistinguishable from that of a human, the masterpiece is perfected. As such, it can now easily pass the TT and a delusion of a sentient AI emerges.

If the reasoning behind substance dualism holds, and we can assume AI cannot in principle be like a human being, then the spread of an AI that can pass the TT should be highly regulated. Since persuasion is defined as an absence of coercion and *deception*, existence of human-like AI

seems problematic, since it may deceive people into believing they are interacting with a sentient being. Jakobsen writes, “We might come to find ourselves persuaded from time to time by “intelligent” artificial artifacts of man. We might form the false belief that they are thinking. We might even be unable to distinguish between it and a human” (2007, p. 79). Also Atkinson’s warning seems appropriate to restate, “Using this concept of social actor uncritically, if we are not careful, will perpetuate an illusion, compound Baudrillard’s Procession of the Simulacra and cause us to fall victim to Rebe Dubo’s warning that humans continue to adapt to maladaptive situations” (Ijsselsteijn et al., 2006, p. 176). To prevent this delusion in PD, in the light of this work it seems appropriate to concur with Atkinson that we do not interact with computers, and computing products are not participants in interaction.

References

- A. Spagnolli et al. (Eds.). (2014). *PERSUASIVE 2014*, LNCS 8462, pp. 302–322. Springer. Switzerland.
- Aagaard, M., Moltsen, L., Øhrstrøm, P. (n.d.). *It might be Kairos*. Department of Communication and Psychology, Aalborg University & Wirtek DK.
- Ananke. (2006) In Encyclopedia Britannica, Inc., *Encyclopedia of World Religions*. Chicago, US. Can be retrieved also at <http://www.britannica.com/topic/Ananke-Greek-mythology> (23/6/2016).
- Anderson, S. L., Anderson, M. (2009). How Machines Can Advance Ethics. *Philosophy Now: A Magazine of ideas*. Retrieved from https://philosophynow.org/issues/72/How_Machines_Can_Advance_Ethics (27/7/2016).
- Anscombe, G. E. M. (1981). *Metaphysics and the Philosophy of Mind*. Basil Blackwell Publisher. Oxford. Also in Reppert, 2003, chapter 3.
- Bang, M., Ragnemalm, E. L. (Eds.). (2012). *PERSUASIVE 2012. Persuasive Technology Considered Harmful? An Exploration of Design Concerns through the TV Companion*, LNCS 7284, pp. 239–250. Springer. Berlin.
- Basamh, S. S., Ali, F., Huq, M., Ibrahim, J. B. (2013). Exploring Persuasive Technology to Enhance Delivery of Professional Services. *International Journal of Humanities and Social Science*. Vol. 3 No. 12. USA.
- Beauregard, M., O'Leary, D. (2008). *The Spiritual Brain: A Neuroscientist's Case for the Existence of the Soul*. HarperCollins. Kindle Edition.
- Beilby, J. (2005). *Epistemology as Theology: An evaluation of Alvin Plantinga's religious epistemology*. Ashgate Publishing Limited, England.
- Ben-Ari, M., Pnueli, A., Manna, Z. (1983). The Temporal Logic of Branching Time. *Acta Informatica*, 20, 207-226.
- Berdichevsky, D., Neuenschwander, E. (1999). Toward an Ethics of Persuasive Technology. *Communications of the ACM* 42, 51–58.
- Berkovsky, S., Freyne, J. (Eds.). (2013). *PERSUASIVE 2013. Three Approaches to Ethical Considerations in the Design of Behavior Change Support Systems*. LNCS 7822, pp. 87–98, Springer. Berlin.
- Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, UK. Kindle Edition.

- Churchland, P. (1987). Epistemology in the Age of Neuroscience. *The Journal of Philosophy*, Vol. 84, No. 10, p. 544-553.
- Clarke, P. G. H. (2015). *All in the Mind?: Does Neuroscience Challenge Faith?* Lion Books, England.
- CNBC. (2016, Mar 16). Hot Robot At SXSW Says She Wants To Destroy Humans | The Pulse | CNBC [Video File]. Retrieved from https://www.youtube.com/watch?v=W0_DPiOPmF0 (11/6/2016).
- Córcoles, E. P., Boutelle, M. G. (2013). *Biosensors and Invasive Monitoring in Clinical Applications*. Springer. London.
- Craig, W. (2013, April 22). Libet's Experiments and Determinism. Reasonable faith with William Lane Craig. Q&A. Retrieved from <http://www.reasonablefaith.org/Libets-Experiments-and-Determinism> (25/6/2016).
- Davis, J. (2009). Design Methods for Ethical Persuasive Computing. In: 4th International Conference on Persuasive Technology, pp. 1–8. ACM, NY.
- Dennett, D. (1984). *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge, MA: MIT Press.
- Dennett, D. C. (2003). The Self as a Responding—and Responsible—Artifact. *Ann. N.Y. Acad. Sci.* 1001:39-50. USA.
- Draper, P. (2007). In Defense of Sensible Naturalism. *The Secular Web: A Drop of Reason in a Pool of Confusion*. Retrieved from http://infidels.org/library/modern/paul_draper/naturalism.html (16/7/2016).
- Driver, J. (2011). Gertrude Elizabeth Margaret Anscombe. *Stanford Encyclopedia of Philosophy*. Retrieved from <http://plato.stanford.edu/entries/anscombe/> (10/7/2016).
- Dubos, R. (1965). Science and Man's Nature. *Daedalus*, Vol. 94, No. 1, Science and Culture, pp. 223-244.
- Fanthorpe, L., Fanthorpe, P. (2014). *Mysteries and secrets: The 16-book complete codex*. Dundurn.
- Fogg, B. J. (2003). *Persuasive technology*. US, Morgan Kaufmann Publishers.
- Fogg, B. J. (2009). *A behavior model for persuasive design*. Persuasive'09. April 26-29, Claremont, California, US
- Fogg, B. J. (2009b) 'The new rules of persuasion,' *RSA Journal*, Vol. 155, No. 5538, pp. 24-29
- Frankfurt, H. (2001). Freedom of the Will and a Concept of a Person, in Gary Watson. ed., 2nd ed., *Free Will*. Oxford University Press, Oxford. Also in Kane, R. (2005). *A contemporary introduction to free will*. Oxford university press, NY. p. 95.

- Frankfurt, H. (2003). *Alternate Possibilities and Moral Responsibilities*, in Gary Watson, ed., 2nd ed., *Free Will*, Oxford University Press, Oxford.
- Friedman, B., Kahn Jr., P.H., Borning, A. (2006). *Value Sensitive Design and Information Systems*. In: Zhang, P., Galletta, D. (eds.) *Human-Computer Interaction and Management Information Systems: Foundations*, pp. 348–372. M.E. Sharpe, New York.
- Gibbons, J. (2006). *Mental Causation without Downward Causation*. *The Philosophical Re-*
- GlySense. (n.a.) *Developing next generation, continuous glucose monitoring systems for people with diabetes*. Retrieved from <http://glysens.com/> (7/6/2016)
- Google. (2013, Nov 5). *Updated Google app for iPhone and iPad [Video File]*. Retrieved from <https://www.youtube.com/watch?v=p8Ey0AufD9g> (8/4/2016)
- Gram-Hansen, S.B. (2009). *Towards an Approach to Ethics and HCI Development Based on Løgstrup's Ideas*. In: Gross, T., Gulliksen, J., Kotzé, P., Oestreicher, L., Palanque, P., Prates, R.O., Winckler, M. (eds.) *INTERACT 2009. LNCS*, vol. 5726, pp. 200–203. Springer, Heidelberg.
- Guthrie, W.K.C. (1965). *A History of Greek Philosophy: II The Presocratic tradition from Parmenides to Democritus*. Cambridge University Press, Cambridge.
- H. Oinas-Kukkonen et al. (Eds.). (2008). *PERSUASIVE 2008*. perFrames:
- Harris, S. (2012). *Free Will*. Free press, NY.
- Hasle, P. (2006). *The Persuasive Expansion – Rhetoric, Information Architecture, and Conceptual Structure*. *ICCS 2006, LNAI 4068*, pp. 2 – 21. Springer.
- Hasle, P. (2011). *Persuasive design: a different approach to information systems (and information)*. Vol. 29 No. 4, pp. 569-572 *Emerald*.
- Hasle, P., Christensen, A.-K.K.: *Persuasive Design*. In: Kelsey, S., St. Amant, K.: *Handbook of Research on Computer-Mediated Communication*. IGI Global, Hershey (in print, 2008)
- Hayward, J., Chansin, G. (2016). *Wearable Sensors 2016-2026: Market Forecasts, Technologies, Players*. *IDTechEx*. Retrieved from <http://www.idtechex.com/research/reports/wearable-sensors-2016-2026-market-forecasts-technologies-players-000470.asp?viewopt=desc> (7/6/2016)
- Heller, J. (1961). *Catch-22*. Simon & Schuster, NY.
- Higgins, C., Walker, R. (2012). *Ethos, logos, pathos: Strategies of persuasion in social/environmental reports*. *Accounting Forum* 36, 194-208.
- Hosch, W. L. (n.d.) *Amir Pnueli: Israeli computer scientist*. *Encyclopædia Britannica*. Retrieved from <http://www.britannica.com/biography/Amir-Pnueli> (17/6/2016)

- Ijsselsteijn, W., et al. (Eds.). (2006). PERSUASIVE 2006, LNCS 3962, pp. 171 – 182.
- In TIME (n.d.) The accelerating pace of change and exponential growth in computing power will lead to the Singularity. Retrieved from <http://content.time.com/time/interactive/0,31813,2048601,00.html> (27/7/2016)
- Inwagen P. V. (1983). *An Essay on Free Will*. Oxford University Press, Clarendon Press, Oxford.
- Jakobsen, D. (2007). The Turing Test and Other Minds. Retrieved from <https://www.researchgate.net/publication/253177127> (27/7/2016).
- Janssens, M. J. P. A, Rooij, M. F. A. M. V., et. al. (2004). Pressure and Coercion in the care for the addicted: ethical perspectives. *J Med Ethics*, 30:453-458.
- Kane, R. (1998). *The Significance of Free Will*. Oxford University Press, NY.
- Kane, R. (2002). Some Neglected Pathways in the Free Will Labyrinth. In Kane, R. (Eds.), *The Oxford Handbook of Free Will* (p. 406-437). Oxford University Press, Oxford.
- Kane, R. (2005). *A Contemporary Introduction to Free Will*. Oxford university press, NY.
- Kane, R. (Eds.) (2002). *The Oxford Handbook of Free Will*. Oxford University Press, Oxford.
- Kapitan, T. (2002). A Master Argument For Incompatibilism? In Kane, R. (Eds.), *The Oxford Handbook of Free Will* (p. 127-157). Oxford University Press, Oxford.
- Keliner, D. (Ed.). (1994). *Baudrillard—a critical reader*, Basil Blackwell, Ltd, Oxford, UK, pp. 119–132.
- Kim, J. (2010). Causation and Mental Causation. In *Essays in the Metaphysics of Mind*. Oxford University Press, Oxford.
- Kurzweil, R. (1999). When Machines Think. *Macleans's* 112.9:54.
- Kurzweil, R. (2005). *The Singularity is Near: When Humans Transcend Biology*. Penguin Books, USA.
- Kurzweil, R. (2005). *The Singularity is Near: When Humans Transcend Biology*. Penguin Books, USA.
- Laing, J. D. (n.d.) Middle Knowledge. Internet Encyclopedia of Philosophy: A Peer-Reviewed Academic Resource. Retrieved from <http://www.iep.utm.edu/middlekn/> 18/6/2016.
- Lewis, C. S. (2001). Miracles, Lewis quoted from J. B. S. Haldane, *Possible Worlds* (1927; reprint, New Brunswick). N.J.: Transaction Publishers.
- Libet, B. (2002). Do We Have Free Will? In Kane, R. (Eds.), *The Oxford Handbook of Free Will* (p. 551-564). Oxford University Press, Oxford.
- Lowe, E. J. (2004). *An introduction to the philosophy of mind*. Cambridge University Press. UK.

- MacTavish, T., Basapur, S. (Eds.). (2015). PERSUASIVE 2015. Ethical Challenges in Emerging Applications of Persuasive Technology. LNCS 9072, pp. 196–201. Springer, Switzerland.
- McKenna, m. (2004). Compatibilism, in Edward N. Zalta, ed., The Stanford Encyclopedia of Philosophy, online edition:
<http://plato.Stanford.edu/archives/sum2004/entries/compatibilism/>.
- McKenna, M., Coates, D. J. (2015, Feb 25th). Compatibilism. The Standford Encyclopedia of Philosophy, Retrieved from <http://plato.stanford.edu/entries/compatibilism/#ConArg> (4/7/2016).
- Menuge, A. (2004). Agents Under Fire: Materialism and the Rationality of Science. Rowman & Littlefield publishers, Inc. USA.
- Menuge, A. (2009). Is Downward Causation Possible? How the Mind can Make a Physical Difference. *Philosophia Christi*. Vol. 11, No. 1.
- Menuge, A. (2011). The Ontological Argument from Reason: Why Compatibilists Accounts of Reasoning Fail. *Philosophia Christi*. Vol. 13, No. 1.
- Menuge, A. (2013). Neuroscience, Rationality, and Free Will: A Critique of John Searle's Libertarian Naturalism. *Philosophia Christi*. Vol. 15, No. 1. Wisconsin.
- Meschtscherjakov et al. (Eds.). (2016). PERSUASIVE 2016, LNCS 9638, pp. 189–196. Springer, Switzerland.
- Morville, P., Rosenfeld, L. (2006). *Information architecture for the World Wide Web*. US, O'Reilly Media, Inc.
- Nick, T. (2014, July 6). Did you know how many different kinds of sensors go inside a smartphone? Phone Arena. Retrieved from http://www.phonearena.com/news/Did-you-know-how-many-different-kinds-of-sensors-go-inside-a-smartphone_id57885 (6/7/2016)
- Øhrstrøm, P., (n.d.). *The Concept of Time – a Philosophical and Logical Perspective*. Department of Communication and Psychology, Aalborg University.
- Øhrstrøm, P., Hasle, P. (1995). *Temporal Logic: From Ancient Ideas to Artificial Intelligence*. Kluwen Academic Publishers. The Netherlands.
- Oinas-Kukkonen, H., Hasle, P., Harjumaa, M., Segerståhl, K., Øhrstrøm, P. (2008). PERSUASIVE 2008. LNCS, vol. 5033, pp. 254–257. Springer, Heidelberg
- Parsons, K. (2000). Further Reflections on the Argument from Reasons. *Philo* 3. No. 1. Also in Reppert. 2003, chapter 3.
- Penrose, R. (1989). *Emperor's New Mind: Concerning Computers, Minds, and The Laws of Physics*. Penguin Books, USA.

- Persuasive Picture Frames for Proper Posture LNCS 5033, pp. 128–139. Springer. Berlin.
- Physical constant. (n.d.) In Encyclopedia Britannica, Inc., Retrieved from <https://www.britannica.com/science/physical-constant> (18/7/2016).
- Plantinga, A. (1993). *Warrant and Proper Function*. Oxford University Press. NY.
- Plantinga, A. (1994). *Naturalism Defeated*. Retrieved from https://www.calvin.edu/academic/philosophy/virtual_library/articles/plantinga_alvin/naturalism_defeated.pdf (8/7/2016).
- Plantinga, A. (2007). Against “Sensible” Naturalism. A Drop of Reason in a Pool of Confusion. Retrieved from http://infidels.org/library/modern/alvin_plantinga/against-naturalism.html (16/7/2016).
- Ploug, T., Øhstrøm, P. (2012). Branching time, indeterminism and tense logic: Unveiling the Prior-Kripke letters. *Synthese*. 188:367-179.
- Ramachandran, V. (1998, Sept 5th). Quoted in *New Scientist*. p .35.
- Reppert, V. (2003). *C.S. Lewis’s Dangerous Idea: In Defense of the Argument from Reason*. InterVarsity Press, Illinois. Kindle Edition.
- Robinson, H. (2016). Dualism. *Stanford Encyclopedia of Philosophy*. Retrieved from <http://plato.stanford.edu/entries/dualism/#Int> (27/6/2016).
- Rosenberg, A. (2011). *The Atheist’s Guide to reality: Enjoying Life without illusions*. Norton & Company, USA.
- Schwitzgebel, E. (2015). Belief. *Standard Encyclopedia of Philosophy*. Retrieved from <http://plato.stanford.edu/entries/belief/#2.2> (19/7/2016).
- Sudduth, M. (n.d.). Defeaters in epistemology. *Internet encyclopedia of philosophy*. Retrieved from <http://www.iep.utm.edu/ep-defea/#SH1b> (13/7/2016)
- Thompson, R. (2000). “Kairos Revisited: An Interview with James Kinneavy.” *Rhetoric Review* 19.1: 73–88.
- Timple, K. (n.d.). Free Will. *Internet Encyclopedia of Philosophy: A Peer Reviewed Academic Resource*. Retrieved from <http://www.iep.utm.edu/freewill/#SH5b> (4/7/2016).
- Truncellito, D. A. (n.d.). *Epistemology*. *Internet encyclopedia of philosophy* Retrieved from <http://www.iep.utm.edu/epistemo/> (13/7/2016)
- Waller, B. (1988). Free Will Gone Out of Control: A Critical Study of R. Kane's Free Will and Values. *Behaviorism* 16: 149-67.
- Watson, G. (1982), ed. *Free Will*, Oxford: Oxford University Press.
- Watson, G. (1987). Free Action and Free Will. *Mind* 96. p. 145-72.

- Wolf, S. (2002). *Sanity and the Metaphysics of Responsibility*. In Robert Kane, ed., *Free Will*. Oxford: Blackwell Publishers.
- Y. De Kort et al. (Eds.). (2007). *PERSUASIVE 2007*, LNCS 4744, pp. 12-17. 2007. Springer. Berlin.
- Yetim, F. (2011). *A Set of Critical Heuristics for Value Sensitive Designers and Users of Persuasive Systems*. In: *ECIS 2011 Proceedings*, Helsinki.

Appendix 1

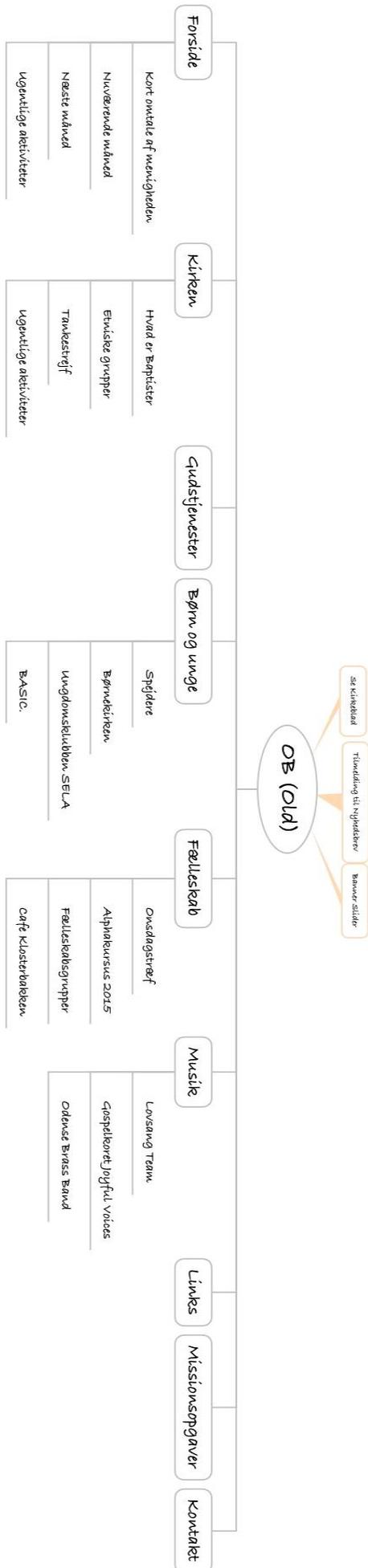
Principle	Example requirement	Example implementation
<p>1. Reduction A system that reduces complex behavior into simple tasks helps users perform the target behavior and it may increase the benefit/cost ratio of a behavior.</p>	System should reduce effort that users have in regard to performing their target behavior.	Mobile application for healthier eating habits lists proper food choices at fast food restaurants [24]. Smoking cessation web site provides an interactive test which measures how much money a user will save with quitting.
<p>2. Tunneling Using the system to guide users through a process or experience provides opportunities to persuade along the way.</p>	System should guide users in the attitude change process by providing means for action that brings closer to the target behavior.	Smoking cessation web site offers information about treatment opportunities after a user has answered an interactive test about how addicted (s)he is on tobacco.
<p>3. Tailoring Information provided by the system will be more persuasive if it is tailored to the potential needs, interests, personality, usage context, or other factors relevant to a user group.</p>	System should provide tailored information for its user groups.	Personal trainer Web site provides different information content for different user groups, e.g. beginners and professionals. Web site for recovering alcoholics presents a user such stories which are close to one's own story.
<p>4. Personalization A system that offers personalized content or services has a greater capability for persuasion.</p>	System should offer personalized content and services for its users.	Users are able to change the graphical layout of an application or the order of information items at a professional Web site.
<p>5. Self-monitoring A system that helps track one's own performance or status supports in achieving goals.</p>	System should provide means for users to track their performance or status.	Heart rate monitor presents a user's heart rate and the duration of the exercise. Mobile phone application presents daily step count [3].
<p>6. Simulation Systems that provide simulations can persuade by enabling them to observe immediately the link between the cause and its effect.</p>	System should provide means for observing the link between the cause and effect in regard to their behavior.	Before and after pictures of people who have lost weight are presented on a Web site.
<p>7. Rehearsal A system providing means with which to rehearse a behavior can enable people to change their attitudes or behavior in the real world.</p>	System should provide means for rehearsing a target behavior.	A flying simulator.

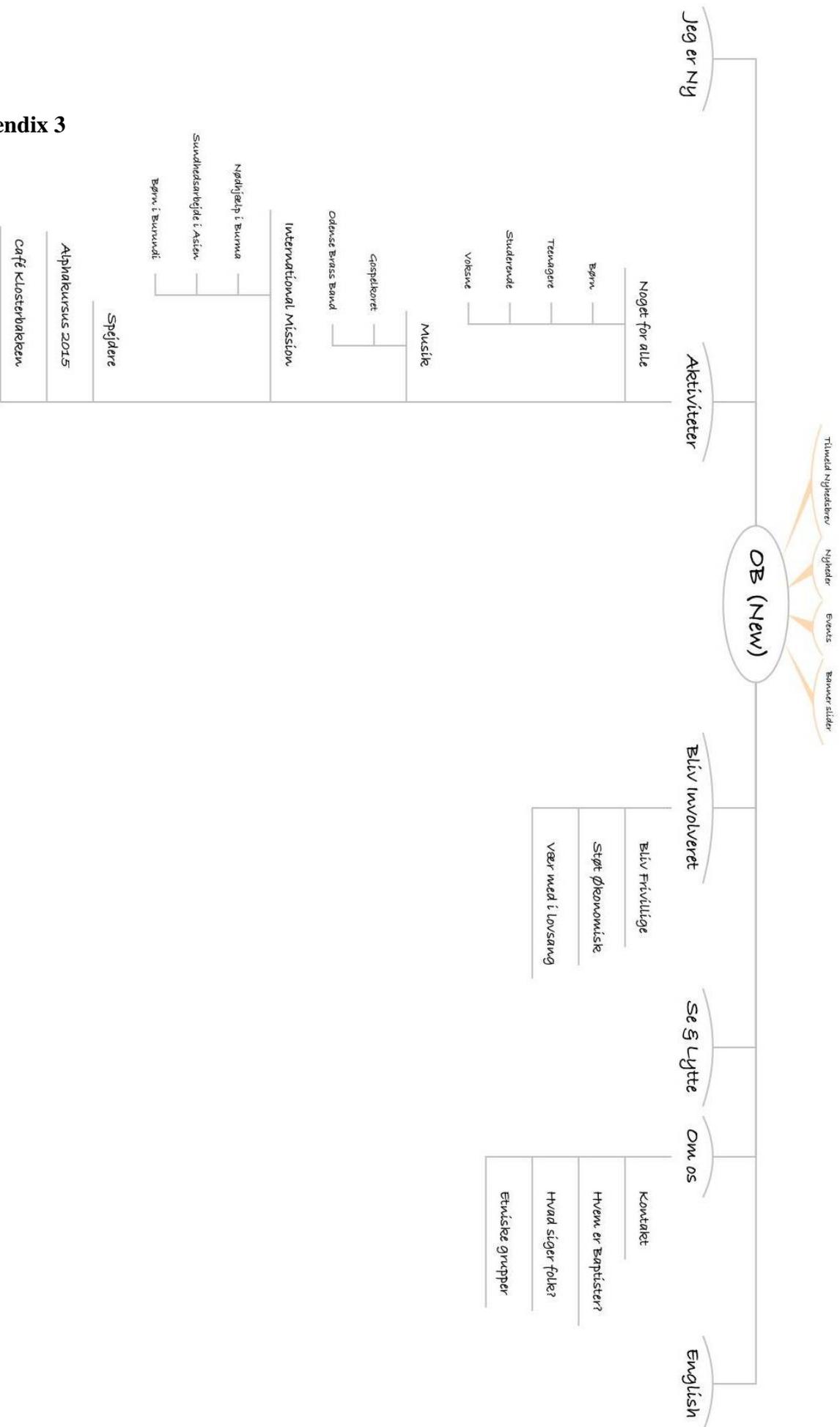
Principle	Example requirement	Example implementation
8. Praise By offering praise a system can make users more open to persuasion.	System should use praise via words, images, symbols, or sounds as a way to give positive feedback for a user.	Mobile application which aims at motivating teenagers to exercise praises user by sending automated text-messages for reaching individual goals. [24]
9. Rewards Systems that reward target may have great persuasive powers.	System should provide virtual rewards for users in order to give credit for performing the target behavior.	Heart rate monitor gives a user a virtual trophy if they follow their fitness program. Game rewards users by altering media items, such as sounds, background skin, or a user's avatar according to user's performance. [21]
10. Reminders If a system reminds users of their target behavior, the users will more likely achieve their goals.	System should remind users of their target behavior during the use of the system.	Caloric balance monitoring application sends text-messages for their users as daily reminders. [10]
11. Suggestion Systems offering suggestions at opportune moments will have greater persuasive powers.	System should suggest users certain behaviors during the system use process.	Application for healthier eating habits suggests children to eat fruits instead of candy at a snack time.
12. Similarity People are more readily persuaded through systems that remind themselves in some meaningful way.	System should imitate its users in some specific way.	Slang names are used in an application which aims at motivating teenagers to exercise. [24]
13. Liking A system that is visually attractive for its users is likely to be more persuasive.	System should have a look and feel that appeals to its users.	Web site which aims at encouraging children to take care of their pets properly has pictures of cute animals.
14. Social role If a system adopts a social role, users will more likely use it for persuasive purposes.	System should adopt a social role.	E-health application has a virtual specialist to support communication between users and health specialists. [19]

Principle	Example requirement	Example implementation
15. Trustworthiness A system that is viewed as trustworthy (truthful, fair, and unbiased) will have increased powers of persuasion.	System should provide information that is truthful, fair and unbiased.	Company Web site provides information related to its products rather than simply providing advertising or marketing information.
16. Expertise A system that is viewed as incorporating expertise (knowledge, experience, and competence) will have increased powers of persuasion.	System should provide information showing expertise.	Company Web site provides information about their core know-how. Company Web site is updated regularly and there are no dangling links or out-of-date information.
17. Surface credibility People make initial assessments of the system credibility based on a firsthand inspection.	System should have competent look and feel.	There are only a limited number of and a logical reason for ads on a company Web site.
18. Real-world feel A system that highlights people or organization behind its content or services will have more credibility.	System should provide information of the organization and/or actual people behind its content and services.	Company Web site provides possibilities to contact specific people through sending feedback or asking questions.
19. Authority A system that leverages roles of authority will have enhanced powers of persuasion.	System should refer to people in the role of authority.	Web site quotes an authority, such as a statement by government health office.
20. Third-party endorsements Third-party endorsements, especially from well-known and respected sources, boost perceptions on system credibility.	System should provide endorsements from respected sources.	E-shop shows a logo of a certificate which assures that they use secure connections. Web site refers to its reward for high usability.
21. Verifiability Credibility perceptions will be enhanced if a system makes it easy to verify the accuracy of site content via outside sources.	System should provide means to verify the accuracy of site content via outside sources.	Claims on a Web site are supported by offering links to other web sites.

Principle	Example requirement	Example implementation
22. Social learning A person will be more motivated to perform a target behavior if he or she can use a system to observe others performing the behavior.	System should provide means to observe other users who are performing their target behaviors and to see the outcomes of their behavior.	A shared fitness journal in a mobile application for encouraging physical activity. [3]
23. Social comparison System users will have a greater motivation to perform the target behavior if they can compare their performance with the performance of others.	System should provide means for comparing performance with the performance of other users.	Users can share and compare information related to their physical health and smoking behavior via instant messaging application. [21]
24. Normative influence A system can leverage normative influence or peer pressure to increase the likelihood that a person will adopt a target behavior.	System should provide means for gathering together people who have the same goal and get them to feel norms.	Possibility to challenge relatives or friends to quit smoking from a web site via email or text message.
25. Social facilitation System users are more likely to perform target behavior if they discern via the system that others are performing the behavior along with them.	System should provide means for discerning other users who are performing the behavior.	A shared fitness journal in a mobile application for encouraging physical activity. [3]
26. Cooperation A system can motivate users to adopt a target attitude or behavior by leveraging human beings' natural drive to co-operate.	System should provide means for co-operation.	The behavioral patterns of overweight patients are studied through a mobile application, which collects data and sends it to a central server where it can be analyzed in detail. [10]
27. Competition A system can motivate users to adopt a target attitude or behavior by leveraging human beings' natural drive to compete.	System should provide means for competing with other users.	Online competition, such as Quit and Win (stop smoking for a month and win a prize).
28. Recognition By offering public recognition (for an individual or a group), a system can increase the likelihood that a person or group will adopt a target attitude or behavior.	System should provide public recognition for users who perform their target behavior.	Personal stories of the people who have succeeded in their goal behavior are published on a Web site. Names of awarded people, such as "quitter of a month", are published on a Web site.

Appendix 2





THE (

Appendix 3

Index

A

Artificial intelligence (AI), 6, 9, 87, 101, 102, 104, 105, 106, 107, 108, 109, 111, 112
 Attitude, 16, 30, 32, 65, 88, 89, 110, 111

B

Behavior, 11, 15, 16, 26, 27, 28, 30, 32, 40, 41, 42, 59, 61, 62, 65, 71, 72, 73, 74, 88, 89, 101, 108, 110, 111, 115
 Belief in existence of other minds (BOM), 6, 105, 107, 108, 109
 Branching Time, 17, 18, 19, 20, 25, 38, 77, 90, 114, 119
 branching tree, 19, 23, 25

C

captology, 14, 15, 30
 Causally closed system (PCC), 6, 44, 58, 92, 97, 110
 Chinese room (CR), 6, 106, 107, 109
 Compatibilism, 5, 43, 47, 59, 74, 76, 77, 78, 82, 83, 87, 89, 100, 110, 118
 Consciousness, 44, 45, 48, 49, 50, 52, 94, 103, 104, 108
 conscious, 48, 49, 50, 51, 64, 65, 75, 92, 96, 99, 102, 104, 105, 108, 109, 111
 consciously, 48, 49, 64
 Control, 37, 40, 41, 42, 43, 44, 47, 51, 76, 77, 80, 82, 85, 86, 92, 99

D

Delusion, 1, 9, 110, 111, 112, 113
 Determinism, 5, 37, 38, 41, 43, 46, 47, 54, 55, 74, 75, 76, 77, 80, 82, 83, 85, 87, 89, 99, 100, 110, 111
 determined, 10, 38, 39, 40, 50, 54, 57, 59, 75, 81, 84, 90, 95, 97, 99

F

Fogg's Behavior Model, 26
 Freedom, 15, 18, 19, 32, 37, 38, 39, 40, 41, 42, 43, 48, 52, 76, 77, 78, 79, 82, 83, 84, 86, 87, 88, 89, 90, 94, 96, 97, 100, 110, 111, 115
 free choice, 17, 18, 38
 free will, 18, 20, 25, 30, 32, 37, 38, 39, 40, 41, 42, 43, 46, 47, 48, 50, 51, 52, 59, 74, 75, 76, 80, 81, 82, 83, 84, 85, 86, 89, 90, 91, 93, 94, 96, 97, 99, 100, 104, 115, 117

G

Garden of forking paths, 18, 90, 100

H

Human computer interaction (HCI), 6, 20, 32, 111, 116
 Hypothetical analysis (HA), 6, 78, 79, 89

I

Illusion, 31, 32, 38, 40, 62, 113
 illusory, 39, 41, 42, 77, 87, 111
 Indeterminism, 18, 90, 94, 98, 119
 indeterministic, 38, 98
 Information architecture (IA), 6, 17, 23, 24, 25, 32
 Intelligent, 21, 29, 101, 102, 106, 109, 113
 Introspection, 48, 52, 53, 94

K

Kairos, 11, 112, 114, 119
 opportune moments, 11, 14
 timing, 11, 112

L

Libertarianism, 43, 89, 92
 libertarian, 81, 89, 90, 118

M

Materialism, 44, 45, 52, 54, 55, 57, 58, 59, 60, 93, 110

N

Naturalism, 44, 45, 54, 57, 58, 60, 63, 67, 70, 72, 74, 108, 110, 111, 115, 119

P

Persuasion, 5, 9, 10, 11, 14, 15, 16, 23, 24, 30, 38, 39, 64, 88, 89, 110, 111, 112, 115, 116
 persuasive, 10, 14, 15, 16, 20, 21, 24, 29, 30, 64, 65, 84, 112, 115
 Persuasive Design (PD), 1, 6, 15, 17, 23, 25, 26, 27, 28, 29, 32, 33, 38, 40, 64, 65, 88, 89, 110, 116
 Persuasive Technology (PT), 6, 15, 16, 27, 30, 111, 114, 115, 118

Principle of Alternative Possibilities (PAP), 6, 79, 80, 81, 82

R

Rationality, 66, 69, 94, 101, 107, 108, 110, 111
rational, 16, 38, 53, 54, 55, 56, 57, 58, 59, 60, 68, 74,
80, 89, 93, 94, 95, 96, 97, 98, 100, 101, 104, 105,
107, 108, 109, 110, 111
Reason, 10, 44, 53, 54, 55, 56, 59, 60, 63, 64, 65, 67, 68,
72, 75, 80, 101, 105, 108, 109, 110
reasoning, 54, 55, 56, 58, 63, 79, 80, 88, 91, 94, 100,
112
Reductionism, 45

S

Self-forming actions (SFAs), 6, 42, 81, 96, 97, 99, 100, 111
sensor(s), 12, 13, 14, 28, 112, 116, 118
Simulation(s), 30, 31, 32, 33, 101

T

The functional triad, 29
Tool(s), 19, 28, 30, 38
Turing test (TT), 6, 102, 103, 104, 105, 106, 107, 108, 109,
112
The Imitation Game, 104, 108